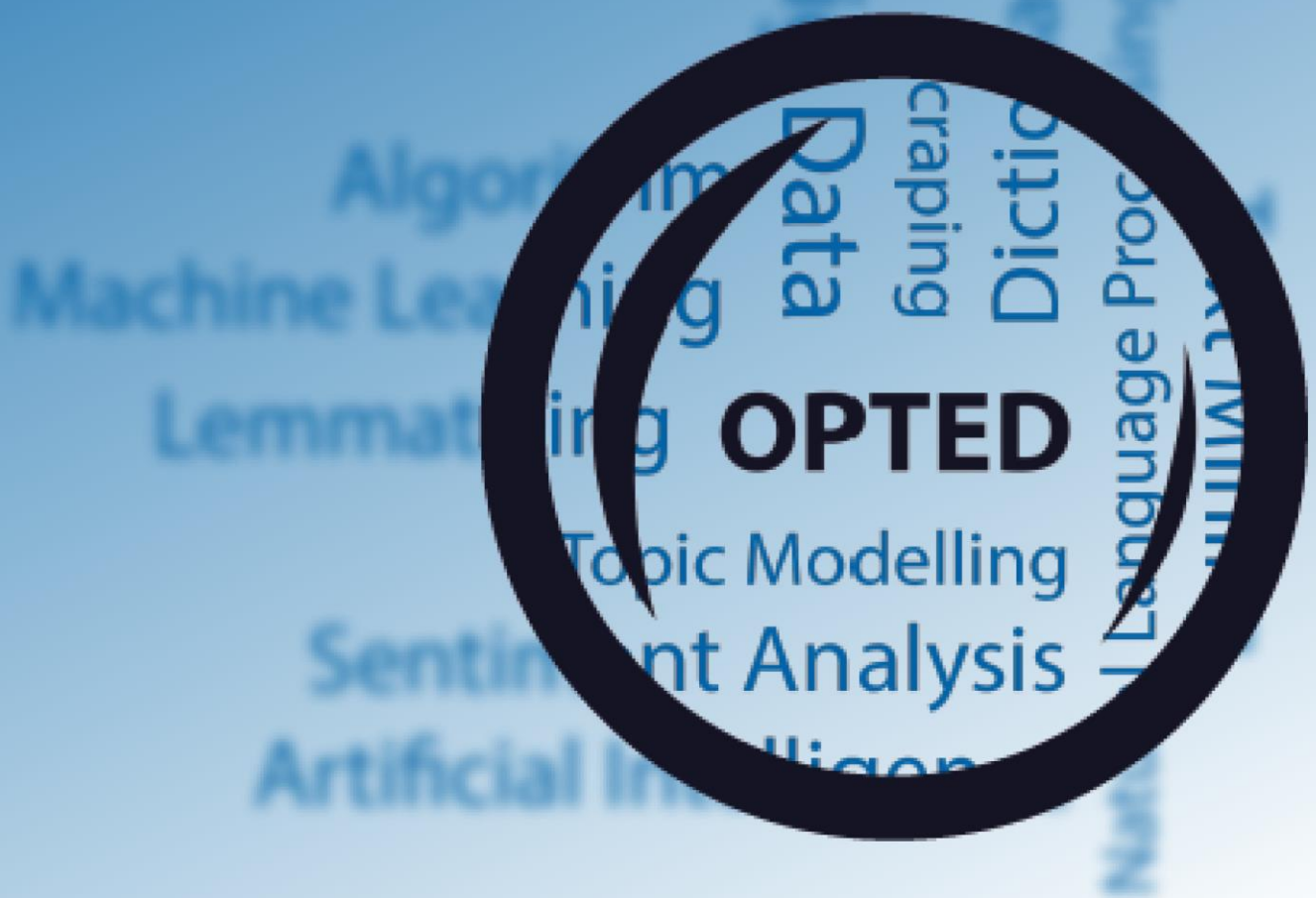
OPTED

Feasibility report: Evaluation of a research infrastructure for the analysis of texts from political organizations

Christoph Ivanusch

WP4 members: Bernhard Weßels, Swen Hutter, Pola Lehmann, Zachary Greene, Heike Klüver, Tobias Burst, Christoph Ivanusch, Sven Regel, Thomas Schober and Lisa Zehnter



Disclaimer

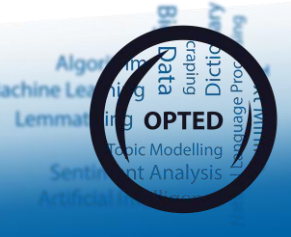
This project has received funding from the European Union's Horizon 2020 Research & Innovation Action under Grant Agreement no. 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report



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

Feasibility report: Evaluation of a research infrastructure for the analysis of texts from political organizations

Deliverable 4.7

Author: Christoph Ivanusch[1]

WP4 members: Bernhard Weßels[1], Swen Hutter[1], Pola Lehmann[1], Zachary Greene[3], Heike Klüver[2], Tobias Burst[1], Christoph Ivanusch[1], Sven Regel[1], Thomas Schober[3] and Lisa Zehnter[1]

[1] WZB Berlin Social Science Center
[2] Humboldt-Universität zu Berlin
[3] University of Strathclyde, Glasgow

Due date: September 2023

Executive summary

This deliverable (D4.7) summarizes our findings of the past 36 months and provides a feasibility report for a research infrastructure in regard to the analysis of texts from political organizations (i.e., political parties and interest groups). First, we summarize our findings on the availability, accessibility, and usability of existing text collections from political organizations. We thereby revisit the actors and texts of interests (D4.1: Alberto & Klüver, 2021), our WP4 inventory of existing text collections from political organizations (Greene, Ivanusch, Lehmann, Schober et al., 2021) as well as the current state of the field in terms of data availability (D4.2: Greene, Ivanusch, Lehmann & Schober, 2021; D4.3: Ivanusch, 2021; D4.4: Ivanusch, Lehmann, Balluff & Scotto di Vettimo, 2023). Over the course of the project, we found that although an increasing number of texts from political organizations becomes available, major gaps remain. Several important types of documents are unavailable or difficult to find and/or access. Therefore, we argue that OPTED can and should play an important role in acting as an infrastructure, advocate and strategic partner to make texts from political organizations (better) available for a variety of users and help them to connect the resources to texts from other actors. Often organizations might not be aware of the potential value of their textual resources for research and OPTED would be an important broker in terms of providing knowledge but also technical support.

Then, we review the insights on the opportunities and challenges for the computer-based analysis of texts from political organizations that we gained from deliverables D4.5 (Ivanusch, Burst & Zehnter, 2022) and D4.6 (Ivanusch, 2023). In a first step, we evaluated different approaches to the computer-based analysis of different types of party texts, providing researchers with insights on existing tools and their respective strengths and weaknesses. Then, we showcased in two case studies how (innovative) methodological approaches can be applied to enable research that covers and compares different types of text relevant to political competition in contemporary democracies. The field of automated text analysis is a fast growing and constantly developing field. This does not always make it easy for individual researchers to stay on top of the methodological developments, be aware of the latest tools and pitfalls one needs to consider when using them. Here, OPTED can offer guidance and assistance to researchers in this field by bringing together experts in the field of political text analysis, who can share their knowledge with the community and/or develop new tools.

Finally, we discuss the role of OPTED in enabling (better) research on political organizations by highlighting the advances made over the past 36 months and also by evaluating current limitations and important future steps. We argue that OPTED has made significant gains in setting up an infrastructure that helps users find resources related to texts from political organizations (e.g., data archives, data sets, tools). Furthermore, OPTED has contributed to more intense networking among the research community, and with stakeholders relevant to the study of political organizations such as data owners (e.g., political foundations) and potential data users (e.g., data journalists). Building on this, we propose that OPTED should continue to develop its infrastructure and further function as a centre of expertise and engage with stakeholders to improve the conditions for research on texts from political organizations.

1. Inventory of existing text collections from political organizations

The first major tasks of our work package (WP4) were first to define the actors (i.e. political parties and interest groups) as well as text types (e.g. manifestos, press releases, social media) of interest when gathering information about political organizations or studying them (see D4.1: Alberto & Klüver, 2021) and second to create an inventory of existing text collections from political organizations (Greene, Ivanusch, Lehmann, Schober et al., 2021). For this inventory, we relied on case-specific knowledge from the WP4 team and keyword searches to locate data archives and data sets that store texts from political organizations in the former EU28 (Belgium, Bulgaria, Denmark, Germany, Estonia, Finland, France, Greece, United Kingdom, Ireland, Italy, Croatia, Latvia, Lithuania, Luxembourg, Malta, the Netherlands, Austria, Poland, Portugal, Romania, Sweden, Slovakia, Slovenia, Spain, Czech Republic, Hungary and Cyprus) and in Norway and Switzerland as well as on the level of the European Union (EU). The resulting inventory (Greene, Ivanusch, Lehmann, Schober et al., 2021) served as an important starting point for a number of important insights that we drew from the project.

In a first step, we used the information on existing text collections contained in the inventory to gain a detailed overview of the current state of the field in terms of data availability, access and usability (see D4.2: Greene, Ivanusch, Lehmann & Schober, 2021; D4.3: Ivanusch, 2021). First, we found that significant differences in data availability exist across the covered countries. While several existing text collections focus on German-speaking countries (i.e. Austria, Germany, Switzerland), France, the United Kingdom, Spain and – to a lesser extent – the Scandinavian countries (Denmark, Finland, Norway, Sweden), Belgium and the Netherlands, data is much scarcer in the case of many Central and Eastern European countries as well as for Italy and Greece.

Second, we showed that few systematic data collections of the internal rules and proceedings of political parties and interest groups exist. While external communication (e.g., manifestos, press releases, public speeches) is more easily available, we have little data on important internal processes. Furthermore, the data are often limited to the behaviour of political organizations in relation to official state targets, particularly so for interest groups. Thus, we lack data to uncover power dynamics when few official records exist outside of more formal environments.

Third, many existing text collections suffer from a recency bias as data from current and recent events are much easier to find than older records from parties and interest groups. This lack of historical data thereby "limits the formation of an accurate and shared public memory and constrains individuals' ability to hold organizations accountable for their past behaviors and messages" (D4.2: Greene, Ivanusch, Lehmann & Schober, 2021, p. 4).

Finally, existing text collections come in multiple different formats and therefore vary considerably in terms of accessibility and usability. The data archives and data sets that we identified in our inventory are stored on various repositories, platforms or individual websites, have different rules in terms of access (free, registration required, restricted), need to be retrieved in different ways (e.g. webscraping, download, API) or

come in different file formats (e.g. text/html, pdf, csv, dta, sav, rds, rdata). Furthermore, they vary in their provision of full and ready-to-use texts, their corpus format as well as the availability of annotations and codebooks or meta data. Thus, clear standards or guidelines regarding the publication of text collections in a user-friendly way are largely missing.

Over the course of the project, we have not only uncovered this current state of the field, but also have taken several steps to mitigate some of the existing limitations and advance data availability and access. In a first step, we have proposed criteria for publishing user-friendly collections of text from political organizations (D4.3: Ivanusch, 2021):

- *Data storage:* Data sets must be "easy-to-find" (see also: FAIR Data Principles).
- *Easy and free access:* Data sets should be easily and freely accessible for users (see also: FAIR Data Principles).
- *Way of text retrieval:* Text collections should be easy to retrieve. Ideally, they can be easily downloaded by users or accessed via a well-documented API.
- *File type:* Textual data should be made available in suitable file types.
- *Availability of full texts*: If possible, full texts should be available.
- *Ready-to-use texts:* We advocate providing ready-to-use texts when publishing text collections.
- *Corpus type:* The type of text provision needs to be user-friendly. The provision (e.g., individual documents, corpus) thereby depends on the type and number of texts included in the data set as well as the respective users.
- *Annotation:* We advocate publishing annotations in addition to the texts whenever they are available.
- *Codebooks:* Codebooks should be published as supplementary material if data sets include annotated texts.

In a second step, we (WP4) have used our inventory of existing text collections from political organizations to contribute to the OPTED platform, which will make relevant data easier to find for a variety of users. The platform will be available for the public on October 1st at https://meteor.opted.eu and will allow to search for different types of data archives, data sets, political organizations, media sources, parliamentary and governmental data, but also scientific publications and (text analysis) tools.[1] All these resources are easy to find and linked to one another in a user-friendly way. The integration of our WP4 inventory allows users to search for political organizations and related data resources.[2] User are then provided with essential information on the resources, such as a brief description of each resource and its properties as well as URLs and/or DOIs to enable quick navigation to the resource. These features allow users to quickly gather information about

[1] The exact platform architecture and the properties of the current prototype are described in more detail in other deliverables from WP3 and WP9 (see D3.2: Balluff, Stecker, Boomgaarden & Waldherr, 2023; D9.4: Banducci, Scotto di Vettimo, Gelovani, Theocharis & Dhamal, 2022; D9.5: Scotto di Vettimo, Banducci, Balluff, Gelovani & Theocharis, 2023).

[2] Deliverable D4.4 describes the integration of the WP4 inventory in the platform in more detail (D4.4: Ivanusch, Lehmann, Balluff & Scotto di Vettimo, 2023).

relevant political organizations in specific countries, get a quick overview over existing text resources and helps them identify the most useful resource for their own query/research. Furthermore, the detailed information provided for each resource and the architecture of the platform also allow users to link different resources based on specific variables, such as country, political actor or period.

In addition, we (WP4 and the broader project) have followed up on the proposed criteria outlined above by getting in touch with data providers and owners to discuss best-practices for publishing text collections from political actors in a user-friendly way. While WP5 kicked-off this discussion by inviting different stakeholder groups (political scientists, data journalists, bureaucrats) to a conference on the challenges and potentials of legislative text data (D5.4: Sebők, Proksch & Rauh, 2022), the final OPTED conference has expanded this discussion to include different types of political texts. Therefore, we invited relevant data owners/providers to discuss the storage and provision of texts from political organizations. For example, we invited Andreas Marquet, the Chief Digital Officer (CDO) and Head of the Infrastructures and Digital Policy Issues Unit at the Archive of Social Democracy at the Friedrich-Ebert-Stiftung in Germany, and Gerrit Voerman, the director of the Documentation Centre for Dutch Political Parties of the University of Groningen, to participate in a panel discussion on the perspective of data owners and the provision of textual data. This intense networking has helped to connect different stakeholders and exchange the respective needs and interests. This is a major step to make text collections better available and more accessible, but more work is needed to implement common practices and standards across countries and domains.

Both the development of the OPTED platform and the community-building and networking with different stakeholders are important steps that will enable (better) research on political texts in the future. As described above, OPTED has made considerable progress in this regard, but more work is certainly needed. On the one hand, developing and implementing common standards and practices for making text collections available across countries and domains is a long-term project. Here, OPTED needs to establish itself as an important advocate and strategic partner offering expertise on data management and publishing. On the other hand, OPTED should further develop its platform to provide easy access to political text collections for different user groups beyond researchers, such as data journalists and policy makers, but also interested citizens. Overall, we are confident that OPTED is in a good position to continue to advocate for and enable better availability, access and usability of texts from political organizations.

2. Opportunities and challenges for the analysis of texts from political organizations

Besides the creation of the WP4 inventory, its integration in the OPTED platform and the mapping of the current state of the field in terms of data availability, we also explored the opportunities and challenges for the (computer-based) analysis of texts from political organizations. The analysis of texts has a long tradition in the social sciences and is a central tool to understand a broad range of political processes. Nowadays, the amount and variety of political texts (e.g., manifestos, press releases, social media) as well as the abundance of

(computer-based) tools offer researchers numerous opportunities, but also a broad range of challenges. We therefore engaged intensively with the computer-based analysis of political texts, thus providing researchers with insights on existing tools and showcasing how (innovative) methodological approaches can be applied to enable research on texts from political organizations.

First, we performed an in-depth analysis of challenges and opportunities for the comparative study of political text types (D4.5: Ivanusch, Burst & Zehnter, 2022). In this report, we provided a review of existing and widely used computer-based text analysis methods in the social sciences and then discussed the challenges arising from the (comparative) study of different text types (e.g., manifestos, press releases, social media, speeches). Here, we focused on the characteristics of the different text types and their implications for the application of computer-based text analysis methods. Against this background, we then evaluated the applicability of tools for computer-based topic classification across text types in a practical application. We evaluated and discussed how unsupervised (LDA), semi-supervised (Newsmap) and supervised (Naïve Bayes, BERT) approaches performed in classifying manifestos, press releases, parliamentary speeches and tweets from political parties in Austria, Germany and Switzerland (01.01.2019-26.09.2021). Based on the findings, we then examined the main advantages and limitations of each approach as well as their applicability to different types of political texts.

This systematic review and test of computer-based tools, their strengths and weaknesses as well as the challenges when working with different types of political text provides researchers with an important point of reference for their own use cases. Even though, we offered important insights, significant challenges remain when it comes to the analysis of different types of political texts. Different types of tools perform differently depending on the type of text and this may be even more pronounced across different languages. Thus, analysing different types of text from political organizations requires substantial technical and case-specific knowledge as well as extensive resources for implementation and validation. Technological developments, such as the advent of large language models, will certainly drive this field forward, but also raise a number of practical and ethical questions. Here, OPTED can serve as an important actor that stays-up-to-date, monitors current developments and provides expertise and guidance for the application of computer-based text analysis tools in the social sciences and beyond. Potential avenues to fulfil these functions are tool development, tutorials and workshops as well as innovative research projects.

Second, we developed two case studies that showcase (innovative) methodological approaches and the value of research covering and comparing different types of political communication (D4.6: Ivanusch, 2023). The first case study addressed the question to what extent and why parties may send different policy signals in different communication channels. Therefore, a state-of-the-art language model (BERT) was trained on labelled manifestos and used for cross-domain topic classification of press releases, parliamentary speeches and tweets from parties and individual party members in Austria, Germany and Switzerland. The results showed that parties indeed address issues to different degrees depending on the used communication channel. The observed variation is thereby moderated by the characteristics of the communication channel. The second case study investigated whether social media transforms the level and nature of issue engagement – i.e., the

extent to which political parties talk about the same issues – compared to more "traditional" forms of political communication, such as press releases. Based on the application of an unsupervised topic model to tweets and press releases from Austrian, German and Swiss parties, the results show that political parties are likelier to perform issue engagement on social media than in press releases. Furthermore, party size seems to be a less crucial factor in issue engagement on social media ("egalitarian effect"), while government parties appear to respond more frequently to other parties there than in press releases. However, no evidence for an increased role of ideological distance ("echo chamber" effect) in issue engagement on social media was found in this study.

Overall, the two case studies highlight the complexity and (changing) nature of political competition in contemporary democracies. The findings show that no single political agenda exists and that political parties send different policy signals in different communication channels. The use of various forms of political communication leads to the simultaneous co-existence of different patterns of party competition and political debate. This can result in different perceptions of parties, the competition between them and the political system. This way, the findings do not only provide important insights on political competition, but also illustrate the need to focus more strongly on the heterogeneity of political agendas to advance our understanding of political processes in contemporary democracies. This is an important avenue for future research and OPTED should play a major role in enabling it. OPTED can offer valuable support in terms of data collection and linking through the OPTED platform as well as in terms of data analysis (e.g., development of tools and best-practices for text analysis).

3. Feasibility: The (future) role of OPTED in enabling (better) research on political organizations

The OPTED project has contributed and will continue to contribute in a number of important ways to enable (better) research on political organizations (i.e., political parties and interest groups). First, we have surveyed the availability of data relevant to the study of texts from political organizations. Therefore, we have searched for and curated an inventory of existing text collections from political organizations. This way, we were able to map out the current state of the field and identified important gaps in terms of data availability, access and usability. To be of continuous use for researchers and other data users, this inventory needs to be constantly updated. Digitalization has revolutionized the field of political text analysis and will continue to do so. More and more resources will become available, just because they are being digitalized or because stakeholders want to make them available or are convinced/urged/forced to make them available. The work of OPTED is thus not over, but has just begun.

Second, we have provided various products and reports that will help researchers both in terms of data collection and data analysis. The OPTED platform allows users to search for and gain information on political organizations and related (data) resources in more than 30 European countries. Furthermore, the platform allows researchers to find resources on further important political actors, such as media sources, parliaments

and governments, as well as on scientific publications and (text analysis) tools. Therefore, this central product developed by the OPTED project will help researchers find and link relevant resources (e.g., data sets, tools) for their own use cases in a systematic and resource-saving way. Moreover, we have produced several reports that focus on different important aspects related to the study of political texts from political organizations. The deliverables created in WP4 can thereby serve as important points of reference for researchers, data owners/providers and data users.

Third, OPTED has contributed significantly to community building among researchers from different fields in the social sciences (e.g., political science, communication research) and to networking with important stakeholders, such as data owners/providers (e.g., political foundations, parliamentary bureaucrats) and (potential) data users (e.g. data journalists). This has helped to connect different stakeholders and exchange the respective needs and interests when it comes to the provision and use of political texts.

Thus, considerable progress has been made over the course of the project in enabling (better) research on political organizations. The core product of the project, namely the OPTED platform, is available for the public from October 1st and will certainly help users identify resources relevant for research on political organizations. Here, the project needs to implement strategies for developing further and maintaining this infrastructure. As also discussed in D3.6 for journalistic mass-mediated political texts (Balluff, Stecker, Boomgaarden & Waldherr, 2023), several challenges persist when it comes to making political text collections findable and/or available through the platform. Ideally, researchers can find all existing text collections from political organizations on the platform and in the long-term are also able to retrieve and/or store text data in a user-friendly way through the platform or other connected OPTED services. To provide these services in the long-term, the platform will require intensive editorial and technical management to keep it up to date. Thus, OPTED needs to develop a clear strategy for finalising and maintaining the platform and connected infrastructure.

In addition to the platform, OPTED is also needed in the future to continue to act as an important advocate, strategic partner and centre of expertise in the area of (social science) text analysis. First, OPTED should further strive to develop and implement guidelines and common standards for making political text data available more widely and in a user-friendly way. Here, the started community-building and networking with stakeholders is crucial to push for better availability, access and usability of texts from political organizations. However, data availability and data sharing remain significant challenges in the case of political parties and interest groups, particularly when it comes to internal documents and social media data. These types of text are often unavailable or data sharing is prohibited. Here, OPTED should continue to connect with strategic partners - such as other research projects, (data) journalists, NGOs or bureaucrats - to push for better data access in order to enable research on political organizations and the functioning of democracies.

Second, OPTED should strive to become a central actor providing technical expertise and assistance in the fast-changing field of computer-based text analysis for the social sciences. The fast-paced development of technology (e.g., large language models) offers several opportunities but also challenges for social science text analysis. OPTED and its members are leading experts in the field, and this allows OPTED to become a central

actor that provides expertise and guidance for the application of computer-based text analysis tools in the social sciences and beyond. Potential avenues to provide common goods for the research community are, for example, tool development for text analysis, the creation of tutorials and workshops or the production of innovative research.

To sum up, OPTED has made considerable progress when it comes to enabling (better) research on political organizations. Several insights and products will help researchers drive the study of political texts from political organizations forwards and gain important insights on their behaviour and the functioning of contemporary democracies. However, more work is needed to further develop data availability and access as well as the products developed within the project. Therefore, we propose that OPTED should continue on the path it has taken over the past 36 months and strive to become an important hub and centre for expertise in the area of computer-based text analysis for the social sciences.

References

Alberto, A., & Klüver, H. (2021). *Defining organizations and texts of interest*. https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D4.1.pdf

Balluff, P., Stecker, M., Boomgaarden, H. G., & Waldherr, A. (2023). *Feasibility of media analysis in OTPED*.

Banducci, S., Scotto di Vettimo, M., Gelovani, S., Theocharis, Y., & Dhamal, P. (2022). *Report on producing a prototype platform for the OPTED project.*

Greene, Z., Ivanusch, C., Lehmann, P., & Schober, T. (2021). *A repository of political party and interest group texts*. https://opted.eu/fileadmin/user_upload/p_compcommlab/OPTED_Deliverable_D4.2.pdf.

Greene, Z., Ivanusch, C., Lehmann, P., Schober, T., Alberto, A., Burst, T., Hutter, S., Klüver, H., Regel, S., Weßels, B., & Zehnter, L. (2021). *Inventory for text corpora by political organizations*. https://opted.eu/results/inventories/.

Ivanusch, C. (2021). *Making political party and interest group texts accessible and usable*. https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D4.3.pdf

Ivanusch, C. (2023). *Case studies: How political parties communicate across different text types.*

Ivanusch, C., Burst, T., & Zehnter, L. (2022). *Challenges and opportunities for the comparative study of political text types: Testing computer-based topic classification approaches across political party texts.*

Ivanusch, C., Lehman, P., Balluff, P. & Scotto di Vettimo, M (2023). *Integrating political organizations into the OPTED platform*.

Scotto di Vettimo, M., Banducci, S., Balluff, P., Gelovani, S., & Theocharis, Y. (2023). *Report on the platform for the OPTED project*.

Sebők, M., Proksch, S.-O. & Rauh, C. (2022). *Report on meeting bureaucrats/data scientists*. https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D5.4.pdf