BRILL

INTERNATIONAL JOURNAL OF PARLIAMENTARY STUDIES
2 (2022) 278–284

INTERNATIONAL JOURNAL OF PARLIAMENTARY STUDIES

Creating an Enhanced Infrastructure of Parliamentary Archives for Better Democratic Transparency and Legislative Research

Report on the OPTED forum in the European Parliament (Brussels, Belgium, 15 June 2022)

Rebeka Kiss | ORCID: 0000-0002-7384-9262
Junior Research Fellow, Institute for Political Science, Centre for Social Sciences; PhD student, Doctoral School of Public Administration Sciences, Ludovika – University of Public Service, Budapest, Hungary
kiss.rebeka@uni-nke.hu

Miklós Sebők | ORCID: 0000-0003-0595-2951
Research Professor, Institute for Political Science, Centre for Social Sciences, Budapest, Hungary
sebok.miklos@tk.hu

Abstract

In this paper, we summarise the key findings of the forum "Data4Parliaments" Parliamentary Data for a Better Democracy, organised by the OPTED network at the European Parliament on 15 June 2022. The aim of the conference was to provide a forum for the exchange of ideas related to sharing and analysing parliamentary text data between official archives and different user groups from the social sciences and the civil society.

Keywords

OPTED forum – parliaments – parliamentary data – big data analysis

 | DOI:10.1163/26668912-BJA10053

Parliaments are the epicentres of European democracies, where policies are debated and potentially shaped.[1] Researchers use legal texts and parliamentary documents to address topics as diverse as representation and responsiveness, parliamentary agendas and parliamentary organisation. While these studies offer many exciting insights into democratic decision-making and the workings of democracy, data constraints regularly force scholars to narrow the scope of their work: they can only examine a short time horizon, a single type of textual document or a specific policy area, as processing written parliamentary documents is resource-intensive.

Parliamentary functions, such as law-making, are supported by procedures that generate a large number of legal documents. The legal data contained in these documents, in a machine-readable format and preferably from multiple sources, can reveal the "black box" of governance to citizens and other public actors, thereby promoting democratic accountability and the rule of law. This is why the development of legal IT systems and platforms for processing and managing aggregated; complex legal data has become focus of recent literature.

Access to parliamentary speech and documentary data remains challenging, as parliaments continuously update their website structures without seeking to leverage the potential of big data analysis as a prime goal. In many countries today, implementing information systems for the reliable creation, storage, retrieval and dissemination of legislative documents such as bills, debate transcripts, committee reports, arguments, parliamentary minutes and laws in order to ensure the efficient and effective functioning of the legislature is still a challenge to. Even today, most parliamentary documents are still not "easily retrievable", thus constituting a direct backsliding in transparency. The availability and re-use of legal data are therefore essential for promoting transparency, accountability and trust in government institutions. This seriously hampers replication, leads to unnecessary duplication of work, prevents data linkages, and limits opportunities to learn about political decision-making.

Work Package 5 (WP5) on Parliamentary, government and legal texts of the Observatory for Political Texts in European Democracies (OPTED) project[2] engage relevant officials in national parliaments in active exchange with data scientists to achieve mutually beneficial data access opportunities. The OPTED project's activity stems from the realisation that representatives from

1 Prepared with the professional support of the Doctoral Student Scholarship Program of the Co-operative Doctoral Program of the Ministry of Innovation and Technology financed from the National Research, Development and Innovation Fund.

2 The EU-funded H2020 project involves 17 research institutions organized in 10 work packages, which collaboratively work on designing the building blocks of the infrastructure.

different disciplines have been working on all kinds of policy texts over the last 10–15 years, covering parliamentary debate texts, media texts, social media data and so on. Many of them have worked with similar texts, but not in a very coordinated way, often forming clusters that have focused on particular types of languages or countries. Hence, the OPTED network has from the outset, argued that research, not only academic research but also research in the public domain, that supports the resilience and sustainability of liberal democracy of contemporary Europe should and can rely on comprehensive assessments of communication processes. Almost all phases and areas of the political process are now if not characterised, at least signified by sorts of political texts that we have that are increasingly available in digital form and that we can analyse more systematically and comprehensively than ever before. The OPTED network aims to understand democracy better and support it, to create a more effective, sustainable, resource-efficient, innovative, democratic, non-discriminatory and essentially meaningful research environment for policy texts, placing the European research area, it is hoped, at the forefront of global development.

WP5 pursues four objectives:
– creating an inventory of parliamentary texts;
– building an integrated database of parliamentary speeches and legislation;
– establishing a better infrastructure for scientific research on parliaments;
– enhancing data access.[3]

In addition to providing a curated inventory of available parliamentary text corpora in Europe, WP5 focuses on designing infrastructures that foster the collection, harmonisation, publication and analysis of textual data related to parliamentary speeches, legislation and bill amendments and laws. The goal is to provide scholars, practitioners and the broader public with enhanced electronic access to parliamentary corpora.

To contribute to scientific research in parliamentary archives and to develop better infrastructure and knowledge transfer, as well as to identify and share good practices, a scientific meeting called "Data4Parliaments" was held by the OPTED network on 15th June 2022 in Brussels, attracting bureaucrats and data scientists from the EU and national parliaments.

The Opening Plenary stressed that the conference aimed to bring together stakeholders with different perspectives from different countries to discuss the status of analysing accessing parliamentary text data, priorities for moving ahead and the obstacles to achieving that. Finally, it identified the key action points that still need to be implemented.

3 Work package 5 of the OPTED network available at: https://opted.eu/.

The welcome addresses were followed by a panel on *The Possibilities of Utilizing Parliamentary Data,* with keynote speeches and discussions. The panellists presented the possibilities of using parliamentary data and some successful high-impact projects in each area. The findings were presented through three critical stakeholder groups: academia, data journalists and parliamentary archivists.

Miklós Sebők, Sven-Oliver Proksch, Christian Rauh and Jan Schwalbach presented the academic perspectives, introducing the so-called ParlLawSpeech (an integrated, multilingual database of parliamentary speech and legislation) database,[4] which has been published as an open-source database. However, it is only available in a form that requires experience in using the R programming language, creating entry barriers for scholars working in other environments, public officials, journalists and the interested public who wish to gain insights from the information trove of parliamentary text data. Supporting the application and development of text-as-data approaches in the political sciences, ParlLawSpeech provides large full-text vectors of parliamentary (plenary) speeches in the key legislative chambers of selected representative democracies. Previous research had used past speech data to examine how critical specific topics are in domestic politics and how government opposition dynamics look like in legislative speech, but it has not been possible so far to link those debates to actual legal change. The database produced addresses this gap.

Andrea Abellán, a member of the European Journalism Centre's Datajournalism.com project, presented the data journalism perspective. She stressed that data journalism is now an established part of the media ecosystem, with many newsrooms having a dedicated data team and others looking to create one. The formalisation of the field is less than a decade old, and the practices, skill sets, and technologies used are rapidly evolving. Its rapid evolution indicates the need for continuous snapshots to understand how data journalism is conducted and how it is changing over time. Currently, the field lacks a common and systematic approach that can help to make sense of the role, modus operandi and industry composition of data journalism. Data analysts across the globe are spending more of their time trying to gather data and cleaning it. Access to quality data is one of the biggest challenges. Moreover, the complexity of the European political jargon is another key challenge that data journalists are seeing today. In addition, data are often unsearchable and not interconnected, data journalists often have to rely on personal contacts to obtain data and there is a culture of non-transparency. This reduces public interest and increases distrust in public institutions. Access to data is limited

4 The ParlLawSpeech database is available at: https://dataverse.harvard.edu/dataverse/ParlSpeech.

and, even when available, it is not in the best format. One possible solution is to try to involve journalists in the decisions and help them to produce better, more communicative and more engaging reports and pieces by providing them with clear and regular data.

Christian Heyer and Monika Jantsch, members of the Directorate-General Information and Documentation of the Administration of the German Bundestag, discussed their parliamentary archives. The Parliamentary Documentation Division in Germany catalogues and indexes the publicly accessible materials of the German Bundestag and Bundesrat according to formal and thematic aspects (including plenary minutes, bills, motions, interpellations and various reports), so it provides information about all parliamentary processes to all members of parliament and the general public. The so-called Documentation and Information System for Parliamentary Material (DIP) primarily focuses on enabling a specific or discrete relationship or deliberative process to be the identified. The presenters pointed out that the DIP is a well-established system. Nevertheless, there are further directions for improvement, such as improving and developing the XML structure and document type definition (DTD) for plenary minutes and further developing the DIP to allow structured XML documents to be uploaded to the system and downloaded via the Application programming interfaces (APIs), as well as developing an improved, extended and, where possible, uniform XML structure for each of the nearly 80 document types.

The panel was followed by a roundtable on the intersection between parliamentary data and democracy. The discussion on *Better Parliamentary Data, More Transparency and Democracy?* focused on how greater transparency of legislative speeches, procedures and outcomes (such as laws) can contribute to the renewal of public discourse and representative democracy.

Attila Bátorfy, an investigative and visual journalist at Átlátszó/ATLO, pointed in Hungary out the lack of awareness of tools and data sources. In most cases, the Átlátszó/ATLO projects produce media content for the general public and the data they have available for analysis comes from Google Analytics. The largest project was Koronamonitor in connection with COVID-19, which reached 5.5 million users (almost half of Hungarian society). The fact that the government did not provide adequate data, which the civil organisation replaced, led to the success of this project. The project's analyses and data visualisations were also used for research information by medics and health care researchers.

Ieva Dunčikaitė, a member of the Lithuanian Manoseimas project, highlighted that full text data or quantification or system notification of text data, speech data, and legal data is helpful for data journalists. In addition to

determining who attends parliamentary sessions, the ManoSeimas.lt project also tracks the legislative footprint. The project aims to make their data available to politicians and decision-makers and to create a habit of transparency.

The second half of the conference included a showcase of data users and a roundtable of data providers. Among them, we can highlight the presentation of Michal Ovádek, the maintainer of the eurlex R package on European law. The presentation argued that, currently, research based on observational data is generally difficult to replicate but APIs can help overcome the "replication crisis". Good public APIs can help to improve data collection practices and, down the line, the replicability of scientific research. The presentation highlighted the two main ways to improve data collection with high-quality open data APIs. The first is for everyone to work from the same data. Even if the analysis requires some manipulation, at least we know that the input is the same for everyone who touches the data. The second is transparency: good quality APIs or good APIs are based on how data are put together and connect to the real world. It is also easier for researchers to use the data if they are in the correct format and if they do not have to spend days collecting data from websites in various forms or converting them into different formats, and so on. Instead of people locking away their data sets and computers, these things could be in public (e. g. on the GitHub platform), primarily when they are maintained by a public institution, such as a parliament.

The key funding of the conference is that the availability of parliamentary documents and legal texts is still very much dependent on official authorities. The operation of the three branches of power in modern democracies generates enormous amounts of data. As a result, and to promote transparency, many legal documents are now available on the internet. However, there is a limitation because such information is usually published in an unstructured format. Infrastructure in the form of databases and various tools is often needed to identify meaningful patterns and promote new insights. One of the biggest challenges in this context is building the infrastructure. Overall, the fact that different legal documents are openly available does not mean that a database is easily accessible. Most parliamentary documents are still stored in an image or scanned PDF format. The text does not need to be converted (using optical character recognition (OCR) into machine-encoded text formats) for full use in research.

APIs are a core technology solution that needs attention in digital government agendas. APIs can help governments manage organisational change for digitalisation and facilitate innovation in public processes and public service delivery. Governments should be encouraged to incorporate APIs into their digital strategies to support these policy goals. Awareness of the digital

ecosystem should also be promoted. The involvement of EU governments and the private sector is essential to developing and building interoperable government IT platforms connecting multiple stakeholders.

A joint legislative portal in the European Parliament, a legislative observatory with more and better manageable data than is currently available, should be created to increase transparency and data availability, opening up the possibility for proper data analysis. The policy should have a strategic approach to accessibility, transparency and openness. A proactive approach to the publication of data and the involvement of civil society and the public in consultations or other participatory mechanisms is also necessary. Strengthening interaction with the public can be enhanced through communication activities by producing simple, digestible reports, explanations and other materials.

The best and forward-looking practices and suggestions collected during the conference will be summarised by WP5 members in a White Paper. This report will shed light on rise good approaches and action points for further dissemination and cooperation in the expectation that the legislative authority will also take the current shortcomings into account and implement the proposed initiatives.