Mapping micro data to aggregate data for EVS with DDI 3.0 and SDMX

Rob Grim

Research Data Specialist/e-Science Coordinator

Library & IT-Services

Tilburg University

2nd SDMX-Conference, Perspectives on Metadata

1. Linking library developments to SDMX.
2. Business case for the use of the SDMX framework.
3. European Values Study (EVS) case study: mapping DDI 3.0 to SDMX.
   - EVS.
   - What is DDI 3.0?
   - How to link DDI 3.0 to SDMX?
4. The Dream: use of SDMX in a library environment.
1. The published refereed article is no longer the only or even the prime source for research information (Brown, 2007).

2. Impact of e-Research on statistical information services:

   ✓ More data available.

     - Consequence: improved search mechanisms, that need more structured and granular rich metadata.

   ✓ Data sharing.

   ✓ Research is becoming distributed, collaborative, multidisciplinary.

   ✓ Consequence: researchers want to share data across disciplines. E.g. researchers want to add meso- and macro level economic indicators to microdata and timeseries data.

   ✓ Data lifecycle

     - Relevance to capture (index) data provenance information from statistical data disseminators (See also: ESFRI roadmap).
Four data management tasks in a digital library: store, find and identify, provide access to and deliver quality content.

- Finding and identifying ready-to-use comparative data items is perceived as difficult by researchers, phd’s, students.
- Only sparse metadata are often available for datasets. Only a limited set of metadata fields is often available allowing only coarse grain discovery.

Enhanced Publications: linking publications with datasets:

- Network of European Economists online (NEEO).
  - 16 NEEO-partners in the project.
- DatapluS (SURF-funded project, partners: EVS, DANS, GESIS, CentERData, Twente University).
- Standards used: ORE, MPEG-21 DIDL framework, DDI 3.0 (SDMX).
Business case for the use of SDMX

Policy level

1. The infrastructure for making data available is far less mature than is the case for publishing reports of research (Borgman, 2008).

2. OECD principles and guidelines for access to research data from public fundings (OECD, 2007).

3. Enhance access to publicly funded research results (Janez Potocnik, EU Commissioner, 2008).

Researcher:

✓ Facilitate discovery, identification and delivery of data items and datasets.

Practical:

✓ We are approached by government organisations and research centers, what can SDMX do for us?
The European Values Study (EVS) is a large-scale, cross-national and longitudinal survey research program.

First wave held in 1981, repeated every 9 years, latest wave (2008) is fielded in 45 countries, not only in Europe.

Topics covered: work and leisure time, family and sexuality, religion, politics and ethics, solidarity, social capital, democracy, work ethics.

Data are analyzed at both at the individual (Europeans) and at the aggregate level (Europe at large, European countries and regional clusters).

- EVS research group want to add meso and macro level economic indicators to their datasets.

- Machine actionable data provision?

More information on the European Values Study can be found at: www.europeanvalues.nl
How important is God in your life? (Wave 1999/2000).

<table>
<thead>
<tr>
<th>Country code</th>
<th>how important is God in your life (not at all)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>very</th>
<th>Total</th>
<th>N=</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>30.4</td>
<td>6.1</td>
<td>7.8</td>
<td>5.8</td>
<td>16</td>
<td>7.5</td>
<td>7</td>
<td>7.2</td>
<td>3.9</td>
<td>8.5</td>
<td>100</td>
<td>1580</td>
</tr>
<tr>
<td>Great Britain</td>
<td>21.5</td>
<td>10.1</td>
<td>9.3</td>
<td>7.5</td>
<td>10.1</td>
<td>7.8</td>
<td>8.6</td>
<td>8.5</td>
<td>4.5</td>
<td>14.4</td>
<td>100</td>
<td>960</td>
</tr>
<tr>
<td>Germany</td>
<td>21.9</td>
<td>7.2</td>
<td>8.5</td>
<td>7.3</td>
<td>7.6</td>
<td>8.3</td>
<td>10.9</td>
<td>13.9</td>
<td>8.3</td>
<td>3.3</td>
<td>100</td>
<td>1988</td>
</tr>
<tr>
<td>Austria</td>
<td>8.9</td>
<td>4.2</td>
<td>6.2</td>
<td>2.7</td>
<td>12.7</td>
<td>11.3</td>
<td>7.8</td>
<td>12.1</td>
<td>9.2</td>
<td>24.8</td>
<td>100</td>
<td>1505</td>
</tr>
<tr>
<td>Italy</td>
<td>5.1</td>
<td>1.9</td>
<td>3.5</td>
<td>3.8</td>
<td>7.8</td>
<td>9.4</td>
<td>11</td>
<td>14.9</td>
<td>9.5</td>
<td>33.1</td>
<td>100</td>
<td>1951</td>
</tr>
<tr>
<td>Spain</td>
<td>11.6</td>
<td>6.5</td>
<td>6.7</td>
<td>5.5</td>
<td>12.4</td>
<td>9.7</td>
<td>13.5</td>
<td>9.6</td>
<td>6.4</td>
<td>18.1</td>
<td>100</td>
<td>1176</td>
</tr>
<tr>
<td>Portugal</td>
<td>3.1</td>
<td>0.8</td>
<td>1.6</td>
<td>2.8</td>
<td>7.2</td>
<td>10.2</td>
<td>11.3</td>
<td>12.3</td>
<td>14.6</td>
<td>36.1</td>
<td>100</td>
<td>988</td>
</tr>
<tr>
<td>Netherlands</td>
<td>25.9</td>
<td>4.8</td>
<td>8.6</td>
<td>2.7</td>
<td>9.1</td>
<td>10.2</td>
<td>11.1</td>
<td>9.8</td>
<td>3.2</td>
<td>11.7</td>
<td>100</td>
<td>999</td>
</tr>
<tr>
<td>Belgium</td>
<td>22.6</td>
<td>5.8</td>
<td>6.8</td>
<td>4.6</td>
<td>13.6</td>
<td>9.3</td>
<td>10.1</td>
<td>8.4</td>
<td>3.9</td>
<td>15.7</td>
<td>100</td>
<td>1881</td>
</tr>
<tr>
<td>Denmark</td>
<td>25.5</td>
<td>13.5</td>
<td>12.8</td>
<td>5.9</td>
<td>16.4</td>
<td>5.5</td>
<td>6.3</td>
<td>5.1</td>
<td>2.5</td>
<td>6.6</td>
<td>100</td>
<td>1001</td>
</tr>
<tr>
<td>Sweden</td>
<td>28.9</td>
<td>11.4</td>
<td>11.7</td>
<td>7.7</td>
<td>12.2</td>
<td>6.1</td>
<td>5.6</td>
<td>5.4</td>
<td>2.8</td>
<td>9</td>
<td>100</td>
<td>994</td>
</tr>
<tr>
<td>Finland</td>
<td>13.1</td>
<td>5</td>
<td>6.4</td>
<td>6.9</td>
<td>10</td>
<td>9.4</td>
<td>13.1</td>
<td>13.4</td>
<td>5.6</td>
<td>17</td>
<td>100</td>
<td>989</td>
</tr>
<tr>
<td>Iceland</td>
<td>9</td>
<td>3.3</td>
<td>6.1</td>
<td>7.2</td>
<td>11.6</td>
<td>12</td>
<td>13.6</td>
<td>14.6</td>
<td>7.2</td>
<td>15.4</td>
<td>100</td>
<td>952</td>
</tr>
<tr>
<td>Northern Ireland</td>
<td>5</td>
<td>4</td>
<td>5.4</td>
<td>4.9</td>
<td>8.4</td>
<td>8.8</td>
<td>12.2</td>
<td>12.6</td>
<td>10.1</td>
<td>28.7</td>
<td>100</td>
<td>995</td>
</tr>
<tr>
<td>Ireland</td>
<td>3.9</td>
<td>1.6</td>
<td>3.8</td>
<td>4.8</td>
<td>10.1</td>
<td>8.6</td>
<td>10.9</td>
<td>12.3</td>
<td>7.9</td>
<td>36.2</td>
<td>100</td>
<td>1009</td>
</tr>
<tr>
<td>Estonia</td>
<td>21.8</td>
<td>12.7</td>
<td>14.4</td>
<td>9.8</td>
<td>10.6</td>
<td>8.4</td>
<td>5.8</td>
<td>5.7</td>
<td>3.7</td>
<td>7.1</td>
<td>100</td>
<td>959</td>
</tr>
<tr>
<td>Latvia</td>
<td>14</td>
<td>5</td>
<td>4.8</td>
<td>6.8</td>
<td>6.5</td>
<td>18.6</td>
<td>8.4</td>
<td>8</td>
<td>11.6</td>
<td>5.8</td>
<td>100</td>
<td>968</td>
</tr>
<tr>
<td>Lithuania</td>
<td>10.2</td>
<td>4.3</td>
<td>6.3</td>
<td>2.3</td>
<td>12.7</td>
<td>5.4</td>
<td>6.1</td>
<td>13.1</td>
<td>9.6</td>
<td>29.6</td>
<td>100</td>
<td>927</td>
</tr>
<tr>
<td>Poland</td>
<td>2.1</td>
<td>1.3</td>
<td>1.2</td>
<td>1.8</td>
<td>7</td>
<td>5</td>
<td>8.7</td>
<td>10.9</td>
<td>9.9</td>
<td>52.1</td>
<td>100</td>
<td>1078</td>
</tr>
<tr>
<td>Czech Republic</td>
<td>42.1</td>
<td>10.4</td>
<td>7.2</td>
<td>3.6</td>
<td>11.1</td>
<td>5.4</td>
<td>4.6</td>
<td>4.3</td>
<td>2.8</td>
<td>8.6</td>
<td>100</td>
<td>1846</td>
</tr>
<tr>
<td>Slovakia</td>
<td>12.7</td>
<td>5.7</td>
<td>4.3</td>
<td>2.6</td>
<td>11.5</td>
<td>5.5</td>
<td>7</td>
<td>10.1</td>
<td>10.5</td>
<td>30.1</td>
<td>100</td>
<td>1273</td>
</tr>
<tr>
<td>Hungary</td>
<td>23.1</td>
<td>6.8</td>
<td>6.8</td>
<td>3.5</td>
<td>14.2</td>
<td>6.2</td>
<td>5.5</td>
<td>7.8</td>
<td>5</td>
<td>21</td>
<td>100</td>
<td>982</td>
</tr>
<tr>
<td>Romania</td>
<td>2.7</td>
<td>0.9</td>
<td>1.1</td>
<td>1.5</td>
<td>4.2</td>
<td>3.9</td>
<td>6.1</td>
<td>10.9</td>
<td>12.7</td>
<td>56</td>
<td>100</td>
<td>1124</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>19.7</td>
<td>9.6</td>
<td>8</td>
<td>5.6</td>
<td>11.1</td>
<td>8.8</td>
<td>9.1</td>
<td>9.6</td>
<td>3.6</td>
<td>14.9</td>
<td>100</td>
<td>965</td>
</tr>
<tr>
<td>Croatia</td>
<td>4.6</td>
<td>2.7</td>
<td>5.6</td>
<td>2.4</td>
<td>11.8</td>
<td>6.7</td>
<td>11.4</td>
<td>10.7</td>
<td>10.7</td>
<td>33.5</td>
<td>100</td>
<td>986</td>
</tr>
<tr>
<td>Greece</td>
<td>6.1</td>
<td>3</td>
<td>3.1</td>
<td>4.2</td>
<td>7.3</td>
<td>8.7</td>
<td>10</td>
<td>14.8</td>
<td>12.2</td>
<td>30.6</td>
<td>100</td>
<td>1123</td>
</tr>
<tr>
<td>Russia</td>
<td>20.2</td>
<td>7.5</td>
<td>8.1</td>
<td>4.8</td>
<td>15.8</td>
<td>5.8</td>
<td>7.5</td>
<td>8.3</td>
<td>5.8</td>
<td>16.1</td>
<td>100</td>
<td>2392</td>
</tr>
<tr>
<td>Malta</td>
<td>0.7</td>
<td>0.4</td>
<td>0.2</td>
<td>0.4</td>
<td>2.7</td>
<td>2.7</td>
<td>5.5</td>
<td>9.4</td>
<td>11.2</td>
<td>66.8</td>
<td>100</td>
<td>1000</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>18.5</td>
<td>6.4</td>
<td>6.5</td>
<td>6.2</td>
<td>13.6</td>
<td>8.5</td>
<td>9.7</td>
<td>10.7</td>
<td>4.6</td>
<td>15.2</td>
<td>100</td>
<td>1178</td>
</tr>
<tr>
<td>Slovenia</td>
<td>24.7</td>
<td>4.8</td>
<td>8.4</td>
<td>5</td>
<td>15.6</td>
<td>6.8</td>
<td>7.8</td>
<td>8.5</td>
<td>3.6</td>
<td>14.9</td>
<td>100</td>
<td>980</td>
</tr>
<tr>
<td>Ukraine</td>
<td>13.2</td>
<td>4.8</td>
<td>7.3</td>
<td>4.3</td>
<td>15.1</td>
<td>6.5</td>
<td>6.6</td>
<td>10.5</td>
<td>6.3</td>
<td>25.4</td>
<td>100</td>
<td>1107</td>
</tr>
<tr>
<td>Belarus</td>
<td>9.2</td>
<td>6.5</td>
<td>8.6</td>
<td>7.3</td>
<td>18.2</td>
<td>6.5</td>
<td>6.4</td>
<td>10.9</td>
<td>6</td>
<td>20.3</td>
<td>100</td>
<td>905</td>
</tr>
</tbody>
</table>
What is DDI 3.0?

- Data Documentation Initiative (DDI).
- Documenting microdata.
- Pre DDI 3.0 versions mainly used by data archives, for after-the-fact coding of surveys.
- Core DDI 3.0: data life-cycle approach, modular approach:
  - Capturing metadata during the whole lifecycle of data.
  - Machine actionable (e.g. binding metadata and data during data collection phase).
  - Reuse of modules, components, metadata.
  - Knowledge of statistical data lifecycle (SDMX) was heavily used for the spec.
Ad 3. DDI 3.0 data life-cycle model

The statistical production process → Archiving → (Secondary) use of data

Source: Chris Nelson ©
Identifiable, versionable, maintainable objects

Source: Chris Nelson©
Source: Wendy Thomas
Arofan Gregory
1. **Researcher workflow:** create ‘enriched’ datasets.
   - We want to capture comparable data semantics for a limited set of elements.
     - Dimensions (time-period, geographical information).
     - Universe information (e.g. persons aged 16-64).

2. **Library system:** content aggregation, information discovery:
   - Mark up all tables in the publication in DDI 3.0 (NCubes): dimensions, concepts, attributes.
   - Identify SDMX metadata and dataflows for these elements.
The Dream: SDMX as a local service to researchers

Collaborative approach:

✓ Implement and populate a global SDMX metadata registry for the SDMX disseminators with detailed information on statistical data items for the academic community (which indicators are available, data provenance).

Local implementation of SDMX service:

✓ Pull SDMX metadata and data from SDMX implementations.

✓ Use facet indexing of DSD-dimensions and metadata artefacts in our local search engine.
  ▪ Facets: UNECE statistical subject-matter domain list, community - and subject based portlets.

✓ Create ORE-resource maps with comparable micro macro and (time series) data items.

✓ Contribute to domain specific DSD and MSD definitions(?).
<table>
<thead>
<tr>
<th>Title</th>
<th>Year</th>
<th>Collection</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media coverage of political scandals</td>
<td>2008</td>
<td>Tilburg University (37)</td>
<td>workingPaper</td>
</tr>
<tr>
<td>On the Sum of Laplacian Eigenvalues of Graphs</td>
<td>2008</td>
<td>Maastricht University (126)</td>
<td>workingPaper</td>
</tr>
<tr>
<td>Interval Game Theoretic Division Rules</td>
<td>2008</td>
<td>ULB Brussels (11)</td>
<td>workingPaper</td>
</tr>
<tr>
<td>On Semidefinite Programming Relaxations of the Traveling Salesman Problem</td>
<td>2008</td>
<td></td>
<td>workingPaper</td>
</tr>
</tbody>
</table>
Questions?

m: rob.grim@uvt.nl