A user-driven and flexible procedure for data linking

by Cees van der Eijk and Eliyahu V. Sapir[1]

PIREDEU

Abstract

Over the past decades, social scientists enjoyed a rapid increase in the availability of various types of data. Many of these pertain to different aspects of the same multifaceted phenomenon. In the field of electoral studies, for example, there are data about citizens, political elites, party manifestos, media and the context within which elections take place. Researching such multifaceted phenomena requires this diverse information to be analysed jointly, rather than separately. That, in turn, requires the linking of separate datasets (also known as data fusion, or conflation), which is largely a relational database (RDB) management problem. In spite of its large and rewarding potential for empirical research, data-linking is practiced relatively little in the social sciences. This is largely caused by lack of relevant training amongst social scientists to cope with the methodological and technical difficulties involved in data linking. This article presents an approach for facilitating data linking without requiring additional training of researchers. It defines necessary RDB operations as a structured series of user-choices, to be included in a user interface that generates and implements the RDB operations once all choices have been made. The advantage of this approach over the public dissemination of integrated datasets constructed by 'experts' is that it does not assume that 'one-size-fits-all'; it is flexible and tailored to the needs of end-users. This approach can be applied in a wide variety of contexts. An implementation is under development for the PIREDEU project in the field of comparative electoral research. .

Keywords: data linking, data integration, relational database, PIREDEU, electoral research.

Introduction: An embarrassment of riches

Compared to only a few decades ago, comparative social researchers now enjoy the availability of a wealth of datasets. If they are interested in behaviours, attitudes and orientations of citizens, they will find that in addition to sundry ad hoc surveys, many regularly conducted national election studies, general social surveys, household, labour-market, crime, and other surveys are available in an increasing number of countries. Moreover, they will find an ever growing number of explicitly comparative surveys that are repeatedly conducted in multiple countries, thus enabling comparisons across national contexts as well as over time. These include, amongst many others[2] , the Comparative Study of Electoral Systems (CSES, covering up to 38 countries), the World Value Studies (WVS, up to 87 countries) and European Value Studies (EVS, up to 45 countries), the International Studies of Political Psychology (ISPP, up to 45 countries), the European Social Surveys (ESS, up to 31 countries), and the European Election Studies (EES, up to 27 countries).

> the promise that these data hold is much greater if they can be linked to each other

This wealth of empirical material is not restricted to any particular social discipline, nor is it only made up of mass surveys. For reasons that will become apparent later in this paper, we are particularly interested in studies relating to elections, parties and public opinion, but our observations of that domain can easily be generalised to other large areas of social scientific inquiry. When reviewing our own field of interest, we find, increasingly, that in addition to the many available mass surveys, political and social elites of various kinds are surveyed as well, in single countries or comparatively across a number of countries. Beyond the domain of surveys, we find data derived from party manifestos covering almost all parties that ever competed in democratic general elections after World War II. These have more recently been complemented by the Euro-manifesto program that codes the contents of the manifestos of all parties that have ever competed in direct elections of the European Parliament (EP). Content analyses are also increasingly used to generate systematic data about media communications, and include projects

that yield data comparable over time and across countries. Other extensive data sets have become available for yet other organisations and institutions, such as social movements and pressure groups. At the level of states, there is also an abundance of data pertaining to sundry economic indicators, political and social indicators, formal institutional arrangements, government performance, and so forth.

Each of these data collections by itself provides rich possibilities for empirical research, and indeed we see increasing numbers of publications making use of this potential. Yet the promise that these data hold is much greater if they can be linked to each other. In recent years, many of the principal investigators of these large data collection efforts have referred to this larger potential as part of the justification for investing in these costly enterprises.

Indeed, many of the most interesting questions in social research do not pertain only to citizens, or only to elites, or only to media, and so forth. Rather, they have to do with the interactions between various kinds of actors, organisations and institutions, which are affected by the characteristics of different contexts, or with the social, political and economic consequences of these contextualised interactions. As a case in point, questions about the quality and functioning of representative democracy pertain simultaneously to citizens, political parties, political elites, and mass media, among other things. Seemingly simple concepts such as accountability and representation relate to the interaction between citizens and (political) elites, as well as various types of processes that involve the media (for example, the effects on citizens and elites of agenda-setting, framing, priming, and spin and hype). To the extent that the functioning of representative democracy is affected by economic developments, all of these interactions and relations have to be contextualised in economic terms. From a dynamic perspective this leads to questions as to whether 'the economy' is an autonomous factor affecting the behaviours and interactions of the various actors in democratic processes, or whether it is endogenous and the consequence of these behaviours and interactions.

Important questions that require empirical information from a variety of different actors, groups, organisations, institutions and contexts exist in all social sciences. They may focus on, for example, social integration, crime, traffic and mobility, or the efficiency of markets, but they all have in common that they cannot be adequately addressed using information pertaining to only one of the interacting actors and institutions.

Our ability to address important multi-facetted questions has not only been increased by the availability of abundant relevant data, but also by advances in multivariate analytical methods, software and affordable computing power. Complex models that until recently were beyond the computing infrastructure available to most researchers can currently be estimated on standard personal computers using generally available software. Of particular importance for empirical social research are the advances in ordinal- and nominal-level multivariate analysis, latent structure modelling, structural equation modelling, dynamic modelling and multi-level methods.

In view of the wealth of relevant data and the availability of tools to analyse them, one might expect current social science literature to abound with publications that join together information from multiple data sources in order to more effectively address the important and broad-ranging questions referred to previously. Yet, such publications are fewer than one would expect,[3] which creates a somewhat puzzling and embarrassing situation.

Diagnosis

In principle, many separate datasets can be linked in ways that would allow important research questions to be addressed in more powerful ways than analysing each of these resources separately. If these separate datasets relate (by way of their units or their variables) to the same real-world objects, such as countries, political parties, media outlets, and the like, they can be seen as component parts of relational databases (RDBs). The methodology of RDBs is well developed and provides a multitude of ways to generate joint information from different components that provides a richer base for analyses than the sum of the separate parts. Moreover, RDB software is widely available. Why, then, do we see so very few efforts of integrating or linking different datasets? A number of factors contribute to this state of affairs, including the following (without claiming to be exhaustive):

a) **Lack of harmonization**. Linking of separate datasets in a RDB requires the same objects being identified in the same way in each of them. Frequently this is not the case. As a case in point, identification codes of political parties differ more often than not between successive editions of a series of national election studies, each of which pertains to a different election. Even data infrastructures that pride themselves on their over-time comparability of coding often fall short in this respect.[4] This problem is not limited to the identification of parties, but also to countries, regions, media outlets, and so on.[5] As a consequence, any linking has to be preceded by a complex and costly data harmonization stage. Without dedicated resources it is impossible for most researchers to undertake such projects, and funding agencies see little glory in providing grants to produce such 'continuity' datasets. The few that do exist are not extended when new studies are released and become outdated, thus losing their relevance. Analysts that do aspire to the simultaneous use of data from different sources thus have to make their own tailored 'solution', which is often too narrowly focused to suit the needs of others. It is therefore not surprising that, when faced with such obstacles and with publication requirements from their own universities, many opt for the short-term option to analyse a single dataset and forego the potential riches that could be gained in the long term from data-linking.

b) **Limitations of 'statpacks' and other statistical software.** Much of the statistical software used in the social sciences is bundled in so-called statpacks such as SPSS, SAS, STATA and R. These packages contain a wealth of statistical procedures, as well as extensive procedures for data management, such as recoding and creation of new variables. Yet, they are fundamentally geared towards the analysis of 'flat' rectangular data matrices, and their capabilities for managing RDB information range from non-existent to extremely limited and cumbersome. As a consequence, after separate databases have been harmonised and linked in a RDB structure, dedicated RDB management tools must be used to generate the (rectangular) data matrices that lend themselves to analysis with the analytical software social researchers have been trained to use. This not only requires additional work, but requires working with software that is often unfamiliar to many researchers in the social sciences.

c) **Lacunae in social science research training**. Research training in the empirical social sciences traditionally focuses on questions of general research design, data collection methods, and multivariate statistical data analysis with statpacks and similar software. Many researchers, therefore, are well-versed in advanced multivariate modelling procedures, yet woefully untrained in recognising the potential benefits of RDB in managing data productively. When confronted with multiple datasets, it seems that many otherwise excellent researchers see only a collection of separate datasets, each of which can be analysed in sophisticated ways, but individually. What they often do not see are the ways in which separate datasets

can be linked in a RDB, which, in turn, can be used to generate new rectangular data matrices that provide better platforms for addressing the substantive research questions they wish to pursue.

None of these three obstacles to linking and merging data from different sources is insurmountable, yet they are not easily overcome as they are rooted in entrenched traditions of training and acquired routines. Unleashing the full potential of linkable data thus requires more than just pointing out the benefits to be gained. It also requires infrastructure, in the form of software that offsets the lack of RDB familiarity amongst social scientists and that can be used without extensive further training.

This paper describes an attempt to work around the obstacles that often prevent analysts from linking data from various sources. Although this attempt is, in principle, not limited to any particular kind of research problem, or to any particular collection of datasets, we nevertheless present it in the context of its development, the PIREDEU program. Moreover, as our work is still ongoing, our presentation in this paper relates to 'work in progress', with some parts having been developed already, and others still under development (see also van der Eijk and Sapir 2010).

PIREDEU

PIREDEU is a program of research funded by the European Union (EU) under the Seventh Framework Programme from 2008 to 2011. This three-year design study assesses the feasibility of upgrading the existing European Election Studies to a research infrastructure for studies into citizenship, political participation, and electoral democracy in the European Union. The scientific and technical feasibility of this infrastructure is elaborated by means of a pilot study conducted in the context of the 2009 elections to the European Parliament.[6] In contrast to many other comparative research programs about elections that only investigate voters, or only parties, PIREDEU considers its subject matter, electoral democracy, as a complex set of interactions between voters, parties, candidates, media and relevant institutions (such as election rules). It therefore collects empirical information concerning different units and it does so for each one in the most appropriate manner. The various data components of PIREDEU are:

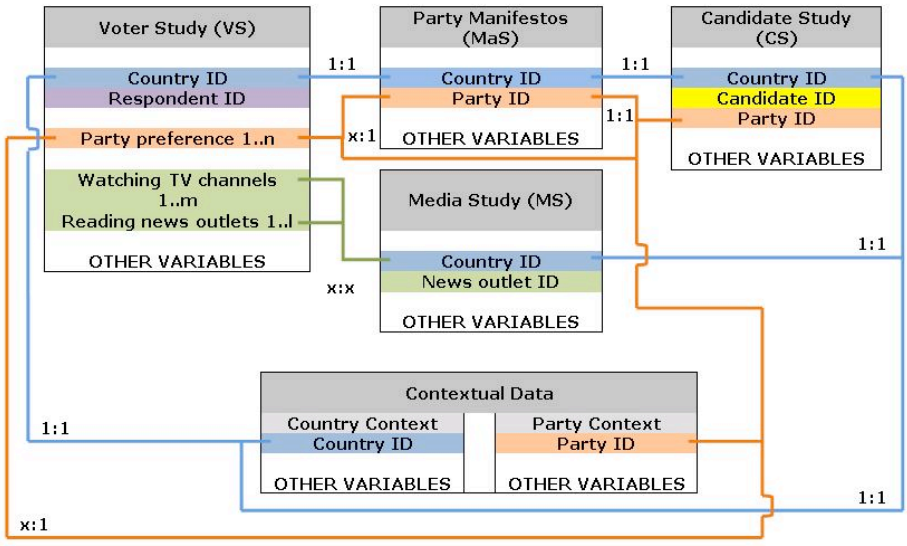

Figure 1. The PIREDEU Relational Data Base

- A **voter study** that conducts surveys of representative samples of approximately 1,000 respondents each from the electorates of the 27 member states of the EU.[7] Apart from translation and reference to country-specific institutions, the questionnaires were the same for all countries. The questionnaires covered three main themes: 1) electoral behaviour and party preferences; 2) political attitudes and orientations; and 3) background characteristics and media usage.
- A **candidate study** that conducts surveys of the candidates listed on the ballots of the European Parliament elections of 2009 in each of the 27 member states of the EU.[8] With the same provisos as mentioned for the voter survey, the questionnaires were identical for the candidates from each country and from each party.
- A **manifesto study** that consists of coding the content of the election manifestos of the political parties vying for votes in the 2009 European Parliament elections.[9] The coding units are sentences or quasi-sentences, each of which is classified in one of many content categories relating to a wide range of policy domains: external relations, freedom and democracy, political system, economy, welfare and quality of life, fabric of society and social groups, and European integration (cf. Wüst and Volkens 2003). After aggregation to the level of parties, this provides data on the relative emphasis that parties place on the substantive topics reflected in the coding categories.
- A **media study** that consists of coding the content of media items during the three weeks leading up to the European Parliament election.[10] All news items were coded for the most important TV news programs and the most important newspapers in each of the 27 EU member states. Each news item was coded on a large range of characteristics, including topics, actors displayed, use of frames, and physical characteristics (length, placement, embellishments, etc.).
- **A contextual information study** that brings together information at the level of the 27 member states of the EU, regarding the results of the European Parliament election and other recent elections, electoral procedures, voting rules and other relevant institutions, the incumbent government, and economic conditions.[11]

In line with PIREDEU's perspective on electoral democracy, these various datasets are seen as providing complementary information about a complex and multifaceted reality. And although it is perfectly feasible to analyse each of the datasets in isolation, one of the main tasks of the program is to link or integrate them in user-friendly ways, thereby promoting a fuller utilization of the joint potential of the data, and, ultimately, better research on electoral democracy in Europe.

In order to achieve this objective, great care was given to assuring that the variables by which the data from the different components can be linked – the 'keys' in RDB terminology – were coded in exactly the same way in each component. This relates in particular to the identification of countries, political parties and media outlets. All components are structured by country and all relate to political parties. The voter study includes questions about party choice, party preference and perception of parties; the candidate study explores the relationship between candidates and the party for which they are listed on the ballot, and other questions about their own and other parties; the

manifesto study explores the relationship between each manifesto and its author (i.e., party); the media study examines coded news items in terms of parties being mentioned and evaluated; and, lastly, the contextual data include the identification of the party/parties that form the incumbent government, and the results of various elections in each of the countries. The identity of media outlets is crucial in the media study to identify the outlet from which each coded item was taken, and in the voter study to identify media outlets that respondents use for their information. These keys define the primary relationships in the PIREDEU RDB, as illustrated in Figure 1.

The solution that we chose for allowing data from the various components to be linked in a rectangular data matrix that can be analysed with the statistical software mostly used in social science research will be described below, but to clarify the task at hand we first present an example of a substantive research question that requires information from all components to be merged into a single analysable file.

A substantive example of the need for linking data from PIREDEU components

Suppose we are interested in the orientations and behaviour of candidates, and, more particularly, in how salient various issues are for candidates. This information is specific for the candidates, and available from the candidate survey.

Were we to analyse issue salience for candidates only from the data in the candidate survey, we would model the variance in the dependent variable (salience of issues for candidates) in a multi-level model with candidates nested within parties, which, in turn, are nested within countries. As independent variables at the candidate level we can use all other variables collected in the candidate survey, including various political orientations and background characteristics, and the identity of the party and the country of each candidate. In the absence of further information about the traits of these parties and countries, their impact would be modelled in a random effects specification.

Such a random effects model would be unsatisfactory because it would only tell us that some of the variance in issue salience at the level of individual candidates can be attributed to parties and to countries, but we would be in the dark as to the form of this relationship. That is, which kinds of parties and which kinds of countries have a positive or negative impact on issue salience?

The model could be made more informative by adding information about parties and about countries, thus allowing a mixed model specification in which the explicated characteristics of parties and countries are modelled as fixed effects and the remaining variance at that level as random effects. This additional information can be derived from the other data components of PIREDEU.[12] In the absence of any infrastructure or specific tools, such information would have to be added manually by the analyst. From the manifesto study one could, for example, derive how much emphasis parties place on each of the issues in their manifestos. This information would be added to the candidate dataset via a tedious procedure involving a large number of conditional statements. With some 200 political parties this procedure would be prone to error, and would take several hours to accomplish even for an experienced data analyst. In a similar way, one can add country information to the candidate file (which would be somewhat less onerous as there are only 27 countries). Additional information can be added that originates from the voter study or from the media study. The work involved becomes increasingly more complex if the theories that we want to test involve a wider set of relationships between candidates, parties, voters, media and contexts. Consider the following elaboration, which, although used here for its illustrative value, would substantively be neither unrealistic nor excessively complex.

The dependent variable – the salience of various issues for candidates – can be seen as a function of:

- The difference between the candidates' personal views on issues and those of his/her party as expressed in its manifesto. This expectation reflects partly the tendency to reduce cognitive dissonance and partly the political expediency of downplaying differences in views between oneself and one's party. To test this, it would be necessary to add information from the manifesto data to the candidate dataset.
- The salience of the various issues for various media outlets, moderated by the extent to which potential voters for the candidate's party are exposed to those media outlets. This expectation could be based on the notion that, politically, candidates cannot afford to ignore issues that are played up in the media, particularly if their own potential voters are exposed to the contents of those media. To test this, one has first to arrive at a measure of issue salience for media outlets, which has to be derived by some form of aggregation from the news items that have been coded. This outlet-issue-salience then has to be added to the candidate survey using country as key, as candidates are not directly linked to media outlets. Subsequently, the voter survey has to be used to distinguish potential voters for each of the parties, and then to determine for each of these groups the extent to which they are exposed to media outlets. This information has to be added to the candidate dataset using media outlet and party as keys.
- The salience of various issues for voters, moderated by their propensity to vote for the party of the candidate in question. This would be based on the expectation of candidates being responsive to voters in general, and in particular to voters who are likely to vote for their party. To implement this, one would have to use the voter study to distinguish voters according to their propensity to vote for each of the parties, and then to assess how salient the various issues are to each of these groups. This aggregated information would then have to be added to the candidate survey using party and country as keys.
- The effects of the factors listed in the previous bullet points are potentially moderated by country-contextual factors, such as the (temporal) location of the EP election in the domestic electoral cycle (for theoretical foundations of this expectation see van der Eijk and Franklin, 1996).This would require retrieving the relevant information from the contextual information dataset, and then adding said information to the candidate dataset using country as key

After performing all these data operations, the candidate dataset, extended with information from the datasets pertaining to manifestos, voters, media and contexts, can be used to perform the desired multi-level mixed effects regression with cross-level interactions.
Again, without specific tools or infrastructure to help accomplish these tasks, the required data management would easily take days of error-prone and tedious work. Moreover, all investments in that work would only be relevant for this particular question, and similar, but substantively different operations, would have to be performed for other substantive questions.

Managing the linking problem

In our view, any attempt to facilitate productive linking of data from different sources (or in our particular case, from different components of the PIREDEU program) has to recognise the following considerations and constraints:

- In terms of the outcome of the linking process – a data matrix that can be analysed by the kind of statistical software used by social scientists – there is no single or one-size-fits-all solution. What has

to be linked, and how exactly, is different for various substantive research questions. Any attempt to impose a single solution would be futile, as researchers will not use it if it does not fit their aims and theoretical and conceptual perspectives.[13]

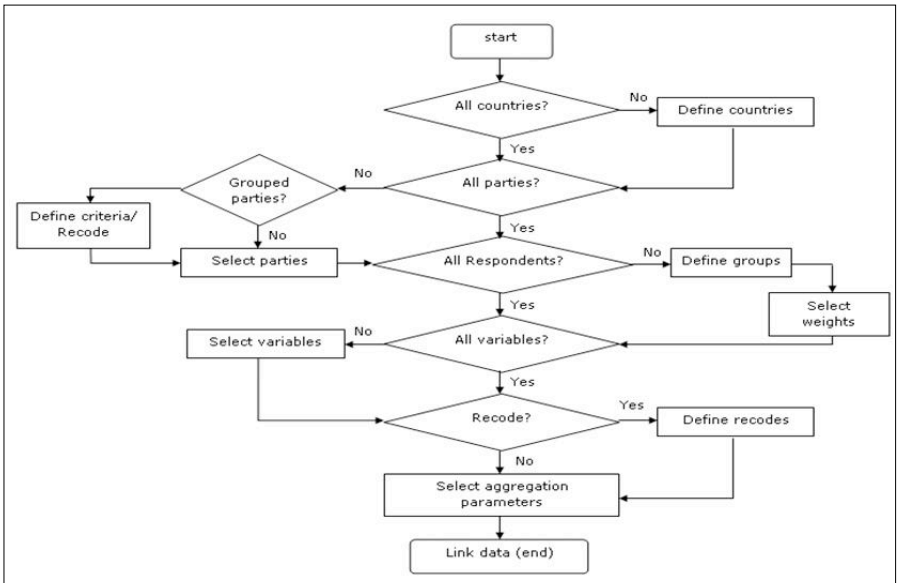

Figure 2. Flowchart of user decisions involved in merging voter study data into manifesto study data — [MaS & ◄ VS]

As a consequence, **facilitating linking has to take the form of providing tools to accomplish the task, and not providing a linked dataset**. This holds equally for linking data pertaining to different observation units, as for linking data pertaining to the same kind of observation units (e.g., repeated studies of the same kind).[14]

- The outcome of the linking process must be a datafile that lends itself to analysis with the statpacks and other statistical software that is ubiquitous in social science research. As such statistical software is generally not capable of handling RDB structures, **the linking process must result in a flat, rectangular data matrix**.
- Many excellent empirical social science data analysts have not been trained in RDB management. Therefore, tools to facilitate data linking should not assume such familiarity and must provide, in a structured manner, the kinds of options available. As a consequence, **relevant tools must disaggregate the linking process into successive tasks of limited complexity and clarify the options available at each for the user.**

In accordance with these considerations, we chose to develop tools for linking data across the various PIREDEU data components in the form of a structured user interface that guides the user through a set of choices. The entire sequence of choices generates a syntax that specifies the required RDB operations and, when implemented, yields as an outcome the desired dataset with linked data. Actually, in view of the software considerations referred to above, it produces a tailored dataset into which information from other datasets is merged.

We will illustrate this approach by focusing on only two of the PIREDEU data components, the voter study (VS) and the manifesto study (MaS) (see Figure 1). Linking other combinations of data components operates along the same lines and will not be elaborated in this paper.[15]

Linking and merging the Voter Study (VS) and Manifesto Study (MaS)

The units in the VS are individual respondents. The primary key in the VS is the respondent-ID. The units in the MaS are political parties, and the primary key in the MaS is the party-ID. The relationship between these two datasets is defined by a number of foreign keys in the VS that relate to survey questions about parties (each of these foreign keys is coded in the same way as the primary key in the MaS). These questions are about different matters, including actual party choice made in particular elections, attractiveness of parties as options to choose in a particular election, generalised affect, and perceptions of parties, among other things. What they all have in common is that their possible responses are defined in terms of parties. These foreign keys can be used for merging information from both datasets. In view of the considerations discussed in the previous section, this merging has to result in a 'flat' data matrix that can be analysed by statistical software packages. This merging can therefore take two different forms, resulting in two different linked datasets: (1) a VS dataset (units are individual respondents) with information from the MaS merged into it, or (2) a MaS dataset (units are political parties) into which information from the VS is merged. Owing to the difference in the character of their units, these two forms cater to different research questions, and thus to different groups of researchers. Moreover, because the keyed link between the VS and the MaS is of a one-to-many type, the actual merging process is somewhat different when going from VS to MaS than the other way around.

Merging MaS data into the VS. The MaS offers information about the political parties that respondents mention in their answers to survey questions. As this information is not respondent-specific, it is identical for all respondents who refer to the same party in response to a question. Thus, for this information, one can regard the respondents as being nested, so to speak, in the parties they mention in their responses.

Merging is in this case a simple operation, as each mention of a party by a respondent relates to only a single case in the MaS data. When new variables are added to the VS, they contain the desired information from the MaS, linked by the correspondence between the chosen foreign key in the VS and the primary key in the MaS.

Merging VS data into the MaS. This kind of merging provides information about the composition of the groups of respondents who relate to the various parties in terms of choice, affect, particular perceptions, and so forth. This linked information may include anything available in the VS, such as respondents' views on political issues, their social characteristics, media usage, political behaviour, and so on.

Merging in this case is somewhat more complex than in the previous case, as the relationship between each party and respondents generally involves multiple respondents. The variables to be added to the MaS file, therefore, have to contain summarizing information about the relevant group of cases in the VS. The analyst has to decide which of various possibilities is most desirable. Obviously, this is partly dependent on the measurement level of the relevant information in the VS. Means and variances would be relevant for interval level

variables,[16] but that level of information is rarely available in survey data. For ordinal level data the ordinal equivalents of these summarizing parameters are available. At nominal level, summarisation is limited to proportions in all (or only in some) of the categories. The user interface for linking thus contains a set of choices for the analyst that specify the desired mode (or modes) of summarising VS data to be merged into the MaS.

Flow of end-user decisions. In order for the user linking interface to generate the desired tailored dataset, the analyst will be guided through a series of structured choices which are reflected in the flowchart in Figure 2. As a very first step, the user has to decide on the units that are to populate the desired merged file. In the example presented here, where we consider only the VS and the MaS, the question is whether we want to obtain a file of voters with merged data from party manifestos, or, alternatively, a file of parties and their manifesto data, into which voter information is merged. Figure 2 has been specified for the situation where the integrated file has parties as units; obviously a very similar, yet in detail, somewhat different flowchart would specify the decisions to be taken for the choice of individual respondents as units in the integrated file. Once the choice of units has been made, one should decide on the number of countries one would like to include in the integrated file. This number could range from 1 (i.e., a single country database as the final product) to 27 (i.e., all member states included). Next, the user should determine which parties will be included in the integrated file. This number is in the range of 1 (a single party dataset) to all parties participating in the EP elections' (over 200). The next step is to define which respondents to include in the information to be merged (all respondents can be used to provide the information to be merged, or a selection that can be specified in terms of group criteria and weights). Then, the user should determine whether or not the variables to be merged should be recoded and, if so, how. Once these steps are completed, the user needs to define the aggregation parameters he/she would like to use in aggregating the VS data into the MaS data. In the user interface, the possible choices will be presented in the form of drop down lists or of user-specifiable values.

Linking and merging data between more than two sources

In the previous section we described the linking between two sources of different data, one with survey respondents as units, the other with party manifestos as units. As illustrated in Figure 1, however, the PIREDEU data consist of five different components. Between each pair of these, the linking and merging of information proceeds along the logic described in the previous section, possibly with minor modifications necessitated by unique characteristics of each of these five components. When merging data from more than two sources we can distinguish two possibilities, one of which presents its own challenges. We use the notation [Primary (or recipient) & Secondary (or donor)] to denote the merging of information from the secondary into the primary dataset.

The easiest way to merge data between more than two sources consists of **parallel merging**, which is using the same source as primary dataset vis-a-vis all other ones as successive secondary datasets. In the case of the PIREDEU data, this could, for example, involve the MaS as primary data, into which information is merged, in successive rounds, from other sources such as the VS, the CS and the MS. In other words:

[MaS & VS] + [MaS & CS] + [MaS & MS].

The merging operation for each of these successive rounds is similar, and proceeds along the lines described in the previous section. In which order the successive rounds of merging are executed is immaterial for the final result. This example will result in a MaS with a great number of additional variables which contain information from the other data components.

A more complex way consists of **sequential merging,** where, for example, 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:[CS & [MaS & VS]].

This implies that 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, as discussed earlier:

$[MaS \& VS] \neq [VS \& MaS]$.

The design of the user interface for multiple parallel merging is not intrinsically more complex than it is for single merging operations. However, in the case of the interface for multiple sequential merging, it is more difficult. The additional difficulty is not technical, but didactic in nature: the user interface is intended to allow users who are not used to RDB management to merge data from different components of a RDB by guiding them through a series of structured questions, the answers to which generate the syntax of the required RDB operations in the background. The challenge will thus be to implement possibilities for sequential merging, while keeping the interface simple and comprehensible

Concluding remarks

We believe that our approach to producing 'integrated' datasets has a great advantage over other approaches. It does not invest in the production of a specific end-product, but rather in the tools to be used that will allow such a product to be achieved. The implication is that, once the tools are produced, different 'integrated' datasets can easily be produced from the same original empirical material. Or, stated differently, it does not produce a straightjacket to which researchers have to adapt themselves, but rather it allows the production of datasets that are tailored to the specific needs of individual researchers.

This approach to linking and merging data is in principle also applicable for other data than those collected in the context of the PIREDEU program. As a case in point, many studies that are conducted repeatedly – such as national election studies – strive to unleash the potential for longitudinal comparison by making available longitudinally integrated datafiles, often referred to as so-called continuity studies. But the construction of such datasets is never straightforward, as they invariably require many decisions being made to cope with unavoidable differences in operationalisations, coding schemes, and the like. Irrespective of what decisions are made, they can never be optimal for all the different research projects that would require such over-time comparable data. In these contexts too, it may be advantageous not to invest in the production of datafiles that will not be well suited for at least some researchers, but rather in a flexible interface that allows analysts to tailor the longitudinal data integration to the specific needs of their research.

References

Blumler, Jay G. 1983. Communicating to voters: Television in the first European Parliament elections. London: Sage.

Braun, Daniela, Slava Mikhaylov and Hermann Schmitt. 2010. EES (2009) Manifesto Study Documentation Advance Release. Mannheim: MZES. [http://www.piredeu.eu/].

van der Brug, Wouter, and Cees van der Eijk, eds. 2007. European elections and domestic politics: Lessons from the past and scenarios for the future. Notre Dame, Ind.: University of Notre Dame Press.

Czesnik, Mikolaj, Michal Kotnarowski and Radoslaw Markowski. 2010. EES (2009) Contextual dataset Codebook, Advance Release. Warsaw: SWPS. [http://www.piredeu.eu/].

Connolly, William E. 1999. The terms of political discourse. Princeton: Princeton University Press.

EES. 2009a. European Parliament Election Study 2009, Candidate Study, Advance Release. July/2010. [http://www.piredeu.eu/].

EES. 2009b. European Parliament Election Study 2009, Contextual Data, Advance Release. 16/05/2010. [http://www.piredeu.eu/].

EES. 2009c. European Parliament Election Study 2009, Manifesto Study, Advance Release. 22/07/2010. [http://www.piredeu.eu/].

EES. 2009d. European Parliament Election Study 2009, Media Study Data, Advance Release. 31/03/2010. [http://www.piredeu.eu/].

EES. 2009e. European Parliament Election Study 2009, Voter Study, Advance Release. 7/4/2010. [http://www.piredeu.eu/].

van der Eijk, Cees, and Mark N. Franklin, eds. 1996. Choosing Europe? The European electorate and national politics in the face of union. Ann Arbor: University of Michigan Press.

van der Eijk, Cees, and Eliyahu V. Sapir. 2010. Linking Electoral Data about Citizens, Political Parties, Mass Media and Countries. A General Approach with a User Application for the 2009 European Election Studies. Paper for PIREDEU User Community Conference 'Auditing Electoral Democracy in the European Union', Brussels, 18-20 November 2010 [available at http://bit.ly/emVWJ0].

van Egmond, Marcel H., Eliyahu V. Sapir, Wouter van der Brug, Sara B. Hobolt and Mark N. Franklin. 2010. EES 2009 Voter Study Advance Release Notes. Amsterdam: University of Amsterdam. [http://www.piredeu.eu/].

Giebler, Heiko, Elmar Haus and Bernhard Weßels 2010. 2009 European Election Candidate Study – Codebook, Advance Release. Berlin: WZB. [http://www.piredeu.eu/].

Reif, Karlheinz, and Hermann Schmitt. 1980. Nine second-order national elections. A conceptual framework for the analysis of European election results. European Journal for Political Research Vol. 8, pp 3–44.

Schmitt, Hermann, and Jacques Thomassen, eds. 1999. Political representation and legitimacy in the European Union. Oxford: Oxford University Press.

Schuck, Andreas, Georgios Xezonakis, Susan Banducci, and Claes H de Vreese. 2010, EES (2009) Media Study Data Advance Release Documentation. Exeter: University of Exeter. [http://www.piredeu.eu/].

Thomassen, Jacques, ed. 2009. The Legitimacy of the European Union after enlargement. Oxford: Oxford University Press.

Wüst, Andreas M., and Andrea Volkens. 2003. Euromanifesto coding instructions. MZES Working Paper 64, Mannheim: MZES [http://www.mzes.uni-mannheim.de/publications/wp/wp-64.pdf].

Notes

1. Paper presented at 36th Annual Conference of IASSIST in the panel 'Virtual Research Environments: Tools for Presenting and Storing Data', Cornell University, Ithaca (NY), June 1-4, 2010.

Please address all correspondence to:
Methods and Data Institute
University of Nottingham
University Park, Law & Social Science Building
Nottingham NG7 2RD, UK
cees.vandereijk@nottingham.ac.uk
eliyahu.sapir@nottingham.ac.uk

2. The website of the CSES lists a large number of comparative data collection projects in the general field of elections, parties and public opinion, with their respective URL's; see: http://www.umich.edu/~cses/about.htm.
3. Quite common, however, are publications in which separate and unconnected analyses from single datasets are 'linked' in a narrative fashion. Some of these are excellent and generate important insights. Yet, as will become obvious below, this nevertheless falls far short from linking the diverse data before the analysis stage and then analysing the merged data.
4. As a case in point, coding of parties is not fully comparable across successive editions of the European Social Survey (ESS), mainly by not anticipating that national party systems change over time. In 2008, code 11 for the variable asking about the party the respondent voted for in the last general elections in the Netherlands (PRTVNL) pertains to the PVV (Freedom Party), while in 2002 the same code pertains to 'other party'. Such incomparabilities are particularly large in countries with instable party systems.
5. We do not want to suggest that no useful efforts at harmonization are undertaken at all. Some of the most productive ones pertain to harmonization efforts aimed at making the coding of, e.g., educational attainments comparable across countries (the UNESCO initiated ISCED codes).
6. Detailed information about the PIREDEU program, including questionnaires and coding schemes, can be obtained from its website: http://www.piredeu.eu/.

As the focus of this paper is on data linking, we refrain here from presenting substantive information about the specific character of European Parliamentary elections, and we refer to the relevant literature, e.g., Reif and Schmitt (1980), van der Eijk and Franklin (1996), Schmitt and Thomassen (1999), van der Brug and van der Eijk (2007), and Thomassen (2009).

7. EES (2009e); van Egmond et al. (2010).
8. EES (2009a); Giebler, Haus & Weßels (2010).
9. EES (2009c); Braun, Mikhaylov & Schmitt (2010.
10. EES (2009d; Schucket al. (2010).
11. EES (2009b); Czesnik, Kotnarowski & Markowski (2010.
12. Such additional information can, of course, also be derived from external sources, such as the World Bank, OECD, EuroStat, and so on. For our example however we focus only on various PIREDEU data components.
13. This is similar to the futility of occasional proposals to 'standardise' the observation of essentially contested concepts (cf. Connolly 1999), or to standardise questionnaire items in survey research.
14. This implies that many of the attempts to provide 'continuity' files for, e.g., national election studies, are suboptimal at best. In the process of producing such files many operational decisions have to be made which are not of a technical and innocuous nature, but which have conceptual and theoretical implications. If analysts do not subscribe to these implications, the resulting continuity file will be less desirable, and they have to either repeat the same work on their own terms or, more frequently, abandon the project that required such linking.
15. For full elaboration of the linking solution, see van der Eijk and Sapir (2010).
16. Of course, many other summarising measures could also be relevant for interval level variables, such as x-tiles and x-tile ranges, skew and kurtosis, proportions above/below specified cut-off values, etc.