Why should official statistics care about data integration?

First experiences in linking business microdata in Germany

Christopher Gürke
RDC of the Federal Statistical Office of Germany
Integration of data from different sources:

**motivation**

Major reasons in official statistics:

- Can be a faster and/or cheaper way to compile certain information/results
- Can be a possible way to reduce respondents reporting duties
- Can enable novel and more comprehensive analyses
Integration of data from different sources: Projects

- „Amtliche Firmendaten für Deutschland“ (AFiD)
  - merge official business micro data over time and different fields of official statistical business micro data
  - advance the analytical potential and to offer information about enterprises and their local units with respect to different official statistics, different times and different thematically issues

- “Integrated Enterprise Data for Germany” (KombiFiD)
  - merge business data of the German Federal Statistical Office, the Land Statistical Offices, the Institute for Employment Research (IAB) and the Deutsche Bundesbank.
Integration of data from different sources:
Timeline

2008 Start
2009 Survey
2010 Linkage
2011 Release
2012 End
Integration of data from different sources: Start

Projectpartner:
- Federal Statistical Office
- Institute for Employment Research of the Federal Employment Agency
- Deutsche Bundesbank
- Leuphana University of Lüneburg
- Cooperation with the RDC of the Statistical Offices of the Länder

Funded by:
- Federal Ministry of Education and Research
Integration of data from different sources: Start

- Generating a unique business micro dataset
- Providing an enhanced information background for entrepreneurial decision-making
- Reducing respondent burden for businesses in official surveys and notification procedures
- Reason to think about a permanent matching of the data
Integration of data from different sources:
Timeline

- 2008: Start
- 2009: Survey
- 2010: Linkage
- 2011: Release
- 2012: End
Legal situation and the consequences

- Business statistics can only be matched if the data producer is the same (§13a Statistical Law)

- Limits the possibilities with regard to the integration of enterprise data, collected by different public institutions
  - Enterprises have to give their approval to this procedure
  - Time and cost consuming
Survey

- In 2009, a sample of 54,960 enterprises were asked whether they agree with the combination of their data

- Response rates

  - First survey phase (April – August 2009)
    - Response rate 27.6%
    - Agreement rate: 17.8%

  - First reminder
    - Response rate 43.6%
    - Agreement rate: 24.9%

  - Second reminder
    - Response rate 55.6%
    - Agreement rate: 29.6%
Survey
Response behaviour by economic sector and region as %

<table>
<thead>
<tr>
<th>Headquarters of enterprise</th>
<th>Economic sector</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Manufacturing</td>
<td></td>
</tr>
<tr>
<td>Western Germany</td>
<td>39.9</td>
<td>30.6</td>
</tr>
<tr>
<td>Eastern Germany</td>
<td>39.1</td>
<td>28.2</td>
</tr>
<tr>
<td></td>
<td>39.9</td>
<td>30.7</td>
</tr>
<tr>
<td>Total</td>
<td>39.9</td>
<td>30.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Construction</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Western Germany</td>
<td>32.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eastern Germany</td>
<td>26.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>30.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Survey
Response behaviour by economic sector and enterprise size class as %

<table>
<thead>
<tr>
<th>Enterprise size class</th>
<th>Economic sector</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Manufacturing</td>
<td>Construction</td>
<td>Trade</td>
<td>Services</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/20 – 49</td>
<td>29.2</td>
<td>21.4</td>
<td>19.3</td>
<td>16.7</td>
<td></td>
<td>19.6</td>
</tr>
<tr>
<td>50 – 99</td>
<td>33.3</td>
<td>28.9</td>
<td>23.8</td>
<td>21.5</td>
<td></td>
<td>27.0</td>
</tr>
<tr>
<td>100 – 249</td>
<td>36.7</td>
<td>35.2</td>
<td>29.1</td>
<td>21.8</td>
<td></td>
<td>30.8</td>
</tr>
<tr>
<td>250 – 499</td>
<td>37.6</td>
<td>40.6</td>
<td>32.4</td>
<td>28.0</td>
<td></td>
<td>34.2</td>
</tr>
<tr>
<td>500 – 999</td>
<td>42.2</td>
<td>34.0</td>
<td>34.3</td>
<td>33.1</td>
<td></td>
<td>38.4</td>
</tr>
<tr>
<td>&gt;= 1000</td>
<td>42.0</td>
<td>19.2</td>
<td>38.9</td>
<td>32.4</td>
<td></td>
<td>38.5</td>
</tr>
<tr>
<td>Total</td>
<td>34.2</td>
<td>25.4</td>
<td>22.1</td>
<td>18.7</td>
<td></td>
<td>30.7</td>
</tr>
</tbody>
</table>
Integration of data from different sources: Timeline

- 2008: Start
- 2009: Survey
- 2010: Linkage
- 2011: Release
- 2012: End
Record Linkage – Data sets

- **Federal Statistical Office**
  - Business register system 95 (Masterfile)
  - 11 Statistics of the sectors
    - Industry and Manufacturing
    - Construction
    - Services
  - Turnover tax statistic

- **Institute for Employment Research**
  - Establishment-History-Panel

- **Deutsche Bundesbank**
  - Microdatabase Direct Investment (MiDi)
  - Corporate balance sheets (USTAN)
Record Linkage Methods

- Linkage between IAB and Federal Statistical Office
  - Direct identifier of local units
  - Challenge: Inconsistent number of local units

- Linkage between IAB and Deutsche Bundesbank
  - No direct identifier of enterprises
  - Name and address used as key identifier
  - Challenge: Inconsistent spelling or special characters
  - Software used: record linkage software Merge-Toolbox (University of Duisburg)
Record Linkage
IAB and Federal Statical Office

Number of complete enterprises identified in the BHP and single-establishment enterprises

<table>
<thead>
<tr>
<th>year</th>
<th>total number of enterprises</th>
<th>Complete enterprises identified in the BHP</th>
<th>Single-establishment enterprises</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>absolute</td>
<td>percentage of all URS enterprises</td>
</tr>
<tr>
<td>2003</td>
<td>15,812</td>
<td>13,722</td>
<td>86.8</td>
</tr>
<tr>
<td>2004</td>
<td>15,888</td>
<td>13,653</td>
<td>85.9</td>
</tr>
<tr>
<td>2005</td>
<td>15,924</td>
<td>13,580</td>
<td>85.3</td>
</tr>
<tr>
<td>2006</td>
<td>15,942</td>
<td>13,549</td>
<td>85.0</td>
</tr>
</tbody>
</table>
Record Linkage
IAB and Deutsche Bundesbank

Results of the precise comparison of names and addresses

<table>
<thead>
<tr>
<th>Model</th>
<th>Blocking variable</th>
<th>Variables</th>
<th>Number of valid links</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3-digit postcode</td>
<td>Enterprise name, legal form, town, street, street number</td>
<td>22,080</td>
</tr>
<tr>
<td>2</td>
<td>3-digit postcode</td>
<td>Enterprise name, legal form, 5-digit postcode, street, street number</td>
<td>631</td>
</tr>
<tr>
<td>3</td>
<td>2-digit postcode</td>
<td>Enterprise name, legal form, town, street, street number</td>
<td>20</td>
</tr>
</tbody>
</table>
## Results of the error-tolerant comparison of names

<table>
<thead>
<tr>
<th>Model</th>
<th>Blocking variables</th>
<th>Variables</th>
<th>Similarity function</th>
<th>No. of valid links</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5-digit postcode, street, street no.</td>
<td>Enterprise name</td>
<td>Bigrams</td>
<td>15,435</td>
</tr>
<tr>
<td>5</td>
<td>2-digit postcode, town, street, street no.</td>
<td>Enterprise name</td>
<td>Bigrams</td>
<td>114</td>
</tr>
<tr>
<td>6</td>
<td>5-digit postcode, street</td>
<td>Enterprise name</td>
<td>Bigrams</td>
<td>2,976</td>
</tr>
</tbody>
</table>
Integration of data from different sources: Timeline

- 2008: Start
- 2009: Survey
- 2010: Linkage
- 2011: Release
- 2012: End
Release of KombiFiD 2.0

- Panel covering the years 2003-2006
- about 14,000 firms per year
- Available since autumn 2011
- Metadata available for all variables
- Data access via Safe Centre and remote execution
Release of KombiFiD 2.0

- **Challenges before the release:**
  - Harmonisation of different confidentiality rules
  - Harmonisation of organisational issues

- **Data Structure File**
  - Building a Pseudopanel
  - Preservation of the participation pattern over time and statistics
  - Building an absolutely anonymised micro dataset
  - Development of a data swapping macro
Integration of data from different sources: Timeline

- 2008: Start
- 2009: Survey
- 2010: Linkage
- 2011: Release
- 2012: End
Results of KombiFiD

- KombiFiD shows the possibilities for linking micro data across the borders of data producers

- combination of different data sources offers new possibilities in research activities

- KombiFiD Results
  - discussed in the Statistical Advisory Committee of the FSO
  - discussed in the context of the revision of the Statistical Law (BStatG)
Thank you for your attention

Christopher Gürke
Research Data Centre
Federal Statistical Office Germany
Tel. +49 611/75 2378
christopher.guerke@destatis.de
http://www.forschungsdatenzentrum.de
http://www.destatis.de