European University Institute

MAX WEBER PROGRAMME

EUI Working Papers

MWP 2010/21
MAX WEBER PROGRAMME

COLLABORATIVE DATA COLLECTION IN POLITICAL SCIENCE: A NEW DATA INFRASTRUCTURE ON PARTIES, ELECTIONS AND GOVERNMENTS

Holger Döring

EUROPEAN UNIVERSITY INSTITUTE, FLORENCE

MAX WEBER PROGRAMME

Collaborative Data Collection in Political Science: a New Data Infrastructure on Parties, Elections and Governments

HOLGER DÖRING

EUI Working Paper **MWP** 2010/21

ISSN 1830-7728

Printed in Italy
European University Institute
Badia Fiesolana
I – 50014 San Domenico di Fiesole (FI)
Italy
www.eui.eu
cadmus.eui.eu

Abstract

Information on political institutions, data on parties, elections, and governments, has yet to be provided in a format which makes it easily accessible for empirical research. Contemporary data on political institutions is scattered, limited to some countries or time periods only and difficult to combine, so that quantitative studies of political institutions have no systematic data infrastructure available which is equivalent to survey research or OECD data. As a consequence, work on political institutions rests on very heterogeneous information sources and the quality of data does not match standards of replication for empirical research. Political scientists are in need of a modern replacement for data handbooks and recent technological innovations have broadened the opportunities to develop such an infrastructure within the political science community.

I discuss existing approaches towards collaborative data collection in political science and highlight contemporary shortcomings. In the paper, I propose a novel approach towards data collection in comparative research and present a new data infrastructure on parties, elections and governments, the Parliament and Government Composition Database (ParlGov). The data infrastructure combines a database, data presentation in webpages and software scripts in order to generate more dynamic datasets. So far, it includes information about more than one thousand parties, around five hundred elections and almost one thousand governments. This infrastructure allows us to derive a wide range of datasets for studies in political science and can be easily extended. Hopefully, the paper will encourage rethinking about contemporary ways of collecting data on legislatures and executives.

Keywords

Comparative politics, dataset, political instituions, parties, elections, governments.

The ParlGov database (www.parlgov.org) that I introduce in this paper is based on joint work with Philip Manow, whom I wish to thank for his support. Our work on this infrastructure started at the Max Planck Institute for the Study of Societies (MPIfG) from 2005 to 2007 and continued at the University of Konstanz from 2007 to 2009. At the European University Institute, my work on the project has been significantly enhanced through support by Mark Franklin, Peter Mair and Alexander Trechsel. In addition, I would like to thank Laurie Anderson, Fabio Franchino, Alexia Katsanidou, Alyson Price, Julia Sievers and Luca Verzichelli.

Holger Döring
Max Weber Fellow, 2009–2010

Introduction

An "enterprise of madness" was the label that colleagues put on Peter Flora's efforts to collect data for comparative research in a systematic fashion (Flora 1983, 5). Three decades later, students of political institutions are still spending an enormous amount of time creating datasets for their empirical work. Here, I am concerned with data and parameters on election outcomes, government compositions and party positions. This information is available in printed and machine readable form, but it is heterogeneous, of different quality and hard to combine. We are still in need of more up-to-date information about political institutions that suit our empirical research, a modern equivalent for data handbooks and yearly reports of political events. Unfortunately, as of today there is no systematic approach to overcome contemporary limitations in providing adequate data sources on parties, legislatures and executives. Here, I will introduce such a data infrastructure for comparative research and address open questions of collective data accumulation.

Other scientists have faced similar challenges in harnessing the new potential from information technologies in order to improve data sources for scientific research. Biologists, for example, are trying to combine and to improve information about species through the Encyclopedia of Life (EOL), a free, collaborative online encyclopedia documenting known species. In doing so biologists are trying to find a modern way of presenting information that was previously documented in printed form. The challenge for biologists is similar to problems with data about political institutions. How to combine existing sources, revise and improve them collaboratively?

There are well established ways to collect data on elections and cabinets in printed form such as Mackie and Rose (1991) and the yearly political data section of the European Journal of Political Research (EJPR). Unfortunately, these well-respected data sources are the basis of many different digital datasets on political institutions. Establishing a coordinated effort to provide better and more readily available data on parties, elections and governments has largely failed, causing a collective action problem in gathering data for comparative research. As a consequence, empirical studies of political institutions are difficult to replicate, with respect to the data sources they make use of. This is surprising, as there are now high standards when it comes to replicate the statistical analyses of quantitative work (King 1995).

By discussing the state of the art, I hope to foster more cooperative approaches towards data collection in political science. Currently, it is very cumbersome to answer even simple questions of comparative politics due to the limited availability of necessary data sources. As a consequence, it is also difficult to communicate many of the existing findings of comparative politics about the functioning of political institutions across disciplines and to teach these insights to university students.

In this paper, I introduce a new approach towards data collection in political science and a new data infrastructure on parties, election results and governmental compositions. The new infrastructure is named ParlGov – Parliament and Government composition database – and a first version was released in February 2010 (Döring and Manow 2010). The infrastructure makes use of recent technological innovations and has four components: a database to file the information non-redundantly, computer scripts to calculate institutional parameters, a web interface to present coded observations in a more accessible manner and a feedback system that allows other researchers to contribute their country expertise in an open and transparent way. The latest version of ParlGov includes more than a thousand parties, about five hundred elections and a thousand cabinets. These observations are linked to existing data sources such as information about party positions.

In order to present the new approach towards data collection and its innovative potential, I proceed in three steps. First, I provide a discussion of previous approaches to the collection of empirical data about political institutions. In this part, I also discuss existing sources of information about political institutions such as party positions, election results and government compositions. I conclude this part by summarising the shortcomings of contemporary approaches. In a second part, I introduce my ideas towards data collection in political science and present ParlGov, a new database with information on political institutions. I conclude by providing some guidelines and best practices on how to create datasets that are easier to link.

Large scale data collection for comparative research

Approaches in the discipline

Several systematic attempts to collect empirical data for political research have been developed in sub-disciplines of political science over the last decades. But why are there well established practices for the collection of data for work in political behaviour and political economy but not for research on political institutions? For studies of political behaviour, the cost of large surveys has forced researchers to develop institutionalised ways of collecting and archiving opinion data. As a result, there are national election studies that regularly run large scale opinion polls and archive their results. For these studies, there are well established rules about how to conduct, document and file the collected information. As a consequence, students of political behaviour have a large set of archived studies that they can base their analysis on. There are difficulties in combining national election studies across countries and time, but researchers have a wide range of datasets available in digital form upon which to base their empirical work and there are collective efforts to link these existing sources across countries. In the field of political economy, data is provided through national statistical offices, international organisations such as the OECD and the World Bank, or research institutes. Again, there are institutionalised ways of collecting economic data, updating and archiving it. Empirical studies of political behaviour and political economy start by deriving datasets from institutionally provided sources which are available in a format that makes it possible to apply the information without independent data collection.

The situation is very different for information about political institutions such as election results, government compositions and observations about political parties, the types of data fundamental for comparative politics. There are established ways to collect, combine and archive this information, but the data is often not suitable for (quantitative) empirical work without major revisions. Currently, students of political institutions spend an enormous amount of time on data preparations. In my view, there is significant room for improving contemporary approaches toward data collection in political science.

What are the empirical sources of information on parties, elections and governments that are available today? What are contemporary approaches towards data collection? Why do I think there is significant room for improvements? Mackie and Rose (1991) and its successor the EJPR data yearbooks are probably the most authoritative sources of data on election results in advanced democracies. In addition, Nohlen (2005); Nohlen et al. (2001, 1999) has collected election results for most elections around the world in the last century. These sources provide carefully collected information about election results in printed form and the library system guarantees that this data is available to all scholars, but it is difficult or time consuming to draw on them because they are not available in digital form. As a consequence, scholars use different datasets derived from these sources and there is no shared data source that forms the basis of empirical work.

A significant improvement in terms of providing access to empirical information has been to accompany data handbooks books with CD-ROMs. The two volumes of the Comparative Manifesto Project (CMP) are the shining examples of very thoroughly developed information about party positions and election results (Budge et al. 2001; Klingemann et al. 2006). Caramani (2000) is a another unprecedented effort to offer better empirical data about parties and elections at the sub-national level. These datasets have formed the basis of many empirical studies. Unfortunately, this information is not updated regularly so that many scholars extend these data sources with their own information, with the result that different datasets are derived from one source and there is no collective approach towards updating and improving these sources.

Over the last decade, the Internet has offered new opportunities for researchers to present and distribute their data and there is now an almost unlimited amount of information on the web. Müller and Strøm (2000) is a good example of work on political institutions that was first published in a format similar to data handbooks but is now accompanied by an online source, the Comparative Parliamentary Democracy Data Archive (CCPD). In comparative politics, the Armingeon et al. (2009) Comparative Political Dataset (CPDS), the Comparative Study of Electoral Systems (CSES) and the Kollman et al. (2010) Constituency-Level Elections Archive (CLEA) are only some of the more important datasets that are available online. These online sources follow a traditional format: they combine a dataset with a codebook similar to survey research. The codebook documents the data, its variables and sources. There are certain shortcomings of online information such as data format problems, a lack of documentation, and difficulties in archiving this information in the long term. The problem of archiving has been reduced by encouraging scholars to submit their datasets to data archives such as the Interuniversity Consortium for Political and Social Research (ICPSR) or the Economic and Social Data Service (ESDS). Providing political science data online has significantly broadened the opportunities for empirical research but it is often difficult to access, combine or update these sources.

Finally, there are some more recent approaches towards generating data for political science research that draw on new computer techniques. Høyland et al. (2009) for example, suggest creating automated databases for political research based on official online presentations. They give an example for the Members of the European Parliament (MEPs). Information on MEPs, such as biographical data or committee assignments, are available on the webpage of the European Parliament. However, this information is not presented in a way that makes it suitable for comparative research without modifications and has to be transformed into a data matrix. Hoyland ea. suggest to applying computer techniques in order to automatically convert these official sources into a data matrix that can be used for empirical work in political science. By running these computer conversions at regular intervals, they provide data for researchers that is up to date and includes the most recent official information. With this approach, students of comparative politics do not have to collect and update data themselves but make use of computer tools to convert existing sources into data for political analysis. However, these approaches are limited to information that is prepared and made available by other agencies and does not include the type of data that I am concerned about in this paper.

Having summarised the evolution of datasets for comparative research, we can conclude that comparativists have always been able to make use of technological advances to improve data sources for empirical research: from carefully collected information in data handbooks over data on digital disks to the recent usage of the Internet. Innovation in gathering empirical information over time has allowed political scientists to investigate a broader set of research questions. Nevertheless, as I will argue in the next sections, there is even more potential ahead of us.

Contemporary shortcomings

There is now a broad set of empirical information and data about political institutions (esp. parties, elections and governments) available. However, this information is still difficult to combine, hard to correct and cumbersome to work with. Let me be more specific about the critique that I address towards contemporary approaches of collecting data on parties, election results and cabinets. First, data sources are often very difficult to combine. This is a result of the fact that different IDs are used across datasets, a problem that may not be solved totally. For parties, it may be difficult to find one unique identifier to link all information about parties across various datasets. Parties split, change their names, or form alliances and we may disagree how to code these changes over a party's life cycle. However, we should be able to find overlapping information for most parties and are in need of sources that link existing observations on political parties. The difficulties of linking observations does also apply to elections and governments but this data is easier to combine by technical means. Hence, we are facing the challenge to find ways to better link existing data sources and documentation about the problems of linking these sources.

The second critique concerns the enormous number of variables that are often combined into one data matrix at the coding stage. I am not concerned about the amount of information but the lack of distinction between different types of data and the difficulties in comprehending the vast amount of content. Take for example an election result: There is some information that is unique to every election. Other observations have to be coded at the party level such as the number of seats a party won. There may be data about party positions in a different source and we might want to calculate some institutional parameters from these observations such as the effective number of parties. All this data is often entered manually or semi-manually into one rectangular data matrix, thereby duplicating a lot of observations. Technically, this information should be kept separated in different data tables and be combined by merge scripts or a database design. I will propose four different types of data later in this paper by introducing a novel approach towards data collection in political science. By distinguishing these data types information can be coded more coherently and consistently.

Third, there is no systematic way of improving the information that is provided in different datasets. Sometimes, the exact coding of an election result or a government termination may be controversial, but it is easy to agree on most of the observations. Today, researchers often correct errors they find in their personal copy of a dataset as it was downloaded or generated from a data handbook. They may inform the original collector about a data bug but only rarely is this information included in an updated version of the original data or communicated separately in a list of known data bugs. Once data is published in a handbook, on a CD or online, this data is fixed forever. Providing stable versions of a dataset is necessary for the replication of analysis. Nevertheless, there could still be updated information in succeeding versions of a dataset or a list of known errors and later releases should inform us about changes and include received feedback. As of today, there are hardly any institutionalised approaches to create regularly updated digital datasets on legislatures and executives.[1]

Finally, I argue that data on political institutions can often be presented in a more accessible format. Political institutional data differs from mass level survey data by providing information at different levels of observations. A dataset may contain variables at the country, election or party level. For most of these observations, we know the 'true' values and coding errors should be corrected. However, presenting all observations in a large combined data matrix and a codebook reduces the likelihood of identifying potential coding errors. Traditional data handbooks have presented empirical observations in a more accessible manner by combining data observations, notes and comments. Hence, we should try to find a modern equivalent to present our empirical observations in a more accessible format. Presenting information in different forms may make data errors more easily identifiable and facilitate collaborative data revision.

To sum up, most of the contemporary approaches to collecting data about political institutions no longer match the demands of empirical analysis. Data handbooks, yearly political reports and static data sets offer the information needed for data analysis, but do not present them in a format that can serve as a consistent basis for empirical analysis. These existing sources have yet to be transformed and extended in a way that makes it possible to address a particular research question. As a consequence, most of the current data collections for political analysis are heterogeneous, not up to date and difficult to combine. Hence, questions of reliability are a major concern for empirical work on political institutions due to differences in the underlying data collections. How can these challenges be overcome and what may a new infrastructure for data on parties, elections and governments look like?

A new data infrastructure

ParlGov is a new data infrastructure to foster empirical work on parties, legislatures and executives. The infrastructure makes use of recent innovations in information technologies and

[1] There are well established practices to provide this information in printed form such as EJPR political data yearbooks. The Constituency-Level Elections Archive (CLEA) is an attempt to establish such an infrastructure for district level electoral results Kollman et al. (2010).

provides an example for new types of collaborative data collection in political science. The new approach towards data collection has four components:

- a database to store empirical observation and coded information
- a presentation of data content in webpages
- feedback mechanisms for collaborative data enhancement
- programmed scripts to calculate institutional parameters and to link external datasets to the database

The data can be accessed via an online interface, but can also be downloaded and used on personal computers. All observations are visualised in webpages and can be accessed as data tables. Users can provide feedback and observations are updated regularly. Yearly releases of static versions of the data guarantee a stable set of information for replication purposes. The following paragraphs describe each of the components of the integrated data infrastructure in more detail.

Empirical information collected

Table 1 summarises the empirical information from the first release of the ParlGov database (Döring and Manow 2010). For *34 countries* it includes observations about all parties, electoral results and cabinets that followed democratic elections in the post-war period. The countries include all EU and most OECD members. In ParlGov, all information can be easily combined by making use of unique identifiers.

The latest ParlGov version includes observations for *1116 parties* and classifies them into party families. It records a party's name in the original language (native and Latin characters), its English name and the official abbreviation. In addition, all name changes over a party's history are coded. Observations on parties are also linked to those parties that were formed by merging or by splitting up. This coding scheme makes it possible to track the evolution of a party system over time.

Parties in ParlGov are linked to a set of well known datasets with information about party positions at a particular point in time. The major party expert surveys from Castles and Mair (1984), Huber and Inglehart (1995), Ray (1999), Benoit and Laver (2006), and the Chapel Hill Expert Survey Series (Hooghe et al. 2010; Steenbergen and Marks 2007) are connected to ParlGov. Party observations are also connected to the CMP data (Budge et al. 2001; Klingemann et al. 2006) and the EU Profiler (Trechsel and Mair 2009). This allows users to add observations about the political positions of parties from various external sources to all observations in

Table 1: Summary of observations in ParlGov database (Version 10/02)

	parties	elections	cabinets		parties	elections	cabinets
Australia	13	25	33	Japan	25	18	39
Austria	11	20	27	Latvia	41	6	17
Belgium	44	20	45	Lithuania	42	6	16
Bulgaria	42	6	9	Luxembourg	18	15	20
Canada	14	21	22	Malta	16	17	14
Cyprus	20	9	15	Netherlands	39	19	27
Czech Rep.	37	6	12	New Zealand	18	22	29
Denmark	29	25	36	Norway	20	17	32
Estonia	31	5	11	Poland	50	7	21
Finland	35	24	41	Portugal	30	13	19
France	60	18	62	Romania	37	6	16
Germany	40	17	21	Slovakia	36	6	13
Greece	25	13	18	Slovenia	24	5	12
Hungary	25	5	9	Spain	44	10	11
Iceland	22	20	32	Sweden	18	20	29
Ireland	20	19	25	United Kingdom	42	17	22
Italy	131	17	57	GDR	17	1	1
				Total	1116	475	813

ParlGov. The party table makes it also possible to combine external datasets in order to cross-validate party positions or to derive positional parameters from this information.

The data infrastructure includes all democratic elections for the post-war period and information about the party make up of governments. The latest version of ParlGov includes, *475 elections with 3779 election results* at the party level. Most of the information is based on official electoral results and all parties with seats in national parliaments are coded. For some of the countries, the number of votes for all parties that won more than 0.5% of the national vote are included. The coding scheme distinguishes parties that form electoral alliances and run on a joint list from the parliamentary groups these parties join in the legislature. For the former the percentage of votes is recorded whereas the number of seats are coded for the latter. This approach makes it feasible to compare party systems in the electoral arena and in the legislature with the help of the data.

To record the party composition of governments, data about cabinets and the parties represented in them has been collected. Cabinets are coded in line with a definition of a change in government proposed by Budge and Keman (1993, 10): any change in the set of parties holding cabinet membership, any change in the identity of the prime minister, any official resignation of a government and any general election. The latest released version includes *813 cabinets with 1899 governing parties*. Again, this data on cabinets can be linked to previously presented information about parties and legislatures.

Finally, ParlGov also includes the results of European Parliament elections into the database. Here, information for 127 elections with 939 electoral results at the party level has been collected and for most of the elections party alliances have also been recorded. The observations in the database have unique identifiers for all parties, election results and cabinets. By including these unique identifiers into all observations, we can combine information in ParlGov with the help of a database, to which I turn now.

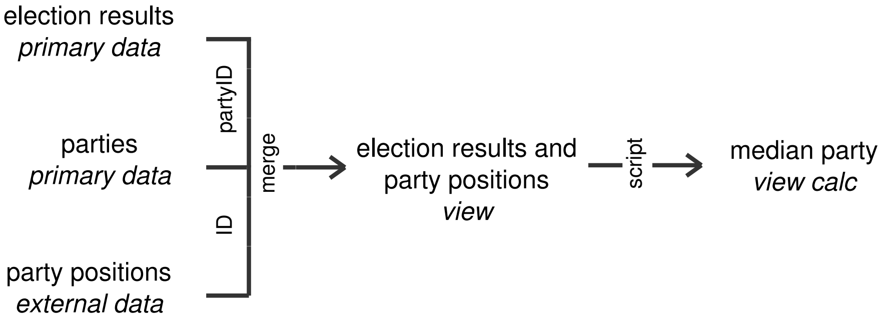
Database

Making use of a database allows us to separate data about political institutions more carefully. Databases come in various forms and relational databases are optimised to store data non-redundantly according to a defined table schema. Take for example an election result: in a relational database, we would create at least two tables for electoral results. One table includes data about each election, such as date and turnout. A second table gathers observations about each party that took part in the election, for example the number of votes and seats it won. For each election, the first type of information is observed at the election level, the second at the party level. In a database, we store this data separately and combine it at a later stage. We could also add more information about the parties in a third table such as left/right positions or link to external data sets with this type of observation. For our empirical analysis, we can create a dataset based on these three tables by combining the data sources. In ParlGov, the original tables that record observations are called *primary information* in order to highlight the fact that this data is collected or coded information and can not be derived from other observations nor be calculated. I elaborate on this distinction in the next paragraph.

Making use of a database allows it to integrate other datasets more easily. For parties many different datasets about party positions can be combined by creating a primary table with party IDs of the different datasets. Having all IDs in one data matrix allows us to combine the different datasets. Previously, I have listed the set of party position data such as expert surveys and manifesto based sources that are connected in ParlGov. Other types of *external observations* may include turnout data for every election from a different source or economic data for a country. Keeping this information in separate tables allows potential users to link the external information to core data as needed. Hence scholars can decide if they want to link election results with one of the expert surveys or CMP data. Using this approach makes it possible to distribute our database without including the external datasets. The database includes example scripts that demonstrate how to link ParlGov data and external observations.

Another type of table in the database is generated dynamically, based on primary and external observations. These are virtual tables, *views* in technical terminology, generated through database operations by combining primary and external tables via defined queries. The Parl-

Figure 1: Data types and combining different sources



Gov database creates a virtual table for election results that provides information for each party but adds party positions from a different table. Virtual tables can be very powerful. In the database there are also more complex views, for example a data matrix about government formation linking cabinet parties and election results. This table gives information about all parties in parliament at every instance of government formation and indicates if a party becomes a government member or not. This data matrix can be extended in order to determine if a party was ever in government before, its seat share in the preceding parliamentary term or party system parameters by creating the required database queries and including them in the view. If any of the primary data is changed, information in the virtual table is updated instantly. Virtual tables are permanently saved database queries based on primary and external data. Instead of working with database views, ParlGov data tables can also be combined via merge operations in a statistical software package. Virtual tables are most likely to form the basis for empirical work based on ParlGov data. The latest release includes two major views, one on election results and a second on government formation.

Some variables that are of interest to political scientists are also logically based on primary and external information but are difficult to calculate with merge or database operations only. These may be complex institutional parameters that have to be calculated by programmed functions. Determining the position of the median party in parliament is one example or various power indexes. These observations are calculated by software routines from statistical software packages and are based on primary and external data. Because this information is still virtual, based on other coded observations, it is called *calculated views*. In the latest version of the database, there is for example one table that calculates parameters of electoral and party systems (disproportionality, advantage ratio, effective number of parties etc.) based on the vote and seat share of parties in parliament.

Figure 1 provides an example of how the different data types are interrelated. The figure shows how to determine the median party based on election results from primary data and party positions from an external source. These two types of information are merged, joined in database terminology, through a table that contains information about parties linking IDs from different datasets. As a result, there is a new table (view) with the electoral results of parties and their policy positions for every observation. Based on this information the median party for each election can be calculated with a computer script.

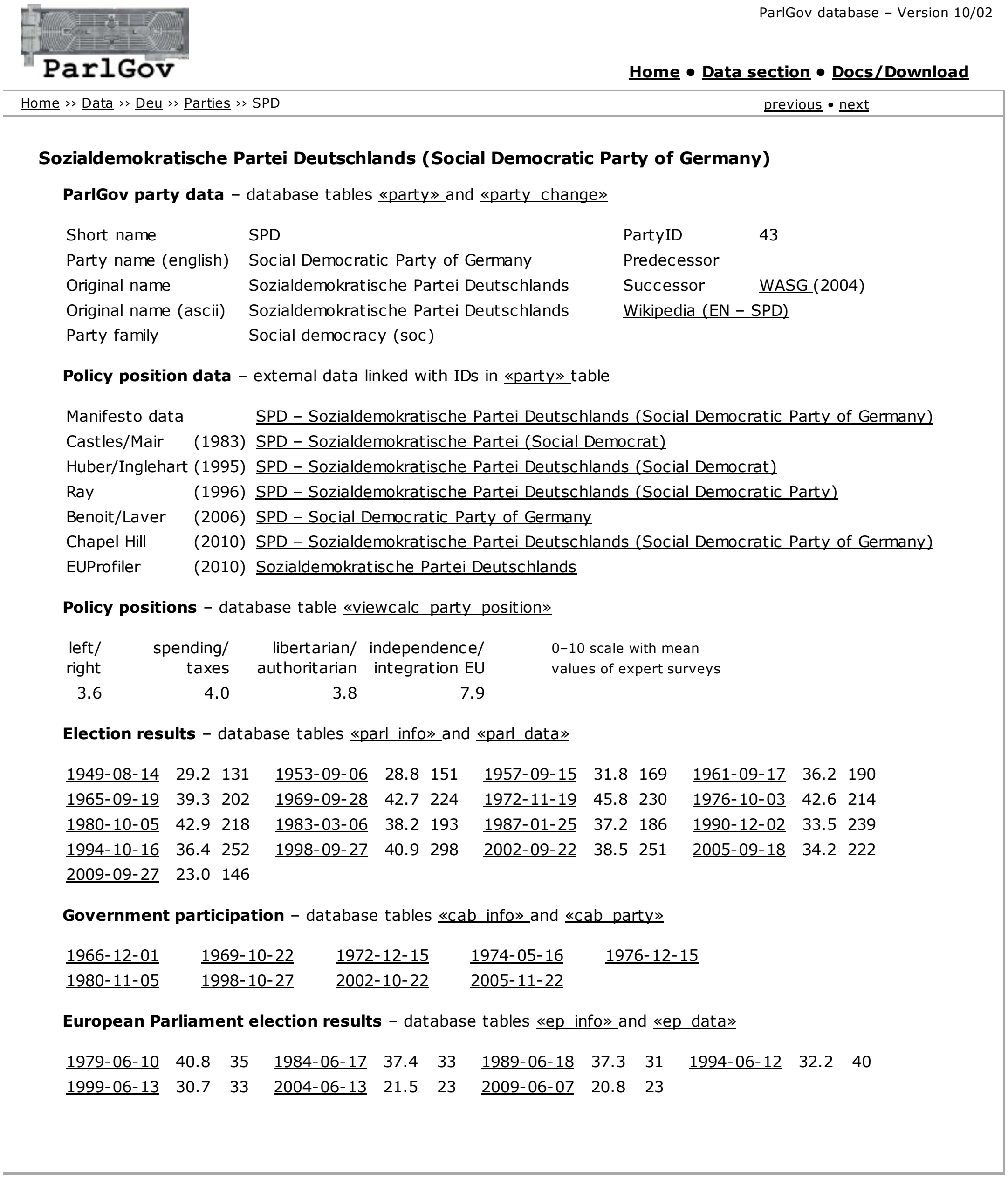
The database in ParlGov has multiple data tables that are combined to produce datasets for empirical research. This approach makes it possible to combine a wide set of existing information with observations on parties, elections and governments. However, combining such a wide set of sources leads to a data structure that may be difficult to understand at the beginning. It is not a problem of the approach per se, but simply the result of integrating an enormous amount of data that already exists. After collecting information about a certain number of variables a dataset can become difficult to understand. Hence, we have to think about alternative ways to present that content in order to make it more accessible. How can we save highly structured observations in a database and present it in an accessible format?

Data presentation in webpages

Data handbooks have the advantage of presenting empirical information in a very comprehensible way. A description of observations, introductory chapters and footnotes provide very detailed summaries about all aspects of the empirical information in these sources. However, preparing information in such a format makes it difficult to use this information in machine readable form or to include it into a dataset. For our contemporary work, we need information in a data matrix, which is often difficult for human beings to read. In ParlGov, empirical information from data tables is presented in webpages to overcome these limitations. These webpages are available online on the Internet as well as offline in a local version and are a modern equivalent for data handbooks.

Webpages are a powerful way of presenting information from databases and they offer an alternative form of data visualisation. In ParlGov, all information about parties, elections and governments is presented through these pages. For example there is one page for each party in the dataset and Figure 2 shows the page of the German Social Democrats (SPD). This page lists all information about the party that is included in the ParlGov database, as well as a list of all the names of parties in external datasets that are linked to the observation. In our example, we find that the SPD is included in all data sets on party positions linked to ParlGov. If available, the page lists elections and electoral alliances a party took part in, its government participation as well as renamings of a party. For the German SPD, the page summarises election results for 17 national and 7 EP elections in addition to 9 cabinet participations. No name changes or electoral alliances are recorded for the SPD and several entries on the webpage are connected to other

Figure 2: Example of data presentation in a webpage



ParlGov database – Version 10/02

ParlGov

Home ● **Data section** ● **Docs/Download**

Home ›› Data ›› Deu ›› Parties ›› SPD

previous • next

Sozialdemokratische Partei Deutschlands (Social Democratic Party of Germany)

ParlGov party data – database tables «party» and «party_change»

Short name	SPD	PartyID	43
Party name (english)	Social Democratic Party of Germany	Predecessor	
Original name	Sozialdemokratische Partei Deutschlands	Successor	WASG (2004)
Original name (ascii)	Sozialdemokratische Partei Deutschlands	Wikipedia (EN – SPD)	
Party family	Social democracy (soc)		

Policy position data – external data linked with IDs in «party» table

Manifesto data		SPD – Sozialdemokratische Partei Deutschlands (Social Democratic Party of Germany)
Castles/Mair	(1983)	SPD – Sozialdemokratische Partei (Social Democrat)
Huber/Inglehart	(1995)	SPD – Sozialdemokratische Partei Deutschlands (Social Democrat)
Ray	(1996)	SPD – Sozialdemokratische Partei Deutschlands (Social Democratic Party)
Benoit/Laver	(2006)	SPD – Social Democratic Party of Germany
Chapel Hill	(2010)	SPD – Sozialdemokratische Partei Deutschlands (Social Democratic Party of Germany)
EUProfiler	(2010)	Sozialdemokratische Partei Deutschlands

Policy positions – database table «viewcalc_party_position»

left/ right	spending/ taxes	libertarian/ authoritarian	independence/ integration EU	0–10 scale with mean values of expert surveys
3.6	4.0	3.8	7.9	

Election results – database tables «parl_info» and «parl_data»

1949-08-14	29.2	131	1953-09-06	28.8	151	1957-09-15	31.8	169	1961-09-17	36.2	190
1965-09-19	39.3	202	1969-09-28	42.7	224	1972-11-19	45.8	230	1976-10-03	42.6	214
1980-10-05	42.9	218	1983-03-06	38.2	193	1987-01-25	37.2	186	1990-12-02	33.5	239
1994-10-16	36.4	252	1998-09-27	40.9	298	2002-09-22	38.5	251	2005-09-18	34.2	222
2009-09-27	23.0	146									

Government participation – database tables «cab_info» and «cab_party»

1966-12-01	1969-10-22	1972-12-15	1974-05-16	1976-12-15
1980-11-05	1998-10-27	2002-10-22	2005-11-22	

European Parliament election results – database tables «ep_info» and «ep_data»

1979-06-10	40.8	35	1984-06-17	37.4	33	1989-06-18	37.3	31	1994-06-12	32.2	40
1999-06-13	30.7	33	2004-06-13	21.5	23	2009-06-07	20.8	23			

Data source: ParlGov database Version 10/02 (Döring and Manow 2010) – released 25 February 2010. Some minor modifications to the layout of the original page have been applied and underlined entries indicate links to other pages.

pages where further information can be derived. The list of elections that a party took part in on the page link to pages that show all parties that took part in that particular election with their respective electoral results. On this page, information about the governments that formed after this election is given and links to separate pages listing the cabinet members and information about the cabinet are provided. Again, these pages are based on the same database that is used to generate the data tables for empirical research and the pages are available online as well as offline.

By providing such an alternative format to present empirical information, the quality of our coding becomes more transparent and open for close scrutiny by country experts. Later in this paper, I will describe the release strategy and demonstrate how updated data is offered at regular intervals. Here, I want to note only, that the webpages are available online for the most recent version and are also included in the dataset released as a static version. The online presentation of the data on the Internet does also allows users to offer feedback on empirical information and I will now describe the feedback system more generally.

A feedback system to improve cooperation

Most of the observations on parties, elections and governments have defined values for all variables. There is an official election result, an official party name and a date a government is sworn into office. Explicit coding rules may further narrow down coding ambiguities. However, collecting all this information in detail is time consuming and may leave mistakes uncovered. Reasonable effort is sufficient to collect data on the number of seats for parliamentary parties and government participation. Nevertheless, having more detailed information and integrating official sources requires the support of country experts with detailed knowledge about the institutional structure of a country. Often it is time consuming for non-natives to find out details about a particular electoral alliance or about the causes of a specific government breakdown. Hence, giving users and country experts an easy way to access the data and to update it can significantly improve the quality of empirical information in the long term. Similar to scientific publishing, we are in need of platforms to improve our data over time and to debate about the coding of ambiguous cases.

New computer techniques can help to integrate the feedback of users and experts. Modern software development techniques have significantly enhanced the potential to collect error reports (referred to as bugs among programmers), feature requests, user comments and documentation. Modern software development offers many valuable ideas for new approaches towards data collection in comparative research and some of these tools are integrated into the ParlGov infrastructure.

Some of these practices are rather straight forward: encourage and provide a mechanism for feedback, document known errors first and fix them later. In its most simple form this can be done by encouraging suggestions in the documentation and by listing known problems on an Internet page. There are some more advanced techniques that foster cooperation and feedback mechanisms. The ParlGov project makes use of an online project management software that includes a wiki.[2] Users can file error reports as well as suggestions in such a tool, assess development progress and add data or software scripts. This openly available information allows users to closely follow and to participate in the evolution of the data project.

Finding more collaborative approaches towards collective data generation in comparative research is an important challenge for research in political science. Better tools to foster cooperation are available now through the Internet. At this stage, there are almost no attempts to incorporate these technological innovations into our scientific work. There is still too much valuable research time spent on simply collecting and combining existing information about political institutions.

Versions and archiving

The previous section has highlighted the fact that errors in datasets about political institutions should and can be corrected through feedback mechanisms. These tools encourage experts to provide their knowledge to facilitate a continuous evolution of data collection. As a consequence, the content of the data structures proposed here changes regularly. In addition, including new data and recent political events such as elections or government formations also alters the observations in the database. Hence, the exact same mechanisms that improve the data in the long term undermine standards of replication. Approaches designed to overcome these shortcomings are well established. Nowadays, researchers are encouraged to file their datasets in data archives such as the ICPSR or the ESDS. These agencies guarantee the archiving and long term distribution of social science data. In this way, they ensure that empirical information is available and accessible for future researchers.

For ParlGov, there will be two types of released datasets. First, there is a stable version, that is well documented. In this version, the quality of all observations has been double checked and all details of the data are documented. This version should provide the basis for empirical work because it gives a fixed and replicable amount of information. This data will also be archived. Second, there is a development version, that includes all recent changes, user feedback and corrected error reports. It may also contain some variables and observations that have to be

[2] Wikis are online tools to easily edit and create web pages in a web browser. Changes to documents can be performed by all users and a history of previous page versions is stored in the wiki. Wikipedia is the most widely known online collaboration platform that is based such a technology.

documented in more detail. Nevertheless, it contains the most recent events (elections and new governments) with data errors corrected and some scholars may want to rely on this more up to date information. The development version provides the basis for the next stable release and there will be at least one of these stable versions every year. Again, this is an equivalent for a well established practice: yearly data reports in political science journals that document recent political events and make the information available in the long run.

How to improve data collection for comparative research

I would like to conclude with some recommendations for data collection in political science. My experience has grown from setting up the ParlGov database previously described and some preliminary work with data on political careers. Some of the suggestions I make are a summary of best practices, other ideas are aimed at encouraging new approaches to collect and present empirical information. Based on my experience, I believe that some of these suggestions can significantly reduce the costs of setting up and linking different data sources.

First, datasets that make use of information on parties and politicians should contain a unique numeric identifier for each party and politician. Mackie and Rose (1991) and the CMP project (Budge et al. 2001; Klingemann et al. 2006) provide such a unique numeric identifier for every party in their datasets, but other sources contain only party names. Preferably, this identifier should be based or linked to a well documented widely accepted dataset such as the CMP or the ParlGov database introduced here. Providing unique identifiers allows others to link different datasets more easily and offers a standard for comments on observations.

Scholars should also offer a way for users to give feedback and error reports about publicly available datasets. Most of the contemporary datasets are made available over the Internet so that it is easy to encourage users to submit short error reports in an e-mail. However, it is also important to add a short list of these known errors to the dataset. By publishing such a list of known problems, dataset maintainers may reduce repeated misapplication of their empirical information. Ideally, once a significant number of errors have been reported, a second version of the dataset should be made publicly available. This new version may include a list of changes to the original version and provide an updated empirical basis for succeeding studies.

Nowadays, it is best practice to accompany a dataset with a codebook which lists information about variables, their coding, a list of references and usage instructions. The practice is derived from survey research, where it is necessary to document survey questions and their coding with a dataset. I have had mixed experience with codebooks while working on data about political institutions. On the one hand, these codebooks make it easy to understand the structure of the dataset at hand. On the other hand, codebooks often include data that may be accessed more directly, especially observations about parties and politicians. For parties, their IDs, original names, English translations and abbreviations are often listed carefully in these codebooks. However, sometimes it is preferable to work with this data in a more accessible way. Often it is upon users to turn this information into a separate table. Once this information is converted into a data table, record linkage techniques can be used to connect different datasets (Christen 2006). These linkage techniques aim to link differently spelt names and may require some manual revision. However, these techniques reduce the amount of time to merge datasets significantly.

These critical comments on locking important information into codebooks leads to a more general point. I think that many contemporary data sets in comparative politics would profit from keeping separate data separate. Recording a dataset in several tables, maybe later merging them with a script into a large data matrix, would allow researchers to link different datasets more easily. This does not always require making use of a fully normalised relational database design. The latter approach comes with its own overhead. Nevertheless, it would be helpful if future datasets are accompanied by a script that links the different data sources a study is based upon. The dataset on roll call votes in the European Parliament by Hix et al. (2007) provide a well done example. The data is presented without a codebook in one file only. Nevertheless, all information, such as sources of data and funding, publications etc. are presented on a webpage. The coding of the data is documented by including several data tables in the spreadsheet. As a consequence, this information can be linked far easier with other data sources.

The base line of these recommendation is to consider administrating a dataset as a continuous and long term endeavour that evolves. Printed data handbooks often come in several editions and the same practice should be implemented for digital empirical data sources. If researchers decide not to maintain their datasets any more, they may consider allowing others to take on this role. Political scientists need a certain set of information about legislatures and executives for most of their studies and more institutionalised approaches to collect this information within the political science community should be established.

Conclusion

The purpose of this paper was threefold. First of all, I wanted to give an overview of the evolution of data collection in political science; especially data about parties, elections and governments. I discussed the evolution from data handbooks to digitally collected datasets. Second, I provided a summary of the shortcomings of contemporary approaches towards data collection in political science. I emphasised the fact that data collection in political science is facing its own collective action problem. Most of the information we need for empirical work on political institutions is available but it is cumbersome to combine existing data sources. Finally, I proposed a number of techniques to improve collective data collection in political science.

In this paper, I have also introduced a new data infrastructure on parties, elections and governments, the ParlGov database. For this data infrastructure, the techniques I recommend for collective data generation have been applied in a novel way. Many of the concepts are based on social and software techniques applied in software development. ParlGov offers an infrastructure for empirical research that overcomes many of the shortcomings of contemporary data collection approaches. With the help of a database design, it can combine information on electoral outcomes and cabinet compositions with a wide range of external data sources (e.g. party positions) and ParlGov offers ways to calculate institutional parameters from these observations. Providing collected empirical information in webpages offers a more accessible way of presenting data and links between data sources. Presenting empirical information in such a format should facilitate the integration of detailed country expertise for future revisions of the data.

The data infrastructure described offers a modern and innovative approach towards data collection for comparative research. It may mark the next step in the evolution of collecting empirical information. Modern datasets for comparative research should encourage collective data gathering and reduce barriers of cooperation. In the paper, I have discussed some recent technologies that significantly lower the cost of collective data gathering. The ParlGov infrastructure provides an example of how to make use of these techniques. Gone should be the days of manually typing information from codebooks and data handbooks into spreadsheets to link existing sources on parties, elections and governments.

References

Armingeon, K., R. Careja, S. Engler, M. Gerber, P. Leimgruber, and P. Potolidis (2009). Comparative political data set III, 1990–2007. Institute of Political Science, University of Berne.

Benoit, K. and M. Laver (2006). *Party policy in modern democracies*. London: Routledge.

Budge, I. and H. Keman (1993). *Parties and democracy: Coalition formation and government functioning in twenty states*. Oxford: Oxford University Press.

Budge, I., H.-D. Klingemann, A. Volkens, J. Bara, and E. Tanenbaum (2001). *Mapping policy preferences: Estimates for parties, electors, and governments, 1945–1998*. Oxford: Oxford University Press.

Caramani, D. (2000). *Elections in Western Europe since 1815: Electoral results by constituencies*. London; New York: Palgrave.

Castles, F. G. and P. Mair (1984). Left right political scales: Some expert judgments. *European Journal of Political Research 12*(1), 73–88.

Christen, P. (2006). A comparison of personal name matching: Techniques and practical issues. In *ICDMW '06: Proceedings of the Sixth IEEE International Conference on Data Mining – Workshops*, Washington, DC, USA, pp. 290–294. IEEE Computer Society.

Döring, H. and P. Manow (2010). Parliament and government composition database (ParlGov) – Version 10/02. http://www.parlgov.org.

Flora, P. (1983). *State, economy, and society in Western Europe 1815–1975: The growth of mass democracies and welfare states*. Frankfurt: Campus Verlag.

Hix, S., A. G. Noury, and G. Roland (2007). *Democratic politics in the European Parliament*. Cambridge, UK: Cambridge University Press.

Hooghe, L., R. Bakker, A. Brigevich, C. de Vries, E. Edwards, G. Marks, J. Rovny, and M. Steenbergen (2010). Reliability and validity of measuring party positions: The Chapel Hill expert surveys of 2002 and 2006. *European Journal of Political Research 49*(5), 687–703.

Huber, J. D. and R. Inglehart (1995). Expert interpretations of party space and party locations in 42 societies. *Party Politics 1*(1), 73–111.

Høyland, B., I. Sircar, and S. Hix (2009). An automated database of the European Parliament. *European Union Politics 10*(1), 143–152.

King, G. (1995). Replication, replication. *PS: Political Science & Politics 28*(3), 444–452.

Klingemann, H.-D., A. Volkens, J. Bara, I. Budge, and M. D. McDonald (2006). *Mapping policy preferences II: Estimates for parties, electors and governments in Central and Eastern Europe, European Union and OECD 1990–2003*. Oxford: Oxford University Press.

Kollman, K., A. Hicken, D. Caramani, and D. Backer (2010). Constituency-level elections archive (CLEA; www.electiondataarchive.org). February 3, 2010 version [dataset]. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor].

Mackie, T. and R. Rose (1991). *The international almanac of electoral history*. Washington, DC: Congressional Quarterly.

Müller, W. C. and K. Strøm (Eds.) (2000). *Coalition governments in Western Europe*. Oxford: Oxford University Press.

Nohlen, D. (2005). *Elections in the Americas: A data handbook*. New York: Oxford University Press.

Nohlen, D., F. Grotz, and C. Hartmann (2001). *Elections in Asia and the Pacific: A data handbook*. Oxford: Oxford University Press.

Nohlen, D., M. Krennerich, and B. Thibaut (1999). *Elections in Africa: A data handbook*. Oxford: Oxford University Press.

Ray, L. (1999). Measuring party orientations towards European integration: Results from an expert survey. *European Journal of Political Research 36*(2), 283–306.

Steenbergen, M. R. and G. Marks (2007). Evaluating expert judgments. *European Journal of Political Research 46*(3), 347–366.

Trechsel, A. H. and P. Mair (2009). When parties (also) position themselves: An introduction to the EU Profiler. EUI Working Papers RSCAS 2009/65; EUDO – European Union Democracy Observatory.

European
University
Institute

European University Press