Facilitating access to data on European Union laws

Article *in* Political Research Exchange · January 2021

DOI: 10.1080/2474736X.2020.1870150

Michal Ovádek
University of Gothenburg

POLITICAL RESEARCH EXCHANGE
PRX

Political Research Exchange
An ECPR Journal

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/prxx20

Facilitating access to data on European Union laws

Michal Ovádek

To cite this article: Michal Ovádek (2021) Facilitating access to data on European Union laws, Political Research Exchange, 3:1, e1870150, DOI: 10.1080/2474736X.2020.1870150

To link to this article: https://doi.org/10.1080/2474736X.2020.1870150

Published online: 14 Jan 2021.

POLITICAL RESEARCH EXCHANGE
2021, VOL. 3, 1870150
https://doi.org/10.1080/2474736X.2020.1870150

RESEARCH NOTE

OPEN ACCESS

Facilitating access to data on European Union laws

Michal Ovádek

Centre for Empirical Jurisprudence, KU Leuven, Leuven, Belgium

ABSTRACT
Much of empirical research on European integration relies on data published by European Union (EU) institutions concerning the EU's laws and policies. However, despite wide disciplinary appeal and increasing regularization of data repositories such as Eur-Lex, researchers have so far not developed a standardized tool for accessing the main EU law databases. This research note presents the benefits of creating an open-source data collection infrastructure that takes advantage of the structured way in which data is published by the EU. I exemplify how software implementing this idea can be used by researchers.

ARTICLE HISTORY
Received 12 October 2020
Accepted 23 December 2020

KEYWORDS
Database; R-programming; European Union; law and courts; political methodology

Introduction

Countless studies across a wide range of disciplines exploit data on European Union (EU) laws. From legislative politics (e.g. Rauh 2020; Hurka and Steinbach 2020; Blom-Hansen 2019; Kardasheva 2013) through law and economics (e.g. Lampach, Wijtvliet, and Dyevre 2020; Marciano and Josselin 2002) to computational linguistics (e.g. Caled et al. 2019; Chalkidis et al. 2019; Quaresma and Gonçalves 2010), data on EU laws and policy documents represent core empirical material for the study of European integration. Although usually each study comes with its own bespoke dataset, the data is frequently collected from the same source, the Eur-Lex website,[1] which aggregates documents from EU institutions (Düro 2009; Bernet and Berteloot 2006).[2] The result is a sort of imperfect duplication, whereby researchers create local copies of different parts of the Eur-Lex (or related) database, using data collection methods that are (un)specified to various degrees.[3]

This research note presents an alternative approach to collecting data on EU laws and policies. By leveraging dedicated application programming interfaces (APIs), the open-source software outlined below makes access to data easier, faster and more transparent. The next section describes the main contribution – the *eurlex* R package – followed by a discussion of its advantages over the status quo and a reflection on the relationship between the research community and data providers. The final section shows concrete examples of how the software package can be used in applied research.

CONTACT Michal Ovádek michal.ovadek@kuleuven.be

The *eurlex* package

The *eurlex* package was developed to facilitate access to vast amounts of EU law data for researchers. It is written in and as an extension to the popular statistical programming environment R.[4] The main contribution of the package is to function as a curated, simplifying 'gateway' to APIs and data provided by EU institutions and bodies, including the EU Publications Office which maintains the Eur-Lex portal.[5] The APIs are written using a semantic web technology called SPARQL, which few applied researchers will have the time to learn. The *eurlex* package allows users to benefit from the efficiency and structure of the APIs without having to learn the SPARQL language and other technicalities related to data retrieval. In contrast, knowledge of R is comparatively widespread among researchers and the package could be ported without major difficulties into other popular high-level languages such as Python. There are currently no other packages providing comparable functionality. Although the primary beneficiary of the package is academic research, it is also suitable to be used in teaching, in particular as part of methods or data science courses. The package makes it easier for instructors to teach data manipulation, analysis and visualization in R with domain-relevant datasets without having to manually curate them.

The package operates atop a vast infrastructure created by the EU to organize and publish its data. Although each EU institution has its own system and website for publishing documents (sometimes with structured metadata), a central role in the publication of official EU documents is played by the EU Publications Office, a unique inter-institutional office charged with publishing, archiving and bibliographic tasks. The Eur-Lex portal is the most visible access point for the majority of EU publications, and it sits on top of a complex 'content and metadata repository' called CELLAR (Francesconi et al. 2015). The *eurlex* package therefore provides access to data collated and indexed by the Publications Office, but it is important to remember that key decisions – what is published and in what form – remain in the hands of the authoring institutions.[6]

The core functionality of the package revolves around SPARQL queries requesting data from CELLAR. The queries were written with the help of the Eur-Lex helpdesk to ensure they return valid results. They queries are created automatically based on the user's input. For example, the function call 'elx_make_query(resource_type = 'directive', include_lbs = TRUE)' creates a SPARQL query that will request all directives and their legal bases from CELLAR. Subsequently, the user simply executes the query with 'elx_run_query' and obtains a 'data.frame' (an R spreadsheet) with results in return. The results are generally returned orders of magnitude faster than if the same data is scraped from the website. The package comes with documentation explaining the full gamut of options on offer, some of which are exemplified below.[7] In addition to downloading essential metadata about EU legislation and policy documents, the *eurlex* package facilitates access to legislative texts, votes in the Council and the complete list of judicial cases from the Court of Justice, the General Court and the Civil Service Tribunal. Development of the open-source code is continuing, and more features will be added in the future. For example, the package could include functions to access data on internal decision-making processes inside the Parliament, the Commission and the Council. Anyone can review code or contribute ideas for improvement through the associated Github page.[8]

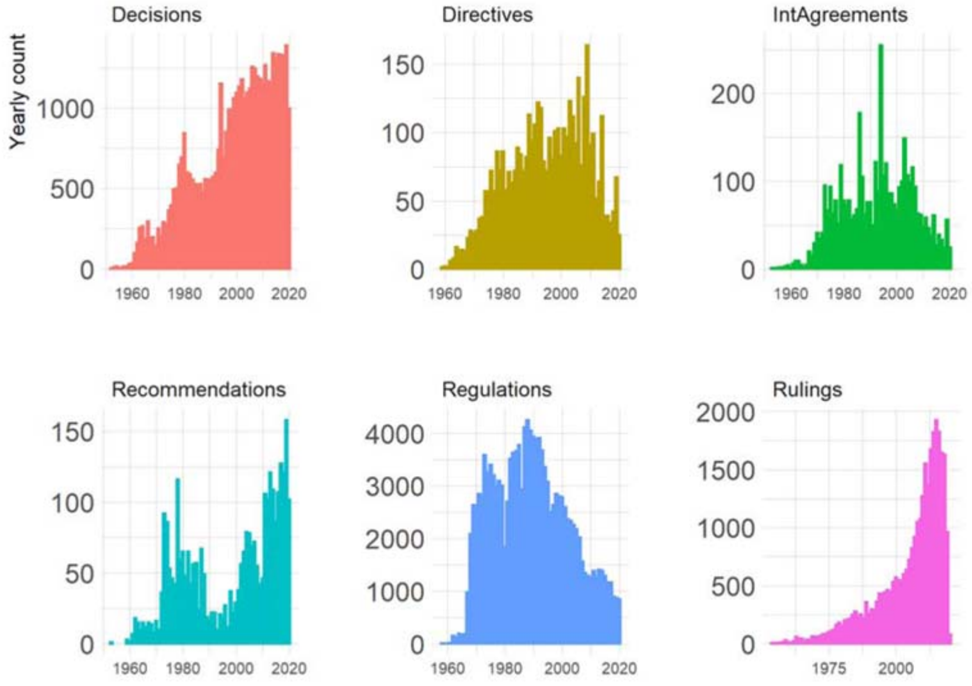

Figure 1. Number of legal acts and court rulings accessible through the eurlex R package

In its current version (0.3.4), the package gives access to hundreds of thousands of records which cover all core output of EU law and policy making, as well as judicial proceedings. To give a sense of what is on offer, as of late 2020, users can obtain data on approximately 52 000 decisions, 4300 directives, 3150 recommendations, 140 000 regulations, 5000 international agreements and 31 000 court rulings (Figure 1). These as well as additional data sources accessible through the package are further discussed below.

A better way to collect data

Using the aforementioned infrastructure for data collection brings five distinct advantages over current data collection practices: efficiency, dynamism, openness, replicability and commonality.

Leveraging designated APIs is efficient for both users and providers. Being able to rely on existing open-source software means not only that coding effort is not unnecessarily duplicated, but also that applied researchers have more time to focus on substantive questions rather than learning how to write SPARQL queries. Specialization helps with breaking down complexity. Moreover, data collection through APIs is faster than web scraping (to say nothing of manual collection) – the *eurlex* package cuts the time needed to collect certain data from days to hours. Importantly, the data provider can benefit as well through optimizing performance to the different demands of API and website access. In addition, this method is also more efficient in the sense that researchers automatically obtain the most up-to-date information at the moment of data collection, something existing datasets cannot provide (Fjelstul 2019). In contrast, applying web scraping techniques typically relies on first (inefficiently) creating an updated list of data identifiers. The ability to tap into a continuously updated stream of data opens new possibilities in both research workflow and paper publication. Using tools such as RMarkdown it is possible to fully integrate computation into the paper-writing process. With programmatic data collection, scientific papers, which normally take months or years to finalize, can be published with fully up-to-date data and analysis, reducing the significant lag that arises between the moment data collection is completed (normally early in the workflow) and the paper being published. The potential implications for scientific publishing of analysing a continuous stream of data, rather than time-bound snapshots, are considerable.

Second, open-source software, such as the *eurlex* R package, bring a much-needed dose of openness to data collection. Because every user can see precisely what each function does, the data collection process becomes transparent. As a result, any mistakes in the code are similarly visible and discoverable by any person at any time. In contrast, data collection not based on open-source code calls for a larger degree of trust in the processes underlying the creation of a new dataset. While all data collection should be transparent, only open-source, programmatic data collection makes the entire process visible outside the research team. What is more, the new infrastructure advances the worthy cause of data transparency without additional investment of resources on the part of the researchers (Wuttke 2019, 6). Transparency is baked in.

Closely linked to openness is the question of replicability. The proposed approach to data collection reduces some of the costs associated with replication and greatly enhances the possibility to inspect the data collection stage. Because the available data is more or less fixed – and ideally transparently managed – by the data providers, data is more likely to be compromised by researchers, be it through oversight or malpractice. With the *eurlex* package it is now possible to include the verification of data collection as part of a basic research integrity check conducted by journals.

Fourth, the proposed data infrastructure would create a *common* basis for empirical research in EU studies. Currently, considerable resources need to be spent on bridging related datasets. A switch to a common data infrastructure would make datasets created by different researchers more interoperable. Perhaps even more importantly, the use and development of open-source software fosters community exchange among researchers. Open-source software, such as the *eurlex* package, creates a point of reference for debates on the quality of data, faulty code, errors and omissions on the part of the data provider and more. These debates, in turn, can improve the data infrastructure for all researchers, as both the open-source software providing access to data and the open data themselves are public goods.

Relationship between researchers and data providers

Making reliance on open data APIs explicit calls for additional reflection about the relationship between the research community and data providers – the EU Publication Office, the Council, the Court of Justice, etc. The idea of more closely connecting data collection in research with data publication by providers should not be accompanied by uncritical acceptance of said data. On the contrary, researchers should insist on increasing transparency about how the data are generated and who controls their flow. The EU's open data practices rank it among the more transparent public institutions in the world, but there remains room for improvement. The Eur-Lex dataverse brings together most EU legal and policy resources, but the integration is not complete and individual institutions retain significant control over which, how and when data are published.

The publication of massive amounts of data, some of which depend on human input, inevitably entails errors and omissions. Researchers need to remain vigilant when it comes to the quality of data transmitted from the providers. Data collected through the *eurlex* package can only be as good as the underlying database maintained by the EU Publications Office allows. Nonetheless, a common data infrastructure makes it easier to spot problems. It is not humanly possible for any single researcher to check the completeness and accuracy of the hundreds of thousands of resources.

Data providers should, on their part, establish closer relations with researchers. A positive externality of doing research on the EU is discovering the various problems with official data. From missing CELEX identifiers,[9] through unpublished court cases[10] to outright errors or omissions in titles,[11] legal bases,[12] dates, etc., researchers frequently stumble upon issues that, if left unattended, affect everyone using the data, likely without their knowledge.[13] At the moment, there is no structured way of tapping into this positive externality created by EU research. At the very least, data providers should include a 'report issue' button on their websites to support the establishment of a virtuous cycle between the publication of open data and research.

Examples

Potential use cases of the *eurlex* package are broad. Studies across political science, economics, law and computational linguistics can immediately benefit from the easier and faster access to EU data. Nonetheless, it is worth demonstrating with basic examples what kind of data can be obtained with the package.

For the purposes of the first example I collect all regulations, directives, decisions and recommendations – legal acts according to Article 288 of the Treaty on the Functioning of the European Union (TFEU) – along with some of their basic properties like date and legal basis. Following some minimal data wrangling, obtaining a completely up-to-date and ready-for-use dataset with nearly 200 000 observations of the main EU legal acts is a straightforward task with the *eurlex* package. Table 1 shows an excerpt of the dataset.

Table 1. Excerpt from a dataset of EU legal acts.

CELEX	Type	Date	EuroVoc topics	Legal basis
32020L1151	Directive	2020-07-29	excise duty; approximation of laws~	12016E113
32020L1057	Directive	2020-07-15	carriage of goods; road transport~	12016E091
32020L0876	Directive	2020-06-24	fraud; infectious disease~	12016E115;12016E113
32020L0739	Directive	2020-06-03	infectious disease; health risk~	32000L0054
32020L0700	Directive	2020-05-25	infectious disease; European standard~	12016E091
32020L0612	Directive	2020-05-04	European standard; technical standard~	32006L0126
32020L0432	Directive	2020-03-23	vegetable; plant taxonomy~	32002L0055
32020L0367	Directive	2020-03-04	illness; public health~	32002L0049
32020L0284	Directive	2020-02-18	information storage; VAT~	12016E113

As the Eur-Lex portal itself, the package uses by default CELEX codes to identify laws (including Treaty articles) and policy documents.[14] The CELEX code consists of four properties: a single digit identifying the broad category to which the document belongs (1 = Treaties, 3 = legal acts); the year of publication; the type of act (L = directive, R = regulation, D = decision, H = recommendation); and the number of the act (so 32020L0700 translates to Directive (EU) 2020/700).[15]

The basic dataset of EU acts can be used in various ways. In Figure 2 I show the temporal variation in the type of legal instrument used for the entire history of EU integration (starting in 1952). We can see how initially decisions outnumbered regulations as the most used type of legal act, only for the relationship to be subsequently reversed. Since the 1990s, decisions have been steadily on the rise and today account for half of all legal acts enacted by the EU. Although the different legal implications of each type of act are well-known, it is not clear to what extent they drive the variation in the use of the instruments (Hurka and Steinbach 2020).[16]

Continuing with the same dataset, I look at the thematic content of the legal acts as described by EuroVoc terms. EuroVoc is a multilingual thesaurus providing controlled vocabulary for the indexation of documents. Table 2 shows the ten most common EuroVoc terms for each type of legal act. The resulting picture is rather intuitive: the EU has only limited powers in matters of economic policy and thus issues recommendations on this subject; directives serve predominantly to harmonize technical standards across

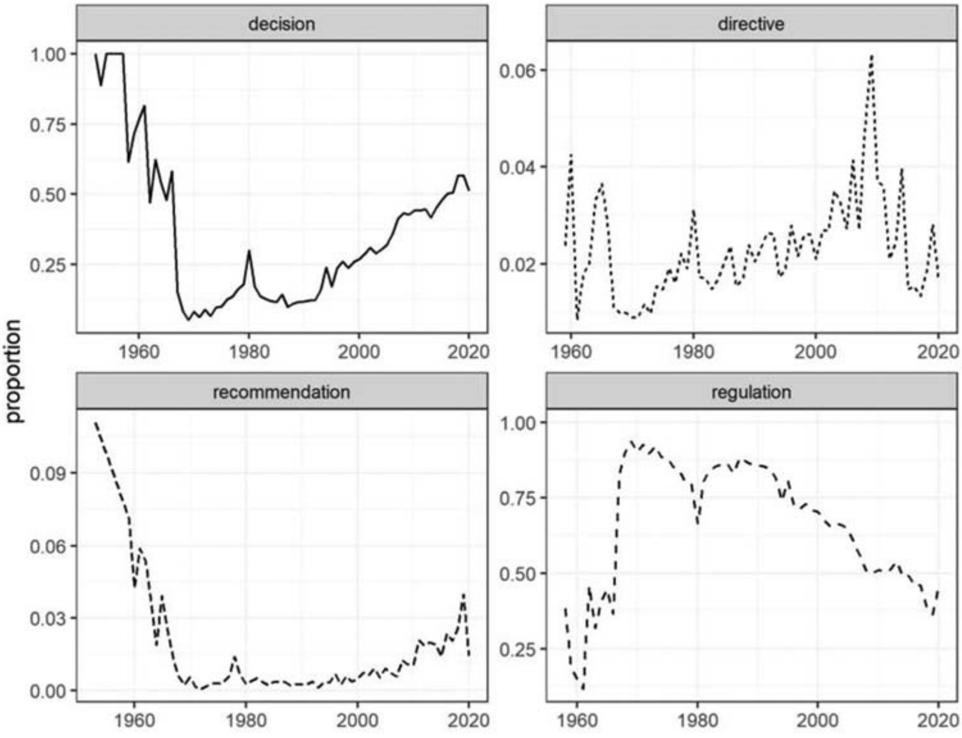

Figure 2. Variation in choice of legal instrument over time. The y-axis captures the proportion of each act type in all acts in any given year.

Table 2. Ten most common EuroVoc terms for each type of legal act.

decision	directive	recommendation	regulation
veterinary inspection	approximation of laws	stability programme	export refund
health control	marketing standard	economic policy	award of contract
State aid	marketing	economic reform	import price
import	motor vehicle	budget deficit	fruit vegetable
financial year	labelling	stability pact	citrus fruit
general budget (EU)	plant health control	employment policy	import
Germany	market approval	EU Member State	pip fruit
agreement (EU)	technical standard	EDF	import licence
appointment of members	plant health product	economic growth	tariff quota
health certificate	technological change	European Commission	import (EU)

Member States; the directly applicable regulations are used to implement the common agricultural policy – a staple area of EU action; while decisions come up in different contexts, from regulating State aid, through signing international agreements, to appointing staff and representatives.

Nonetheless, to leverage more of the over 5000 unique terms we need a more comprehensive numerical representation of the EuroVoc vocabulary. For each year, I construct a matrix counting the number of times each term is associated with one of the four types of legal acts. Using the counts, I then compute the mean correlation among the different types of legal acts. Figure 3 shows that average correlation has been increasing over

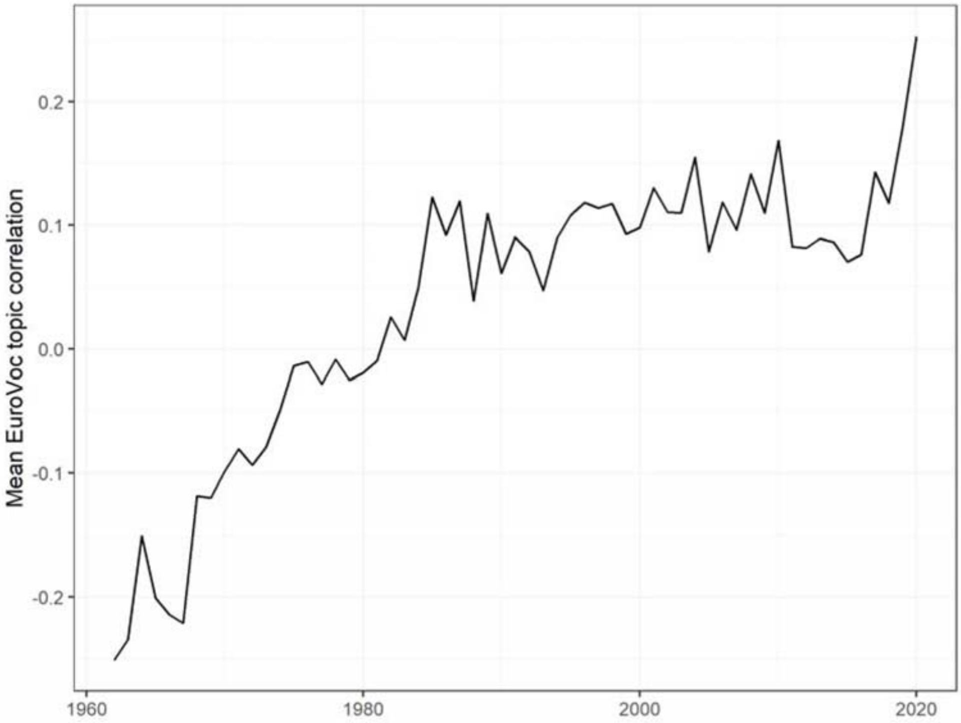

Figure 3. Mean correlation among directives, regulations, decisions and recommendations based on occurrence of EuroVoc terms.

time, which suggests that the four main legal instruments are increasingly deployed in overlapping areas, rather than each having a separate thematic domain. However, this measure of similarity is itself correlated with the rising number of EuroVoc terms describing each legislative act, which should make us cautious about reading too much into the thematic relationship between different types of instruments. On the other hand, in light of the significant decrease in the number of regulations addressing agricultural and trade issues in the past, substantive convergence of EU legal acts is not instinctively implausible.[17]

EuroVoc descriptors appear to be a useful tool for capturing the thematic content of EU legislation. But researchers need to be sure they can be relied upon. I compare the co-occurrence of EuroVoc descriptors and policy areas of legislation as coded by Rauh (2020).[18] The comparison cannot be done in a straightforward manner, because the number of EuroVoc descriptors assigned to documents varies, while Rauh (2020) assigns – as is customary – each document to a single policy area. As we can see in Table 3, the co-occurrence is generally sensible. The connection between fishing and industry might appear confounding at first sight but it is caused by the lack of greater specificity in the policy area coding. This is indicative of the broader difference in how the two classification systems differ: EuroVoc offers by design a richer description of the issue area which naturally comes at the cost of greater parsimony of sorting all EU legislation into mere 15 categories. Which one is more preferable will ultimately depend on research interests and design.

The multitude of EuroVoc terms used to describe documents is more realistic, however. Legislation frequently touches upon more than just a single policy area. The idea that documents consist of a mixture of topics is most fruitfully leveraged in what is known as probabilistic topic modelling. The contents of a document can be classified into several topics based on the co-occurrence of words. Below I show how the *eurlex* package facilitates this type of analysis. But before even getting to the full text of documents, we can similarly exploit the co-occurrence of EuroVoc descriptors to derive a small number of topics. Using non-negative matrix factorization (Greene and Cross 2017), I decompose a sparse matrix of EuroVoc terms and legislative documents into 15 topics

Table 3. Co-occurrence of EuroVoc descriptors (multiple) and policy area (singular) in EU legislation 1985–2016.

EuroVoc descriptor	Rauh (2020) policy area	N documents
common organization of markets	Agriculture	162
general budget (EU)	Budget	50
Cohesion Fund	Cohesion	17
tariff quota	Competition, taxation, customs	541
financial aid	Economic and monetary policy	64
reintegration into working life	Employment and social affairs	101
environmental protection	Environment	105
anti-dumping duty	External relations	631
labelling	Health and consumer protection	58
fishing area	Industry	390
EU statistics	Institutional provisions	112
approximation of laws	Internal market	74
agreement (EU)	Justice and home affairs	119
research and development	Research, Education, Culture	94
air transport	Transport and Energy	104

Table 4. Topics produced by non-negative factorization of a matrix of EuroVoc terms and EU legislation (1985–2016).

Topic	Five EuroVoc terms most associated with topic
1	EU Member State, exchange of information, catch quota, sea fishing, air transport
2	agreement (EU), fishing agreement, financial compensation of an agreement, signature of an agreement, visa policy
3	tariff quota, sea fish, tariff preference, import, Norway
4	originating product, import (EU), administrative cooperation, China, India
5	third country, textile product, quantitative restriction, European official, import
6	import, anti-dumping duty, China, anti-dumping legislation, dumping
7	protocol to an agreement, accession to the European Union, signature of an agreement, association agreement (EU), interim agreement (EU)
8	ratification of an agreement, cooperation agreement (EU), trade agreement (EU), scientific cooperation, technical cooperation
9	fishing area, catch quota, fishery management, authorized catch, conservation of fish stocks
10	EU programme, EU financing, EU aid, action programme, cooperation policy
11	agricultural product, industrial product, suspension of customs duties, CCT duties, fishery product
12	Spain, Portugal, fishing agreement, common organization of markets, France
13	approximation of laws, labelling, marketing standard, motor vehicle, foodstuff
14	VAT, derogation from EU law, tax harmonization, tax relief, tax exemption
15	reintegration into working life, European Globalisation Adjustment Fund, commitment of expenditure, payment appropriation, collective dismissal

to see whether the EuroVoc descriptors organically cluster into similar policy areas as in the classification used by Rauh (2020) (Table 4).

It turns out factorization of EuroVoc terms cannot quite replicate the classification used by Rauh (2020). Areas such as budget and cohesion are missing from the topic model. The reason is simple enough. The 15 different policy areas which we attempted to replicate are not uniformly distributed across EU legislation. In fact, the four most common policy area codes in Rauh's dataset – external relations, agriculture, industry (including free movement of goods) and competition/taxation/customs account for nearly 66% of documents. These are precisely the areas most obviously present in the topic model. More importantly, the topics produced by the model are by and large intuitively understandable and internally coherent. The topics offer greater detail than Rauh's classification about the content of EU legislation. For example, fisheries constitute a recognizable strand of legislation instead of being shoehorned under industry. Similarly, external relations are given more specific contours, highlighting the diversity of EU international action, albeit the vast majority of which deals with trade issues such as anti-dumping policy. It is important to remember that these topics emerge from the data, with the researcher only specifying the number of topics sought, rather than a full classification scheme.

The basic dataset shown in Table 1 offers also a different way of tracking the most common areas of EU legislative activity. In EU law, each piece of legislation needs to specify the competence – so called 'legal basis' – that enables its adoption. The legal basis of secondary law rests typically on one or more provisions of the Treaties (which are primary law), while tertiary law – implementing and delegated acts – are based on a provision of an enabling secondary law. Restricting the dataset to only Treaty legal bases, Figure 4 plots the most common legal basis per type of act in each year. Historically, there have been strong links between certain legal bases and types of acts. The most obvious feature of Figure 4 is the peak of legal bases of regulations around 1980. Article 113 of the Rome Treaty constituted the main competence behind common

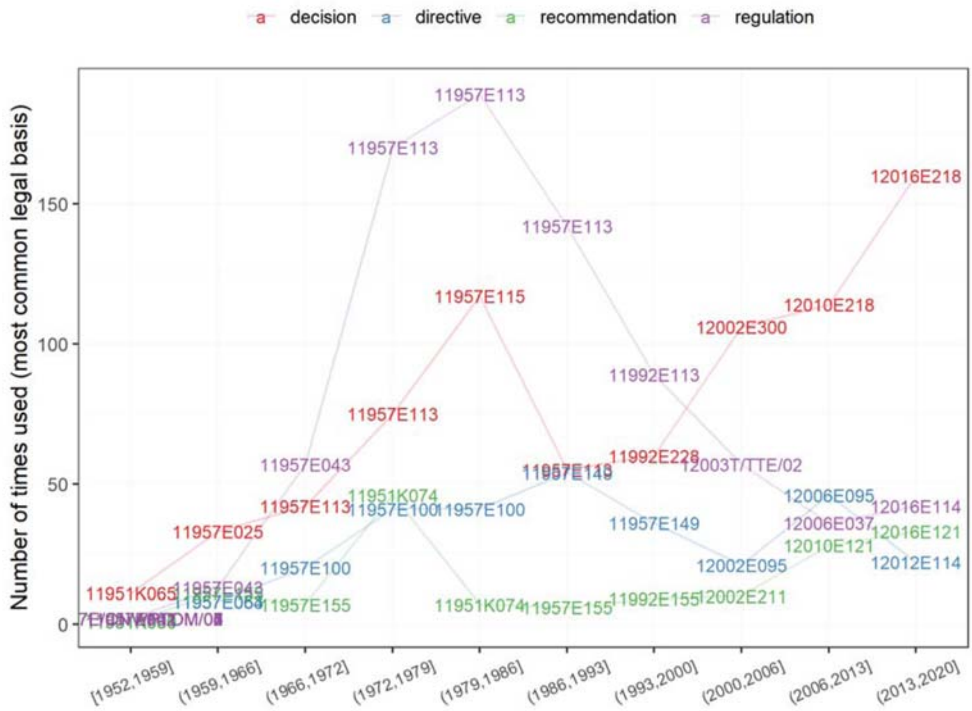

Figure 4. Most common Treaty legal basis per type of act (seven-year aggregate).

commercial policy. The quantitative importance of this legal basis has drastically declined since 1980 (Ovádek and Raina 2019), however, and with it the number of (non-tertiary) regulations. For their part, directives are strongly tied to the harmonization legal basis, which prior to 1987 (Article 100 Rome Treaty) required unanimity in the Council but are also relatively less numerous in recent years. Recommendations used to be associated with the European Coal and Steel Community but today they mainly serve as conduits for the coordination of Member States' economic policies through the European Semester (Article 121 TFEU). Decisions showcase the increase in the EU's international activity, with the competence to conclude agreements and adopt positions (Article 218 TFEU) becoming the most important legal basis since the turn of the millennium.

Possibly the most important resource accessible through the *eurlex* package is text. Both titles and the full text of documents are available for download and easy attachment to metadata shown in Table 1. Access to text opens even more research possibilities amenable to both qualitative and quantitative inquiry. Because some documents are of considerable length, the time needed for data collection is longer for texts. Given the size of the database, the data are particularly interesting for quantitative text analysis. There is a vast array of quantitative text tools, but topic modelling is a common use case (Greene and Cross 2017; Dyevre and Lampach 2021). I use latent Dirichlet allocation (Blei, Ng, and Jordan 2003) to classify the text of (non-implementing) directives into twelve topics. The text was pre-processed by pruning the most and least frequent vocabulary to reduce the dimensionality of the classification task. Figure 5 shows the topic-modelling results. Each topic is described by three terms most likely to

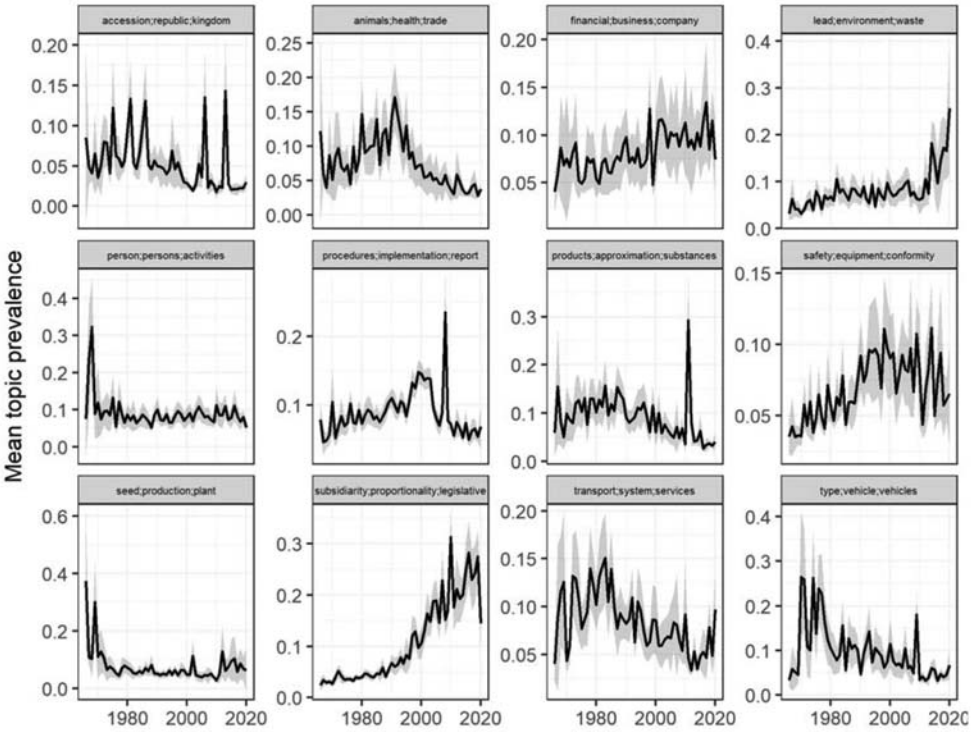

Figure 5. Mean topic prevalence in the text of directives over time with 95% confidence intervals.

be associated with it. The temporal pattern is an aggregation of the probability of topic *k* occurring in document *d*. We can see that in general there is substantial year-on-year fluctuation. Nonetheless, some overall trends are discernible. The regulation of animal health through directives has significantly decreased since 1990. In contrast, environmental standards have become a more important theme. Interesting to note is also the substantial increase in language concerning the principles of subsidiarity and proportionality, likely reflecting the growing politicization of the EU and domestic misgivings about deepening European integration.

Apart from access to data on the Eur-Lex portal, the package contains functions to collect more information from the website of the Court of Justice and the Council. A simple command is available for downloading the list of all cases ever received by the Court (over 40 000), including cases later removed from the docket, which is integral to understanding the Court's decision-making discretion. This resource is also maintained regularly, with updates on the status of cases, while the identifiers retrieved can be used to collect more data about court decisions. Figure 6 makes clear that the combined workload of EU courts has been growing for decades. In 2019 alone, the Court of Justice and the General Court received over 2000 cases.

For researchers interested in exploiting judicial data, it is worth highlighting differences between the dataset obtained from the website of the Court of Justice (Figure 6) and the dataset of court rulings (Figure 1). The most important aspect to be mindful of is that several cases can be joined into a single judgment that is recorded under a single

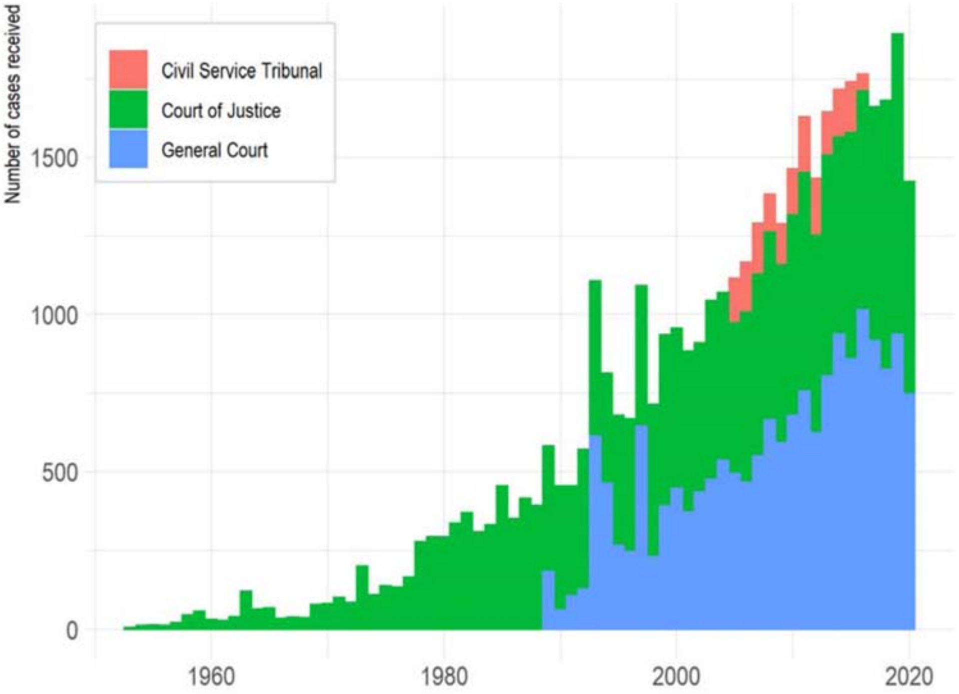

Figure 6. Number of cases received by EU courts yearly (up-to-date as of 14 December 2020).

CELEX identifier. This means that the number of rulings obtained from Eur-Lex will be lower than the number of all cases obtained from the Court's website. Moreover, some cases are closed without a full ruling being issued and these are more comprehensively accounted for on the Court's website than Eur-Lex. On the plus side, the eurlex package allows combining data from both sources where CELEX identifiers are present in the list of all cases.

Moreover, the *eurlex* package taps into a SPARQL API maintained by the Council of the EU. A single function retrieves an up-to-date list of votes taken in the Council along with some basic properties such as date, type of procedure and voting rule. The voting data published by the Council through its API have a more limited temporal span, going back to only 2009. Nonetheless, this dataset, too, will grow over time and already now makes possible inferences about post-Lisbon voting behaviour. I reduce the dimensionality of the voting data by estimating Member States' ideal points along a single latent dimension (Figure 7) using an item-response theory model without covariates (Bafumi et al. 2005; Hagemann 2007). Even this overly simplistic model of Member States preferences reveals the one Member State that left the EU – the UK – to stand perceptibly apart from the rest of the Member States. This would suggest that the latent dimension captures at least to some extent preference for European integration. However, a more sophisticated model would be required to unpack Council voting data in practice, in light of the complex and frequently opaque ways in which Member States reach agreements (Kardasheva 2013). The *eurlex* package makes this type of investigation readily accessible to a greater pool of researchers.

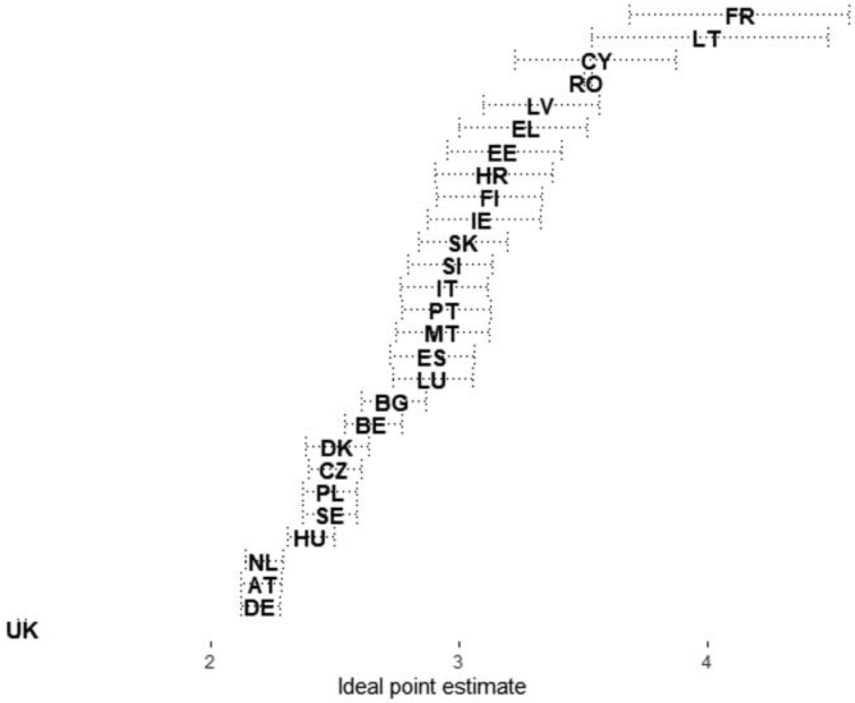

Figure 7. Unidimensional ideal point estimates of EU Member States based on their voting behaviour in the Council.

Conclusion

In light of the increasing availability of an ever-growing volume of data, researchers and data providers need to invest in the creation of infrastructure that will make data collection open, replicable and efficient. The *eurlex* R package attempts to partially fulfil this role – building on open data systems maintained by EU institutions – for a community of political scientists, lawyers, economists and others who are interested in data on European Union laws and policies.

For those who already use data on EU laws and policies, the *eurlex* package reduces time spent on data collection. For those who were considering using such data but were dissuaded by the technical hurdles involved, the barriers to entry are significantly reduced. At the end of the day, the main goal of the package is to empower applied researchers to exploit EU data in novel ways and focus on theoretical and methodological contributions instead of learning about technical protocols for accessing programming interfaces. In addition, wide adoption of the package (or similar open-source software) could enhance the integrity of data collection practices and interoperability of datasets constructed by different authors. There is also further room for improving and extending the data collection software with anyone having the possibility to contribute.

Notes

1. https://eur-lex.europa.eu/

2. As a general rule, Eur-Lex contains legal documents published by EU institutions as listed in Article 13(1) of the Treaty on European Union. In addition, it publishes documents by the European Ombudsman, the European Economic and Social Committee and the Committee of Regions. Insofar as they are not published in the Official Journal of the European Union, most documents produced by EU agencies do not appear on Eur-Lex.
3. For example, both Rauh (2020) and Hurka and Steinbach (2020) collect similar datasets of Commission proposals which should be the same size for the overlapping period 2010-2016. In practice, the size of the datasets diverges by some 50 observations. The datasets cannot even be readily compared because the respective authors use different identifiers.
4. The *eurlex* package is embedded in and requires the R software environment to run. The stable release of the package has been checked in accordance with the policies of the Comprehensive R Archive Network (CRAN) which enables its swift installation from within the R environment using the function 'install.packages("eurlex")'. As of 15 December 2020, the package has been downloaded 2035 times from CRAN.
5. APIs are in this context themselves dedicated gateways for programmers to gain access to raw data that we can ordinarily see displayed on websites such as Eur-Lex.
6. Article 4(1) of Decision 2009/496/EC, Euratom of the European Parliament, the Council, the Commission, the Court of Justice, the Court of Auditors, the European Economic and Social Committee and the Committee of the Regions of 26 June 2009 on the organization and operation of the Publications Office of the European Union 'Each institution shall have exclusive competence to take decisions on the publishing of its own publications.'
7. The package vignette provides a step-by-step introduction to using the package: https://michalovadek.github.io/eurlex/articles/eurlexpkg.html.
8. https://github.com/michalovadek/eurlex
9. Dozens of examples can be found browsing early editions of the Official Journal. See, for example, Décision n° 1–61 du 16 janvier 1961 concernant l'octroi de subventions à des entreprises charbonnières belges en 1961.
10. Some less important rulings of the Court cannot be found on Eur-Lex but can be found on the Court's own website, Curia.eu. See, for example, Judgment of the Court of First Instance (Second Chamber) of 14 June 1995 in *Henri de Compte v European Parliament* (CELEX: 61992TJ0061). When approached for an explanation of these discrepancies, the Eur-Lex helpdesk explained that they have no control over what the Court decides to publish through Eur-Lex.
11. For example, at the time of writing, for document 62013TJ0656 the English version of the Eur-Lex website displays the title in Bulgarian and no English or French title can be retrieved via the API.
12. For example, the CELEX-encoded data on the legal basis of Council Directive 81/855/EEC of 19 October 1981 is missing. The missing entries are 11957E100 and 11957E235 (Articles 100 and 235 of the Treaty establishing the European Economic Community).
13. In general, the quality and completeness of data has improved over time as a function of increasing standards on the part of the EU institutions. Omissions and errors are more likely to be found among older documents when database technologies were less developed.
14. Vauchez's deep dive into the history of creating an EU law database gives additional background to the importance of the CELEX classification, in addition to making clear that the open data status quo is a relatively recent innovation.
15. See https://eur-lex.europa.eu/content/help/faq/celex-number.html for further details on CELEX numbers.
16. It would appear that even some core assumptions concerning the types of legal acts used should be revisited. Williams and Bevan 2019, 615 for example argue that '[d]ue to the unique nature and relatively narrow scope of decisions, it seems unlikely that EU-wide public opinion would influence decision adoption.' But arguably one of the most controversial acts ever adopted by the EU was the decision to introduce a relocation mechanism during the 2015 refugee crisis. See Council Decision (EU) 2015/1601 of 22 September 2015 establishing provisional measures in the area of international protection for the benefit of Italy and

Greece. Decisions are also used – frequently, as will be shown below – for the adoption of international agreements, which has become an increasingly contested field.

17. Hurka and Steinbach (2020) document the significant degree of discretion enjoyed by the EU legislator in deciding what type of legal act to use.
18. Rauh's (2020) classification builds on Biesenbender (2011).

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Michal Ovádek http://orcid.org/0000-0002-2552-2580

References

Bafumi, J., A. Gelman, D. K. Park, and N. Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13: 171–187. doi:10.1093/pan/mpi010.

Bernet, H., and P. Berteloot. 2006. "EUR-Lex: A Multilingual on-line Website for European Union law." *International Review of Law, Computers & Technology* 3: 337–339. doi:10.1080/13600860600947109.

Biesenbender, J. 2011. "The Dynamics of Treaty Change – Measuring the Distribution of Powerin the European Union." *European Integration Online Papers* 15: 1–24.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Blom-Hansen, J. 2019. "Studying Power and Influence in the European Union: Exploiting the Complexity of Post-Lisbon Legislation with EUR-Lex." *European Union Politics* 20: 692–706. doi:10.1177/1465116519851181.

Caled, D., M. Won, B. Martins, and M. J. Silva. 2019. "A Hierarchical Label Network for Multi-label EuroVoc Classification of Legislative Contents." In *Digital Libraries for Open Knowledge*, edited by A. Doucet, A. Isaac, K. Golub, T. Aalberg, and A. Jatowt, 238–252. Cham: Springer.

Chalkidis, I., M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. 2019. "Extreme Multi-Label Legal Text Classification: A case study in EU Legislation." *NAACL-HLT*.

Düro, M. 2009. *Crosswalking EUR-Lex: a Proposal for a Metadata Mapping to Improve Access to EU Documents*. Luxembourg: Office for Official Publications of the European Communities.

Dyevre, A., and N. Lampach. 2021. "Issue Attention on International Courts: Evidence from the European Court of Justice." *Review of International Organizations* 17: 1–25.

Fjelstul, J. C. 2019. "The Evolution of European Union Law: A New Data set on the Acquis Communautaire." *European Union Politics* 20: 670–691. doi:10.1177/1465116519842947.

Francesconi, E., M. W. Küster, P. Gratz, and S. Thelen. 2015. "The Ontology-based Approach of the Publications Office of the EU for Document Accessibility and Open Data Services." In *Electronic Government and the Information Systems Perspective*, 29-39. Springer International Publishing. doi:10.1007/978-3-319-22389-6_3.

Greene, D., and J. P. Cross. 2017. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *Political Analysis* 25: 77–94. doi:10.1017/pan.2016.7.

Hagemann, S. 2007. "Applying Ideal Point Estimation Methods to the Council of Ministers." *European Union Politics* 8: 279–296. doi:10.1177/1465116507076433.

Hurka, S., and Y. Steinbach. 2020. "Legal Instrument Choice in the European Union." *Journal of Common Market Studies* Early View: 1–19.

Kardasheva, R. 2013. "Package Deals in EU Legislative Politics." *American Journal of Political Science* 57: 858–874.

Lampach, N., W. Wijtvliet, and A. Dyevre. 2020. “Merchant Hubs and Spatial Disparities in the Private Enforcement of International Trade Regimes.” *International Review of Law and Economics* 64: 1–10.

Marciano, A., and J.-M. Josselin. 2002. *The Economics of Harmonizing European Law*. Cheltenham: Edward Elgar Publishing.

Ovádek, M., and A. Raina. 2019. “The Evolution of EU Trade Law Through the Prism of Competence: A Quantitative, Longitudinal Perspective.” *Journal of World Trade* 53 (3): 489–508.

Quaresma, P., and T. Gonçalves. 2010. “Using Linguistic Information and Machine Learning Techniques to Identify Entities from Juridical Documents.” In *Semantic Processing of Legal Texts*, edited by E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, 44–59. Berlin: Springer.

Rauh, C. 2020. “One Agenda-Setter or Many? The Varying Success of Policy Initiatives by Individual Directorates-General of the European Commission 1994–2016.” *European Union Politics* OnlineFirst: 1–22.

Williams, C. J., and S. Bevan. 2019. “The Effect of Public Attitudes Toward the European Union on European Commission Policy Activity.” *European Union Politics* 20: 608–628. doi:10.1177/1465116519857161.

Wuttke, A. 2019. “Why Too Many Political Science Findings Cannot be Trusted and What We Can Do About it: A Review of Meta-scientific Research and a Call for Institutional Reform.” *German Political Science Quarterly*, 1–22.