Article

EUP

European Union Politics
12(3) 455–477

DOI: 10.1177/1465116511398739
eup.sagepub.com
SAGE

The European Union Policy-Making dataset

Frank M. Häge
University of Limerick, Ireland

Abstract
This article introduces the European Union Policy - Making (EUPOL) dataset. The dataset contains the complete records of the European Commission's PreLex database, which tracks the interactions between the European institutions on legislative proposals and non-legislative policy documents over time. To be of maximum use to the research community, the dataset is both comprehensive and replicable. It relies on 2600 variables to describe the detailed event histories of more than 29,000 inter-institutional decision-making processes between 1975 and 2009. The data collection has been completely automated, enabling scholars to scrutinize and replicate the generation of the dataset. To illustrate the dataset's general utility and discuss specific pitfalls, I present a descriptive analysis of the outcome and duration of Council decision-making.

Keywords
Council of Ministers, dataset, decision-making, efficiency, policy-making, PreLex

Recent years have seen a growing number of quantitative studies on European Union (EU) legislative politics, often basing their analysis at least partly on data gathered from the EU's online databases. To give just a few examples, researchers have used this approach to study the selection of European Parliament (EP) rapporteurs (Høyland, 2006; Kaeding, 2004), the vertical division of labour in the Council structure (Häge, 2007, 2008), and the influence of the EP under the consultation procedure (Kardasheva, 2009). Also, the efficiency of decision-making has received widespread attention (Golub, 1999, 2007; König, 2007; Schulz and König, 2000). Although these and other quantitative studies pursue very different research questions, they often have similar data requirements in that they take

Corresponding author:
Dr Frank Häge, Lecturer in Politics, Department of Politics and Public Administration, University of Limerick, Limerick, Ireland
Email: frank.haege@ul.ie

individual decision-making processes as their unit of analysis and rely on an overlapping set of independent and dependent variables. However, given the lack of a comprehensive, publicly accessible dataset, most researchers engage in their own data collection efforts, which are usually tailored to meet the needs of their particular research question. The resulting datasets are thus of limited use beyond their initially intended purposes. This practice is not only onerous for individual researchers; the resulting duplication of work is also collectively inefficient for the research community.

This article introduces the European Union Policy-Making (EUPOL) dataset. The dataset includes virtually all information contained in the European Commission's online database PreLex.[1] PreLex is maintained by the European Commission and its mission is to monitor the inter-institutional decision-making process. It tracks the progress of legislative proposals and other policy documents submitted by the Commission to the other EU institutions, providing detailed and comprehensive information about various events and actors involved during all stages of the decision-making process, as well as cross-references to documents contained in other online databases. In combination with its long-term coverage since the mid-1970s, PreLex is arguably the most useful online database for studies of EU politics. The EUPOL dataset provides the complete information contained in PreLex in a standardized and machine-readable format. Overall, the dataset covers 29,366 decision-making processes, whose events and event features are described by 2600 variables. Next to the replicability of the data collection process, ensuring the comprehensiveness of data coverage has been the main goal in the development of the dataset. The comprehensive inclusion of all available information ensures maximally effective data provision by avoiding the need to duplicate data collection efforts.

Developing such a comprehensive dataset is almost impossible without automating the data collection process. In the next section, I elaborate on the goal of developing a comprehensive and replicable dataset and the strategies used to achieve that goal. I give a brief overview of how the computer script extracts the information from PreLex and how it represents and saves that information in a format suitable for further statistical processing. Subsequently, I describe the features and coverage of EUPOL in more detail. A comparison with the EU-Lex dataset developed by König and colleagues (König et al., 2006a, 2006b) illustrates the respective strengths and weaknesses of the two datasets. The comparison aims to help researchers in identifying the dataset most suitable for their own purposes. EUPOL's major strength lies in its comprehensive coverage, but it provides only the raw information given in PreLex. In most instances, additional data management and data manipulation will be required to generate and code substantively meaningful variables from that raw information before the actual data analysis can proceed. To illustrate that process and to point to potential pitfalls, I describe the generation of variables for the type of legal act, for the type of legislative procedure, and for the outcome and the duration of Council decision-making.[2]

Finally, a descriptive and exploratory analysis of Council decision-making between 1976 and 2009 sheds new light on the outcome and the speed of negotiations between member states. With respect to the outcome of Council decision-making, the analysis shows that Council negotiations on legislative dossiers hardly fail. In about 89 percent of all cases, negotiations conclude with the adoption of a Council decision. Regarding the process of decision-making, the typical duration of Council decision-making on legislative dossiers has increased considerably over time, and Council decision-making under procedures that grant the EP substantial law-making powers takes considerably longer than decision-making under the consultation procedure. Although the associations of duration with the type of legislative procedure do not necessarily imply a causal connection between the two variables, they raise some interesting questions about the potential effect of EP empowerment that goes beyond the delay caused simply by the formal institutional requirement to reach agreement with an additional veto player.

Replicability and comprehensiveness

In the development of the dataset, I pursued two main goals: replicability and comprehensiveness. To be of general use to the research community, a dataset should contain all available information and it has to be clear how that information was collected. Researchers cannot rely on a dataset whose generation is not fully documented. The replication standard demands that sufficient information is provided so that the results of an empirical study can at least in principle be reproduced (King, 1995). Although the provision of replication datasets on which the results of statistical analyses are based is becoming more and more common, the generation of these datasets themselves is still often insufficiently documented. The use of computer technology to generate datasets by extracting information from online databases provides unprecedented opportunities for a gap-less documentation of the research process, starting with the collection of the data and ending with the presentation of the results of the statistical analysis. In contrast to data collected and coded by humans, the automated extraction process is completely reliable and any repetition of the process in the future will result in exactly the same dataset. All that is required is that the original data sources are permanently stored and that the computer script used to extract the information is made publicly available. When generating the dataset, the proposal pages in PreLex were not just temporarily accessed to extract their informational content; their HTML source code was actually downloaded and locally stored to allow for subsequent replications of the data collection process. Furthermore, all software used during the data collection process is open source software and freely available on the web. Thus, the replication of the data collection process is not only possible in principle but also made practically easy by the use of free and publicly available software tools.[3]

At the same time, the reliability and replicability of the automated extraction process do not guarantee that the information is extracted correctly. The validity of

the extracted information depends crucially on the way the computer scripts extract that information. Just as there are plenty of ways in which the formulation and structure of a questionnaire can bias measurement in survey research, there are plenty of ways in which errors in the way that online information is extracted by a computer script can result in systematic distortions of the extracted information. Many political scientists might not be familiar with automated information extraction procedures and may find the description of those procedures rather technical. Yet they need to be documented transparently, just like any other data collection procedure, if we are to have any confidence in the validity of the generated data. Thus, the remainder of this section gives a brief overview of the download and extraction procedure; the supporting information (SI) in the online appendix (available at the website of this journal) provides more details.

To guarantee the comprehensiveness of the dataset, the information extraction procedure builds the dataset from the bottom up. Crucially, the procedure does not require prior knowledge about the number and type of variables for which information should be extracted. Specifying all possible events and event characteristics contained in PreLex in advance is impossible. Therefore, only a flexible procedure that generates the variables and develops the structure of the dataset during the extraction process and in response to the extracted information can ensure complete coverage. The procedure proceeds in three steps, implemented in the form of three computer scripts written in the programming language Python. These scripts are run sequentially.

The first script downloads the proposal pages of the PreLex database and saves their HTML source code in text files on the local hard drive. The PreLex database is continuously updated by Commission officials. Thus, saving and preserving the proposal pages in their current form ensures that the information extraction process can be replicated in the future even if the online database has been modified in the meantime.

The second script extracts the information contained in the proposal page text files and temporarily stores it in a Python dictionary. The extraction script relies on the structure of the HTML code to identify different decision-making events (for example, 'Adoption by Commission' or 'First reading approval by Council') and event characteristics (for example, 'Primary responsible' or 'Council agenda' – the first characteristic referring to the Directorate General [DG] of the Commission responsible for drafting the proposal, and the second to the type of item the proposal formed on the Council's agenda). The script uses the abbreviated titles of these events and event characteristics to generate variable names and then extracts the information associated with those events and event characteristics to generate variable values. Because events can occur several times and because event characteristics can have several descriptors, the script adds a running number to each of them. For example, if both the event and the event characteristic occurred for the first time in the context of a specific proposal page, the event 'Adoption by Commission' results in the variable name 'adopByComm_date_0', with the corresponding variable value providing the date on which the College of Commissioners

formally adopted the proposal; the associated event characteristic 'Primary responsible' would result in the variable name 'adobByComm_0_primResp_0', with the corresponding variable value providing the name of the Commission DG responsible for drafting the proposal.

Finally, the third script writes the extracted information from the Python dictionary to a comma-separated text file. In the latter format, the data can then be imported into any statistical software package for further processing.

Scope and content

As mentioned above, a major goal in the generation of EUPOL was comprehensive coverage. As a result, the dataset contains all information contained in the PreLex database. Its 2600 variables describe 29,366 decision-making processes relating to legislative (for example, directives, regulations or decisions), non-legislative (for example, working papers, communications or reports) and budgetary proposals (for example, transfer of appropriations) submitted by the Commission between 1975 and the end of 2009.[4] The large number of variables is partly owing to the fact that the same type of event (for example, 'Discussion by Council') can occur several times during the decision-making process, and that a single event descriptor (for example, 'Mandatory Consultation') can refer to several characteristics (for example, 'European Parliament' and 'Economic and Social Committee'). Although all variables present valuable information, it is noteworthy that they cover 680 unique types of events and event descriptors (see the online appendix for a complete list).

EUPOL's comprehensive coverage distinguishes it from the EU-Lex dataset developed by König and colleagues (2006a, 2006b), but the latter has other advantages. To assist potential users in assessing the usefulness of the two datasets for their particular purposes, Table 1 summarizes their main characteristics. EU-Lex is a cross-validated dataset relying on extracted information from two online

Table 1. Comparison of EU-Lex and EUPOL data sets

	EU-Lex	EUPOL
Time period	1 Jan. 1984 – 1 Feb. 2003	~1975 – 31 Dec. 2009
No. of observations	8475	29,366
Total no. of variables	47	2600
Types of document	Decisions, regulations, directives	Decisions, regulations, directives, communications, reports, transfer of appropriations, working papers, opinions, assent, 24 other types of document
Types of variable	Fully coded and labelled	Raw information
Data sources	Celex and PreLex	PreLex

databases, PreLex and CELEX. EU-Lex includes fully coded and labelled variables that can be directly used in a statistical analysis without much additional data-processing. Furthermore, EU-Lex allows the leveraging of the information from both datasets to reduce the number of missing values on particularly important institutional variables such as the legislative procedure or the voting rule. The dataset includes 47 variables related to the following features of the legislative decision-making process (König et al., 2006a): the date of adoption of the proposal by the Commission, the dates of amendments made to the proposal by the Commission, the name of the responsible Commission DG, the date of the conclusion of the legislative process and the type of outcome (adoption, withdrawal, rejection or pending), the type of policy sector, the type of legislative act (directive, regulation or decision), the type of voting rule in the Council, the type of legislative procedure, and the type of agenda item at the last meeting of the Council (A-item, B-item or written procedure). In instances where these variables are sufficient to investigate a research question, EU-Lex has obvious advantages. However, in instances where these variables are not sufficient and further information from PreLex is required, EUPOL is a useful alternative or at least a complementary source.[5]

The information in EUPOL opens up a number of new research opportunities. First, the temporal coverage of EUPOL includes proposals introduced between 1975 and 2009, adding almost 15 years to the 1984 to 2003 coverage of EU-Lex and allowing us to track temporal changes in the late 1970s and early 1980s, as well as in the more recent period after the entry into force of the Nice Treaty. Second, next to legislative dossiers, EUPOL covers all types of document submitted by the Commission to the Council or the European Parliament and tracks their progress over time. Scholars of agenda-setting might find the information on non-binding policy documents most interesting, whereas students of budget politics might find the information on budgetary procedures especially useful. Third, EUPOL includes references to associated documents in the public registers of different institutions, the Official Journal of the EU, press releases, and other databases such as EUR-Lex (the successor to CELEX). For more recent years, these documents are often directly accessible online. Thus, the rather procedural information contained in EUPOL can easily be linked to information about the substance of decision-making derived from automated or manual content analyses of those documents.

Most importantly, EUPOL provides more detailed information on individual events and event characteristics. Amongst other things, it describes all stages of the legislative decision-making process, not just its start- and end-points. To give a few examples, the information contained in EUPOL might be a useful resource for studies on the endogenous institutional choice of legislative procedures, on the influence of rapporteur characteristics on EP decision-making, and on ministerial involvement in Council decision-making. Regarding the institutional choice of legislative procedure, EUPOL contains information about the legislative procedure under which the Commission originally submitted the proposal, any demands for

changes in the procedure made by the Council or the EP, and the final procedure under which the proposal was eventually adopted. This information could be used to study the conditions under which disagreements form about the appropriate legislative procedure and why those disagreements are solved in favour of one or the other institutional actor (for example, Jupille, 2004). EUPOL also contains the name of the EP rapporteur, a link to the document containing the rapporteur's report, the date when the EP adopted its opinion, and information on the type of outcome of the EP decision. By linking this information with information on the rapporteur's party group affiliation and other personal attributes available from other datasets (for example, Høyland et al., 2009), the effect of rapporteur characteristics on the efficiency and outcome of EP decision-making could be investigated. Finally, EUPOL can also be used to shed more light on ministerial involvement in Council decision-making (for example, Häge, 2007). The dataset includes information on all Council meetings at which a formal decision has been taken, including the session number, the Council configuration, links to relevant Council documents and, most importantly, the type of item the proposal formed on the ministers' agenda. To assess the questions of whether or not and how often ministers are directly involved in Council decision-making, we need to examine all meetings during the entire legislative process, not just the meeting at the end of the process. EUPOL provides that kind of information.

Although EUPOL does not consist of readily coded variables and requires additional data manipulation in a statistics programme, it makes the data collection step of the research process obsolete. Most quantitative political science researchers will have the basic data management skills necessary to generate substantively meaningful variables from the information contained in EUPOL, whereas few will have the skills to programme their own extraction procedure. Ideally, we would like to have a comprehensive dataset that consists of a set of fully coded and documented variables. However, a combination of the enormous amount of information available in PreLex and the idiosyncratic information needs of researchers makes the generation of such a dataset practically impossible. Thus, a second-best solution is to provide the complete raw information, which omits the need for data collection while allowing researchers to construct variables tailored to their own specific research needs. The EUPOL dataset contains the information from the PreLex database in a comma-separated text file. This observation-by-variable format can easily be read by any statistical software package. In the following, I first depict the contents of EUPOL and then present a descriptive analysis of the outcome and duration of Council decision-making. In the process, I discuss problematic issues related to the selection of appropriate cases and to the generation of substantively meaningful variables. As the next section will show, many of these data management tasks involve more or less contestable decisions about how best to restrict the temporal and policy domain of the study sample, and how to code and combine information from EUPOL to generate the variables of interest for the analysis. The many uncertainties involved in making these data management decisions illustrate why it is important to provide access to the full raw

information contained in PreLex. Researchers can easily check whether alternative coding options make a difference and construct variables in a way that makes most sense in light of their own research questions.[6]

EU policy-making: A quantitative assessment

The dataset contains information on a large number of legislative, non-legislative and budgetary documents, the overwhelming majority of which have been submitted by the European Commission. Figure 1 presents an overview. Not surprisingly, legislative proposals are amongst the most frequent types of file. The dataset includes 8994 proposals for regulations, accounting for roughly 31 percent of all proposals. Proposals for decisions follow suit, with a frequency of 5850 (20 percent). The dataset also includes 1990 (7 percent) proposals for directives and 294 (1 percent) requests for Council assent. Among the non-legislative documents, communications (4082; 14 percent) and reports (3441; 12 percent) are the most common, followed by working papers (1363; 5 percent) and proposals for opinions (349; 1 percent). Finally, transfer of appropriations is the most frequent budgetary document (1498; 5 percent). The 'other' category comprises the 24 remaining types of file, each individually accounting for less than 1 percent and collectively for less than 4 percent of all documents (see Table SI-1 for more details).

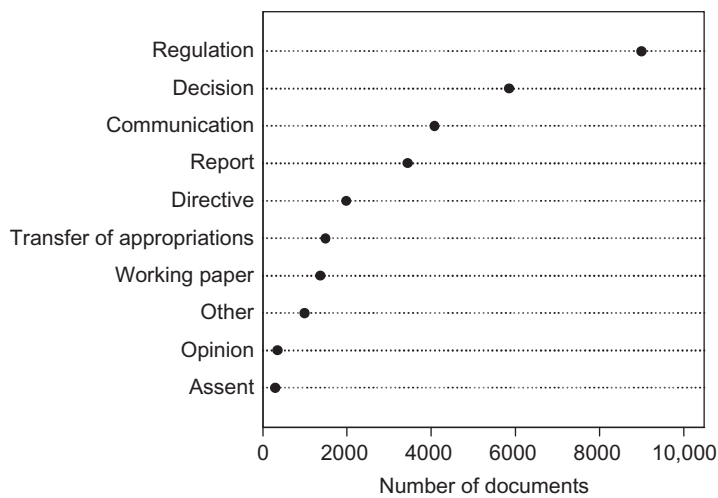

Figure 1. Number of documents by type of file.
Note: The 'other' category comprises 24 types of document, each individually accounting for less than 1 percent of the total number of proposals ($N = 28{,}846$). See Table SI-1 in the online appendix for the detailed statistics underlying this figure.

The categorical 'Type of file' variable used to generate Figure 1 is primarily based on information from the type of file descriptor of the 'Adoption by Commission' event (27,667; 95 percent). Where such information was not available, I replaced the missing values with information from the 'Transmission to Council' event (1425; 5 percent). Yet, before merging the information of the two event variables into a single type of file variable, their descriptor values needed to be recoded. Often, several slightly different descriptor values are used to refer to the same event characteristic. For example, the descriptor values 'regulation', 'proposal for a regulation', recommendation for a regulation' and 'draft regulation' all refer to a proposal for a regulation and had to be recoded to reflect this fact. Similar corrections had to be made to almost all types of file descriptor. Indeed, such inconsistent usage of descriptor values is quite common in PreLex. Finally, after identifying the start date of the policy process by merging the information from the 'Adoption by Commission' date with the 'Transmission to Council' date in a manner similar to the type of file information, I left-censored the dataset. Although the dataset includes a large number of proposals that started in 1975, some indications exist that the coverage for 1975 is not quite complete. In particular, the number of documents submitted during the first half of 1975 seems to be unusually low compared to the following years. Also, years before 1975 are clearly not systematically covered (see also endnote 4 on this point). Thus, for the purposes of this analysis, I excluded all cases that were introduced before 1976. This selection reduces the number of observations from 29,367 to 28,846.

Figure 2 presents another set of statistics of interest to scholars and practitioners of EU legislative politics: the number of legislative proposals submitted under different legislative procedures. Interestingly, only 9303 (32 percent) out of 28,846 PreLex documents were examined under a legislative procedure (see Table SI-2.2). Of those 9303 proposals, 6166 (66 percent) were processed through the consultation procedure, 1170 (13 percent) through the agreement procedure, 1112 (12 percent) through the codecision procedure, 580 (6 percent) through the cooperation procedure and 251 (3 percent) through the assent procedure. The 'other' category comprises 24 cases processed through four very rarely used procedures (see Table SI-2.1 for details).

For generating the procedure variable underlying Figure 2, I relied primarily on the procedure information given by the 'Adoption by Commission' event. When that information was missing, I fell back on the procedure information provided by the 'Transmission to Council' event or the header of the proposal pages. However, this strategy resulted in only 195 additional changes. Another way to find out about the involvement of the EP is to check each case for the occurrence of an 'EP single reading' or an 'EP first reading' event, the former signalling that the proposal was a consultation file and the latter that the proposal was either a cooperation or a codecision file. Cooperation and codecision files can then be distinguished by how they were eventually adopted. The cooperation procedure usually ends through a 'Formal adoption by Council' event, whereas the codecision procedure usually ends through a 'Signature by EP and Council' event. Overall,

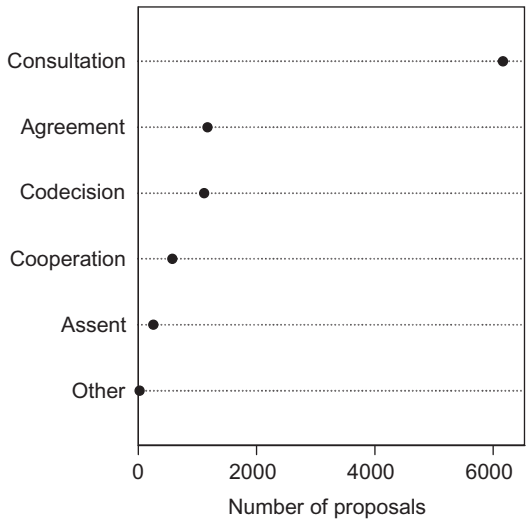

Figure 2. Number of proposals by type of legislative procedure.
Note: The figure plots the number of proposals examined through a legislative procedures ($N = 9303$), which constitute about 32 percent of the proposals in the dataset (total $N = 28{,}846$). The 'other' category comprises 'Consultation European Central Bank', 'Consultation Court of Auditors', 'Social protocol' and 'Special legislative procedure (EP consent required)'. Each of those categories accounts for less than 1 percent of all legislative procedures. See Table SI-2.1 in the online appendix for the detailed statistics underlying this figure.

this recoding procedure resulted in 1328 additional consultation cases and 10 additional cooperation cases.

Finally, I also corrected some obvious errors in the variable's values. Several proposals indicated a type of legislative procedure that did not exist at the time the decision-making process had started. I recoded all codecision and cooperation procedure files to consultation procedure files if they were introduced before the Single European Act came into force in July 1987, and all codecision procedure files to cooperation procedure files if they were introduced before the Treaty of Maastricht came into force in November 1993. These inconsistencies probably occurred because the proposals' legal basis changed sometime during the decision-making process and the corresponding information was 'updated' after the fact. However, since the current analysis is interested in the type of procedure at the start of the decision-making process, these changes are required.

The preceding discussion exposes a weakness of the PreLex database. When a certain descriptor is missing, it is often hard to ascertain whether that descriptor is missing because it is not applicable to the case at hand or because of an oversight when Commission staff entered the information into PreLex. The best that can be done is to utilize all relevant information contained in a proposal page to check

and, if appropriate, to adjust the coding of the variable in question. Another weakness of PreLex is the inconsistent usage and uneven coverage of events and event characteristics. In this case, missing information is the result not of a mistake when entering the information but of systematic differences in the inclusion and coding of different events and event descriptors. Although the current analysis is not directly affected by this problem, much interesting information in PreLex is available for only certain time periods. For example, information about the field of activity is available only before the year 2005, and information about Commission DGs associated with (rather than being responsible for) drafting the proposal is available only before 2001. On the other hand, information about meetings in which the Council discussed but did not decide on a proposal is available only since mid-2000. Thus, the fact that a certain variable exists in the dataset does not imply that it includes information for the entire time period. Any analysis using parts of the data must be preceded by a systematic examination of missing values on the relevant variables.[7]

With these caveats in mind, Figures 3 and 4 present the changes in the volume of non-legislative documents and legislative proposals over time. Figure 3 shows the

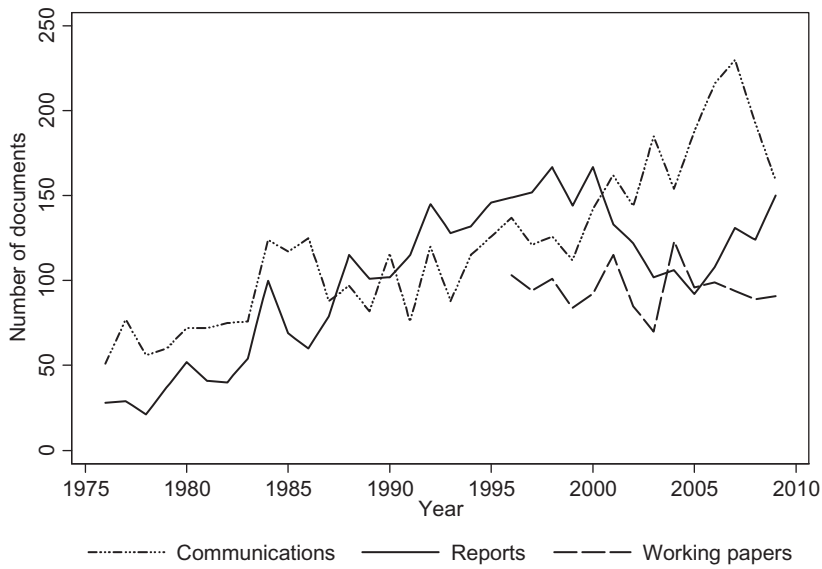

Figure 3. Number of non-legislative documents by type of file, 1976–2009.
Note: The figure shows changes over time in the volume of the three most common non-legislative types of document: communications, reports and working papers ($N = 8886$; 31 percent of all documents in dataset). The yearly number of working papers is essentially zero before 1996, which indicates that this type of document had previously not been documented in the database. See Table SI-3 in the online appendix for the detailed statistics underlying this figure.

yearly number of communications, reports and working papers submitted by the Commission between 1976 and 2009. The number of communications shows a clearly increasing trend over time, with a minimum number of 51 in 1976 and a maximum number of 230 in 2007. The number of reports first steadily increased as well, from a minimum of 21 in 1978 to a maximum of 167 in 2000, but subsequently dropped to 92 in 2005. Only recent years have seen a re-surge to 150 reports in 2009. Before 1996, PreLex records 27 working papers in total, indicating that this type of document was not systematically covered during that period. Starting in 1996, the number of working papers remained relatively stable, fluctuating around a mean of 95 with a standard deviation of 13, reaching a minimum of 70 in 2003 and a maximum of 123 in the subsequent year.

Figure 4 shows the yearly number of proposals for regulations, decisions and directives during the study period. This figure and the remainder of the analysis in this section are based on legislative proposals submitted under the consultation, cooperation or codecision procedure. This selection further reduces the sample from 9303 to 7858 proposals. The number of proposals for regulations shows a negative trend over time, from a high of 198 proposals in 1977 and falling to its

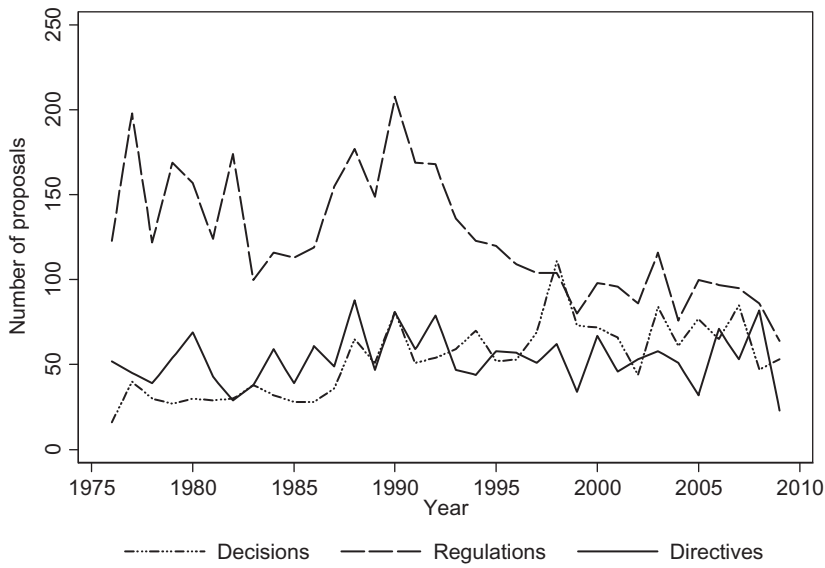

Figure 4. Number of legislative proposals by type of file, 1976–2009.
Note: The figure shows changes over time in the volume of legislative proposals for decisions, regulations and directives ($N = 7858$; 27 percent of all documents in dataset). Note that these figures exclude decisions, regulations and directives that were not adopted through the consultation, cooperation or codecision procedure. See Table SI-4 in the online appendix for the detailed statistics underlying this figure.

current minimum value of 64 in 2009. Only the late 1980s and early 1990s saw somewhat of a reversal of this trend, with the number of proposals for regulations briefly reaching a peak of 208 in 1990. In contrast to the number of proposals for regulations, the number of proposals for decisions actually increased over time – from its minimum number of 16 in 1976 to its current number of 53 in 2009, with a maximum of 111 proposals in 1998. Finally, the number of proposals for directives shows no trend in either direction, varying around a mean of 54 proposals with a standard deviation of 15 proposals. The directives time series reached its maximum value of 88 proposals in 1988 and its minimum value of 23 proposals in 2009. Interestingly, the latter value indicates a huge reduction after the second-highest value of 82 proposals in 2008. In contrast to the visible increase in the number of non-legislative documents, no clear-cut trend exists in the supply of legislative proposals. The number of regulations decreased, the number of decisions increased and the number of directives remained largely the same, although with large fluctuations around its time average (the changes over time of the last two instruments are better visible when the time series are plotted individually; see Figures SI-4.1 and SI-4.2).

Figure 5 looks at changes over time in the frequency of the type of procedure through which the legislative proposals have been considered. By far the largest

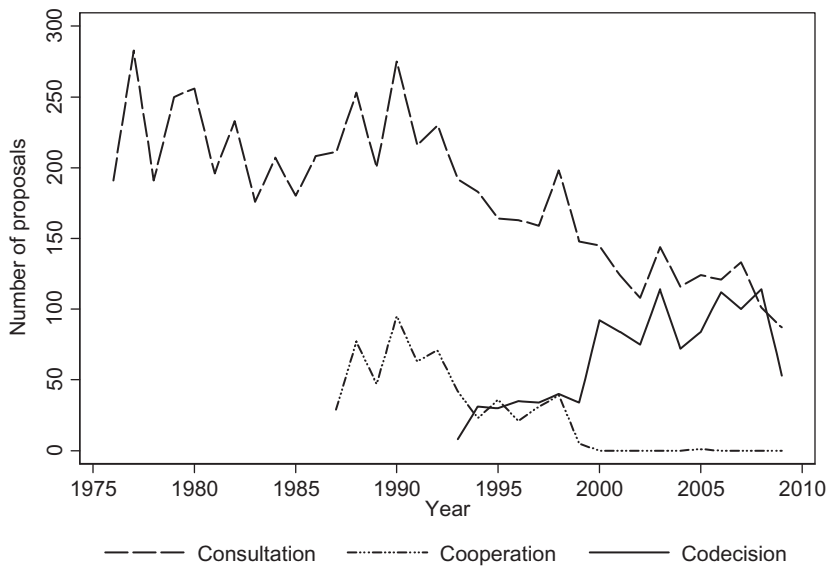

Figure 5. Number of legislative proposals by type of procedure, 1976–2009.
Note: The figure shows changes over time in the volume of legislative proposals submitted under the consultation, cooperation and codecision procedure ($N = 7858$; 27 percent of all proposals in the dataset). See in the online appendix for the detailed statistics underlying this figure.

number of proposals used to be submitted under the consultation procedure, with a maximum number of 283 in 1977. The Single European Act introduced the cooperation procedure in July 1987. The number of proposals examined under this procedure quickly rose to a peak of 95 in 1990. In November 1993, the Treaty of Maastricht introduced the codecision procedure, replacing the cooperation procedure in many policy areas. A corresponding substitution effect is clearly visible in Figure 5, with the decline in cooperation files from 71 in 1992 to 36 in 1995 being roughly similar in magnitude to the rise in codecision files from 0 in 1992 to 30 in 1995.

Finally, the Treaty of Amsterdam, which came into force in November 1999, almost completely replaced the cooperation procedure by the codecision procedure and also expanded the latter's applicability to areas previously governed by the consultation procedure. As a result, the number of proposals considered under the cooperation procedure dropped to zero in 2000 and remained there for most of the remaining time period. The only exception was the year 2005, in which a single proposal was examined under this procedure. In contrast, the number of codecision files rose steadily to its maximum number of 114 in 2008. In this year, the number of codecision files for the first time surpassed the number of consultation files. The year 2009 saw a strong reduction in the number of legislative proposals in general, but the drop in codecision files was larger than the drop in consultation files. Whether this reduction indicates the start of a negative trend or represents only a once-off fluctuation remains to be seen.

The outcome and duration of Council decision-making

In order to give the reader an impression of the breadth of EUPOL's coverage, the figures in the previous section shed light on the supply of proposals and the legislative procedures through which those proposals were processed. This section illustrates EUPOL's added value by examining a topic that has not received much attention in the quantitative literature on EU politics so far: the outcome and duration of Council decision-making.[8] By Council decision-making, I refer to the first formal decision of the Council during the legislative procedure. In the case of consultation, this coincides with the adoption of the legislative act. In the case of procedures that grant additional powers to the EP, this usually refers to the adoption of the Council's common position. In the remaining stages of those procedures, this common position then acts as the collective bargaining position of the Council in negotiations with the EP. Thus, the negotiations among member states in the Council largely take place during the first reading stage; the latter stages of the procedure are mostly concerned with finding a compromise with the Parliament (for example, Bostock, 2002: 219–22).

Figure 6 plots the relative frequency of different types of Council decision-making outcomes. More precisely, the figure distinguishes between the adoption, the partial adoption (which implies partial withdrawal or partial replacement), the withdrawal and the replacement of the proposal. It also indicates the proportion of proposals that are still pending. Proposals have been coded as having been adopted

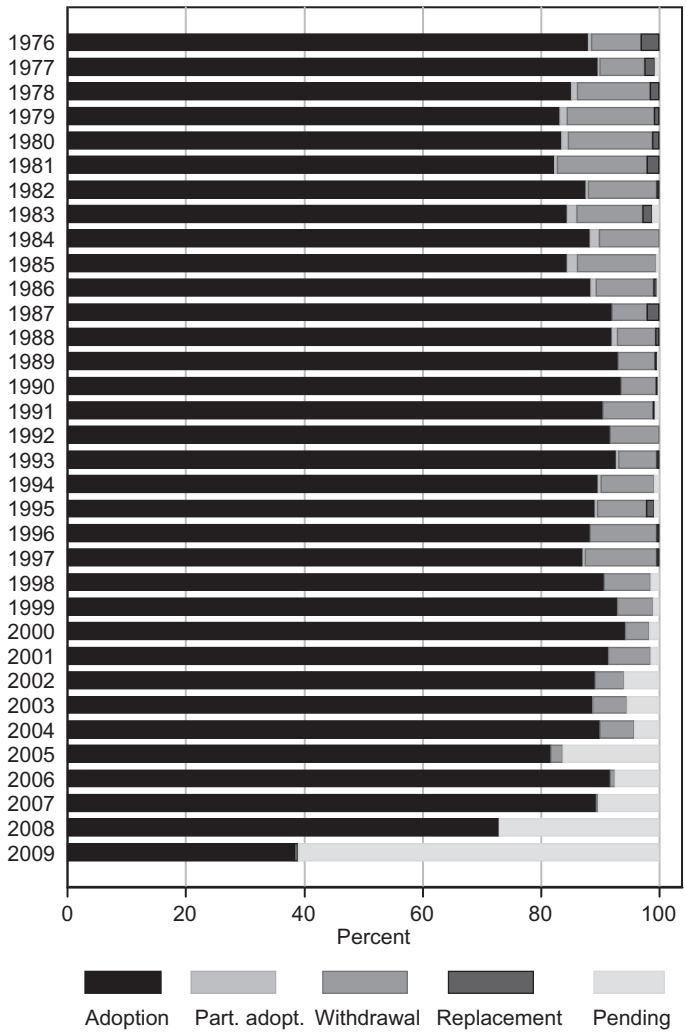

Figure 6. Outcome of Council decision-making by start year, 1976–2009.
Note: The figure shows the percentage of legislative proposals that were adopted by the Council, partially adopted by the Council, withdrawn by the Commission, replaced by the Commission, and still pending. Council adoption includes formal adoption of the act under the consultation procedure, approval at first reading under the codecision procedure, and adoption of a common position under the codecision or cooperation procedure. Legislative proposals refer to proposals for directives, regulations and decisions submitted under the consultation, cooperation or codecision procedure ($N = 7858$). See Table SI-6.1 in the online appendix for the detailed statistics underlying this figure. Table SI-6.2 provides the absolute frequencies.

by the Council if there was a 'Formal adoption by Council' (under the consultation procedure), a 'Council approval 1st reading' (under the codecision procedure) or an 'Adoption common position' event (under the cooperation or codecision procedure). In the absence of all of these events, the proposal has been coded as partially adopted if there was a 'Partial adoption by Council' event, as withdrawn if there was a 'Withdrawal by Commission' event, and as replaced if there was a 'Replacement' event. Proposals that indicated none of those events were treated as still pending.

The figure displays a very large success rate of Council negotiations on legislative dossiers. Note that the success of Council negotiations refers not only to an agreement amongst member states but also to an agreement between member states and the Commission. The Commission attends all Council meetings and can withdraw its proposal at any stage of the process before the Council has adopted a formal decision. Thus, the adoption of a formal Council decision also implies that the Commission has not objected to that decision. The fact that most withdrawals occur in groups at periodic intervals indicates that withdrawals are the result of gridlock amongst member states rather than disagreements with the Commission. However, in general, the withdrawal by the Commission might reflect either a blockage in the Council or a genuine objection by the Commission.

Figure 6 shows that many of the proposals introduced during the last two years of the study period are still pending. However, looking at the period up to and including the year 2007, the Council and EP adopted 89.2 percent of all proposals completely and 0.4 percent partially. Only 8.1 percent were withdrawn and 0.6 percent replaced by the Commission, and another 1.8 percent had not been concluded yet (see Table SI-6.3). During that period, the lowest success rate of 82.8 percent was reached in 1981, and the highest success rate of 94.3 percent in 2000. The figure indicates that the success rate might have increased somewhat in the long term but, if at all, this tendency is quite weak compared with short- and medium-term fluctuations.

Relying on the same event information from PreLex as for the construction of the type of decision-making outcome variable, I constructed a variable indicating the end-date of the Council decision-making process. In cases where the same type of event occurred several times, I used the date of the last event as the end-date. The reasoning behind this coding choice is that Council decision-making cannot have been completed if there were further adoption events at later points in time. Subtracting the start-date from the end-date variable yielded a variable indicating the duration of Council decision-making. Unfortunately, proposals that fail to be adopted are often not immediately withdrawn by the Commission but are left dormant until a general review of pending proposals at a later date finds that they are no longer topical. Thus, the withdrawal date in PreLex will often not closely correspond to the date of the actual failure of the proposal in the decision-making process. Although decision-making processes that ended with a withdrawn proposal might be expected to be somewhat longer than those that ended with an adopted proposal, the data suggest that the median duration for processes

that ended with a withdrawn proposal is about eight times longer than the median duration for processes that ended with the adoption of the proposal (see Figure SI-7.1). Such a large difference seems rather implausible.

Thus, Figure 7 plots only the duration of Council decision-making for proposals that have actually been adopted, completely or partially. Even with the restriction to adopted proposals, the duration variable is extremely right-skewed (see Figure SI-7.2). To enhance readability, Figure 7 does not show a number of extremely large values lying outside the range of the box-plot's whiskers (see Figure SI-7.3 for a plot including the outside values). Because of the large proportion of still pending proposals that were introduced during 2008 and 2009, I restrict the further analysis to proposals introduced by the end of 2007. Although Figure 7 shows a consistently large variability in duration values between 1976 and 2007, it also indicates that the typical duration of Council decision-making, in the form of its median value, has considerably increased over time.

Figure 8 illustrates this change in the median duration of Council decision-making more clearly. Over the entire time period studied, the median duration of Council decision-making more than doubled, from 145 days in 1976 to 303.5 days in 2007. Its maximum value of 430 days was reached in 2006, the second-last year of the study period. To make it easier to distinguish the short- to medium-term fluctuations from long-term developments, the observed medians are overlaid with a median spline scatter-plot smoother. The scatter-plot smoother indicates that the increase in duration did not occur in the form of a constant trend over time. Despite some large fluctuations, the median duration stayed relatively stable during the late 1970s and throughout the 1980s. In fact, the year 1987 had the lowest median duration of just 125 days. Much of the growth in the duration of Council decision-making started only in the early to mid 1990s and gained renewed impetus around the turn of the millennium.

Figure 9 gives an indication of the relative contribution to this increase in the aggregate duration of Council decision-making by files examined under different legislative procedures. The dashed line in the figure represents the aggregate median duration of all files. The aggregate duration is compared with the duration of subsets of files decided according to the consultation, cooperation and codecision procedures, respectively. Again, it has to be stressed that the duration under the last two procedures refers to the Council's first-reading decision, not to the procedure as a whole.

The duration of Council decision-making under the cooperation and codecision procedure does not indicate a clear trend over time in either direction. In a comparative perspective, Council decision-making under procedures with EP involvement takes consistently longer than under the consultation procedure, contributing to a consistently larger aggregate duration since the entry into force of the Single European Act in 1987. Yet this contribution remains largely constant for most of the time period. In contrast to procedures that grant more substantial powers to the EP, the duration of files decided under the consultation procedure has been generally on the rise since around 1993. With the exception of the years 2004 and 2005,

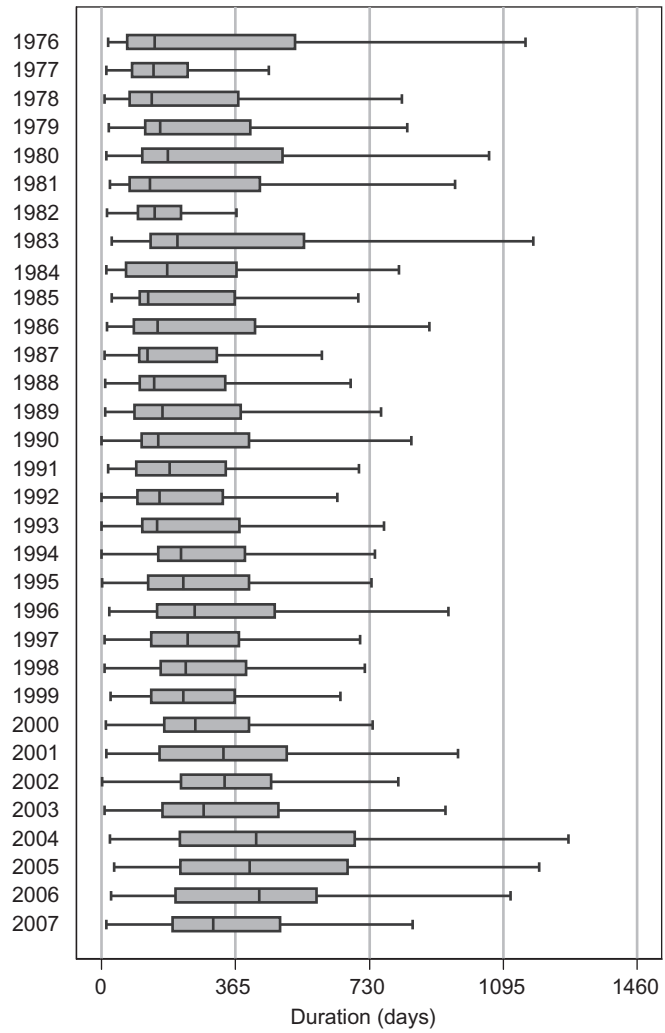

Figure 7. Duration of Council decision-making by start year, 1976–2007.
Note: The figure shows box-plots of the distribution of the duration in days of Council decision-making on legislative dossiers, conditioned by the year in which the proposal was submitted. The figure is based on proposals that have been adopted or partially adopted ($N = 6720$; 89.6 percent of all legislative dossiers [$N = 7503$] submitted during that time period). Proposals that the Commission has withdrawn or replaced are not included ($N = 648$; 8.6 percent); neither are proposals that were still pending at the time of the data extraction in April 2010 ($N = 135$; 1.8 percent). Many of the conditional distributions include a number of extreme outliers that are larger than the adjacent value used to determine the end of the right whisker. These outside values have been omitted from the box-plots to increase the readability of the graph. See Figure SI-7.3 in the online appendix for a graph including the outside values and Table SI-7 for the detailed statistics underlying Figure 7.

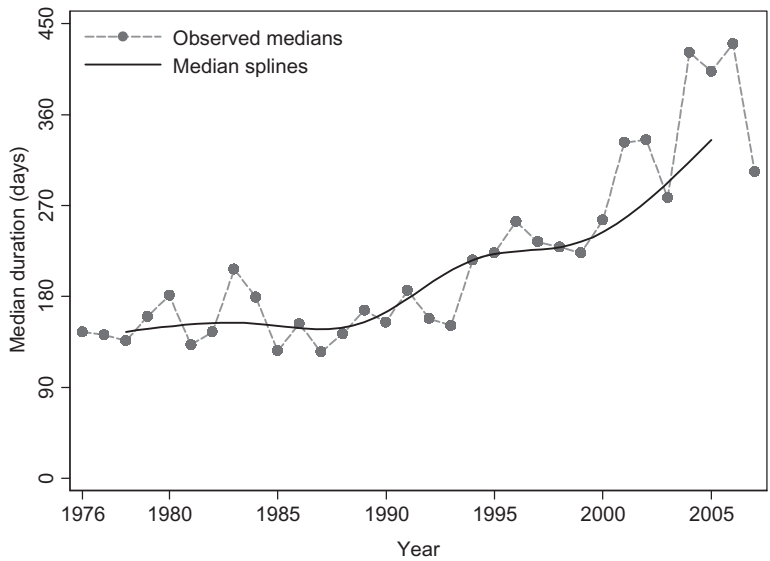

Figure 8. Median duration of Council decision-making by start year, 1976–2007.
Note: The figure plots the median duration of Council decision-making by start year over time. The curve is a median splines scatter-plot smoother. For further information about the sample, see note to Figure 7.

which also saw a large jump in the duration of codecision files, much of the increase in the aggregate duration seems to mirror this increase in the duration of consultation files. To be very clear, the associations of the level of and the change in the duration of Council decision-making with the type of procedure do not necessarily imply the existence of causal relationships. Indeed, as the institutional rules governing interactions under the consultation procedure did not experience any change over time, the increase in Council decision-making duration under this procedure cannot possibly be causally related to features of the procedure itself. However, these associations raise a number of interesting questions for further research that will be discussed in the concluding section.

Conclusion

This article introduces the EU Policy-Making (EUPOL) dataset. The dataset includes virtually all information contained in the Commission's online database PreLex, which monitors the inter-institutional decision-making process of the EU. Next to the replicability of the data collection process, comprehensive coverage was a major goal in the development of the dataset. Only a comprehensive and transparently generated dataset can provide maximum value to the research community. In line with the goal of transparency, the paper first described the automated generation of the dataset, including the download of the relevant PreLex proposal

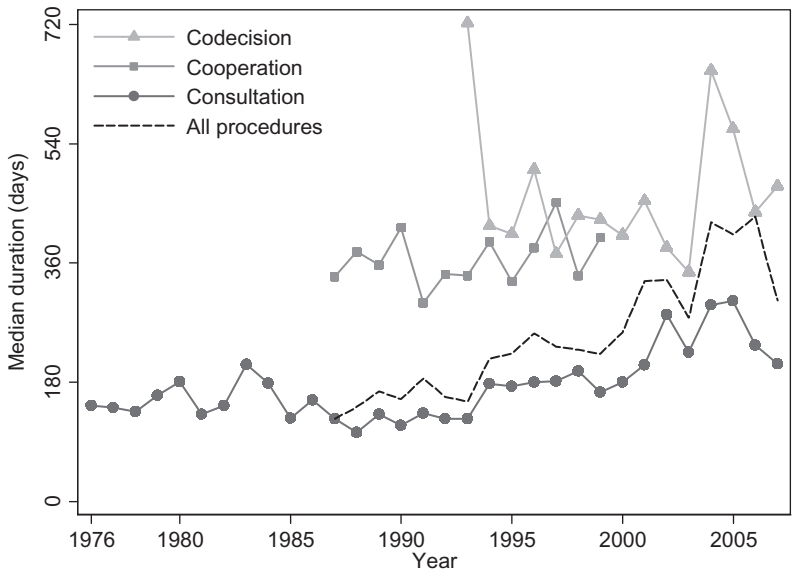

Figure 9. Median duration of Council decision-making under different legislative procedures, 1976–2007.

Note: The figure compares the median duration of Council decision-making under different procedures (solid lines with different marker symbols) with the aggregate median duration across all procedures (dashed line). Note that the extremely high median value for the codecision procedure in 1993 is based on only 8 cases. For further information about the sample, see note to Figure 7.

pages, the information extraction process, and the storage of the final dataset. Each of these steps can be fraught with error, so it is important to explicate them. The second section discussed the content of the dataset and its possible uses in more detail, arguing that the new information provided by EUPOL will be useful for studying a wide variety of questions interesting to students of EU politics.

The EUPOL dataset is not an off-the-shelf dataset ready to use. It provides the raw information as recorded in PreLex. Further data-processing will often be needed to transform that information into substantively interesting variables suitable for a statistical analysis. Nevertheless, EUPOL offers the potential to remove the rather resource-consuming data collection step from the research process of quantitative analysts requiring information from PreLex. Given substantive knowledge about EU politics and basic data management skills, meaningful variables can be computed from the information contained in EUPOL with relative ease. To illustrate this process, the third and fourth sections of the paper not only presented descriptive statistics of the content of EUPOL and examined some basic features of Council decision-making, but also described how the relevant variables were generated from the raw data.

To demonstrate the usefulness of the detailed event information contained in EUPOL, I examined the outcome and duration of Council decision-making. The description of the type of decision-making outcome highlighted the high success rate of Council negotiations. Overall, the Council was able to reach an agreement on about 89 percent of all legislative proposals submitted by the Commission between 1976 and 2007. Council negotiations failed in less than 10 percent of all cases. The analysis of the time it takes the Council to reach its first formal decision in the legislative procedure revealed a substantial increase in duration over time, starting in the early to mid 1990s. Although Council decision-making under the codecision and cooperation procedures takes consistently longer than under the consultation procedure, the increase over time in the aggregate median duration of Council decision-making seems to be mainly related to factors increasing the length of Council decision-making under the consultation procedure. The median duration of Council decision-making under the codecision procedure showed a marked increase only in the years 2004 and 2005.

These patterns raise a number of questions: First, why would the Council need more time to reach a decision under the codecision or cooperation procedures than under the consultation procedure? Despite the recent increase in so-called 'early agreements', the EP did not directly engage in negotiations with the Council during the first-reading stage of those procedures for much of the time period studied. Thus, the need to negotiate a compromise with the EP cannot be directly responsible for the consistently longer duration of Council decision-making under procedures with EP involvement. Is the longer duration an indirect result of a more subtle feature of those procedures (for example, differences in political and public scrutiny; see Häge, 2011) or is it caused by an underlying third factor that is associated with both the type of procedure and Council decision-making duration (for example, differences in the salience of policy issues)?

Another interesting point relates to the differential development of decision-making duration over time. Why has the duration of Council decision-making increased under the consultation procedure but generally not under the cooperation and codecision procedures? Obviously, the increase under consultation cannot simply be due to features of the procedure itself, because they remained constant over time. Yet many temporally changing factors, such as the increasing number of member states through successive enlargements, should have had a similar effect on Council decision-making regardless of the procedure. The differential pattern in temporal development raises questions about the extent to which Council decision-making speed is governed by a homogeneous set of causal forces that combine in an additive manner. Maybe some temporally changing factors influence only certain types of procedure? At the very least, the differential pattern points to the possibility of interaction effects between temporal changes and features directly or indirectly related to the type of procedure. Clearly, without a more systematic study of the potential causes, we can only speculate about the relevant factors. This short exploratory study thus raises more questions than it can possibly answer. However, at the same time, it illustrates the potential of EUPOL to

create novel factual insights and generate a host of opportunities for future research.

Funding

This work was supported by a postdoctoral fellowship grant by the Netherlands Institute of Government (2008).

Notes

1. PreLex can be accessed at http://ec.europa.eu/prelex/apcnet.cfm?CL=en (accessed 2 July 2010).
2. The figures and tables based on those variables provide basic but surprisingly hard-to-come-by statistics (for example, for teaching purposes) about important features of EU policy-making. The paper includes only the figures, but the corresponding tables are printed in the supporting information in the online appendix and can be downloaded as Excel files from http://www.frankhaege.eu.
3. The PreLex HTML files, the extracted dataset and the Python download, extraction and storage scripts are available for download at http://www.frankhaege.eu. The download and extraction scripts were written in Python 2.6.5, using ActiveState's PythonWin editor (http://www.activestate.com/activepython/downloads [accessed 2 July 2010]). The scripts relied on the following external Python modules:
   BeautifulSoup (http://www.crummy.com/software/BeautifulSoup/ [accessed 2 July 2010]),
   ClientForm (http://wwwsearchsourceforge.net/old/ClientForm/ [accessed 2 July 2010]),
   Mechanize (http://wwwsearchsourceforge.net/mechanize/ [accessed 2 July 2010])
4. The exact starting date for the coverage of PreLex is unknown; the database includes a few proposals that were submitted before January 1975, but there is a very clear jump in the number of proposals at that point in time. In the past, the online documentation of PreLex stated that the database is complete as of 1976 (see also König et al., 2006b: 556), so this year might be a more conservative cut-off point than 1975 for using the data in an analysis. The HTML file download was conducted on 8 and 9 April 2010. Although the download did not include proposals introduced later than 31 December 2009, the downloaded files include information on the progress of proposals introduced before that date until the download date in April 2010.
5. The two datasets can easily be merged to enjoy the best of both worlds.
6. The data management and analysis for this part of the paper were conducted in R 2.10.1 (R Development Core Team, 2009) and Stata/SE 11.1 (StataCorp, 2009). All datasets, the R-script and the Stata do-files are available at http://www.frankhaege.eu.
7. I thank one of the anonymous reviewers for reminding me of this important point.
8. A partial exception is the study by Toshkov and Rasmussen (2010), which relies on data derived from the EP's legislative observatory database. Their study's substantial and temporal scope is narrower in that it focuses exclusively on first-reading decisions under the codecision procedure during the time period 1997–2009.

References

Bostock D (2002) Coreper revisited. *Journal of Common Market Studies* 40(2): 215–234.

Golub J (1999) In the shadow of the vote? Decision making in the European Community. *International Organization* 53(4): 733–764.

Golub J (2007) Survival analysis and European Union decision-making. *European Union Politics* 8(2): 155–179.
Häge FM (2007) Committee decision-making in the Council of the European Union. *European Union Politics* 8(3): 299–328.
Häge FM (2008) Who decides in the Council of the European Union? *Journal of Common Market Studies* 46(3): 533–558.
Häge FM (2011) Politicising Council decision-making: The effect of European Parliament empowerment. *West European Politics* 34(1): 18–47.
Høyland B (2006) Allocation of codecision reports in the fifth European Parliament. *European Union Politics* 7(1): 30–50.
Høyland B, Sircar I and Hix S (2009) An automated database of the European Parliament. *European Union Politics* 10(1): 143–152.
Jupille J (2004) *Procedural Politics: Issues, Influence, and Institutional Choice in the European Union*. Cambridge: Cambridge University Press.
Kaeding M (2004) Rapporteurship allocation in the European Parliament. *European Union Politics* 5(3): 353–371.
Kardasheva R (2009) The power to delay: The European Parliament's influence in the consultation procedure. *Journal of Common Market Studies* 47(2): 385–409.
King G (1995) Replication, replication. *Political Science and Politics* 28(3): 443–499.
König T (2007) Divergence or convergence? From ever-growing to ever-slowing European legislative decision making. *European Journal of Political Research* 46(3): 417–444.
König T, Luetgert B and Dannwolf T (2006a) EU-Lex: EU legislative databank based on Celex/Prelex. University of Mannheim, Mannheim. Available at: http://www.sowi.uni-mannheim.de/lsp012/08downloads01.html (accessed 20 June 2010).
König T, Luetgert B and Dannwolf T (2006b) Quantifying European legislative research: Using Celex and Prelex in EU legislative studies. *European Union Politics* 7(4): 553–574.
R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Schulz H and König T (2000) Institutional reform and decision-making efficiency in the European Union. *American Journal of Political Science* 44(4): 653–666.
StataCorp (2009) *Stata Statistical Software: Release 11*. College Station: StataCorp.
Toshkov D and Rasmussen A (2010) Duration of European Union co-decision: Myth and reality. Paper presented at the Fifth Pan-European Conference on EU Politics, 23–26 June, Porto.