Digital Parliamentary Data in Action (DiPaDA 2022) – Introduction

Matti La Mela[1,2], Fredrik Norén[3] and Eero Hyvönen[2,4]

[1]Department of ALM, Uppsala University, Sweden
[2]Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland
[3]Humlab, Umeå University, Sweden
[4]Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland

Abstract

The workshop *Digital Parliamentary Data in Action* (DiPaDA 2022) was organised in Uppsala on March 15, 2022, co-located with the 6th *Digital Humanities in the Nordic and Baltic Countries Conference* (DHNB). These workshop proceedings reflect the aims of the workshop to foster interaction and stimulate conversations between humanities, social sciences, and computational sciences – representing scholars from the Nordic region and beyond that work with digital parliamentary data. The contributions in the proceedings present results of ongoing research on creating and using historical and present parliamentary data to study parliamentary culture, politics, language use, and the media. Moreover, the contributions offer novel perspectives on applying, curating, and representing this key societal data, and discuss the future opportunities and challenges in such research.

Keywords

parliamentary data, digital humanities, parliamentary research, interdisciplinary research, Parla-CLARIN

1. Introduction

Parliaments are places of (reactive) actions – actions that are also recorded and fuel research about these assemblies and societies at large [1]. While parliaments have the power to transform a society's future, their documents constitute a democratic resource for the present day that, in turn, can be used to remodel our understanding of the past. Recent years have seen the emergence of growing data collections, services, and user interfaces for accessing and using digital parliamentary documents. These datasets – along with their related infrastructures and available computational tools – have enabled researchers to engage in novel and multidisciplinary approaches to explore and study, for instance, the culture, language, media, polemics, and consensus of parliaments. The research potential is considerable – even exceptional [2].

At the same time, research projects have often signaled the limitations and challenges concerning the digital services related to digital parliamentary data, in their scope, structure, quality, and usability [1]. This has led to demands for collaborative projects about, among other things,

Digital Parliamentary Data in Action (DiPaDA 2022) workshop, Uppsala, Sweden, March 15, 2022.

matti.lamela@abm.uu.se (M. La Mela)
0000-0003-0340-9269 (M. La Mela); 0000-0001-8820-1082 (F. Norén); 0000-0003-1695-5840 (E. Hyvönen)

[1] This was for example a specific topic at one of the Text Mining Parliamentary Democracy seminar, "Practices of Parliament", arranged on March 18 in 2022, https://www.umu.se/en/humlab/program-activities/text-mining/.

OCR-quality, metadata, and possibilities for using uniform standard data formats and structures for representing parliamentary documents and data, such as the XML/TEI-based Parla-CLARIN[2] [3] and linked data [4, 5, 6, 7]. Furthermore, improved quality and accessibility of parliamentary data accentuate issues of democracy, beyond strictly academic interests, raising questions about how such datasets and user interfaces can be used to empower citizen participation as well as enhance the transparency of political work and decision making. As a result, scholarly activities around digital parliamentary documents – be it digitization projects, annotation projects, or building infrastructures for research – are often also embedded in critical societal discussions that have consequences for various sectors and groups in society.

Digital Parliamentary Data in Action (DiPaDa22) was a workshop arranged in Uppsala on March 15, 2022, co-located with the 6th *Digital Humanities in the Nordic and Baltic Countries Conference*, DHNB 2022). The workshop was a response to the interrelated potentials and challenges described above, focusing on the usage of parliamentary data in, foremost, humanities and social science research, including the work of curating data and developing tools and user interfaces through interdisciplinary collaborations. The results of the workshop are published in these proceedings.

The idea behind the workshop originates from the work conducted in the research projects we editors of the proceedings have been involved in, i.e., ParliamentSampo (SEMPARL)[3] and Welfare State Analytics (WeStAc)[4], as well as from the collaborative networks that have supported and facilitated our work, such as Parla-Clarin, ParlaMint[5], and Text Mining Parliamentary Democracy[6]. Based on experiences from these research activities, we identified a need to gather scholars across different disciplines in a joint venture to showcase and discuss results and ongoing work related to creating, publishing, and using digital parliamentary datasets for research purposes. Our interest was to investigate the results of the interdisciplinary work on parliamentary culture and language, media, political concepts and networks in different national and historical contexts, and in this way, also to bring a novel and complementary perspective to the CLARIN meetings and related work on parliamentary corpora and language resources (see [8]). Thus, the aim of the workshop was to foster interaction and stimulate conversations between humanities, social sciences, and computational sciences within the Nordic region and beyond. On the one hand, the contributions in these workshop proceedings present a selection of ongoing research on digital parliamentary data today and, on the other hand, point to future opportunities and challenges in this field.

2. Scholarly approaches to digital parliamentary data

Research that engages with digital parliamentary data is vast in both its disciplinary scope and approaches to using such material. To write a comprehensive state-of-the-art review of such research is a difficult, maybe even an impossible undertaking. The purpose of this preface is therefore rather to present two broad, partly interconnected, scholarly approaches to digital

[2] https://github.com/clarin-eric/parla-clarin

[3] https://seco.cs.aalto.fi/projects/semparl/en/

[4] https://www.westac.se/en/

[5] https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

[6] https://www.umu.se/en/humlab/program-activities/text-mining/

parliamentary data with some examples of related studies. The first is an infrastructure-oriented approach, aiming at digitizing, curating, and tool-building. The second is a research-oriented approach with empirical and methodological focuses. These two approaches also reflect the contributions in the proceedings.

2.1. Infrastructure-oriented approaches to digital parliamentary data

In recent decades, past and contemporary parliamentary work and its related documents – records of the plenary debates, government bills, voting data, biographies of members of parliament et cetera – have been digitized and to various degree annotated, both by the parliaments themselves and by research initiatives and cultural heritage institutions. The growing digital datasets have further attracted scholarly interest in digitization and modelling processes since these not only make documents available but also determine how the material can be published and studied.

Today, several national parliamentary datasets are publicly available.[7] However, these collections are often provided in various formats and data structures (for example, [9, 6]), partly due to nation-specific parliamentary procedures and lack of coordination between actors responsible to digitize the parliamentary documents in different countries. Many national parliaments have responded to these demands of making their documents more accessible by transforming them into standard digital forms. Still, these digitization projects are not always in line with what is desired from scholarly perspectives. Research needs could, for example, demand thorough linguistic annotation of parliamentary debates and comprehensive metadata about the members of parliament or how the parliament works. The awareness of such challenges and potential conflict of interests have created a need for researchers to engage in collaborative projects that involve scholars, parliaments, and cultural heritage institutions. Infrastructural approaches deal with various aspects of data, such as how it is annotated, linked, enriched, and used through human interfaces or machine Application Programming Interfaces (API) for data analysis.

Besides digitization, recent projects have focused on curating and parsing parliamentary data into specific annotation schemes [10, 11, 12]. Many projects, especially those related to the CLARIN infrastructures, have built parliamentary debate corpora that include various linguistic annotations and detailed speaker metadata.[8] Parla-CLARIN is a TEI-based scheme developed to build a common framework for annotating parliamentary corpora in a structured manner for session structure, speaker metadata, different linguistic information et cetera [3]. The ParlaMint initiative is a similar infrastructure, based on Parla-CLARIN's basic structure but with a more detailed annotation scheme. It aims to enhance the homogenization of parliamentary debate corpora for comparative work across national borders [13]. The project includes today several partners from different countries.[9]

Several research projects have also used linked data and semantic web technologies to enrich the parliamentary data both internally and using external data sources. In the project Linked Data of the European Parliament (LinkedEP), for example, the debates of the European Parliament and the biographic information of the speakers were combined into a linked open dataset

[7] See, for example, https://data.parliament.uk or https://www.bundestag.de/services/opendata.

[8] See, for example, https://www.clarin.eu/resource-families/parliamentary-corpora.

[9] https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

with connections to external sources, such as GeoNames and DBpedia [4]. Linked open data has also been employed in national-level projects, such as the Latvian LinkedSaeima, covering the period 1993–2017 and using the solutions of LinkedEP [5]. The Finnish ParliamentSampo project uses both the Parla-CLARIN scheme and linked open data for creating a unified linked open data service and a semantic portal of Parliament of Finland (1907–2022) [7], including a knowledge graph of the speeches [6] as well as of the members of parliament and an ontology of the parliament [14].

In addition to creating and publishing data, another infrastructural focus is the development of applications and user interfaces to allow access to parliamentary datasets and to enhance their use. One example is the Canadian LIPAD[10], with a user interface to search and navigate the recorded debates held in the Canadian parliament, from the early 20th century onwards [9]. A few other examples are the portal of the Italian parliament in linked open data format and with multimedia content[11], the UK Hansard open user interface [15], the German Bundestag documents[12], and ParliamentSampo that includes also data-analytic tools in addition to search and data exploration facilities. A comprehensive list of such applications and interfaces would be too long to be presented here.

2.2. Research-oriented approaches to digital parliamentary data

Another scholarly approach to digital parliamentary data is research-oriented. Here, the focus shifts from digitizing, curating, and making datasets available to using the data collections and infrastructures for research purposes. Such research is often quantitative, but not always, and is combined with qualitative approaches. For scholars aiming for diachronic studies, digitized parliamentary datasets can offer opportunities to write histories with the scope of a *longue durée* [16]. This is especially possible in countries like Sweden, where parliamentary documents date back to the 16th century, which today exist in a digital format. Other countries, such as Slovenia or Poland, do not have the same possibilities because their modern parliaments were established only a few decades ago. However, these datasets could still be considered massive in scale, not least because they often compass different parliamentary document categories, such as debate records, government bills, and various committee reports.

Digitized parliamentary documents constitute a rich source for empirical research of such things as legislation, political culture, language, and democratic development. Some researchers – often political scientists and parliamentary historians – are interested in studying the parliament itself with its procedures, norms, and political behaviour. Other scholars from different disciplines may use parliamentary data as source material to study various cultural and societal phenomena from a parliamentary point of view. While these research approaches often have an empirical orientation and deal with both contemporary and historical periods, yet another category of researchers could be described as those interested in using such material to train, develop, and assess statistical models and alike. These researchers might be less interested in the content and contexts of parliamentary documents. In the proceedings of *Digital Parliamentary Data in Action*, however, the focus lies on empirical research of various kinds.

[10] https://lipad.ca/

[11] https://storia.camera.it/

[12] https://opendiscourse.de

One could divide empirical-oriented research that uses digital parliamentary documents into two kinds of categories, although it can sometimes be difficult to differentiate them. In the first, we find studies that are more driven by research questions and hypotheses, and in the second, there are those that put computational methods in the centre to explore the potential of what these tools can contribute to various research fields. Regarding the first approach, we find examples of studies that, for instance, examine how parties in government versus parties in opposition talk about a certain issue by using topic modelling as a method [17, 18], or how discursive shifts take place when female representation increases in parliament using Pearson correlation [19]. In some cases, it is not the textual content of the documents that are of interest but the generated metadata, for instance, members of parliament's gender, party belonging, constituency, birth et cetera. Such studies could for example highlight factors that influence the likelihood of women taking the parliamentary floor [20]. Moreover, within the field of conceptual history, scholars have used parliamentary documents as one of several sources to explore the emergence and development of so-called key concepts over longer periods of time. Digital parliamentary datasets have recently attracted researchers to use computational methods – also simple and straightforward methods such as n-grams and co-occurrences – to study concepts such as "ideology" [21], "internationalism" [22] and "access rights to nature" [23].

The other category gathers studies geared towards using parliamentary datasets to explore and demonstrate the scholarly potential of employing computational methods and the knowledge that such undertakings can contribute to different fields. Examples of such research questions are: how can topic models be used to study political attention [24] and, similarly, how can dynamic topic models help historians to explore historical change in parliamentary debates about infrastructure policies during the 19th century [16]? What new historical knowledge can more simple statistical methods teach about the far right in parliaments during the World War II [25]? Is it possible to use word embeddings, for instance, to identify ideological positions among poliical parties in parliamentary debates [26]? How can one use information theory to model the degree of innovative political speech in the French revolution's first parliament [27]?

3. DiPaDA22 – a workshop in action

The initiative for setting up the workshop *Digital Parliamentary Data in Action* was taken by Matti La Mela (Uppsala University and University of Helsinki (HELDIG), ParlamentSampo project) in 2021, who formed a workshop organizing committee together with Fredrik Norén (Umeå University, WESTAC project) and Eero Hyvönen (Aalto University, University of Helsinki (HELDIG), ParlamentSampo project). A programme committee (PC) of the workshop was established that, besides the workshop organizing committee, included Kaspar Beelen (The Alan Turing Institute), Kimmo Elo (University of Turku), Tomaž Erjavec (Jožef Stefan Institute), Darja Fišer (University of Ljubljana), Jo Guldi (Southern Methodist University), Laura Hollink (Centrum Wiskunde & Informatica), Pasi Ihalainen (University of Jyväskylä), Måns Magnusson (Uppsala University), Bruno Martins (University of Lisbon), Costanza Navarretta (University of Copenhagen), and Jouni Tuominen (University of Helsinki).

A call for papers was formulated and circulated during the autumn of 2021. In total, 17 papers were submitted to the workshop in early January: 14 long papers and three short papers. The group of reviewers consisted of the programme committee and other scholars with expertise related to parliamentary research from humanities, social sciences, and computer sciences. The organizing committee assigned for each paper two reviewers. After a review process of three weeks, the organizing committee summarized the results for the PC, which convened on February 14. At that meeting, it was decided to accept 13 papers and reject four papers for the workshop proceedings. However, all papers were welcomed to be presented at the workshop.

After communicating the programme committee decisions, the authors were given three weeks time to revise their papers based on the reviewers' comments. A week before the workshop, all revised papers were circulated among the workshop participants. The workshop *Digital Parliamentary Data in Action* took place in Uppsala on March 15, hosted by the Department of ALM and co-located with the 6th *Digital Humanities in the Nordic and Baltic Countries Conference* (15–18 March 2022). The structure of the full-day workshop, as well as these proceedings, was divided into four blocks: Historical concepts and perspectives, Interfaces and transformation of data, Contemporary politics and media, and Language and annotation.

Acknowledgments. We would sincerely like to thank all who have supported the workshop, especially the programme committee, the anonymous peer-reviewers, the session chairs, and all participating authors.

References

[1] C. Benoît, O. Rozenberg (Eds.), Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures, Edward Elgar Publishing, 2020. doi:10.4337/9781789906516.

[2] J. Guldi, The official mind's view of empire, in miniature: Quantifying world geography in Hansard's parliamentary debates, Journal of World History 32 (2021) 345–370. doi:10.1353/jwh.2021.0028.

[3] A. Pancur, T. Erjavec, The siParl corpus of Slovene parliamentary proceedings, in: Proceedings of the Second ParlaCLARIN Workshop. Marseille, France, May 2020, ????

[4] A. Van Aggelen, L. Hollink, M. Kemman, M. Kleppe, H. Beunders, The debates of the European Parliament as Linked Open Data, Semantic Web 8 (2017) 271–281. doi:10.1007/s42001-019-00060-w.

[5] U. Bojārs, R. Darģis, U. Lavrinovičs, P. Paikens, LinkedSaeima: A linked open dataset of Latvia's parliamentary debates, in: Semantic Systems. The Power of AI and Knowledge Graphs. SEMANTiCS 2019, Springer, 2019, pp. 50–56. doi:10.1007/978-3-030-33220-4_4.

[6] L. Sinikallio, S. Drobac, M. Tamper, R. Leal, M. Koho, J. Tuominen, M. L. Mela, E. Hyvönen, Plenary debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN markup, in: 3rd Conference on Language, Data and Knowledge, LDK 2021, Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2021, pp. 8:1–8:17. URL: https://drops.dagstuhl.de/opus/volltexte/2021/14544/pdf/OASIcs-LDK-2021-8.pdf.

[7] E. Hyvönen, L. Sinikallio, P. Leskinen, M. L. Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, Finnish parliament on the semantic web: Using ParliamentSampo data service and semantic portal for studying political culture

and language, in: Digital Parliamentary data in Action (DiPaDa 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, CEUR Workshop Proceedings, 2022. In this volume.

[8] D. Fišer, M. Eskevich, F. de Jong (Eds.), Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II), European Language Resources Association (ELRA), Paris, France, 2020. URL: https://lrec2020.lrec-conf.org/media/proceedings/Workshops/Books/ParlaCLARIN2book.pdf.

[9] K. Beelen, T. A. Thijm, C. Cochrane, K. Halvemaan, G. Hirst, M. Kimmins, S. Lijbrink, M. Marx, N. Naderi, L. Rheault, R. Polyanovsky, T. Whyte, Digitization of the Canadian parliamentary debates, Canadian Journal of Political Science 50 (2017) 849–864. doi:10.1017/S0008423916001165.

[10] E. Lapponi, M. G. Søyland, E. Velldal, S. Oepen, The talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016, Lang Resources & Evaluation 52 (2018) 873–893. doi:10.1007/s10579-018-9411-5.

[11] M. Ogrodniczuk, Polish parliamentary corpus, in: ParlaCLARIN, ????

[12] B. Hladká, M. Kopp, P. Straňák, Compiling Czech parliamentary stenographic protocols into a corpus, in: Proceedings of ParlaCLARIN II Workshop, Language Resources and Evaluation Conference (LREC 2020), Marseille, , 11–16 May 2020, 2020, pp. 18–22. URL: https://aclanthology.org/2020.parlaclarin-1.4.pdf.

[13] T. Erjavec, M. Ogrodniczuk, P. Osenova, et al., The ParlaMint corpora of parliamentary proceedings, Language Resources and Evaluation (2022). doi:10.1007/s10579-021-09574-0.

[14] P. Leskinen, E. Hyvönen, J. Tuominen, Members of Parliament in Finland knowledge graph and its linked open data service, in: Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6–9 September 2021, Amsterdam, The Netherlands, IOS Press, 2021, pp. 255–269. URL: https://ebooks.iospress.nl/volumearticle/57420. doi:10.3233/SSW210049.

[15] J. Guldi, S. Buongiorno, Hansard viewer, 2021. doi:https://github.com/stephbuon/hansard-shiny.

[16] J. Guldi, Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change, Technology and Culture 60 (2019) 1–33. doi:https://muse.jhu.edu/article/719944.

[17] M. Moilanen, S. Østbye, Doublespeak? Sustainability in the Arctic: A text mining analysis of Norwegian parliamentary speeches, Sustainability 13 (2021) 1–15. doi:https://doi.org/10.3390/su13169397.

[18] M. Magnusson, R. Öhrvall, K. Barrling, D. Mimno, Voices from the far right: a text analysis of Swedish parliamentary debates, SocArXiv. April 4. (2018). doi:10.31235/osf.io/jdsqc.

[19] L. Blaxill, K. Beelen, A feminized language of democracy? The representation of women at Westminster since 1945, Twentieth Century British History 27 (2016) 412–449. doi:10.1093/tcbh/hww028.

[20] H. Bäck, M. Debus, When do women speak? A comparative analysis of the role of gender in legislative debates, Political Studies 67 (2019) 576–596. doi:https://doi.org/10.1177/0032321718789358.

[21] J. Kurunmäki, J. Marjanen, How ideology became isms: A history of a conceptual coupling,

in: H. Haara, K. Stapelbroek, M. Immanen (Eds.), Passions, Politics and the Limits of Society, volume 1 in the Helsinki Yearbook of Intellectual History series, De Gruyter Oldenbourg, 2020, pp. 291–318. doi:10.1515/9783110679861.

[22] P. Ihalainen, A. Sahala, Evolving conceptualisations of internationalism in the UK parliament: Collocation analyses from the League to Brexit, in: M. Fridlund, M. Oiva, P. Paju (Eds.), Digital Histories: Emergent Approaches within the New Digital History, Helsinki University Press, 2020, pp. 199–219. doi:10.33134/HUP-5-12.

[23] K. Kettunen, M. La Mela, Semantic tagging and the Nordic tradition of everyman's rights, Digital Scholarship in the Humanities (2021). doi:10.1093/llc/fqab052.

[24] K. M. Quinn, B. L. Monroe, M. Colaresi, M. H. Crespin, D. R. Radev, How to analyze political attention with minimal assumptions and costs, Political Studies 54 (2010) 209–228. doi:https://doi.org/10.1111/j.1540-5907.2009.00427.x.

[25] H. Piersma, I. Tames, L. Buitinck, J. van Doornik, M. Marx, War in parliament: What a digital approach can add to the study of parliamentary history, Digital Humanities Quarterly 8 (2014) 1–13. doi:https://hdl.handle.net/11245/1.439246.

[26] L. Rheault, C. Cochrane, Word embeddings for the analysis of ideological placement in parliamentary corpora, Political Analysis 28 (2020) 112–133. doi:10.1017/pan.2019.26.

[27] A. T. J. Barron, J. Huanga, R. L. Spang, S. DeDeo, Individuals, institutions, and innovation in the debates of the french revolution, PNAS 115 (2018) 4607–4612. doi:https://doi.org/10.1073/pnas.1717729115.