Administrative and Survey Data

DDI-based documentation for a combined analysis

Marcel Hebing (DIW)
David Schiller (IAB)

IASSIST, Washington DC, 2012-06-07
Who we are

**David: Institute for Employment Research IAB (administrative data)**

- Located at the BA Nuremberg
- Data from the German social security system and from internal processes of the German Federal Employment Agency, BA
- Daily information for every employer in Germany since 1975
- Normally 2% samples

**Marcel: German Socio-Economic Panel Study SOEP (survey data)**

- Located at the DIW Berlin
- Wide-ranging representative longitudinal study of private households in Germany
- Every year, since 1984
- 11,000 households and 20,000 individuals a sampled by our fieldwork organisation TNS Infratest
The idea

- **Survey data**
  - Gives you the possibility to decide yourself, what you want to ask.
  - But the quality of the data relies on the willingness and the ability of the respondents to answer your question.

- **Administrative data**
  - Contains only information that are needed within a non-scientific administration process
  - But very powerful (population) and precise (less biased because not voluntary)

- **The idea**: Why don't we combine the best of both worlds?

- **Research question**: How does a good documentation look like that allows substantial scientific research based on the combination of admin and survey data.
Two Perspectives

Our first idea was to focus on the DDI elements you need to combine administrative and survey data. But we realized that need more information than the standard is intended for, mainly information about the business processes.

(1) **Business process view** based on the GLBPM (Generic Longitudinal Business Process Modell) developed at the DDI Dagstuhl 2011 workshop.

(2) **Data product view** on the study and data design as a starting point for our DDI-discussion.

Both perspectives are used to discuss how to enable combined analysis.
The Business Process View

(Perspective 1)
GLBPM

Generic Longitudinal Business Process Model (Dagstuhl 2011)

It's based on:

- GSBPM: http://www.unece.org/stats/gsbpm.html
- DDI Data Life Cycle: http://www.ddialliance.org/
- Tornado Model. Dagstuhl 2010
Generic Longitudinal Business Process Model: Overview

1. Evaluate / Specify Needs
   - 1.1 Define research questions, universe & high-level concepts
   - 1.2 Evaluate existing data & publications
   - 1.3 Establish outputs & needed infrastructure
   - 1.4 Define specific concepts to be measured
   - 1.5 Planning, timetable & needed infrastructure
   - 1.6 Prepare proposal & get funding

2. Design / Redesign
   - 2.1 Identify sources
   - 2.2 Design sampling methods
   - 2.3 Design collection instruments
   - 2.4 Specify data elements
   - 2.5 Specify processing / cleaning methods
   - 2.6 Organize research team
   - 2.7 Design infrastructure

3. Build / Rebuild
   - 3.1 Develop data collection instruments
   - 3.2 Create or enhance infrastructure components
   - 3.3 Validate instruments
   - 3.4 Test production systems
   - 3.5 Finalize production systems

4. Collect
   - 4.1 Select sample
   - 4.2 Set up collection
   - 4.3 Run collection
   - 4.4 Finalize collection

5. Process / Analyse
   - 5.1 Integrate data
   - 5.2 Classify & code
   - 5.3 Explore, validate & clean data
   - 5.4 Impute missing data
   - 5.5 Construct new variables and units
   - 5.6 Calculate weights
   - 5.7 Calculate aggregates
   - 5.8 Anonymize data
   - 5.9 Finalize data outputs

6. Archive / Preserve / Curate
   - 6.1 Ingest data / metadata
   - 6.2 Enhance metadata
   - 6.3 Preserve data / metadata

7. Data Dissemination / Discovery
   - 7.1 Deploy release infrastructure
   - 7.2 Prepare dissemination products
   - 7.3 Deploy access control system / policies
   - 7.4 Promote dissemination products
   - 7.5 Provide data citation support
   - 7.6 Enhance data discovery
   - 7.7 Manage user support

8. Research / Publish
   - 8.1 Discovery of data and relevant research
   - 8.2 Access data
   - 8.3 Prepare data
   - 8.4 Analyse data
   - 8.5 Prepare research papers
   - 8.6 Manage disclosure risk
   - 8.7 Publish research

9. Retrospective Evaluation
   - 9.1 Establish evaluation criteria
   - 9.2 Gather evaluation inputs
   - 9.3 Conduct evaluation
   - 9.4 Determine future actions

---

Project Management / Quality Management

Metadata Management

Use of External Standard Metadata (classifications, concepts, questions, variables)
GLBPM

1. Evaluate / Specify Needs
2. Design / Redesign
3. Build / Rebuild
4. Collect
5. Process / Analysis
6. Archive / Preserve / Curate
7. Data Dissemination / Discovery
8. Research / Publish
9. Retrospective Evaluation
Example 1: Immigrant Sample

- **Basic:** Germany has no central index/register for immigrants.
- **First idea:** use administrative data from the Federal Employment Agency, to get a high quality sample of immigrants (up to 92% coverage of the basic population is expected).
- **Second idea:** combine administrative and survey data to get more sustainable / meaningful data.
- **Four working package:**
  1. sample and weights,
  2. informed consent and willingness/participation,
  3. record linkage and data preparation, and
  4. data privacy and data access.
Example 1: Iterations

Initial iteration:
- Admin: the researcher doesn't have an influence.
- Survey: fundamental design decisions are made

Further iteration:
- Admin: redesign and rebuild of data for scientific use
- Survey (longitudinal): additional waves (primarily collection, but also new samples and instruments)

Combining iteration:
- Integration of admin data to the survey
Example 2: Processes of Mate Choice on Online-Dating. (PPOK)

Main subject of the project is the empirical exploration of Processes of Mate Choice in Online-Dating Services.

The focus of interest is laid on gender-specific preferences and change of these preferences through the dynamics of the online mate-market, aspects of self-presentation of the users, the socialstructural composition of the dating-population and the process of the online mate choice at the timeline.

http://www.partnerwahlforschung.de
Example 2: Processes of Mate Choice on Online-Dating. (PPOK)

Process generated data, very similar to administrative data regarding data generation and data structure.

It started with the original data from a German online-dating-platform, but the data are not sufficient for some research questions. So additional information were collected in an online panel survey.

In this case the panel survey focuses the process generated data from the first iteration on.
Example 2: Iterations

Initial iteration:
• Transactional data are stored as a part of the online-dating-website.

Second iteration:
• Extraction of these data for the usage as research data.

Third iteration:
• Adding a online survey to the website to collect additional information

Fourth iteration:
• Combining process generated and survey data.
Comparison

Comparing the two examples, we find fundamentally different processes of data collection.

It's essential to understand these differences to enable substantial analysis based on such data sources.

The GLBPM seems to be a very good approach to discuss, model, and document these differences as it provides a generic framework and vocabulary for describing business processes.

Nevertheless it is necessary to extend the model for the context of non-survey data.
The Data Product View

(Perspective 2)
Data Collection

BA-Data
No active data collection for scientific research

Data derived from administrative processes:
Notifications to the German Social Security System (e.g. wages and employment on a daily base)

Internal processes of the German Federal Employment Agency, BA (e.g. receipt of unemployment benefits, registered job search)

Ongoing process of data genesis

Participation: mandatory

SOEP
Planed fieldwork carried out by external organisation

Sample: probability sample, random root

Method: questionnaire (interviewer assisted: CAPI / PAPI)

Timeframe: distinct -- once a year, since 1984

Participation: voluntary
Data Structure

BA
- Data derives from different applications and
- is stored within a DWH
- Extracts of the DWH are stored in relational databases
- Result: data on individuals with timestamps

IAB
- Data are edited with SAS (because of data size)
- Samples of data are edited with Stata (because our researchers like it)
- Mostly spell data

SOEP
- Collection-oriented data-structure: almost one table per questionnaire and per year
- Stored in a databases
- Data are delivered as SPSS, Stata and SAS files
- Raw data and generated data
- Wide vs. long, sometimes spell data (only for retrospective questions)
Data Documentation

Admin-Data
- Documentation depends on the involved organisation:
  - Jobcenter
  - Federal Employment Agency
  - IT-Department of IAB
  - Research Data Centre of IAB
- Standardized written (PDF) documentation about data genese including
  - Sample information
  - Data quality
  - Data creation process
  - Changes over time

SOEP
- Basic (written) documentation of the study design
- Web-application for variable documentation and selection (basket)
- Questionnaires

Both are going for a DDI documentation
Disclosure Risk

Admin-Data
- Special law for social insurance data
- Highly detailed data only available via guest stay, remote access and remote execution
- High disclosure risk due to population (no sample participation is mandatory)
- Although IAB provides only sub-samples to the scientific community
- High demands on disclosure control

SOEP
- Data privacy: less strict in comparison to administrative data
- Data access: DVD, remote access, guests, future: download.

To merge both sources the participants have to agree
Conclusion
Need for a standardized documentation

- A standardized data documentation is essential, if others should really be able to use your data
- Otherwise the researcher will not be able to produce reliable results
- Especially when combining data from different sources
- Therefore the documentation must be one of the first things to think about when creating new data
Focus on data collection process

What must be done to extend DDI to the needs of different data products:

Data about Data = Metadata
Admin and survey data do not differ when looking at the edited dataset (Both have: study units, concepts, universes, datasets, etc.)

Data about the Collection of Data = Paradata
In contrast the processes of data collection is quiet different
Next steps

- Look at different kinds of administrative data (there is more out there than employment data)
  - A DDI alliance working group about administrative data is currently starting his work (feel free to join)
- Look at different kinds of data collection (there is more out there than survey and admin data)
  - A DDI alliance working group about paradata is already running
- Modify DDI to fit the different needs (and keep it simple)
Business model

- Need for a standard used during the complete life cycle --> but is this enough?
  - Can a standard like DDI also document the business process?

- The combination of admin and survey data can happen in many ways
  - Documentation of business processes must become a part of the DDI standard

But: keep it simple!
Thanks for listening!

Marcel Hebing, DIW
mhebing@diw.de

David Schiller, IAB
david.schiller@iab.de