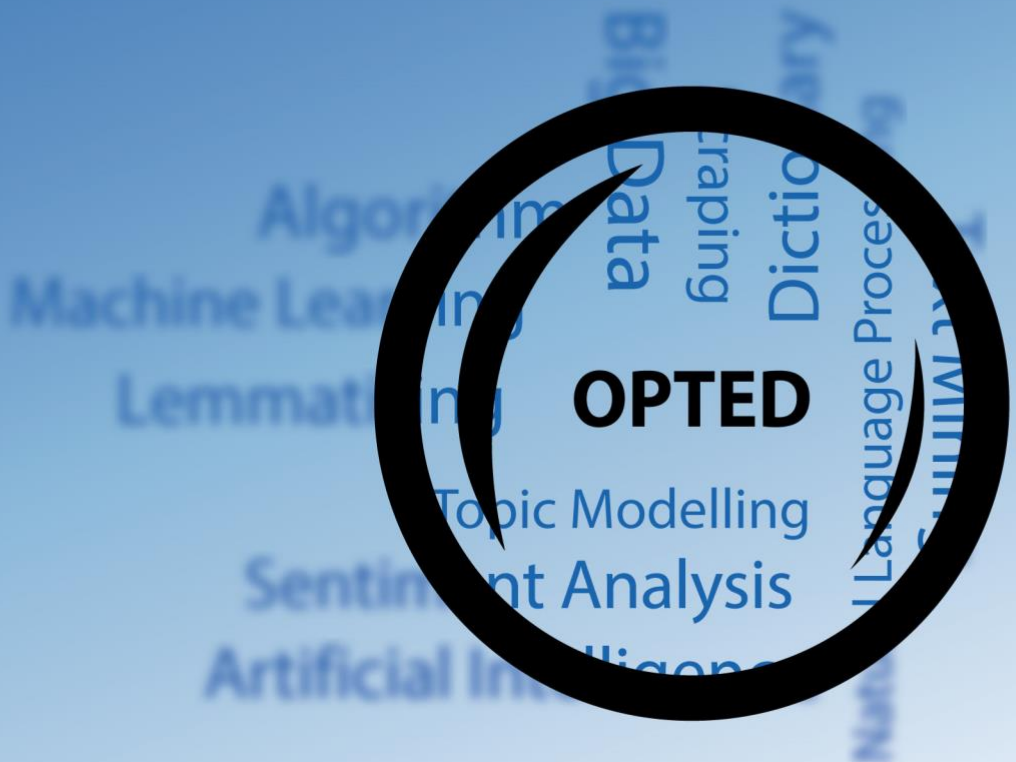
OPTED

Review of available parliamentary corpora

Miklós Sebők, Sven-Oliver Proksch, & Christian Rauh



Disclaimer

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report

OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

Review of available parliamentary corpora

Deliverable D5.1

Authors: Miklós Sebők[1], Sven-Oliver Proksch[2], & Christian Rauh[3]

[1] Institute for Political Science, Centre for Social Sciences, Eötvös Lóránd Research Network
[2] Cologne Center for Comparative Politics, University of Cologne
[3] WZB Berlin Social Science Centre

Due date: March 2021

1 Executive Summary

The long-term objective of WP5 is to compile a database which provides a comprehensive, easy-to-use collection of legislative speeches and legislative documents covering all EU member countries, the most important EU institutions, as well as the United Kingdom and Israel. This comprehensive inventory, to be finalized by the end of the project, is the ParlLawSpeech database. This report reviews our progress in creating this database up until the 6thmonth of the project. The core product of this period is an inventory in the form of a spreadsheet which provides an overview of already existing collections of legislative texts. This sheet has the dimension of 157 sources with 87 variables for speeches, and 139 sources with 91 variables for documents. We collected data from all available sources, such as parliamentary websites and secondary sources created by scholars or NGOs.

The inventory also includes a codebook and server copy of the downloadable databases. We also evaluated and double-checked the quality and usability of these databases. In order to provide a well-rounded analysis, we also prepared a "country report" for all countries and institutions in the sample of at least two written pages. Serving as a general overview, we also wrote a comprehensive research note, which describes the sources we found, the scope and geographic coverage of the sources and their potential usefulness for the future ParlLawSpeech database (see excerpts in the Appendix). Finally, we created a website where our results are accessible in a user-friendly way (see preview in Appendix). Our submission as Deliverable D5.1. comprises all the abovementioned documents.

2 Inventory of the already existing legislative speech and text sources

2.1 Technical information and introduction of the inventory

The inventory is an online, living document which provides a collection of data sources. The inventory consists of five sheets (see Appendix A). The first sheet allows the reader to find technical information and a link to the codebook (see Appendix B). The second sheet gives an overview about the inventory in terms of the number of "importable" (directly usable in a software environment such as R or python) databases are accessible per country. The third and fourth pages present the core of our work. The third page is an inventory of legislative speeches, while the fourth one is an inventory of the legislative documents. Both sheets present our collection in at least 130x90 cells. We collected 87 different variables for the legislative speeches and 91 for the legislative documents. These variables are organized into six main groups: 1) country information, 2) source database information, 3) file format and importability, 4) most important variables regarding the source database 5) quality and content of the available text corpora 6) accessibility and copyright. This information helps the reader to find out how the particular databases can be utilized in a research setting. It also provides an overview of the years covered, which kind of metadata is provided, and, therefore, allows for estimating the amount of work needed to integrate these sources into ParlLawSpeech or quantitative text analysis projects (i.e. it describes whether additional web scraping is needed or plain text is provided in a ready to use format). For all the variables a detailed explanation is given in the codebook (which runs for 31 pages).

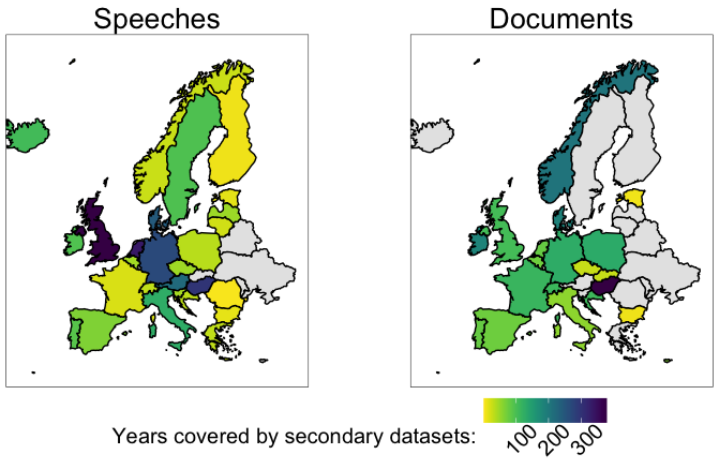
The last sheet of the inventory provides the link and a summary table for the server copies that we collected. Since many of the databases are not in a ready for use format, we uploaded only those corpora where it this possible without additional work (e.g. web-scraping). As a result, we uploaded 53 databases for legislative speeches and 25 for legislative documents, where the representation of different countries is quite diverse: there are countries for which 4-5 databases had been uploaded and there are many for which we couldn't upload any).

Country reports are meant to describe how well each of the countries' legislative text corpora is covered by the sources, and how the most important variables can be extracted from the source databases. Country reports include a screener of the data view of the sources and a comparison of parallel sources' variable availability. The number of written reports currently stands at 36 (see Appendix C.).

2.2 Coverage and extent of the collection

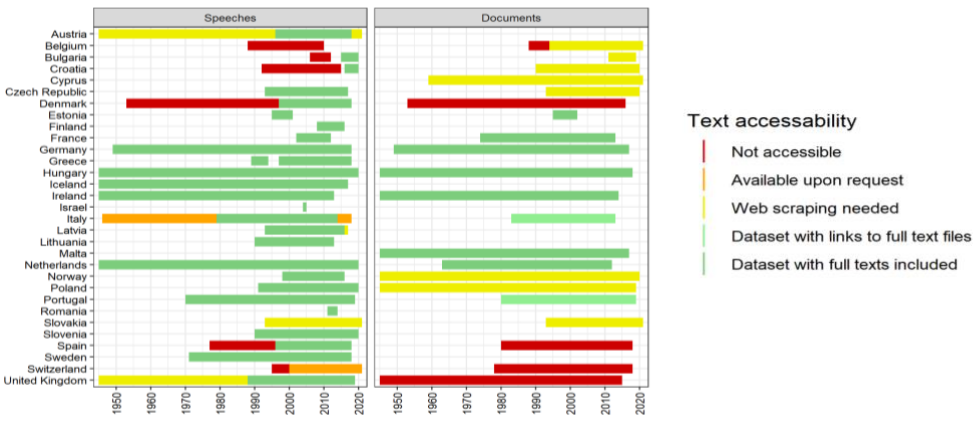
One of the most important quality indicators of our work is the number of years covered by the collected data sources for each country. Figure 1.1 shows this regarding the secondary datasets.[1]

Figure 2.1 GEOGRAPHICAL COVERAGE OF SECONDARY TEXT DATA CORPORA



The length of the coverage shows our current results. Since there are many sources where the texts are not in ready to use format, the second phase of Work Package 5 will undertake the development of ParlLawSpeech based on these sources of varying quality. Figure 1.2 gives an overview about the accessibility of the secondary sources (as primary sources, for the most part, will not offer direct usability for research purposes).

Figure 2.2 SECONDARY DATASET AVAILABILITY OVER TIME



[1] Primary and secondary sources had been differentiated during the work process. We defined a source as primary if it was published by the legislative body itself or its official archive and secondary otherwise (published by researchers, NGO-s, etc.).

Appendix A – Layout of the Inventory

The inventory is publicly available here.

Updated on: March 30th, 2021

		Legislative speeches			Legislative documents		
	Country	Number of databases which can be imported	Number of databases which can be imported with metadata	Number of collected databases	Number of databases which can be imported	Number of databases which can be imported with metadata	Number of collected databases
1	Austria	1	1	6	0	0	3
2	Belgium	0	0	3	0	0	7
3	Bulgaria	1	1	4	0	0	2
4	Council of Europe	0	0	1	0	0	1
5	Croatia	2	2	4	0	0	5
6	Cyprus	0	0	1	0	0	4
7	Czech Republic	3	3	6	0	0	3
8	Denmark	2	2	9	0	0	5
9	Estonia	2	2	3	1	1	2
10	Finland	2	1	2	0	0	3
11	France	3	3	8	5	5	10

Database Server Copy Folder

Country	EU member	Parliamentary speeches (1 N/A available and uploaded, 0 N/A not available)					Legislative texts (1 N/A available and uploaded, 0 N/A not available)				
		text files	metadata	codebook	nr. of uploaded DBs	If there is no uploaded DB, why wasn't it possible to upload?	text files	metadata	codebook	nr. of uploaded DBs	If there is no uploaded DB, why wasn't it
Austria	1	1	1	0	1	N/A	0	0	0	0	only webscra
Belgium	1	0	0	0	0	texts are downl	0	0	0	0	only webscra
Bulgaria	1	1	1	0	1	N/A	0	0	0	0	texts are dow
Croatia	1	1	1	0	2	N/A	0	0	0	0	texts are dow
Cyprus	1	0	0	0	0	texts are downl	0	0	0	0	texts are dow
Czech Re	1	1	1	0	1	N/A	0	0	0	0	only webscra
Denmark	1	1	1	0	2	N/A	0	0	0	0	texts are dow
Estonia	1	1	1	0	1	N/A	1	1	0	1	N/A
Finland	1	0	0	0	0	texts are downl	0	0	0	0	texts are dow
France	1	1	1	1	3	N/A	1	1	1	4	N/A
Germany	1	1	1	1	5	N/A	1	1	1	3	N/A
Greece	1	1	1	0	2	N/A	0	0	0	0	texts are dow
Hungary	1	1	1	1	4	urgent question	0	0	0	4	N/A
Iceland	0	1	1	0	1	N/A	0	0	0	0	texts are dow
Ireland	1	1	1	1	1	N/A	0	0	0	1	N/A
Israel	0	1	0	0	1	N/A	0	0	0	0	texts are dow
Italy	1	1	1	1	2	N/A	1	1	1	2	N/A
Latvia	1	1	1	0	1	N/A	0	0	0	0	texts are dow
Lithuania	1	1	1	1	1	N/A	0	0	0	0	only webscra
Luxembou	1	0	0	0	0	texts are downl	0	0	0	0	only webscra
Malta	1	0	0	0	0	texts are downl	1	0	0	1	N/A
Netherlan	1	1	1	1	4	N/A	1	1	1	2	N/A
Norway	0	0	0	0	0	only webscrapir	0	0	0	0	texts are dow
Poland	1	1	1	0	2	N/A	0	0	0	0	texts are dow
Portugal	1	1	1	1	3		1	1	1	2	N/A
Romania	1	1	0	0	1	N/A	0	0	0	0	only webscra
Slovakia	1	0	0	0	0	texts are downl	0	0	0	0	texts are dow
Slovenia	1	1	1	1	3	N/A	0	0	0	0	texts are dow
Spain	1	1	1	1	2	N/A	1	1	0	3	N/A
Sweden	1	1	1	1	4	N/A	0	0	0	0	only webscra
Switzerlar	0	0	0	0	0	only webscrapir	0	0	0	0	texts are dow
United Kir	0	1	1	1	3	N/A	0	0	0	1	N/A
European	2	0	0	0	0	only webscrapir	0	0	0	0	only webscra
European	2	1	1	0	1	N/A	0	0	0	0	only webscra
European Co	2	0	0	0	0	only webscrapir	0	0	0	0	only webscra
EU Council of	2	0	0	0	0	only webscrapir	0	0	0	0	texts are dow
European Un	2	1	1	1	1	N/A	1	1	1	1	N/A

Variable group		A_Country information			B_Source database information								
Variable description	Database unique identifyer	Country	Geographical region	EU member	Text type	Plenary speeches (excluding questions)	Plenary detail: Additional information for plenary speeches	Questions (excluding interpellations)	Interpellations	Extent of coverage	Source name	Project Name	Source link
var_nr	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*
var_type	*categorical*	*string*	*categorical*	*categorical*	*categorical*	*categorical*	*string*	*categorical*	*categorical*	*categorical*	*string*	*string*	*string*
var_name	***database.id***	***country***	***region***	***eu.member***	***text.type***	***plenary.speeches***	***plenary.detail***	***questions***	***interpellations***	***coverage***	***source.name***	***project.name***	***source.link***
	S1	Austria	1	1	1	1	N/A	0	0	1	Parl Archive (National	N/A	https://www.parlament.gv.at/PAKT/PLENAR/
	S2	Austria	1	1	1	0	N/A	1	0	0	Parl Archive (National	N/A	https://www.parlament.gv.at/PAKT/JMAB/
	S3	Austria	1	1	1	1	N/A	0	0	0	V4	N/A	https://www.v4dplplus.eu/austria-360/digital-library-378
	S4	Austria	1	1	1	1	N/A	0	0	1	Parl Speech	Parl Speech	https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/L4OAKN&version=1.0
	S5	Austria	1	1	1	1	N/A	0	0	1	Clarin	ParlAT	https://www.oeaw.ac.at/acdh/tools/parlat/
	S6	Austria	1	1	1	1	N/A	0	0	1	Clarin	Analyse österreichischer F	s://www.clarin.eu/content/parliamentary-cor
	S7	Belgium	4	1	1	1	N/A	0	0	1	Dekamer	N/A	https://www.dekamer.be/kvvcr/showpage.cfm?section=/cricra&language=nl&cfm=dcricra.cfm?type=plen&cricra=cri&count=all&leg
	S8	Belgium	4	1	1	0	N/A	1	0	1	CAP	CAP	https://www.comparativeagendas.net/datasets_codebooks
	S9	Belgium	4	1	1	0	N/A	0	1	1	CAP	CAP	https://www.comparativeagendas.net/datasets_codebooks
	S10	Bulgaria	1	1	1	1	N/A	0	0	1	Clarin	ParlaMint 1.0	https://www.clarin.si/repository/xmlui/handle/11356/1345
	S11	Bulgaria	1	1	1	1	N/A	0	0	3	WebClark	Corpus of Bulgarian Political and Journalistic	http://www.political.webclark.org/?locale=en
	S12	Bulgaria	1	1	1	1	N/A	0	0	1	Parl Archive (National	N/A	https://www.parliament.bg/bg/plenaryst
	S13	Bulgaria	1	1	1	1	N/A	0	0	1	Clarin	N/A	N/A
	S14	Croatia	1	1	1	1	N/A	0	0	1	Clarin	ParlaMeter-hr 1.0	https://www.clarin.si/repository/xmlui/handle/11356/1209
	S15	Croatia	1	1	1	1	N/A	0	0	1	Clarin	ParlaMint 1.0	https://www.clarin.si/repository/xmlui/handle/11356/1345
	S16	Croatia	1	1	1	1	N/A	0	0	1	Parl Archive (Croatia)	N/A	http://edoc.sabor.hr/Fonogrami.aspx

D_Most important variables												E_Quality and content of the of the available text corpora		F_Accessibility & copyright			
				Available info about the legislative procedure		Available info about the introducing MP						If the database has been collectedfor scientific purposes, is it...					
The topic of law is given	Date of introduction or passing is available	Date of introduction is available	Date of passing (or rejection) is available	Duration from introduction to passing	Vote share available (yes, no, abstention)	Name of the introducer is given	Gender	Leadership position	Single member district (SMD) /party list	Party affiliation	Can people comment on the legislative documents ?	Annotated?	Is there a codebook available?	Dataset and the codebook is accesible free of charge	Dataset and the codebook is accesible with or without registration	Dataset is available for re-use, or need to ask for permission	Owner of the copyright
1	1	1	0	0	0	0	0	0	0	0	0	1	1	1	2	3	agendas@gma
1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2	N/A	erman Assemb
1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2	N/A	erman Assemb
1	1	1	1	0	0	0	0	0	0	0	0	1	1	1	2	3	agendas@gma
N/A	1	1	0	0	0	1	0	0	0	0	0	0	1	1	2	3	orinna; Remsc
1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	2	N/A	llenic Parliame
1	1	0	1	0	0	0	0	0	0	0	0	0	0	1	2	0	ony Lap- és Kö
1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	ony Lap- és Kö
1	1	1	1	0	1	1	0	0	0	1	0	1	1	1	1	2	o@cap.tk.mta.
1	1	1	1	0	1	1	0	0	0	1	0	1	1	1	2	3	agendas@gma
1	1	0	1	0	0	1	0	0	0	0	0	1	1	1	1	2	o@cap.tk.mta.
1	1	0	0	0	0	1	0	0	0	0	0	1	1	1	1	2	o@cap.tk.mta.
1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	2	1	Government of Iceland
1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	2	1	Althingi
1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	2	1	Althingi
1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	2	1	open.data@ oireachtas.ie
1	1	1	1	0	0	0	0	0	0	0	0	0	0	1	2	1	open.data@ oireachtas.ie
0	1	N/A	N/A	0	0	0	0	0	0	0	0	1	0	1	2	3	Government of Ireland

Appendix B – Excerpt from the Codebook

The codebook is publicly available here.

OPTED - Deliverable D1

LEGISLATIVE TEXT CORPORA INVENTORY

CODEBOOK

Version: March 30th, 2021

1.Legislative Speeches

A. Country information

- *database.id*
  - This variable assigns a unique identifier to the inventory item.
- *country*
  - Name of the country or supranational institution (e.g. European Commission).
- *region*
  - Geographical region of the inventory item.
    - 1 - Central Eastern Europe
    - 2 - Northern Europe
    - 3 - Southern Europe + Israel
    - 4 - Western Europe
    - 5 - Supranational
- *eu.member*
  - This variable defines if the *country* in question is an EU member state (excluding EU institutions).
    - 0 - Not an EU member
    - 1 - It is an EU member state
    - 2 - Not applicable as it is a supranational institution

B. Source database information

Appendix C – Excerpt from a Country Report

The country reports are publicly available here.

Portugal

Section 1. – ParlSpeech (Parliamentary Speeches)

1) Available sources

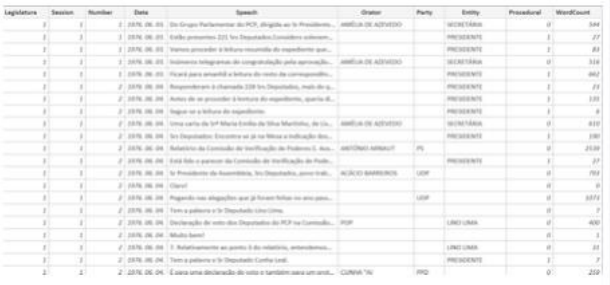
a) POPad - Parliamentary https://popad.org/parliamentary-speeches/
- Secondary dataset
- Years covered: 1980-2019
- File format: csv

b) POPaD - Parliamentary Questions and Requests https://popad.org/parliamentary-questions-and-requests/
- Secondary dataset
- Years covered: 2009-2019
- File format: csv

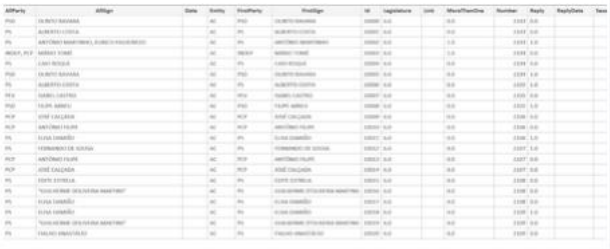
c) PTPARL https://catalog.elra.info/en-us/repository/browse/ELRA-W006
- Secondary dataset
- Years covered: 1970-2008
- File format: txt

d) Parl Archive https://debates.parlamento.pt/catalogo/r3/dar
- Primary dataset
- Years covered: 1976-current
- File format: html, pdf, txt; dataset is accessible throu scraping

a)



b)



c)

d)

3) Variables

1. date:
   a) Available under 'Date'.
   b) Available under 'Date'.
   c) Not available.
   d) Available. Access: 1. Choose the series (Série xy). 2. Choose legislature (Legislatura xy). 3. Choose the legislative session (Sessao Legislativa). In the table, date is available under 'Data' for each document.
2. speech number:
   a) Not available, but 'Session' variable is available.

Appendix D – Preview of the website

The pilot website is available here: www.christian-rauh.eu/opted-wp5-inventory.

OPTED - WP5
Inventory of parliamentary text data sources

Code

OPTED WP5 team (2021-02-25 09:03:19)

Purpose of this website

The project **Observatory for Political Texts in European Democracies** (OPTED; Horizon 2020 Grant agreement 951832) aims to design a European Research Infrastructure facilitating the large-scale computational analysis of political texts in Europe.

Work Package 5 focuses on national and supranational parliaments. We cover textual data on political speeches and debates as well as legislative texts produced in and by these key institutions of European democracy.

Easier access to existing text data collections as well as *identifying the gaps in extant data avilability* are among our key aims. Thus we have initially have assembled an **inventory of available text data sources** covering parliamentary activity. We identified the set of currently available sources - covering both primary archives and secondary data collections - by reviewing relevant academic literature, by scoping extant linguistic infrastructures (such as CLARIN), and by surveying the computational social science community via social media.

This website navigates prospective users and analysts through the inventory. It firstly provides a bird's eye view om the coverage of existing

Find your data source: Parliamentary speeches

Code

	Country ↑	Archive type	Text types	Start Year	End Year	Full covera…	Ready to u…	Access so…
▸								
▸	Austria	Secondary	Plenary spe…	1996	2018	✓	✓	Click
▸	Austria	Secondary	Plenary spe…	1996	2017	✓	✘	Click
▸	Austria	Secondary	Plenary spe…	2013	2015	✓	✘	Click
▸	Belgium	Primary	Plenary spe…	1830	2021	✓	✘	Click
▸	Belgium	Secondary	Questions	1988	2010	✓	✘	Click

Find your data source: Legislative texts

Code

	Country ↑	Archive type	Text types	Start Year	End Year	Full covera…	Ready to u…	Access so…
▸								
▸	Belgium	Secondary	Bills	1988	2010	✓	✘	Click
▸	Belgium	Secondary	Laws, Other…	1994	2021	✘	✘	Click
▸	Bulgaria	Primary	Laws	1990	2021	✓	✘	Click
▸	Bulgaria	Secondary	Laws	2011	2019	✘	✘	Click
▸	Council of E…	Primary	Laws, Other…	1985	2019	✘	✘	Click
▸	Croatia	Secondary	Laws	1991	2020	✘	✘	Click
▸	Croatia	Secondary	Laws	1990	2018	✘	✘	Click