A User-Driven and Flexible Procedure for Data Linking

Cees van der Eijk and Eliyahu V. Sapir

2010 IASSIST conference
Social Data and Social Networking: Connecting Social Science Communities Across the Globe.
June 3, 2010
Ithaca, NY
<table>
<thead>
<tr>
<th>ACE Electoral Knowledge Network</th>
<th>Constituency-Level Elections Dataset (CLE)</th>
<th>International Social Survey Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afrobarometer</td>
<td>Eurobarometer</td>
<td>Latin American Public Opinion Project</td>
</tr>
<tr>
<td>Arab Barometer</td>
<td>European Election Studies</td>
<td>Latinobarometro</td>
</tr>
<tr>
<td>AsiaBarometer</td>
<td>European Social Survey</td>
<td>New Europe Barometer</td>
</tr>
<tr>
<td>Asian Barometer</td>
<td>European Values Study</td>
<td>Survey of Health, Ageing and Retirement in Europe (SHARE)</td>
</tr>
<tr>
<td>Comparative Candidates Survey (CCS)</td>
<td>European Voter Database</td>
<td>Transparency International Global Corruption Barometer</td>
</tr>
<tr>
<td>Constituency-Level Elections Archive (CLEA)</td>
<td>Global Barometer Surveys</td>
<td>World Values Survey (WVS)</td>
</tr>
</tbody>
</table>
The embarrassment of the riches...

- Many separate datasets can be linked to allow addressing important research questions.
- Joint information from different components will provide richer base for analyses than the sum of its separate parts.

Yet – there are very few efforts to integrate different datasets, for various reasons, most importantly:
  * Lack of harmonization
  * Limitations of StatPacks and statistical software
  * Lacunae in Social Science research training

Our mission: working around these data linking obstacles, in the context of PIREDEU
Providing an Infrastructure for Research on Electoral Democracy in the European Union (PIREDEU)

- Pan-European 3 years design study to assess the feasibility of upgrading the European Election Studies and provide an infrastructure for research into citizenship, political participation, and electoral democracy in the EU.
- Involves more than 30 participating researchers from some 19 institutions in western and eastern Europe and the US, and a large community of affiliated experts from all EU member states.
- The PIREDEU infrastructure will provide a comprehensive empirical database regarding voters, candidates for election to the European Parliament, media coverage of the EP elections, party manifestos issued in connection with these elections, and contextual data relevant to these elections.
Providing an Infrastructure for Research on Electoral Democracy in the European Union (PIREDEU)

The precise scientific objectives of PIREDEU are organised around five data collection activities:

- surveys of citizens (voter survey/VS),
- surveys of elites (candidate survey/CS),
- contents of party manifestos (manifesto study/MaS),
- content analyses of news items from national newspapers and TV broadcasts (MS),
- macro data of economic, social, cultural, and political nature.

These datasets provide complementary information on the complex and multifaceted reality of electoral democracy in Europe
Linking the PIREDEU Study Components

Voter Study (VS)
- Country ID
- Respondent ID
- Party preference 1..n
- Watching TV channels 1..m
- Reading news outlets 1..l
- OTHER VARIABLES

Party Manifestos (MaS)
- Country ID
- Party ID
- OTHER VARIABLES

Candidate Study (CS)
- Country ID
- Candidate ID
- Party ID
- OTHER VARIABLES

Media Study (MS)
- Country ID
- News outlet ID
- OTHER VARIABLES

Contextual Data
- Country Context
  - Country ID
  - OTHER VARIABLES
- Party Context
  - Party ID
  - OTHER VARIABLES
Example of data linking required to support substantive research: how salient are various issues for candidates?

**Issue salience could be a function of**

- The distance between the candidate’s personal views on issue j and those of his/her party as expressed in the manifesto
- The salience of the issue for various media outlets, moderated by the extent to which potential voters for the candidate’s party are exposed to those media outlets
- The salience of the issue for voters, moderated by their propensity to vote for the party in question
- All these relations are potentially moderated by country-contextual factors, such as the location of the EP election in the domestic electoral cycle, so that the relationships referred to are more important the later the EP election is in the domestic electoral cycle
Example of data linking required to support substantive research: how salient various issues are for candidates?

Productive linking of the different data components must take into account:

- facilitating linking has to take the form of providing tools to accomplish the task, and not providing a linked dataset
- the linking process must result in a flat, rectangular data matrix
- relevant tools must disaggregate the linking process into successive tasks of limited complexity and clarify for the user the options available at each

The tool we are developing is a user-driven interface structured according to a flowchart that is relevant for the particular kind of merging to be accomplished.
# Example: Linking the Voter Study (VS) and the Party Manifesto Study (MaS)

<table>
<thead>
<tr>
<th>Units of Measurement</th>
<th>Voter Study (VS)</th>
<th>Party Manifesto Study (MaS)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>27,069 Individuals, representing the 2009 electorates in the 27 EU member states</td>
<td>189 National Parties + 7 EU party federations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of Data</th>
<th>Voter Study (VS)</th>
<th>Party Manifesto Study (MaS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Over 200 variables measuring: 1) electoral behaviour, 2) Sundry political attitudes, 3) background/predisposing characteristics</td>
<td>69 content categories Measuring proportion of statements about policy domains</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Themes (not exhaustive)</th>
<th>Voter Study (VS)</th>
<th>Party Manifesto Study (MaS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electoral participation</td>
<td></td>
<td>external relations,</td>
</tr>
<tr>
<td>Electoral choice (past, current, prospective, for national and European Parliaments)</td>
<td></td>
<td>freedom and democracy,</td>
</tr>
<tr>
<td>propensity to support particular parties</td>
<td></td>
<td>political system,</td>
</tr>
<tr>
<td>interest in politics, campaign,</td>
<td></td>
<td>economy, welfare</td>
</tr>
<tr>
<td>perceived most important problems,</td>
<td></td>
<td>quality of life, fabric of society and social groups.</td>
</tr>
<tr>
<td>attitudes regarding EU, and a variety of other issues,</td>
<td></td>
<td>environment</td>
</tr>
<tr>
<td>L/R self placement, L/R perception of parties</td>
<td></td>
<td></td>
</tr>
<tr>
<td>media usage,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demographics</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example: Linking the Voter Study (VS) and the Party Manifesto Study (MaS)

The merged dataset should be a ‘flat’ data matrix, which can be analyzed using StatPacks.

THEREFORE

This merging can take one of two forms, resulting in two different integrated datasets:

Merging MaS data into VS

OR

Merging VS data into MaS
Linking data from more than two sources
Parallel merging

using the same source as primary dataset vis-a-vis all other datasets as successive secondary datasets

Example – MaS is the primary dataset, into which information is merged in successive rounds

\[
[\text{MaS} \& \text{VS}] + [\text{MaS} \& \text{CS}] + [\text{MaS} \& \text{MS}]
\]
Linking data from more than two sources
Sequential merging

in a first round VS data are merged into the MaS, while in the second MaS information (including the data that have their origin in the VS) are merged into the CS

\[\text{CS} \& \lbrack\text{MaS} \& \text{VS}\rbrack\]

the identity of the primary dataset changes between successive rounds of merging. In these situations, the order of the successive rounds of merging is essential, because

\[\lbrack\text{MaS} \& \text{VS}\rbrack \neq \lbrack\text{VS} \& \text{MaS}\rbrack\]
THANK YOU

PIREDEU preliminary-release data is available at the project website: http://www.piredeu.eu/

cees.vandereijk@nottingham.ac.uk

eliyahu.sapir@nottingham.ac.uk