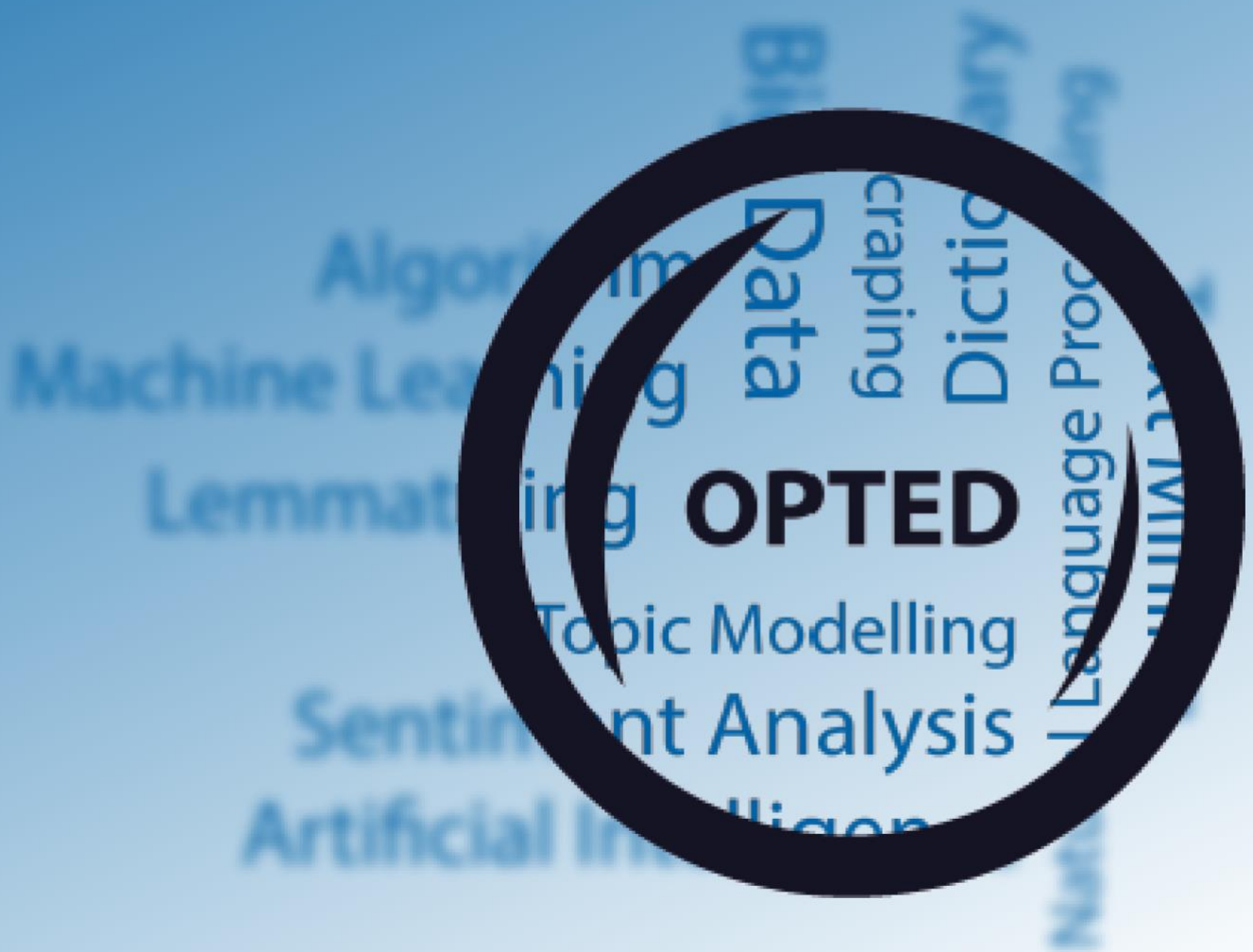
OPTED

Making Political Party and Interest Group Texts Accessible and Usable

Christoph Ivanusch



Dislaimer

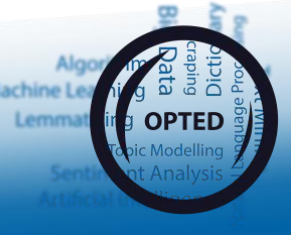
This project has received funding from the European Union's Horizon 2020 Research & Innovation Action under Grant Agreement no. 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report



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

Making Political Party and Interest Group Texts Accessible and Usable

Deliverable D4.3

Author: Christoph Ivanusch[1]

[1] WZB Berlin Social Science Center

Due date: September 2021

1 Introduction

In the context of OPTED, WP4 has created an inventory of texts by political organizations (Greene, Ivanusch, Lehmann, Schober et al., 2021). More precisely, the inventory identifies data sets/corpora that include texts by political parties and interest groups. Furthermore, we have collected variables that store information about the content and quality of these data sets. A previous deliverable (D4.2; Greene, Ivanusch, Lehmann & Schober, 2021) has already introduced the inventory and discussed certain aspects of it. Overall, we have identified more than 200 text collections from 30 countries and the EU level. However, significant differences regarding availability exist between the different types of political texts and geographical regions, as shown in D4.2. Our third deliverable (D4.3) has a different focus and discusses the accessibility and usability of text collections in the inventory.

In a first step, we identify best practices for making text collections easily available and accessible for a variety of users. After discussing different types of users, we identify four criteria for good accessibility and five criteria for good usability. We propose to use these criteria as guidelines for publishing collections of texts from political organizations in a user-friendly way. The criteria for good accessibility focus on: (1) data storage, (2) easy and free access, (3) way of text retrieval, and (4) file type. The criteria for good usability cover the following aspects: (1) availability of full texts, (2) whether texts are ready-to-use, (3) corpus type, (4) annotations, and (5) codebook.

In a second step and based on the proposed criteria, we then analyse the existing data sets included in the inventory with regard to their accessibility and usability. We show to what extent the text collections in our inventory fulfil the outlined criteria and where there are problems with regard to publishing text collections in a user-friendly way.

In a final step, we provide a conclusion and a brief outlook on the further development of the proposed criteria and the role OPTED will play there. First of all, the OPTED infrastructure will help overcome challenges we discovered with regard to accessibility as OPTED will serve as a single access point that makes it easier to find text collections. Secondly, OPTED will also work on establishing standards that will ease usability in the future.

2 Best Practice

2.1 Potential Users

The following sections identify and define best practices for making textual data easily available for a variety of users. We discuss how text collections can be published in a user-friendly way. In order to do that, one has to keep the different types of potential users in mind. In the case of texts from political organizations, we can think of five broad types of potential users. These types as well as their specific needs, interests, skills and preferences are introduced below.

The first two types of potential users both include researchers. However, we differentiate between two groups of researchers. This differentiation arises out of differences between the qualitative and the quantitative research traditions. These two traditions use very different styles and techniques (King et al., 1994; Mahoney & Goertz, 2006). While quantitative research uses numbers and statistical methods, qualitative research usually does not rely on numerical measurements but applies other forms of in-depth analysis (King et al., 1994). This leads to another difference between the two traditions, which is especially relevant with regard to the analysis of political texts. While quantitative researchers often use large-scale data sets, qualitative researchers usually study a small number of cases (King et al., 1994). These differences between the quantitative and qualitative research tradition also apply to the analysis of political texts (e.g., Benoit, 2020). Therefore, qualitative and quantitative researchers have very different needs and preferences when it comes down to the publication and use of textual data. But also within the group of quantitative researchers, we find huge differences with regard to the techniques they are using and the skills they have (e.g., with regard to programming languages). These differences have become even more pronounced with the advent of big data and the use of computational methods in the (social) sciences (e.g., Leonelli, 2020). Hence, the demands of these different groups can vary to great extent and there

might not always be a one-fits-all solution.

Policy-makers and political advisors constitute the third type of potential users. Several texts (e.g., press releases, social media, and manifestos) contain a lot of information about political processes or issues. Hence, policy-makers and political advisors might use and benefit from text collections that include a wide variety of texts from different (political) actors.

The fourth type of potential users can be defined as journalists. As argued above, political texts contain a lot of information and are already used widely by journalists. First and foremost, press releases function as important sources of information for journalists (e.g., Grimmer, 2013; Meyer et al., 2020). However, also other types of texts are of interest to journalists. This in particular applies to social media communication. Political actors use social media to communicate with the public and with journalists (Barberá & Zeitzoff, 2017; Gilardi et al., 2021); journalists in turn closely follow social media communication and use it for their reporting (Jungherr, 2014; McGregor, 2019; Gilardi et al., 2021). Political text collections can store large amount of texts from different actors and therefore are potentially relevant sources of information for journalists.

The fifth type of potential users represents the wider public, such as interested citizens but also commercial companies. In the former case, text collections might, for example, be of relevance for the area of citizen journalism. In the latter case, collections of texts from political organizations might be useful for keeping up-to-date with regard to policy-making or for lobbying and other ways to influence political decision-making.

2.2 Criteria for Publishing User-Friendly Text Collections

As discussed above, the different types of potential users have specific needs, interests and preferences with regard to (political) text collections. Furthermore, all of them have their very own workflows and skill sets. While some might only require a small amount of texts for reading or for a manual analysis, others might need large-scale text corpora for a computer-based analysis. Therefore, text collections have to fulfil certain criteria to be of use for as many potential users as possible. Two main aspects are especially important: data sets have to be easily accessible and well usable.

We have identified best practices for publishing text collections and propose criteria for making them easily accessible and well usable. The proposed criteria are based on three main sources. Firstly, we have gained a comprehensive overview of existing text collections in the process of creating our inventory. This way, we are able to identify best practices with regard to publishing textual data. Secondly, the WP4 team includes several experts when it comes to publishing text collections (e.g., Manifesto Project). Thirdly, we use existing guidelines such as the FAIR Data Principles as points of reference. FAIR data needs to be findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). Based on that, we propose four criteria for good accessibility and five criteria for good usability. These are presented in the following.

2.2.1 Accessibility

Data sets on texts by political organizations should be easily findable and accessible for a variety of potential users. Here, four criteria are important.

- *Data storage:* Data sets have to be “easy-to-find” (see also: FAIR Data Principles). Until now, many data sets containing texts by parties and interest groups are stored on individual project websites or can be found in certain archives. Only few text collections are available via large-scale scientific data repositories such as Harvard Dataverse or GESIS. Hence, these text collections cannot be found in a single place. This makes it harder for users to identify relevant data sets about texts by political parties and interest groups. Therefore, a platform needs to be created that allows researchers to easily identify and find text collections. This platform should contain information on text collections, which are available but stored on several different repositories or archives. OPTED aims to achieve exactly this by creating a database that stores information on and directly links to existing text collections. OPTED thereby does currently not aim to be another repository in the same sense as Harvard Dataverse or GESIS and we also do not suggest that all textual data should be stored on such large-scale repositories. Although it might make sense to publish certain data sets on repositories (e.g., replication data), this is not

always the case. Large-scale (comparative) projects in particular might want to publish their data sets on individual project websites, as they can offer users more options there with regard to data access or analysis. Hence, OPTED will not create another data repository but rather a platform that will make text collections more visible and easier to find and interlink (see D8.1, D8.2) for all types of potential users. The OPTED platform (WP9) will provide for such a searchable database of available text data (raw and annotated). This way, OPTED will do its part in making collections of political texts "easy-to-find" in the future.

- *Easy and free access:* Data sets should be easily and freely accessible for users (see also: FAIR Data Principles). Here, two aspects are relevant. Firstly, access should be easy. Hence, users should be able to use data sets without having to go through a complicated registration process. Ideally, no registration or just a simple registration process should be required. The latter might be needed for reasons of copyright, as the data providers might not be allowed to publicly distribute the texts on the internet. A simple registration process can, however, often solve this problem and still keep the hurdles for accessing the data at a minimum. Secondly, text collections should be freely accessible and therefore not located behind a paywall. This is crucial so that a variety of users can access political text collections. Furthermore, this aspect also fits with the principles laid out by the Open Science and Open Access movement (e.g., Suber, 2012). Therefore, the OPTED platform will contain information about the specific access options for different data sets. We strongly recommend that data is not stored behind paywalls and registration requirements should be kept as easy as possible.
- *Way of text retrieval:* The way in which the actual texts can be retrieved for further use is another important aspect. While some text collections can be retrieved via download, others need to be scraped from websites or retrieved via a so-called API. These different ways of retrieving data heavily influence the workflow of users. While downloading data is comparatively easy, this is not the case when texts need to be scraped from websites. Webscraping requires (advanced) programming skills and can be a time-consuming task. Furthermore, webscraping might not be allowed or possible in some cases. These aspects should be considered, when textual data is made available. In general, text collections should be easy to retrieve. Ideally, they can be easily downloaded by users.
- *File type:* Textual data should be made available in suitable file types. Text collections can come in many different file types. This makes it difficult to establish coherent ways of data storage as well as of workflows for the analysis of political texts. Hence, a consensus about a limited number of well-suited file types is required (e.g., .csv, .RDS). However, we argue that a growing consensus about suitable file types should not result in a single dominant file type for publishing textual data. Rather, data sets should be published in two or three certain formats in parallel such as it is the case for the Manifesto Corpus (Burst et al., 2021) for example. This would allow a wide variety of potential users with different workflows and skills to access text collections. Additionally, the discussion should also focus on the use of suitable encodings (e.g., UTF-8) for publishing textual data. This is important with regard to workflows and multilingual text analysis for example. We therefore advocate a consensus on suitable file types and encodings for storing and sharing textual data. We suggest .csv files with UTF-8 encoding, but this will be subject to further discussions.

2.2.2 Usability

Data sets on political texts come in various forms and quality. Hence, existing text collections show different degrees of usability. We have identified five criteria that are important with regard to the usability of textual data.

- *Availability of full texts:* Ideally, the texts themselves should be fully available. Currently this is, however, not always the case. For certain types of text - especially for websites - no real text collections exist until now. Here, existing data sets most often only include links to the websites as well as some meta data. Therefore, developing text collections on this type of text is a clear research gap that should be tackled in the future. However, also data sets, which actually analyse political texts with regard to their content for example, do not always publish the full texts or do only publish some parts of the texts. This makes it more difficult to perform further analysis on the texts or to replicate the results. Furthermore, some text collections - especially those on social

media data - only include meta data, but no full texts. This is often down to developer agreements or other specific rules. Hence, several data sets on texts by political parties and interest groups do not include the full texts. Therefore, publishing full texts should become the standard in the future (if it is legally and ethically possible). If this is not possible for various legal reasons, at least transparency about the completeness should by provided, e.g. how the corpora were compiled.

- *Ready-to-use texts:* Textual data can be of different quality. While some textual data can be directly used for analysis (maybe with some minor cleaning and pre-processing), some textual data is not ready-to-use. This is the case, for example, when texts are unclean (e.g., incorrect encoding) or have to be transferred into another format before use (e.g., pdf texts in the programming language R). Making such texts ready-to-use can be a time-consuming and sometimes difficult task. Hence, we advocate providing ready-to-use texts when publishing text collections.
- *Corpus type:* There are different ways to provide texts in a data set. The texts can be stored as individual files or as a single file/single corpus. Both approaches can be useful for different tasks and workflows, different research methods and traditions (e.g., qualitative and quantitative methods) as well as for the various types of potential users, which we discussed earlier. Hence, we do not give a definite answer to the question whether texts should be provided as individual files or as a single file/single corpus. We rather argue that it depends on the type and amount of texts included in the data set as well as the respective users. However, it might be best practice to follow the example of the Manifesto Project. They publish the manifestos included in their Manifesto Corpus (Burst et al., 2021) both as individual files and as a single file/single corpus.
- *Annotation:* Annotation is an important aspect with regard to textual data. It can be very useful for researchers and other potential users as it provides direct information on the content of the texts. Furthermore, annotated texts can be highly relevant and valuable with regard to the application of computer-based methods. Annotated texts can be used as training data for supervised classification models (e.g. Anastasopoulos & Bertelli, 2020; Barberá et al., 2021; Osnabrügge et al., 2021; Terechshenko et al. 2020). The creation of training data is usually a very time-consuming and resource-intensive task (Barberá et al., 2021). Using already existing annotated texts can save researchers and other potential users of computer-based methods a lot of time and many resources. Osnabrügge et al. (2021), for example, use annotated documents from the Manifesto Corpus (Burst et al., 2021) as training data and then apply the trained model to parliamentary speech transcripts from the New Zealand Parliament. Similarly, Terechshenko et al. (2020) train a classification model on the annotated US Congressional bills data set provided by the Comparative Agendas Project (Baumgartner et al., 2006) and apply it to a data set containing headlines from the New York Times. Hence, annotated text collections can be very useful in general and increasingly so for computational text analysis. Therefore, we advocate publishing annotations in addition to the texts whenever they are available (for annotation standardization, see e.g. D8.2).
- *Codebook:* Some text collections include annotated texts. If this is the case, it is crucial to publish codebooks as well. Well-written codebooks help users to understand the content of the data set and make the research process more transparent and reliable. Furthermore, they can ensure that data is interoperable and reusable (see also: FAIR Data Principles). Hence, codebooks should be published as supplementary material if data sets include annotated texts.

3 Inventory Analysis

The criteria outlined above represent best practices for publishing collections of political texts. But are text collections already published according to these criteria or not? Which criteria are largely met and which are not?

The inventory created by WP4 identifies text collections that focus on texts by political parties and interest groups (Greene, Ivanusch, Lehmann, Schober et al., 2021). The inventory contains links to the text collections as well as information on their content and quality. This information is captured by several variables. Based on these variables we can analyse the text collections contained in the

inventory with regard to the proposed criteria for accessibility and usability. The results of this analysis are presented over the course of the following sections. Firstly, the focus is on the criteria regarding accessibility; secondly, the criteria regarding usability are analysed.

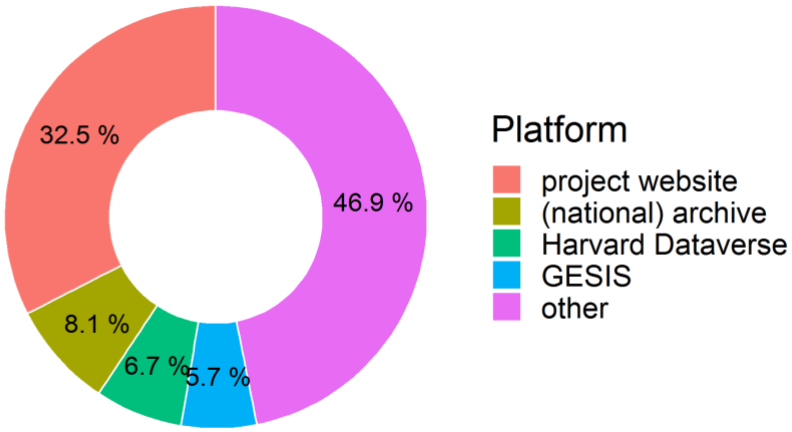
3.1 Accessibility

3.1.1 Data Storage

Scientific data sets can be stored on several different platforms. Many are stored on individual project websites or on large-scale data repositories such as Harvard Dataverse, for example. The variable “storage” captures the type of repository on which the respective text collections are stored. The results are presented in Figure 3.1.

More than 32% of the data sets included in the inventory are stored on individual project websites (e.g., Manifesto Project, Polidoc). Furthermore, several can be found in (national) archives. Only comparatively few are available on large-scale data repositories such as Harvard Dataverse (6.7%) or GESIS (5.7%). In contrast to that, many data sets are available on other types of platforms (e.g., library of the Friedrich-Ebert-Stiftung). Hence, existing text collections are stored on several different individual websites, repositories and platforms and cannot be found in a single place. This leads to a major problem for researchers and other potential users interested in political texts, because they need a lot of time as well as expert knowledge to find relevant data sets. Hence, many existing text collections are not “easy-to-find”. OPTED attempts to overcome this problem by creating a platform that allows users to easily identify and find text collections. As discussed earlier, the platform will provide links to existing text collections and store information on their content.

Figure 3.1 USED DATA STORAGE PLATFORMS



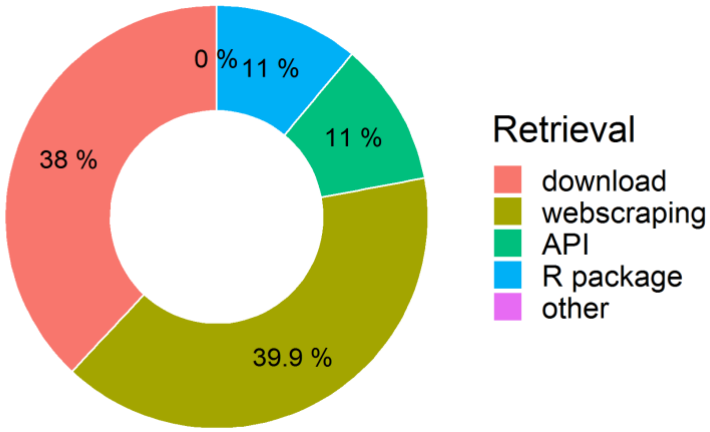
3.1.2 Easy and Free Access

The inventory keeps track of the types and ways of access to the included data sets. One important aspect is, whether the data sets can be accessed easily and freely or not. Overall, most of the existing text collections can be accessed quite easily, as a large majority (64.6%) of the data bases does not require registration. Furthermore and crucially, no data set is behind a paywall. Hence, the data sets can be accessed freely by a wide variety of users. This is in line with the criteria proposed by us and it also conforms to the principles of the Open Access movement (Suber, 2012).

3.1.3 Way of Text Retrieval

There are several ways to retrieve data (e.g., download, webscraping). Figure 3.2 displays the ways in which the text collections included in the inventory can be retrieved for further usage. This analysis shows that several data sets are available via download or an API as well as via some other specific ways (e.g., via R-packages like manifestoR). However, nearly 40% of the data needs to be obtained via webscraping and is therefore not readily available for users. Hence, users often have to invest a lot of effort and time into the (textual) data collection process. Furthermore, (advanced) webscraping skills are needed in many cases. This is a major barrier and does not meet the criteria regarding good accessibility.

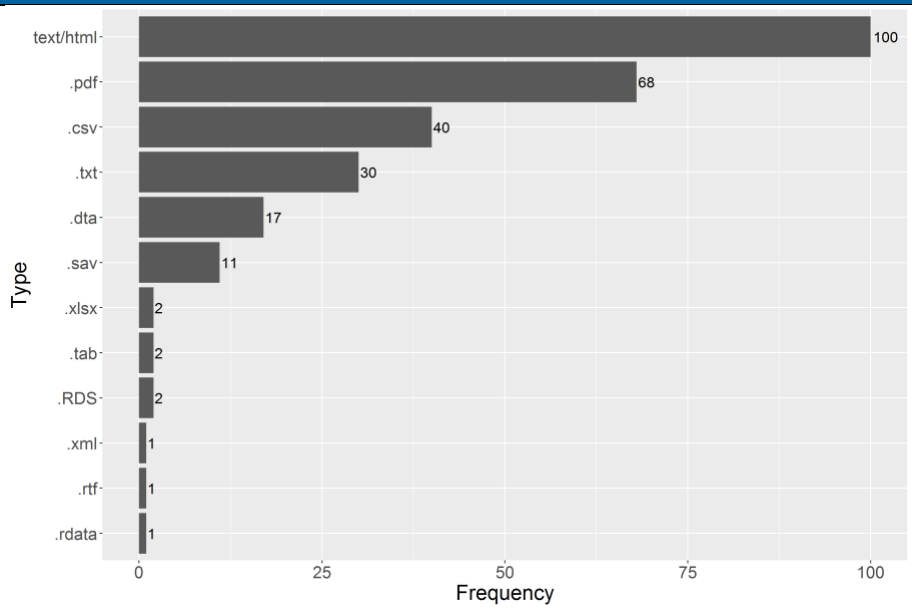
Figure 3.2 WAY OF TEXT RETRIEVAL



3.1.4 File Type

The inventory includes data sets in multiple file formats and types. An overview of the file types is given in Figure 3.3. As can be seen from the displayed results, data on texts by political parties and interest groups comes in many different file types. Hence, there is no clear standard of storing textual data yet. This makes it difficult to establish coherent ways of data storage and workflows for text analysis. Therefore, a consensus on a limited number of suitable file types needs to be developed in order to make text collections easier accessible for a variety of potential users.

Figure 3.3 FILE TYPE



3.2 Usability

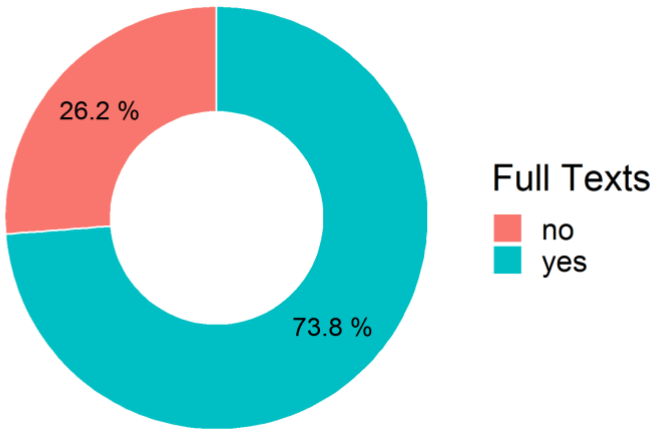
3.2.1 Availability of Full Texts

The following section analyses the usability of the existing text collections that are included in the inventory. Here, we have proposed five criteria for good usability. In a first step, we focus on the actual availability of the full texts (see Figure 3.4).

Overall, full texts are available for the majority of data sets included in the inventory. However, roughly a quarter of the data sets that focus on political texts by parties and interest groups do not

contain full texts. While some text collections only include subsets or certain parts of the full texts (e.g., Müller et al., 2021), other data sets do not contain full texts at all. This mainly applies to social media data and websites. For social media data it is often not allowed or possible to publish the full texts due to developer agreements; for websites we have mainly found collections which provide links to the websites but do not contain the full texts from the websites. This lack of available texts from websites is a clear research gap, which should be tackled in the future.

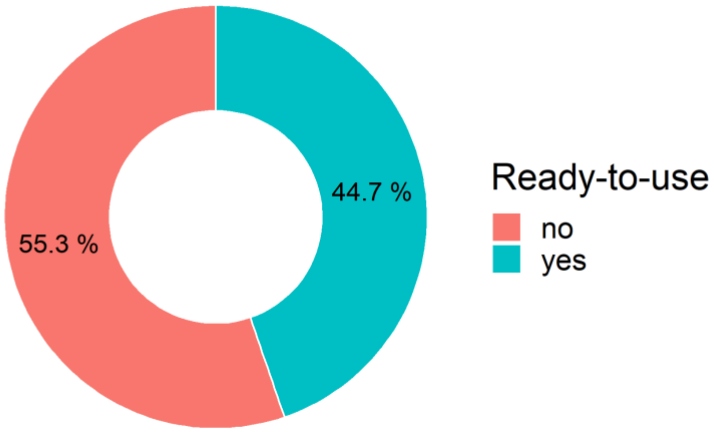
Figure 3.4 AVAILABILITY OF FULL TEXTS



3.2.2 Ready-to-use Texts

Another variable in the inventory analyses whether the texts included in the data sets are ready-to-use. Ready-to-use texts do not require extensive cleaning and do not have to be transferred into another format before they can be used. Figure 3.5 shows whether the text collections included in our inventory mostly store ready-to-use texts or not. The analysis shows mixed results. A slight majority of data sets are not ready-to-use. This means that they need extensive cleaning or have to be transferred into another format before use (e.g., pdf format or text/html). However, there are also several collections containing ready-to-use textual data.

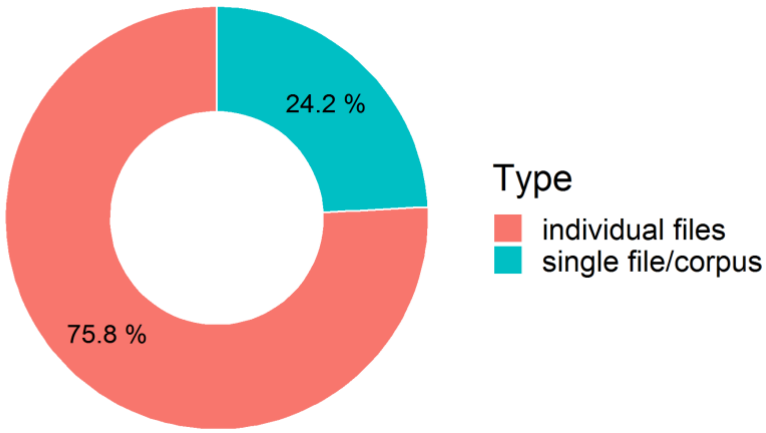
Figure 3.5 READY-TO-USE TEXTS



3.2.3 Corpus Type

Texts can be stored in different ways (as individual files or as a single file/corpus). An overview of the corpus types included in the inventory is given in Figure 3.6. The results show that a clear majority of existing text collections stores texts as individual files. Hence, users have to create a corpus file by themselves in most cases. Only comparatively few data sets provide the corpus as a single file. However, depending on the task at hand, both types can be useful. As discussed earlier, this depends on the type of potential user as well as their needs, preferences and skills.

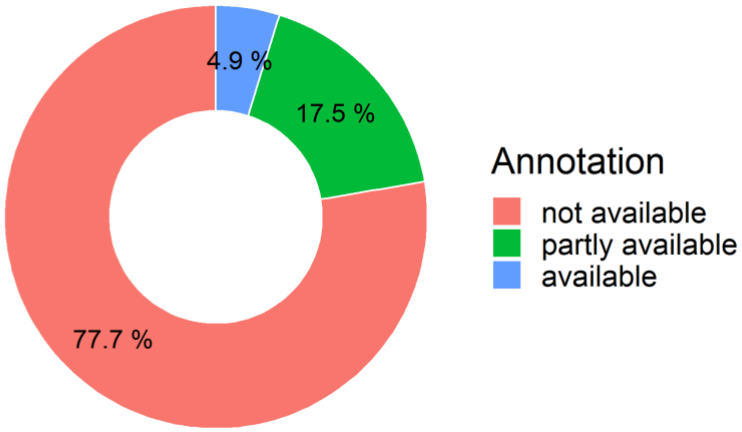
Figure 3.6 TYPE OF CORPUS



3.2.4 Annotation

Another important aspect and criteria with regard to the usability of textual data is annotation. Nearly 78% of the data sets included in the inventory do not contain annotated texts (see Figure 3.7); 22.4% of the data sets store (partly) annotated texts. Annotations are useful in several ways. Firstly, annotations provide direct information on the content of the texts and are therefore valuable for analysis. Secondly, annotated texts can be used as training data for supervised classification models (e.g., Osnabrügge et al., 2021; Terechshenko et al., 2020). Especially the latter aspect becomes increasingly important in academia and also beyond. Hence, we advocate publishing annotations whenever they are available.

Figure 3.7 AVAILABILITY OF ANNOTATION



3.2.5 Availability of Codebook

For those data sets that contain annotations, we can analyse whether a codebook is available or not. The results of this analysis show that all data sets, which contain annotated texts, also provide codebooks. This is in line with the criteria proposed by us. Codebooks help users understand the content of data sets, make the research process more transparent and ensure that data is interoperable and reusable.

4 Conclusion and Way Forward

In our inventory we have collected information on existing data sets/corpora that focus on texts by political parties and interest groups. The analysis of this inventory shows that existing text collections come in various forms. Furthermore, there are clear differences between them with regard to their accessibility and usability. This diversity shows that no clear standards or guidelines with regard to the publication of text collections exist until now. However, exactly this aspect becomes increasingly important. A rapidly growing amount of political texts becomes available. Furthermore, a wide variety of users (e.g., researchers, policy-makers, journalists) are interested in political texts as well as (computer-based) analysis tools for these texts. Hence, it is important to develop certain guidelines for making text collections available in a user-friendly way.

This deliverable (D4.3) makes a first step into this direction by identifying and defining best practices for publishing text collections. Based on that, clear guidelines can be developed in the near future. Overall, we have identified nine criteria for best practice with regard to publishing textual data. More precisely, we have identified four criteria with regard to accessibility and five criteria with regard to usability.

Firstly, text collections have to be easily accessible. Therefore, they have to be “easy-to-find” and easily and freely accessible (see also: FAIR Data Principles). Furthermore, they should be easy to retrieve (e.g., via download) and should be provided in suitable file types. In the latter case, we advocate the development of a consensus within the research community on what file type(s) are well-suited for storing and sharing textual data. The analysis of the data sets/corpora included in our inventory shows that existing text collections can be accessed freely and without any major barriers regarding registration procedures. However, most of them are heavily spread across different platforms and are often difficult to retrieve. Furthermore, they come in various different file types. Hence, there is a need to make text collections on political organizations easier to find and more accessible.

Secondly, text collections have to be “easy-to-use”. Textual data should be fully available and ready-to-use. It is best practice to publish the texts both as individual files and as a single file/corpus.

Furthermore, text collections should ideally include annotation and codebooks, where applicable. Existing text collections show a mixed picture with regard to these criteria. While several are doing quite well in terms of the availability of full texts and their cleanness, this is not the case for all. Furthermore, there is significant room for improvement in terms of publication as a single file/corpus and the inclusion of annotations.

The proposed criteria can serve as points of reference for making text collections easily available and accessible for a variety of users. We regard our deliverable as a starting point for further discussions and the development of clear guidelines for publishing textual data. This discussion will certainly be started and continued within the larger OPTED project, but will hopefully also reach the wider (research) community. Overall, OPTED will contribute to make political party and interest group texts more accessible and better usable. First of all, the OPTED platform will provide a searchable database of available text data. Secondly, OPTED will develop tools for storing, sharing and analysing textual data in a coherent way. Furthermore, training opportunities to foster application of the developed tools will be offered. From the perspective of data users, the OPTED platform will make it easier to find relevant text collections. From the perspective of data producers it will increase the visibility of their work and heighten the chances of getting cited and receiving credit for their work. This will incentivise data producers to adhere to the criteria OPTED specifies.

References

Anastasopoulos, L. J., & Bertelli, A. M. (2020). Understanding delegation through machine learning: A method and application to the European Union. *American Political Science Review*, 114(1), 291-301.

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1), 19-42.

Barberá, P., & Zeitzoff, T. (2017). The new public address system: Why do world leaders adopt social media? *International Studies Quarterly*, 62(1), 121–130.

Baumgartner. F. R., Green-Pedersen, C., & Jones, B.D. (2006) Comparative studies of policy agendas. *Journal of European Public Policy*, 13(7), 959-974.

Benoit, K. (2020). Text as data: An overview. In L. Curini, & R. Franzese (Eds.), *The SAGE Handbook of Research Methods in Political Science and International Relations* (pp. 461-497). Sage.

Burst, T., Krause, W., Lehmann, P., Lewandowski, J., Matthieß, T., Merz, N., Regel, S. & Zehnter, L. (2021a). Manifesto Corpus. Version: 2021-1. WZB Berlin Social Science Center.

Gilardi, F., Gessler, T., Kubli, M., & Müller, S. (2021). Social media and political agenda setting. *Political Communication*, 1-22.

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/.

Grimmer, J. (2013). *Representational style in Congress: What legislators say and why it matters*. Cambridge University Press.

Jungherr, A. (2014). The logic of political coverage on Twitter: Temporal dynamics and content. *Journal of communication*, 64(2), 239-259.

King, G., Keohane, R. O., & Verba, S. (1994). 1. The Science in Social Science. In G. King, R. O. Keohane, & S. Verba (Eds.), *Designing Social Inquiry* (pp. 1-33). Princeton University Press.

Leonelli, S. (2020). Scientific research and big data. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition). https://plato.stanford.edu/archives/sum2020/entries/science-big-data.

Mahoney, J., & Goertz, G. (2006). A tale of two cultures: Contrasting quantitative and qualitative research. *Political analysis*, 14(3), 227-249.

McGregor, S. C. (2019). Social media as public opinion: How journalists use social media to represent public opinion. *Journalism*, 20(8), 1070-1086.

Meyer, T. M., Haselmayer, M., & Wagner, M. (2020). Who gets into the papers? Party campaign messages and the media. *British Journal of Political Science*, 50(1), 281-302.
Müller, W. C., Bodlos, A., Ennser-Jedenastik, L., Gahn, C., Graf, E., Haselmayer, M., Haudum, T., Huber, L. M., Meyer, T. M. & Reidinger, V. (2021). AUTNES Content Analysis of Party Press Releases 2017 (SUF Edition). *AUSSDA*. V1. https://doi.org/10.11587/EVSU6G.
Osnabrügge, M., Ash, E., & Morelli, M. (2021). Cross-Domain Topic Classification for Political Texts. *Center for Law & Economics Working Paper Series*, 2020(04).
Suber, P. (2012). Open Access. MIT Press.
Terechshenko, Z., Linder, F., Padmakumar, V., Liu, M., Nagler, J., Tucker, J. A., & Bonneau, R. (2020). A comparison of methods in political science text classification: Transfer learning language models for politics. *Available at SSRN*. http://dx.doi.org/10.2139/ssrn.3724644
Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1), 1-9.