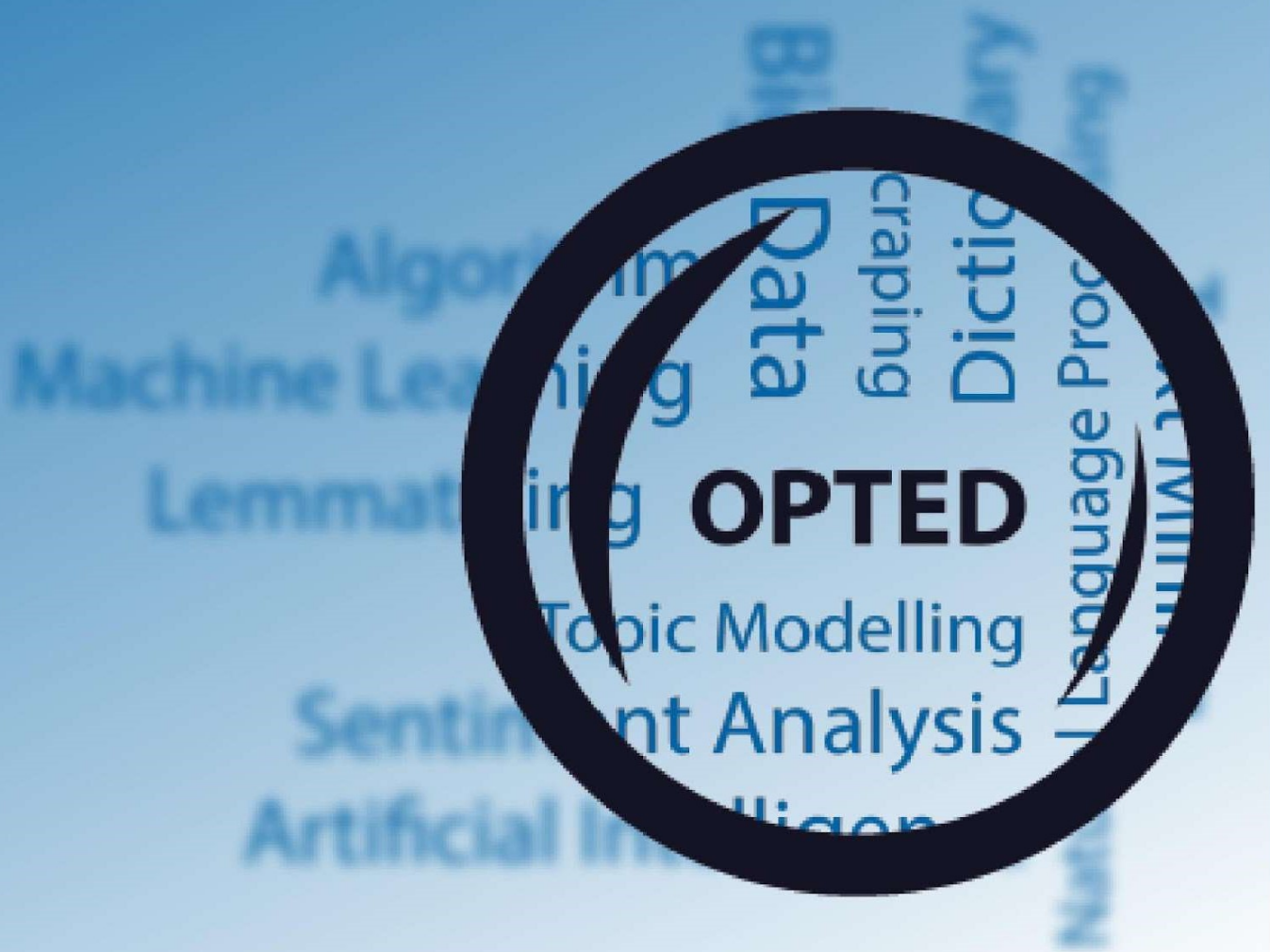
OPTED

Guidelines for good practice and proposal for developing a data repository of CPPT in collaboration with WP10 and WP9

Amanda Haraldsson, Shota Gelovani, Bente Kalsnes, Karolina Koc-Michalska & Yannis Theocharis



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

Guidelines for good practice and proposal for developing a data repository of CPPT in collaboration with WP10 and WP9

Deliverable D2.5

Amanda Haraldsson[1], Shota Gelovani[2], Bente Kalsnes[3], Karolina Koc-Michalska[1] & Yannis Theocharis[2]

[1] Audencia Business School
[2] TU Munich
[3] Kristiania University College

Due date: July 2023

1 Executive Summary

Citizen-produced political text (CPPT) – or text produced by citizens, either offline or online, relating to a political process, a policy or a civic issue (Gelovani et al. 2022) – comes with particular difficulties relating to researchers' sharing their data. These difficulties include additional privacy concerns given that citizens are not public figures, and that CPPT data is often from social media platforms that encompass certain legal and technical issues compared to e.g. political speeches or news articles. In this deliverable, we firstly give a glimpse of the current practices engaged in by CPPT scholars for sharing data, and then outline some concerns of researchers from interviews. We then describe the intended actions to build an infrastructure by OPTED and to improve the possibilities and best practices for sharing CPPT data, keeping in mind the needs expressed by researchers in a survey and the recommendations from academic literature.

Our data show that half of CPPT researchers (and text-as-data researchers generally) feel that a dedicated online platform for storing datasets and other resources is very important for their research activities. Our data also show an extremely low propensity of scholars who have published CPPT studies from 2014-2022 to provide links to data that they have used. Only 47 out of more than 3,000 studies (1.3%) provided valid data links for CPPT data. Quantitative studies using English text data were most likely to share data, though the small number of studies sharing data makes it difficult to draw conclusions.

We find from our in-depth interviews with scholars, supported by the academic literature, that there are four primary reasons why scholars using CPPT data or social media data in general are reluctant to share data. Firstly, even when it would be legally possible to share data, researchers face high uncertainty about whether they actually can legally share data, due to the difficulty navigating the often changing terms and services of platforms and regulations on the national or supranational level. Secondly, researchers take the ethical concerns when using citizen data very seriously, and (as with legal barriers) feel uncertain about what/how much CPPT data can be shared without compromising privacy and anonymity of citizens. Thirdly, CPPT data requires a high degree of contextual information to be properly interpreted, leading to scepticism towards data sharing (or using others' shared data). Finally, properly storing, managing or sharing data requires resources in terms of finances, skills and time that many researchers are unable or unwilling to devote to the cause.

Our survey data asking researchers if and what they potentially need from an online platform that may serve as a hub for text analysis research (including CPPT) shows that about half of respondents felt such a platform could be crucial for improving their research. Important considerations for such a platform is to allow for data linking, to expand the geographical and temporal coverage of data available, to be a repository where researchers can themselves contribute their data and a hub to find collaborators. These and other considerations are important priorities within OPTED, as showcased by our actions relating to dissemination and community-building described herein and our website.

In the next section, we present the methods used to collect data in this deliverable: interviews, a survey and an inventory of CPPT publications from 2014-2022. We then analyse the data with the aim of uncovering current practices CPPT scholars engage in and obstacles to sharing data. Finally, we propose a CPPT data repository by engaging with the expressed needs of CPPT scholars and the recommendations of academic literature on the topic, a collaborative effort between WP2, WP3, WP7 and WP9 of OPTED.

2 Purpose and methods

2.1 Best practices and data repository

Sharing of CPPT data, in social media research in particular, is a vital part of improving access to text-based data, equality in research possibilities and research quality. As summarised by Weller and Kinder-Kurlanda (2016), data sharing is necessary for three main reasons: "To support validity by advancing reproducibility and comparability", "To avoid 'digital divides' in data accessibility", and "To save time and money in data collection processes" (p.166). Moreover, sharing data allows for collaboration opportunities (Van Atteveldt et al. 2021) and is part of accurately archiving human interactions during the 'algorithmic age' (Acker and Kreisberg 2020). Not only the ability to use datasets for new questions or to replicate studies is impacted by data sharing standards in CPPT research, but also the important task of creating benchmarking datasets on which computational methods can be tested and compared (Assenmacher et al. 2022).

Yet, the integrity of citizens themselves and their preferences cannot be overlooked when it comes to researchers sharing citizens' online posts, their letters to newspapers or their other forms of politically

expressing themselves. This is not only to respect their right to privacy, but because of the real potential of harm - please see also the report by WP10, D10.1. For example, data may inadvertently be shared with others who use the data to determine the health status of participants or the location of citizens under repressive regimes (Asiedu et al. 2021). Survey and interview data with researchers find that researchers typically agree with the importance of sharing data, but do not share data personally (Zhu 2020; Weller and Kinder-Kurlanda 2015), and it is therefore difficult to determine what data sharing practices – for CPPT data in particular – are legally and ethically sound, as well as leading to high quality datasets with the necessary documentation.

Our initial goal was to use the CPPT inventory to analyse the best practices among those researchers in our inventory who share data. However, out of 3,738 articles in our inventory, just 82 (2.2%) provide a link to their data, of which only 47 are valid. Rather than an accurate analysis of current best practices, these 82 cases allow us to provide a glimpse of patterns among those who currently share data. The low propensity to share CPPT data also invigorates the need for more or better known repositories where researchers can share CPPT data. Additionally, we employ different data sources from our WP to document the reasons why CPPT data sharing is so sparse.

2.2 Methods

In this report, as well as reviewing the academic literature on the topic, we reference the interview and survey data collected for a previous report, D2.3. In that report, a more detailed description of the selection procedures and questionnaires can be found (the survey methodology can also be found in WP9's report D9.3).

Firstly, we use the inventory of CPPT publications, collected as our first report D2.1 and updated in D2.4, to examine those publications sharing CPPT data (for detailed descriptions of methodology, see D2.1 and D2.4). Focusing on peer-reviewed journal articles published between 2014-2022, we identified articles that used one of our pre-defined search terms (e.g. "comments section" political; instagram "political text"; user-generated AND "political text" OR "political comments"). Trained coders coded the articles, including whether or not the authors of the article provide a link where they share the data they used.

Secondly, we use our interview data (see D2.3, also for its appendix material: the interview questionnaire and summary table of interviewees). The in-depth interviews[1] were intended to get a more nuanced understanding of some of the primary challenges faced by CPPT researchers, or others with expertise relating to CPPT research (e.g. editors, methods teachers or experts on research ethics). Recruitment focused on ensuring that respondents were diverse in terms of their stage of career, regional variation and academic field. Twenty-one interviews were conducted, between May and July 2022. Interviewee quotes are attributed below following the same numbering as in D2.3.

Finally, we use our survey data (see D2.3 or D9.3, also for the appendix material: the survey questionnaire and a summary table of survey respondents). The survey[2] was created in order to understand issues and needs of researchers engaging in text analysis research via various methods, or interested in using text analysis in the future. The survey was sent to attendees of various workshops and conferences relating to text analysis, as well as to authors in our inventory of CPPT publications. In total, 295 responses were collected between February and August 2022, of which 163 are CPPT users.

3 Sharing CPPT data

3.1 Current practices within CPPT literature

In our inventory of CPPT publications, a total of 82 articles share data. 69 of these are English-language publications, and 13 are publications in other languages. The likelihood of sharing data does seem to increase by year: from only 2 articles in 2014, to 17 in 2022 (Figure 1, the blue line).

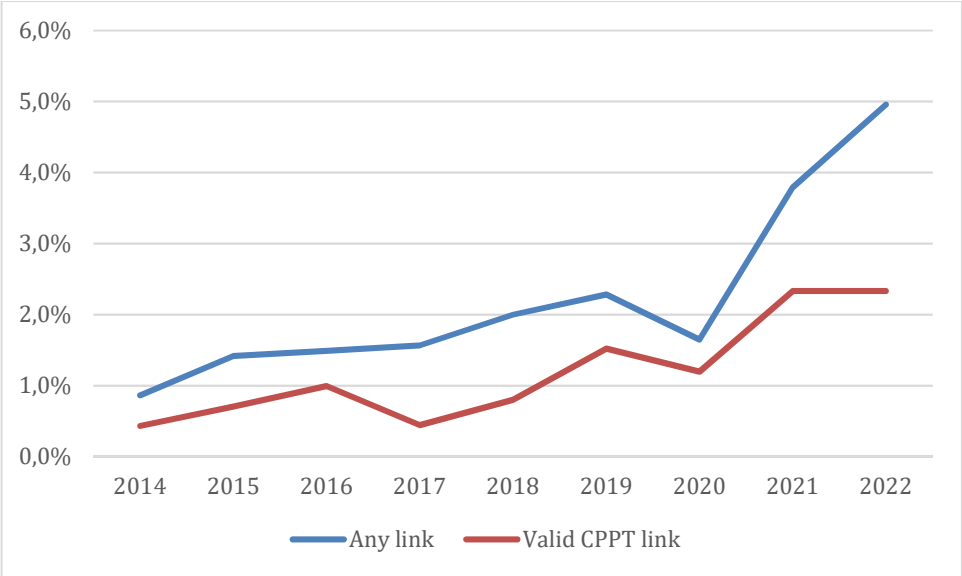
Of those articles sharing data, the majority (47) use only quantitative/computational methods. However, 19 use a mix of qualitative and quantitative methods, and 16 use qualitative methods only. While some interviewees suggested that the fear of respondents being identifiable may be greater with typical qualitative methods such as interview or focus groups, there are still CPPT publications sharing data which qualitatively analyse online posts of citizens, e.g. using screenshots of tweets. Indeed, for those CPPT publications that used

[1] The interviews received ethics approval from Audencia Business School in April 2022.

[2] The survey received ethics approval from Exeter University in January 2022.

mixed or qualitative-only methods and shared data, data collection methods were similar to the collection method in quantitative studies (e.g. copy-pasting online posts), rather than interview or observational data.

Figure 1 Number of articles sharing data by year



Sharing of data appears to be more commonplace among those who study English or multi-language text. Fifty-two of the articles sharing data include English language text in that shared data, and 39 of these share only English text. As for the region where the text is from (which could be coded as multiple regions), the text being shared originates from: Europe (33), North America (23), Asia (11), Australia (5), South America (5), MENA (4) and Central America (2) – 14 articles do not specify the region text is from. At first glance this may suggest that CPPT scholars are more likely to use (and provide data sharing links to) pre-existing text collections predominantly in English and originating from Europe or North America. However, only three of the 82 articles are coded as using pre-existing data collections rather than self-collected data. Instead, it therefore seems likely that the barriers to sharing data are stronger in some parts of the world than others (as explored in section 3.2.4).

The majority of articles sharing data have used social media platforms to gather their data. Where multiple sources of data could be coded for one article, the following were the sources of data that was shared by authors: Twitter (35), Facebook (15), forums (10), Youtube (8), blogs (8), online newspapers (8), surveys (4), Reddit (4) and Instagram (2).

By closer examination, only 36 of the articles indicating data-sharing had at the time of this writing (June 2023) still functioning links that lead to data description and files. These are primarily shared via an author's GitHub page, a university-linked repository or another repository such as OSF or Kaggle. Additionally, 11 articles provide links to the original data instead of a dataset link (e.g. links to YouTube videos or Facebook pages), which could potentially allow repeating the data download process of the authors. Providing links to original data may be more likely to comply with terms and conditions of social media platforms, but complicates replicability given that many pages, posts or user accounts are deleted over time. This relates to a third category of articles sharing data: 12 articles have URLs that no longer work or where data has already been deleted. This could also be caused by, for example, authors feeling unsure whether it is legal or ethical to provide Tweets or other forms of CPPT data, or perhaps choosing to keep the dataset available for a limited time period. A further 23 articles are coded as providing data links, but upon opening these links it appears either a) it is an online supplementary material with e.g. example tweets that are provided, rather than datasets that could plausibly be used by other researchers or b) the data provided is not CPPT data, but rather some other data also analysed in the article (e.g. an article analysing both politicians' speeches and reactions to these by citizens online, but only sharing the politicians' speeches in the data provided).

This closer inspection highlights that CPPT data sharing via sustainable pathways only occurred for 47 of the 3,738 articles in the dataset (1.3%). Promisingly, as seen in the red line in Figure 1, the number of CPPT articles actually sharing data has also increased by year (though the small count makes drawing strong

conclusions unfeasible). Moreover, that 12 articles shared data but have links that no longer work, suggests that more sustainable data sharing may be possible by providing greater support for authors in where to deposit data and how to do so legally, technically and ethically.

3.2 Primary obstacles to sharing data

Sharing CPPT data is severely limited, for several reasons. In many cases, regardless of researchers' own preferences and intentions to share data, they may be prohibited from doing so due to the terms and conditions of social media platforms (Bruns 2019), because sharing of data was not agreed to by citizens who were interviewed or observed, or other reasons outside the control of researchers themselves - for example, due to requirements by ethical boards on a national or university level. When researchers are legally able to share data, their reservations about doing so appear to fall under four main categories: 1) feelings of uncertainty about legality of sharing CPPT data; 2) a sense that it is not in citizens' best interests to share the data, for ethical reasons; 3) a belief that the data could not reliably be used by others due to not being able to share necessary contextual information alongside the text data; or 4) lack of resources needed to properly store and share data. Each of these factors is explored below.

3.2.1 Uncertainty regarding legality (including GDPR)

With the onset of stricter regulatory guidelines for the storing, sharing and using of (online) data, researchers may struggle more than ever to understand when they can or cannot make CPPT data they collect available to others. In particular, the fear that data platforms will 'legally punish' those who share data has become worse as API access has become more restricted (Bruns 2019).

> *"There's even restrictions to sharing Tweet ID's to, you know, hydrate them[3]. I see the concern and it's obviously problematic to share large datasets. I don't have a solution for it, or I cannot think of a possible solution for it but yeah, obviously it has some problems when it comes to replicability and transparency." (Interviewee_19)*

National level guidelines offering specific advice are rare, and may, for example, highlight that researchers "must make an independent judgement of what is ethically appropriate" (NESH 2019) given the type of data they have collected. Not only are copyright and terms of service constantly changing, but they are hard to interpret and vary between national contexts, even between EU countries (Van Atteveldt et al. 2021), making such independent judgements difficult.

On the one hand, following regulations such as the EU's General Data Protection Regulation (GDPR) to the best of researchers' abilities leads researchers to take more steps in ensuring anonymisation and protection of study subjects, including those whose online posts are studied:

> *"[After GDPR] regarding the storage, what we have done is we anonymize all the users. We conserve the first version of the raw data, but the data with which we are working as a team, it's already anonymised. Yeah, that's how we have changed. We have added a step more in the research." (Interviewee_16)*

On the other hand, these regulations can be applied in ways that, rather than protecting citizens' data, create more uncertainty for researchers that could complicate storing CPPT data in the first place (let alone sharing it with others), for example in the case of international cooperation between researchers:

> *"The thing is, it's a debate within the global project how to deal with data privacy? Would respondents sign or not sign [consent forms]? For example, how would you store the data in a safe place that is not traceable, especially with the security and the surveillance. So those are issues we're discussing." (Interviewee_2)*

> *"Twitter asks you to sign this ToS. It used to be like that, that if someone has deleted a tweet, you can no longer you know, keep it also in your records. But how do we know whether or not it has been*

[3] Hydrating refers to the process wherein researchers store (or share) the IDs associated with each individual Tweet in a Twitter dataset (in this case), rather than the text of tweets itself. Another user of the Tweet ID list can then '(re)hydrate' the dataset by calling for the Tweets using this list.

deleted is an issue all researchers struggle with. So sometimes, even if a tweet has been deleted, if I have a copy of it in my machine, how do I know that it has been deleted? So I don't keep tracking the IDs and all. And this is not something that only I've struggled with, all researchers." (Interviewee_18)

These statements regarding storing data echo the findings of Weller and Kinder-Kurlanda (2015), whose interview, observation and survey data with social media researchers reveals the same concerns relating to sharing data:

> "Although many researchers expressed a desire to share, few social media datasets are in fact publicly available for reuse as secondary data. Mainly this seemed to be due to an insurmountable lack of clarity with regard to the legality of sharing social media data. Most researchers were unsure whether they were allowed to share the data collected or what repercussions they would face in the case that data was accidentally or deliberately shared" (Weller and Kinder-Kurlanda 2015)

The Association of Internet Researchers (AoIR) highlight that within social media big data research, researchers are themselves responsible for ensuring that, whether legally allowed or not, data sharing practices are not harmful (Franzke 2020). And yet, researchers may have difficulty turning to their own institution for guidance: "there may be no single authority at your local institution who actually knows what the rules are, and many institutions lack a clear authority for making decisions about sharing of sensitive data" (Van Atteveldt et al. 2021). Uncertainty may therefore push researchers towards erring on the side of caution, and choosing not to share data.

3.2.2 Ethical limitations regarding citizens' privacy

When using citizen data, interviewees shared a general sense that special care is needed to discern potential ethical issues. Compared to public figures, citizens need more privacy and for researchers to have more concern with the intentions behind the text citizens create, before researchers use, publish or share this data:

> *"Understanding the consequences of what you put online, it requires a lot of cognitive skill and technical skill. And we know that people generally don't have these skills [...] It's the researchers responsibility to consider: Is this going to be good for that person also in the future? Will it have good consequences, which is an important ethical principle. And if that perhaps could create problems for that person in 5-10 years' time, it's the researchers responsibility to solve that problem." (Interviewee_12)*

> *"It rings to me rather hollow, if a colleague right now would say 'All our problems in computational social science are solved if we, as researchers, get access to all the data'. Because sometimes we as researchers are the problem. And sometimes people need to be protected from us as researchers." (Interviewee_14)*

Moreover, leading experts such as Rebekah Tromble (2021) caution researchers not to be too quick to admonish social media platforms for limiting API access in light of researchers' misuse scandals, given the unfortunate development of norms of data collection practices that propagate "amassing as much digital data as possible", justifying any illegal or unethical practices by only considering the potential harm done to the platforms' profit rather than the potential for citizens to be harmed. With new ways to communicate online, new challenges emerge that are not yet given enough consideration in guidelines or common practices; for example, how to adhere to the right to be forgotten, in research of platforms like Snapchat that "collects, stores, and analyzes data intended to disappear". Tromble goes on to explain how these concerns can multiply when data is shared with academics and non-academics without proper guidelines for ethically doing so. Tromble also addresses our general need to question our current data sharing practices: "How do we know whom to trust? How do we keep track of whom is using data and how they are doing so? How do we more quickly and effectively identify abuses? How do we stop those abuses?" (Tromble 2021).

Other research has likewise found that social media researchers are generally concerned about the privacy and security of social media users, when choosing whether to share this data with others (Weller and Kinder-Kurlanda 2015). One common practice for improving data sharing ethically is to share only metadata, such as Tweet IDs rather than the text of tweets themselves (Van Atteveldt et al. 2021). Yet such solutions mean that incomplete datasets will be called for by future researchers with the list of Tweet IDs (due to data being deleted) or changes to API access could mean that the Tweet IDs can no longer be accessed by future researchers (Acker and Kreisberg 2020). As an interviewee quote in section 3.1.1 showed, this concern may have been

especially strong in the case of Twitter, which indeed has now become more or less closed off to academic research without paying exorbitant fees: "Twitter will soon charge for API access, with plans starting at an astounding $42,000 per month for access to 50 million tweets. (By contrast, researchers could previously collect more than 3 million tweets per day for free.)" (Opsal 2023). Another common solution for other types of social media data than text, is to engage in 'differential privacy': adding random noise to datasets before sharing, to minimise the risk of de-anonymisation. Yet, this solution is particularly difficult with text data:

> *"So differential privacy techniques is usually you put a little bit of the noise, [...] But at the aggregate level, all the random numbers cancel out, so aggregate levels you'll still be able to figure out the mean of the number of the clicks on Trump posts or something on average, correctly. [...] It's just a random number plus your real number, so we never know the real numbers of clicks, likes, on a particular post. [...] What kind of noise [could be] put onto the text, right? I don't think there's a method out there for now." (Interviewee_8)*

Researchers have been bringing attention to the ethical issues when sharing CPPT data for more than a decade, for example by examining the Harvard-based study using Facebook data from 2006, when supposedly anonymous data was deanonymized by others after the dataset was shared (Boyd and Crawford 2012). Without satisfactory solutions to ensure citizens are protected when their data is shared, it is understandable that researchers may choose not to share data at all.

3.2.3 Contextual information required with CPPT data

When sharing CPPT data – or using CPPT data shared by others – there is a concern about how the importance of context may be lost:

> *"You have to understand the contextual integrity. So if you post something at some point, two years later, in a different context, it might, you know, it might mean something different [...] And researchers haven't sort of understood the context of the information, so they never sort of grasp the fact that this was a joke or this was a humorous exchange that, through time, doesn't travel well, but you know it was for the person, for instance, it was the opposite of what they actually meant." (Interviewee_12)*

> *"Data have to be represented as they are, and you have to give as much context as you can to, you know, give the possibility to the reader to have a 100 or close to 100 percent accurate view of the context." (Interviewee_20)*

Keeping in mind the previous issue of proper ethical care when sharing CPPT data, researchers may struggle balancing the need to provide context for the CPPT data they share, and the need to ensure anonymity and privacy of citizens who created the text.

Moreover, with large social media datasets, there may be no satisfactory context to share; i.e., a dataset of Tweets collected during a certain electoral period may encompass text that relate to other simultaneous events that are difficult to exclude from the dataset due to the collection method. Such scenarios lead to "a general skepticism towards datasets collected by others" (Weller and Kinder-Kurlanda 2015), with fear that data is not collected or documented correctly, uses a different data format, does not document tools used for data collection and potential reasons for data loss, or that there are different method standards in different disciplines of the researchers who collected the data.

3.2.4 Resource intensity

Ethically and sustainably storing data – a necessary precondition to sharing data – is resource-intensive, becoming even more so for researchers not focusing on Western Europe or North America:

> *"The archivability of those texts is, I would say, a disaster in the North African context because there are no organisations behind that that archive that. There are individuals who would save an archive, but it means also funding again, and someone who believes in that cause. And that is a major obstacle. And it simply means that that memory or that material or that moment can get lost." (Interviewee_2)*

> *"Within our consortium, when we are talking about open science, then we kind of agreed that there is no open science. It's just so expensive that to think about open science - it's open for the top players, so it's absolutely related to how much infrastructure and money is involved that some countries, even in Europe, are simply excluded. I would say that from what we've seen from our researchers, if one could get funding, then of course this person would prefer to publish or to use repositories or the*

preprints or whatever, if this person would have the funding for this. If you're just like a standalone, independent researcher, then we know that the cost might be too high for it." (Interviewee_1)

Weller and Kinder-Kurlanda (2015) state that the cost (in terms of time, money, skills or otherwise) needed to collect social media data make data sharing practices a way to "alleviate inequalities in data access", especially if the data that is shared is high quality (meaning already cleaned and well documented). Yet, converting data into a format that is reusable to others is a time consuming task requiring skills in ensuring the data is easily understood by others (Zhu 2020), if not also costly monetarily. There are examples of teams of researchers organising formally to manage datasets that can be used by other researchers, for example the Pushshift Reddit data (Baumgartner et al. 2020), although the dataset is no longer available, yet such collaborations are rare[4]. Moreover, not all CPPT researchers use social media data, and they may find it difficult to find possible collaborators interested in the same data type in order to share the cost of managing datasets and allowing third parties access. Even when collaboration is possible, data ownership assumptions of collaborators may vary, putting data sharing at risk:

"The weirdest things happen [during collaboration], right? Like people who then later tell you that you can't use the data or pitch about stuff like this and... I believe in open science, collaborative science, and you can't do this if you have people who have funny territorial or imperialist kind of claims about what belongs to whom." (Interviewee_5)

The resource-intense nature of not only collecting the data but of properly storing, managing and sharing data lead to another unfortunate consequence, in that it may exacerbate competitiveness between researchers who expended valuable resources and do not necessarily want others to have equal use of the data freely: "Sharing of data among researchers may also give competitors access to the data and result in them profiting from it" (Assenmacher et al. 2022). In particular, junior scholars may feel a lack of incentive to data sharing when this leads to little recognition or acclaim, especially when junior researchers fear that more experienced others might be able to publish quicker and therefore get more 'credit' for the data than the original data collector (Zhu 2020).

4 Proposal for data repository

In order to support the data sharing practices within CPPT research – in terms of increasing its occurrence and ensuring researchers are able to share data legally, ethically, with proper documentation and context, and minimising unequal access to data sharing practices – OPTED is proposing to allow CPPT researchers to upload datasets to an AmCAT4 server. It is linked to the official OPTED Meteor platform, whereby other users can find datasets and researchers' network building hub. Please see WP7's D7.1 report for greater details about the infrastructure. AmCAT4 is a system allowing the storing and running of analyses on text data "without giving access to the underlying data using non-consumptive research"[5]. However, this will be done in such a way that the dataset cannot be accessed, only browsed. Moreover, we propose to make the CPPT inventory collected by WP2 searchable for platform users, also without being directly accessible but with metadata, keyword search and link to publishers available.

While existing platforms for data sharing (e.g. Harvard Dataverse, Clarin VLO, OSF, Zenodo) exist and provide a necessary service for many researchers, they are less appropriate for CPPT scholars specifically because they do not allow for non-consumptive research. Given the specificities of CPPT data sharing, non-consumptive research is a necessary condition for sharing of datasets.

Before outlining our intended (or realised) actions, jointly with WP2, WP3, WP7 and WP9, we briefly provide some comments on what CPPT scholars have expressed as needs and preferences for the OPTED platform, and some recommendations for best practice from the academic literature.

[4] Another example is Netvizz by Bernhard Rieder (http://thepoliticsofsystems.net/permafiles/rieder_websci.pdf), although Facebook's change in API access has made this tool defunct.

[5] https://amcat.nl/book/about.html

4.1 CPPT scholars' needs and preferences

In our survey with researchers using (or interested in using) text data of various kinds, respondents were asked questions relating to their needs and preferences for an online platform. The respondents clearly indicate that, in the current research environment, there is a need for a dedicated online space for text analysis resources: 49% of respondents who use CPPT research indicated that a single platform for text analysis tools and resources was 'very' or 'extremely' important for their research activities.

This single platform will be most useful if it reduces hindrances that text analysis researchers currently face. Table 1[6] shows the percentage of respondents overall, and CPPT researchers specifically, who felt each given response option was 'very' or 'extremely' important for databases to address in the future in descending order by the CPPT researcher column – with other answer options (not at all important, slightly important or moderately important) not shown. The most important aspect is to facilitate linkages between datasets. In the open responses, one respondent elaborated an example: "data linking news articles & social media data". The second most important aspect is to expand both the geographical and temporal coverage of datasets – section 3.2.4 examined how this may be needed partially because of differences in archiving between regions, but an open survey response also suggested the need for "more possibility for Arabic language users".

Other open responses to this question indicated that researchers are interested in being able to access private Facebook groups (or other less-open social media groups/spaces), and "fair use/legal exemptions for copyright protections under non-profit research".

Table 1: Respondents' priorities for the database

Thinking about the future, how important are the following improvements to these types of databases for researchers?		
	All (N = 295)	**CPPT (N = 163)**
Facilitating the linkage of existing datasets		
Very important	176 (81.48%)	117 (82.39%)
Not very important	40 (18.52%)	25 (17.61%)
Expanding the temporal coverage of existing datasets		
Very important	168 (77.42%)	104 (73.24%)
Not very important	49 (22.58%)	38 (26.76%)
Expanding the geographical coverage of existing datasets		
Very important	161 (74.19%)	102 (71.83%)
Not very important	56 (25.81%)	40 (28.17%)
Develop public web portals		
Very important	142 (66.05%)	96 (68.09%)
Not very important	73 (33.95%)	45 (31.91%)
Develop API-web services		
Very important	139 (64.35%)	96 (68.09%)
Not very important	77 (35.65%)	45 (31.91%)
Develop new datasets		
Very important	149 (68.66%)	93 (65.49%)
Not very important	68 (31.34%)	49 (34.51%)
Develop procedures for local installations of datasets		
Very important	116 (54.21%)	76 (54.68%)
Not very important	98 (45.79%)	63 (45.32%)

[6] Respondents could choose which questions to answer or not, therefore total N by question may vary.

Respondents were also asked to evaluate which of several possible aspects of an OPTED database would be important to prioritise. Among all those aspects mentioned, three are the most central to the possibility of data sharing, and are shown in Table 2. Firstly, the use of the database as a repository for data sources and documentation. Among the respondents as a whole (and CPPT users specifically), more than 90% believe this should definitely be a priority, with about two thirds indicating that they would personally make use of the repository. Secondly, more than 80% (both overall and for CPPT users specifically) indicate that it should be a priority for OPTED to allow users to contribute with their own resources (whether this be their own datasets, software or packages they have developed), though only about one third of users indicate that they would personally contribute their own resources. Finally, about 80% also indicate that it should be a priority for OPTED to allow for collaborative work, including in the sharing of text analysis resources – with just under 50% indicating they would personally engage in these collaborative practices via the platform.

In open-ended responses for this question about which aspects the OPTED platform should prioritise, the following themes emerged: user-friendliness, even for non-experts; a 'Stackoverflow' platform for text analysis; a repository that is 'living', in that it is updated regularly rather than data being uploaded and then no further updates made; clear indexing and proper documentation of both texts and tools.

Table 2: Researchers' needs for the database

Which aspects do you think deserve to be prioritised to best satisfy the users' community needs, including your own?		
	All (N = 295)	**CPPT (N = 163)**
An open repository for different types of data sources and relevant documentation		
Definitely a priority and I would make use of it	126 (62.07%)	84 (62.69%)
Definitely a priority	61 (30.05%)	40 (29.85%)
Not a priority at all	16 (7.88%)	10 (7.46%)
A platform that can be used to work collaboratively on the discovery, creation and sharing of text analysis resources		
Definitely a priority and I would make use of it	97 (47.78%)	66 (49.25%)
Definitely a priority	66 (32.51%)	42 (31.34%)
Not a priority at all	40 (19.70%)	26 (19.40%)
A platform users can contribute to with their own resources		
Definitely a priority and I would make use of it	75 (36.95%)	48 (35.82%)
Definitely a priority	92 (45.32%)	64 (47.76%)
Not a priority at all	36 (17.73%)	22 (16.42%)

4.2 Recommendations from literature

There is a great potential for larger collaborations to improve data sharing practices via providing well-structured observatories, networks, consortiums or other formalised communities (Tromble 2021; Van Atteveldt et al. 2021; Bruns 2019; Weller and Kinder-Kurlanda 2016) – potentially with the involvement of private companies (Lazer 2020). The academic literature has made several recommendations about the data sharing practices that are relevant for CPPT scholarship, which OPTED hopes to follow:

1. **Ethical standards for data sharing made clear and signposted**
2. **Variations in level of access, tailored to data type**
3. **Properly documented data and tools**
4. **Allowing recognition for data sharing**

While the recommendations tend to focus on social media based research, many apply equally for other types of CPPT data. Important to keep in mind is that these recommendations are compiled in view of sharing CPPT data in general – it may be that other (or additional) considerations should be taken into account for more specialised situations, such as data sharing for the purpose of creating benchmarking datasets (Assenmacher et al. 2022).

4.2.1 Formalised set of ethical standards, and up-to-date information on legal guidelines

A first recommendation is to provide a set of ethical guidelines for users to follow – from data collection to data sharing. By making the AoIR guidelines more commonplace, there would be a set of standards to adhere to within the community of researchers (Tromble 2021), and users of a data sharing platform could feel confident that data has been collected ethically by the other researcher(s). Alleviating this doubt would promote the use of shared data among sceptics (Weller and Kinder-Kurlanda 2015), and likewise the doubt surrounding legality would be alleviated through providing information on legal guidelines (Weller and Kinder-Kurlanda 2016). **Requiring users to agree to AoIR or other relevant ethical guidelines (for non-internet based research) would ensure better ethical treatment of citizens' whose data is collected and shared, and up-to-date legal guidelines of platforms or other data owners should be pointed to**. The Research Data Alliance's FAIRsharing platform[7] is one such resource that could guide CPPT researchers towards the current guidelines relating to their research when deciding whether to share data.

4.2.2 Allow for a hierarchy in level of access

Within CPPT, there are vast differences in the level of sensitivity of data, which means different requirements for safe data sharing from fully open to highly restrictive (Van Atteveldt et al. 2021). Altering the level of access based on sensitivity could, for example, entail the following types of access: "data may be openly accessible to everyone without registration, accessible only after registration, restricted to be accessed only after the data depositor has accepted a request, accessible only after an embargo period, or even only accessible in safe rooms and after signing complex usage agreements" (Weller and Kinder-Kurlanda 2016). Such a hierarchy is for example in place at the GESIS Secure Data Center[8], where some datasets are available openly, but sensitive datasets must be accessed on-site at the centre after signing a user agreement or off-site after signing a contract. Another example of alternative access to sensitive data is the HathiTrust Digital Library[9]. With signposting to the legal and ethical standards for a particular dataset, users of the platform would be able to determine what level of access is appropriate, as long as **the platform allows for variation in level of access**.

4.2.3 Ensure proper documentation of text and tools

When sharing data, proper documentation is needed for users to be able to understand what the data actually is – relating to the need for contextual information as discussed in section 3.2.3. This way, potential future users of the data can be aware of the data collection method and be confident in its rigor (Weller and Kinder-Kurlanda 2016). **Standards for documenting archived text and tools should be adhered to in a manner that provides confidence in the veracity of the data, while also being intuitive for users to follow**. There are certain difficulties with following standard documentation within archiving when dealing with social media data, and therefore documentation may differ by data type. For example, the standard of 'fixity' (that data will not in any way be altered after being deposited) would mean that social media archiving is only a snapshot of a constantly changing stream of data (Acker and Kreisberg 2020), when in fact updated information may be more useful and/or ethical (to account for deleted platform data).

4.2.4 Provide incentives to share via clearly recognising contributions

As seen in survey data with researchers conducted by Zhu (2020), those who share data often feel little reward comes from this in terms of recognition. Of those surveyed by Zhu, few that had shared data had advertised this fact to others, but those who had previously used data shared by others were more likely to do so themselves – meaning a part of the problem is also awareness. "An incentive system and approach to the

[7] https://fairsharing.org/

[8] https://www.gesis.org/en/services/processing-and-analyzing-data/research-visits/secure-data-center-sdc

[9] https://www.hathitrust.org/htrc

citing of data and databases are needed to promote data sharing in the future" (Zhu 2020). **By ensuring that researchers can be credited for their data sharing within a platform, the incentive to share is greater and transparency is higher**.

4.3 OPTED consortium actions relating to data sharing

WP2, in conjunction with WP3, WP7 and WP9, plans to improve upon the situation of data sharing within CPPT research. We are engaging in two primary actions. Both actions rely on the principle of **non-consumptive research**, which is central also to the other WPs' goals. Our proposed solution is based on two platforms that are developed within the OPTED consortium: Meteor (D3.2, D9.5) and AmCAT4 (D7.1). We propose to use Meteor to integrate our CPPT inventory (D2.1, D2.4) with its entire coding of each element constituting rich and linked metadata. AmCAT4 will serve as a text-data storage solution for researchers who wish to share their data with others.

Firstly, **to make the CPPT inventory public to the degree that is ethically and legally possible**. The CPPT inventory is a useful tool for CPPT researchers to find out about research that has been published across countries and languages - thereby facilitating better networking amongst scholars that could lead to more research being conducted in previously less studied regions or on less studied languages. The inventory will be searchable in the OPTED-linked platform Meteor (see D3.2, D3.3, D9.4, D9.5), wherein users can find tools, datasets, publications, etc. in an easily navigated interface. Importantly, our goal is for users to be able to search for key terms related to their research interests, and find matches for these terms in not only the metadata (or descriptive variables relating to the publications in the inventory), but also within the publications themselves. Platform users would be able **to see how certain key terms are used in various CPPT publications** (e.g. in the publication title, abstract or in the article itself), along with tools, languages and other variables of interest, and get a DOI link to the publication(s) (via e.g. the publisher) to learn more. At this stage:

- The CPPT inventory coding is streamlined to fit the Meteor terminology (see D3.3), through the use of a mapping table to ensure consistent use of terms (for example, in the CPPT inventory the term 'Sentiment scoring' was used, while in Meteor the term 'sentiment' is used). This is required so that the same concept/tool is clearly marked with the same term regardless of whether it is used in the CPPT inventory or in other resources on Meteor in order to properly link the inventory with the remaining resources on the platform and for the sake of consistency. Please see Table A1 and A2 in Appendix A, for the exact variable names that would be required in not only datasets uploaded but also the variable names for publications in the CPPT inventory;
- Encoding issues, spelling mistakes or other issues in the CPPT inventory are being corrected - through a mix of automated and manual cleaning methods performed jointly by the relevant WPs;
- A sample of the CPPT inventory has been uploaded to Meteor in order to test the process, and shows how the inventory (when fully integrated) will appear;
- Platform users can also perform full-text searches in the publication texts themselves and get a small preview of the context where the keyword matched (see Appendix B for screenshots). Due to copyright restrictions, users will not be able to read the entire publication text, but can access the publication via the DOI. We implemented this by leveraging AmCAT4 (see below) and it provides platform users to run their own secondary analysis on the CPPT inventory.

Secondly, **to create a space wherein CPPT researchers themselves can confidently share their datasets**. Similarly to the process described above to share the CPPT inventory, the texts and details within the dataset would not be made accessible to the public. Instead, standardised metadata would be browsable on Meteor and, by linking the dataset's metadata on Meteor with the (non-accessible) dataset on AmCAT4, it would be possible for platform users to search for keywords to see how they appear in the dataset without actually accessing the dataset. It would be possible for an interested user to contact the data collector to discuss potential data sharing. The obstacles listed above in section 3.2 have been taken into account, in our plans to allow researchers to:

1. **Receive recognition** for their data if they choose to include it on the platform;

2. **Find networks** of other researchers engaging in similar research, minimising the need for duplicated data collection processes and improving access to potential collaborators;
3. **Provide informational links** to relevant guidelines before uploading any data, to ensure best practice in data sharing;
4. **Allow control for the data collector**: datasets would not be accessible to other users, but instead be searchable via the AmCAT4 server and the metadata available on Meteor, so that researchers who upload datasets would then be able to decide with whom to share the actual data.

Not only CPPT researchers, but any researcher using text-as-data would potentially wish to upload datasets, and would all use the same procedure within the Meteor/AmCAT4 system. This system is currently being designed and tested (with the interface of Meteor being updated as well). In the planned system, researchers wishing to share datasets would need to:

- Create an account with Meteor and with AmCAT4;
- Structure their datasets in a way that is compatible with the AmCAT4 format (e.g., there must be a variable named 'text', which contains the text data), using easy-to-follow instructions that will be provided to them;
- AmCAT4 and Meteor will exchange meta-information with each other. Researchers who make their datasets available will upload their data in a unified submission form that makes this integrated experience as user-friendly as possible;
- Once the data is submitted, the researcher is in full control of access rights.

5 Summary and Outlook

At the moment, the platform is undergoing great changes with a goal to be available at the end of the OPTED project, as such we also hope that the data sharing hub envisioned within WP2 will be fully functional (please see Appendix B for the preliminary version). Thus, we conclude by highlighting how the aims of WP2 are incorporated in the design:

- The issues CPPT researchers face have been addressed to the highest degree possible (providing better information about legal and ethical guidelines, allowing for more secure and ethical data sharing via making datasets not fully accessible but rather searchable, minimising resources needed by allowing researchers to share datasets within the AmCAT4 system).
- The needs and preferences of researchers are being met, since the inventory sharing and ability to find others studying various regions, platforms and languages allows for more diverse and easily findable networks of scholars.
- The recommendations from the literature are being followed, by allowing researchers control of (and recognition for) their data sharing, by signposting ethical and legal guidelines as well as providing information about how to properly structure data before uploading. Most importantly, the system allows for searching for and viewing datasets (and their metadata) without making them fully accessible.

6 References

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science*, 20(2), 105–123. https://doi.org/10.1007/s10502-019-09325-9

Asiedu, E., Karlan, D., Lambon-Quayefio, M., & Udry, C. (2021). A call for structured ethics appendices in social science papers. *Proceedings of the National Academy of Sciences*, 118(29). https://doi.org/10.1073/pnas.2024570118

Assenmacher, D., Weber, D., Preuss, M., Calero Valdez, A., Bradshaw, A., Ross, B., Cresci, S., Trautmann, H., Neumann, F., & Grimme, C. (2022). Benchmarking Crisis in Social Media Analytics: A Solution for the Data-Sharing Problem. *Social Science Computer Review*, 40(6), 1496–1522. https://doi.org/10.1177/08944393211012268

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14, 830–839. https://doi.org/10.1609/icwsm.v14i1.7347

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, 22(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Franzke, A.S., Bechmann, A., Zimmer, M., Ess, C. and the Association of Internet Researchers. (2020). *Internet Research: Ethical Guidelines 3.0*. https://aoir.org/reports/ethics3.pdf

Gelovani, S., Koc-Michalska, K., Theocharis, Y., Kalsnes, B., & Haraldsson, A. (2022). *Methodologic and Theoretical Approaches to Studying Citizen-Produced Political Text*. Paper presented at COMPTEXT 2022

NESH. (2019). *A Guide to Internet Research Ethics*. Forskningsetikk. Retrieved 17 April 2023, from https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/a-guide-to-internet-research-ethics/

Opsal, E. (2023). *Twitter's Transparency Theater*. (April 18, 2023). NYU's Center for Social Media and Politics. Retrieved 26 April 2023, from https://csmapnyu.org/news-views/news/twitter-s-transparency-theater

Tromble, R. (2021). Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age. *Social Media + Society*, 7(1), 2056305121988929. https://doi.org/10.1177/2056305121988929

Van Atteveldt, W., Althaus, S., & Wessler, H. (2021). The Trouble with Sharing Your Privates: Pursuing Ethical Open Science and Collaborative Research across National Jurisdictions Using Sensitive Data. *Political Communication*, 38(1–2), 192–198. https://doi.org/10.1080/10584609.2020.1744780

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science*, 166–172. https://doi.org/10.1145/2908131.2908172

Weller, K., & Kinder-Kurlanda, K. (2015). Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research? *Proceedings of the International AAAI Conference on Web and Social Media*, 9(4), Article 4. https://doi.org/10.1609/icwsm.v9i4.14687

Zhu, Y. (2020). Open-access policy and data-sharing practice in UK academia. *Journal of Information Science*, 46(1), 41–52. https://doi.org/10.1177/0165551518823174

7 Appendix A

Table A1 Variable name terminology for datasets

Entity Type / Variable	Description	Data Type
authors	Name(s) of the individual(s) or organisation(s) listed as author(s) of the resource	ListRelationship
date_published	Date of publication	datetime
date_modified	Date of any modification	datetime
doi	DOI to the resource	string
url	URL to the resource	string
arxiv	arxiv identifier of resource	string
github	Link to github repository	string
conditions_of_access	How can the user access the archive?	single choice
fulltext_available	Does the dataset contain full text?	bool
sources_included	Sources covered by dataset	ListRelationship
countries	Which countries are covered in the resource?	ListRelationship
geographic_scope	National, supranational or multinational, subnational	multiple choice
languages	Which languages are covered in the dataset?	ListRelationship
temporal_coverage_start	Start date of dataset coverage	datetime
temporal_coverage_end	End date of dataset coverage	datetime
text_types	Text Genres covered by dataset	ListRelationship
documentation	URLs to additional materials for the dataset (e.g., FAQs, documentation, codebooks)	list string
file_formats	File format	ListRelationship
meta_variables	Meta Variables	ListRelationship
concept_variables	Concept variables	ListRelationship
text_units	Text segmentation in the resource (what level of text units are available)	ListRelationship
related_publications	Link to scientific publications using the dataset	ListRelationship

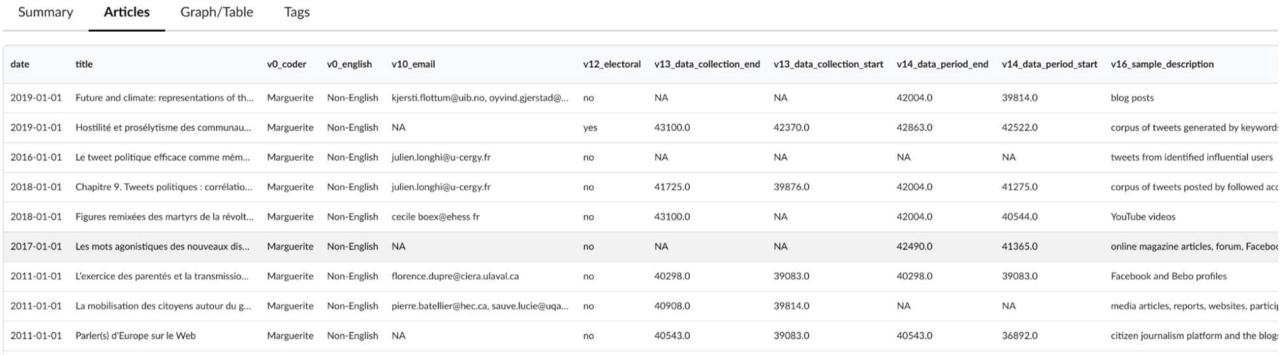
Table A2 Variable name terminology for CPPT publications in the inventory

Entity Type / Variable	Description	Data Type
title	Title of publication	string
authors	Name(s) of the individual(s) or organisation(s) listed as author(s) of the resource	ListRelationship
date_published	Date of publication	datetime
paper_kind	E.g. journal article, book chapter, ...	single choice
venue	Name of the place where the publication can be found. For instance, for academic papers, it is the name of the journal.	string
url	URL link for the publication	string
doi	DOI, if available	string
openalex	ID on openalex	string
arxiv	arxiv identifier of resource	string
tools_used	Tools used in the publication	ListRelationship
methodologies	Methodologies / Operations used in the publication	ListRelationship
concept_variables	Concepts investigated in the publication	ListRelationship
sources_included	Entities covered in the publication	ListRelationship
text_units	Unit of analysis of the text data used in the publication	ListRelationship
datasets_used	Datasets, corpora, or archives used in the publication	ListRelationship
countries	Countries covered in the publication	ListRelationship
geographic_scope	National, supranational or multinational, subnational	single choice
languages	Languages covered in the publication	ListRelationship

8 Appendix B

The below screenshots show the testing version of how Meteor will host the CPPT inventory, as created by WP7.

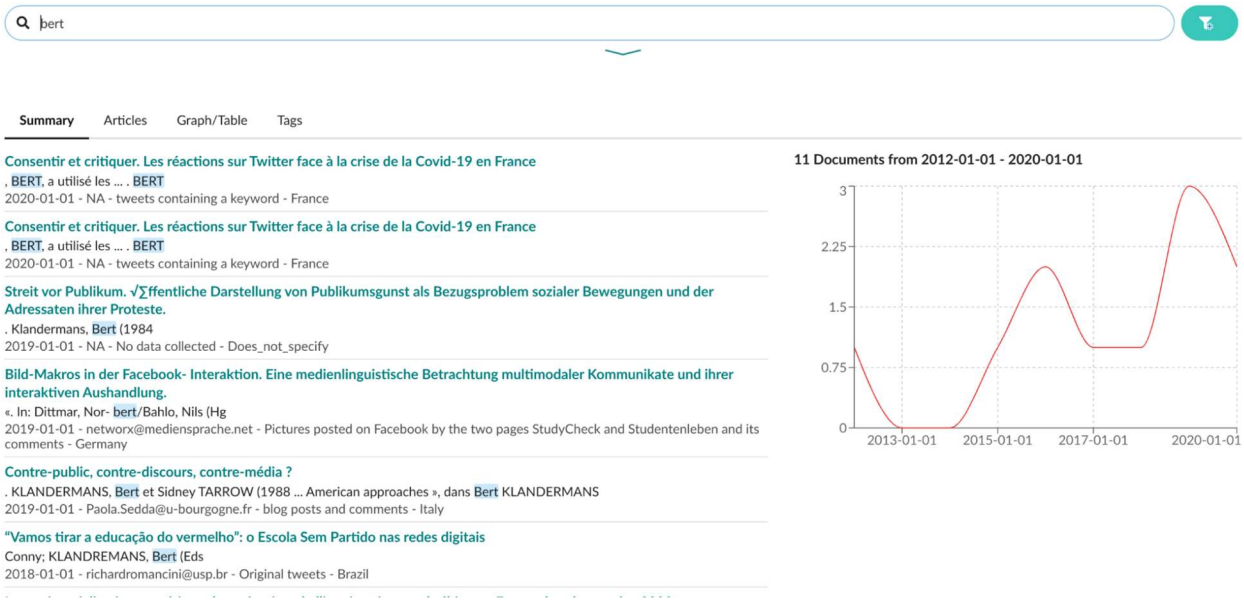
1. It is possible to browse CPPT articles in the dataset.



date	title	v0_coder	v0_english	v10_email	v12_electoral	v13_data_collection_end	v13_data_collection_start	v14_data_period_end	v14_data_period_start	v16_sample_description
2019-01-01	Future and climate: representations of th...	Marguerite	Non-English	kjersti.flottum@uib.no, oyvind.gjerstad@...	no	NA	NA	42004.0	39814.0	blog posts
2019-01-01	Hostilité et prosélytisme des communau...	Marguerite	Non-English	NA	yes	43100.0	42370.0	42863.0	42522.0	corpus of tweets generated by keyword
2016-01-01	Le tweet politique efficace comme mèm...	Marguerite	Non-English	julien.longhi@u-cergy.fr	no	NA	NA	NA	NA	tweets from identified influential users
2018-01-01	Chapitre 9. Tweets politiques : corrélatio...	Marguerite	Non-English	julien.longhi@u-cergy.fr	no	41725.0	39876.0	42004.0	41275.0	corpus of tweets posted by followed acc
2018-01-01	Figures remixées des martyrs de la révolt...	Marguerite	Non-English	cecile boex@ehess fr	no	43100.0	NA	42004.0	40544.0	YouTube videos
2017-01-01	Les mots agonistiques des nouveaux dis...	Marguerite	Non-English	NA	no	NA	NA	42490.0	41365.0	online magazine articles, forum, Faceboo
2011-01-01	L'exercice des parentés et la transmissio...	Marguerite	Non-English	florence.dupre@ciera.ulaval.ca	no	40298.0	39083.0	40298.0	39083.0	Facebook and Bebo profiles
2011-01-01	La mobilisation des citoyens autour du g...	Marguerite	Non-English	pierre.batellier@hec.ca, sauve.lucie@uqa...	no	40908.0	39814.0	NA	NA	media articles, reports, websites, particip
2011-01-01	Parler(s) d'Europe sur le Web	Marguerite	Non-English	NA	no	40543.0	39083.0	40543.0	36892.0	citizen journalism platform and the blog

2. It is possible to get a detailed view of trends and patterns in the inventory. In the example, the line graph shows how many qualitative, how many quantitative and how many mixed methods studies were published each year of the inventory data (please mind that this is based on a sample of the data).

3. It is possible to search for keywords of interest. In the example, the modelling type 'bert' is searched for, showing not only how many publications use this term (and from what year they are published), but also how the term appears in each publication. The usefulness of making texts searchable in this way is made clear, by the fact that some hits are for authors with the name Bert, rather than the model.



4. It is possible to click on any inventory publications to read an abstract, see the metadata and have a link to the publication itself. However, the full text of the article is not possible to access directly within Meteor/AmCAT4.