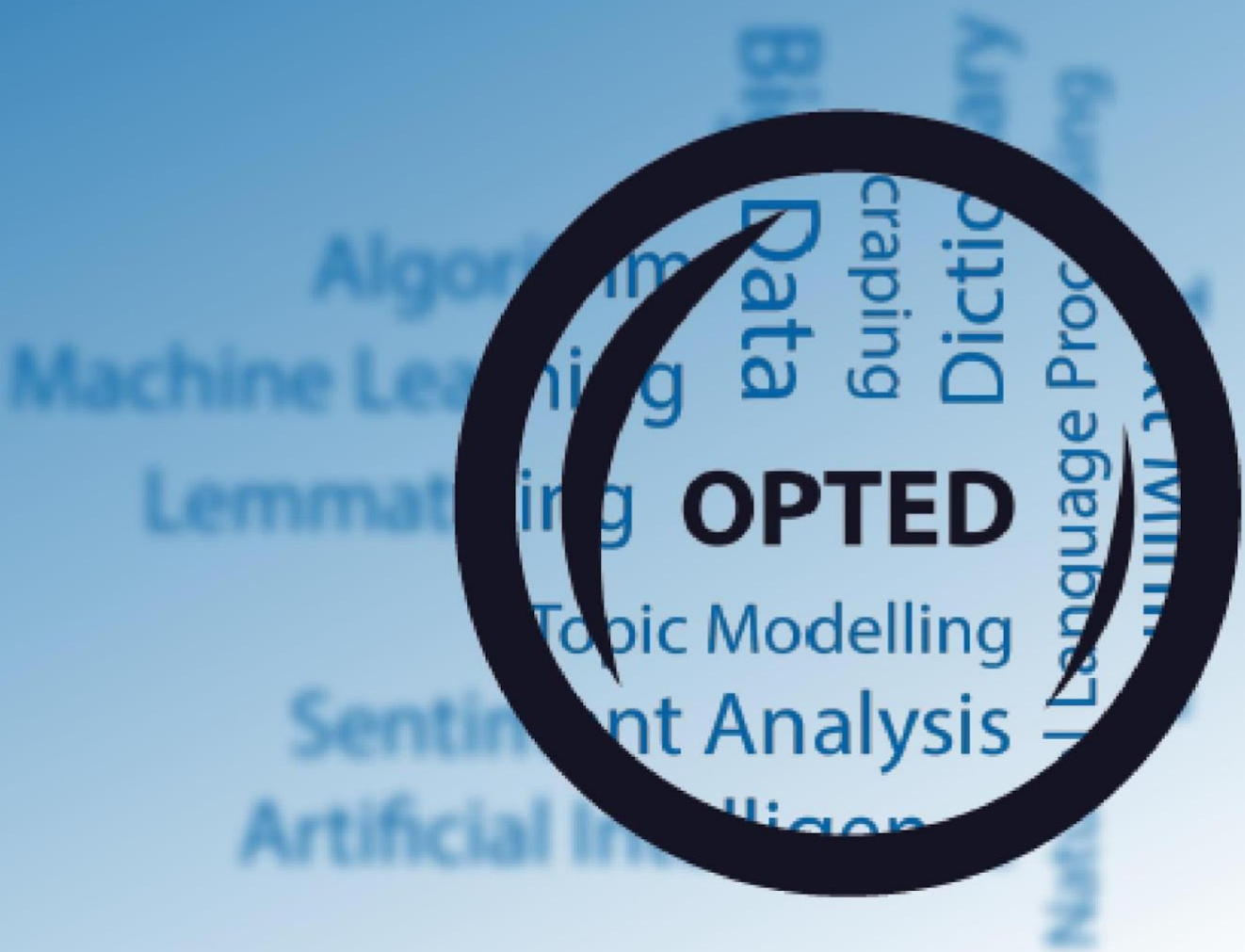
OPTED

Conceptual framework of core challenges in CPPT research

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



Disclaimer

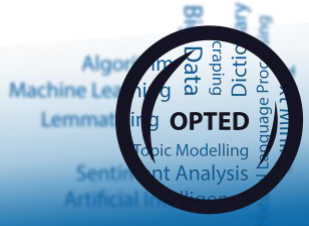
This project has received funding from the European Union's Horizon 2020 Research & Innovation Action under Grant Agreement no. 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report



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

Conceptual framework of core challenges in CPPT research

Deliverable D2.3

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: September 2022

1 Executive Summary

In previous reports, WP2 has outlined the increasing use of citizen produced political text (CPPT) in the political text analysis community, in particular within social media studies. CPPT scholars face many challenges in accessing text data produced by citizens. Moreover, knowing (and following) relevant guidelines relating to ethical and legal restrictions can be uniquely difficult, or at least more prevalent, for those studying text produced by citizens, as compared to those studying texts by elites or public actors. In this report, we analyse survey and semi-structured interview data from experts and CPPT researchers to establish which aspects of CPPT research create the greatest challenges. We highlight those factors that exacerbate challenges for CPPT researchers depending on their language(s) of interest, geographic differences and more. The report thereby provides an overview of those core challenges that impact the ability to access and ethically use CPPT data, with **four main themes: access, ethics, regions and languages, inequalities**. To summarise the major challenges:

- Access to data is restricted
    - Legal and financial barriers to accessing the CPPT data researchers are most interested in often lead researchers to choose data they consider second-best.
    - Identifying and using tools to access social media data are seen as time-consuming.
    - With online data regularly being deleted (by citizens themselves, moderators, companies, or otherwise), researchers are unsure how to (legally and ethically) deal with this issue.
    - Companies imposing restrictions are perceived as the greatest barrier for CPPT data access.
- Ethical guidelines are unclear or hard to follow
    - Guidelines are interpreted differently across and within CPPT disciplines, and their application in specific research projects lead to uncertainty – in particular, when the research topic is sensitive or when legal and ethical objectives are not in line with each other.
    - Researchers are missing fora where they can openly discuss ethical best-practices among peers, such as potential exemptions from gaining informed consent.
    - Institutional knowledge (in particular within ethics committees) of CPPT ethical procedures is perceived as insufficient.
- Non-English and multilingual CPPT research is more resource intensive
    - Data availability differs by region and by language of text.
    - Analytical tools, in particular within computational CPPT research, perform better with English text.
    - Translation costs are high, and the peculiarities of CPPT data make translation particularly difficult.
- Inequalities lead to disproportionate barriers for some researchers
    - Collaboration reduces the individual cost of CPPT research, but networking possibilities are limited for junior researchers and researchers from certain regions.
    - CPPT researchers experience a bias in favour of English, and a bias in favour of quantitative methods.
    - Structural support needed for CPPT research (including funding and ethical guidance) varies by country and by institutions.

In this report, the survey and interview methodologies are briefly described, before giving detailed description of the CPPT challenges. Results are separated into the four main themes: access, ethics, languages and regions, and inequalities.

2 Uncovering core challenges

Research possibilities using text data are growing (Baden et al. 2022), and citizen-produced political text (CPPT) is among the most common text data being analysed. CPPT is here defined as text produced by citizens, either offline or online, relating to a political process, a policy or a civic issue. The questions that can be addressed through the use of such data are broad and nuanced – ranging from examination of millions of search engine queries of netizens in China relating to anti-corruption campaigns (Zhu and Wang 2020) to uncovering the covert political undertones in seemingly neutral memes in Poland (Niekrewicz 2020). Studying CPPT allows for focusing on the views, opinions and experiences of citizens rather than elites, and thereby contributes an invaluable perspective to political communication research. Despite being a rich and important data to engage with, there are many issues with using CPPT that require better understanding. We use survey and interview data to address two primary questions:

1. What challenges are faced by CPPT researchers relating to the themes of access, ethics, languages and resources?
2. How do these challenges vary between researchers due to stage of career, country of affiliation, region or language of interest, or other personal factors?

2.1 Survey

The survey was run jointly with WP9 of OPTED and received approval from the Research Ethics committee of the University of Exeter – the survey questionnaire can be found in the appendix of D9.3, wherein more details about the survey method and selection procedures can also be found. In this deliverable, survey responses as of August 15th 2022 are analysed including 295 respondents, however respondents could skip questions. Therefore, total N for the question being addressed is noted in results. Respondents were reached by different channels – email groups, email addresses from journal publications and advertising through the OPTED network. Of the 251 respondents who indicated which text data they use, 163 report currently or previously using CPPT data, while a further 59 indicate that they have not yet used CPPT data but "would maybe use in future", and 29 do not intend to use CPPT. Respondents come from a variety of countries across the world and represent a broad array of academic fields within the social sciences. Of those respondents who indicated career stage (83% of all respondents), 20% are PhD students, 18% are early-career researchers (less than 5 years post-PhD), 35% are mid-career (5 to 15 years post-PhD) and 24% are senior (more than 15 years post-PhD). 58% of respondents indicated being male while only 40% female (the remainder preferring not to indicate gender or identifying as neither male nor female).

2.2 Interviews

The in-depth interviews study was conceptualised within WP2 and was approved by The Ethics Committee of Audencia Business School. 21 interviews with CPPT researchers and experts were conducted. Interviewees were recruited via purposeful sampling to fit the perspectives described in Table 1.

Potential interviewees were carefully identified through a combination of methods (including reviewing the CPPT dataset, contacting leading methods teachers in the relevant fields, or otherwise identifying experts for the main themes of data access, ethics and inequalities in text analysis of citizen data). Interviewees, like survey respondents, represent several disciplines: 8 from political science, 5 from communication, and the remaining from computer science, international relations, linguistics, law and data companies. 16 interviewees come from Europe, representing Eastern, Northern and Southern Europe as well as Western Europe, and 5 interviewees report being from countries outside of Europe (Africa, Asia, South America). As defined above, 8 interviewees are PhD students or junior, 10 are mid-career and 3 are senior researchers. 13 interviewees were women compared to 8 men. Interviewees were assured of their anonymity, and are below given a random identification number when referenced[1]. Interviews were semi-structured, with a list of primary questions asked to all interviewees relating to the four themes (access, ethics, regions and languages, inequalities), and additional questions specific to their expertise (for details, please see appendix). Interviews were conducted over Zoom in the months of May-July 2022, with interviews lasting between 35-90 minutes. Interviews were

[1] Interviewee citations are edited for language correction as needed.

audio-recorded and transcribed, before being coded using NVivo.

Table 1 Purposeful interview sampling

Interviewee group	*Motivation*
(Political text) methodology teachers	Experts on the current status of data/methods development and knowledgeable about issues others face.
Ethics review board members	Experts on current restrictions, developments of guidelines and how these affect other researchers.
Editors	Can speak about the 'output' side and have a broad, overview perspective.
Researchers who collaborate with social media	Researchers who have interacted with social media companies can shed light on gatekeeping and collaboration possibilities.
Authors from the CPPT dataset	Studies from the dataset that add diverse perspectives: variety of methods; use CPPT data from diverse sources; at different stages of academic career.
Authors writing about ethics and inequality in research	E.g. researchers who have written about ethical issues in using social media data, or on inequalities in terms of which populations are studied, based on which data can be gathered.

3 Results

3.1 Access challenges

Access to data is experienced as a clear challenge by the survey respondents (N=251). The challenges were characterised as:

- Restricted access to data by companies owning or storing the data (191);
- Restricted access to data due to content removal (154);
- Difficulty identifying or using the tools needed to access data (139);
- Difficulty finding or identifying all relevant text data (e.g., because words that are spelled incorrectly by users are not collected) (134);
- National ethical research rules or guidelines do not enable accessing certain text (113).

The most salient issues – company restrictions and content removal – highlight the importance of social media platforms as a source of data for CPPT researchers. Social media provide (usually) unfiltered first-person accounts from citizens at large, and as our previous deliverables have shown, make up the lion share of CPPT data; 85% of the 3,260 publications in the CPPT inventory use either Facebook or Twitter data (Gelovani et al. 2021). Although the access challenges clearly were salient among the respondents as a whole, all issues were more prevalent among those who use CPPT data. Table 2 shows that (among the 222 respondents who answered access questions) all five access issues were most likely to be perceived as challenges among CPPT users, compared to those who are not interested in CPPT data and those who are interested but have not yet conducted research with CPPT.

Table 2 Access issues

	Non-interested (N = 20)	*Interested (N = 56)*	*Users (N = 146)*
Company restriction	12 (60.00%)	44 (78.57%)	135 (92.47%)
Content removal	8 (40.00%)	34 (60.71%)	112 (76.71%)
Finding text	10 (50.00%)	33 (58.93%)	91 (62.33%)
Finding tools	11 (55.00%)	33 (58.93%)	95 (65.07%)
Ethical restrictions	9 (45.00%)	24 (42.86%)	80 (54.79%)

Note: cell entries are number of responses and column percentages.

3.1.1 Company restrictions

Social media companies, whose data a majority of CPPT researchers wish to analyse, restrict access in a number of ways that the interviewees explain.

Firstly, **companies typically do not allow researchers to access all data**. Often this is to protect the privacy of users; posts within private groups on Facebook, for example, cannot be accessed as freely as public pages can be. This often leaves researchers with what they consider suboptimal data sources, because the potentially more valuable data is not accessible. The typical ways company APIs allow access can also be problematic, for example:

> *"With Twitter API's, there's a limit for how many inquiries that you can make. And also there's a limit for how far you can go back. [...] if you try to replicate exactly the kind of analyses that have been done before, for many, many social media data you can't really go back and get the real data back in. And most of the time for privacy reasons, you are not allowed to store the exact contents of the text as well. So it's actually - there's no way to go back. and then, you can't replicate the analysis. [...] You get only a snapshot each time and then you get wildly different results if you do research today, and then three years ago, and you have no ways to verify where [the difference] is coming from" (Interviewee 8)*

A concern many interviewees shared is that restricted data access by companies leads to less representative data. However, the issue is also that people who post on social media are not representative of the general population (Hargittai 2020), and not everyone using social media post politically relevant content (Boulianne and Hoffman 2021), impacting the ability to infer from social media to 'real world' phenomena:

> *"the stuff that I'm usually interested in is political communication or interaction around political topics. And that is only a fraction of what is happening on Twitter. So you're... You're even talking about a minority of a minority" (Interviewee 14)*

Even when using a third-party company to access data, a potentially costly option, restricted data access can be a problem – although recently, access has increased.

> *"There are lots of companies that offer these kinds of services, so they can access the API and they can download it. Doing something with this. But it's very expensive and they don't give you as much as you want, because I want to work with lots of data, and maybe they can offer to give you, I don't know, the last month of, you know, posts and comments of a page. That was simply not enough for me. and these services as I understood, you know, when I contacted them, what they could offer me to do my research - it was simply not enough" (Interviewee 20)*

Despite these concerns, publicly available social media posts are 'low hanging fruit' (Özkula et al. 2022; Burgess and Bruns 2015) that continue to be relied on when there is no better option available. As summarised by one interviewee:

> *"Twitter is very infrequently the best solution for a problem. It is just usually the easiest solution for a problem." (Interviewee 5)*

Secondly, **companies over time change what is allowed and not allowed to access**, leading to researchers' uncertainty. Interviewees feel that while some platforms (such as Twitter) open up more over time, others[2] (such as Facebook) become more closed due in part to scandals such as Cambridge Analytica. Not only does this restrict what interviewees feel they are able to study because of data becoming unavailable, but some also express frustration that only a sample of data becomes available rather than all relevant posts that are called for with an API, for example:

> *"they are changing the guidelines. Always. Like constantly what you can scrape, what you cannot scrape, changes and – may I say – with zero regard to researchers. Like they do not consider researchers when putting through these guidelines. They consider industry people. Because we use the same data industry people use." (Interviewee 4)*

> *"When I started this was – you could download everything. Everything was scrapable, so of course*

[2] Twitter and Facebook are the most commonly used platforms for interviewees, though other platforms (e.g. Instagram and YouTube) are also perceived as changing terms and conditions over time.

there was no limitation. [...] but now we can't have access to the data, so we have to orientate the research with stuff that we can do" (Interviewee 16)

Interviewees not only feel uncertainty in regards to what they can reliably access at the moment, but in regards to how access will change in the future and impact ongoing projects. Such uncertainty and the power imbalance between researchers and social media companies leads to a distrust in the companies providing data.

> *"Facebook used to be rather open. It isn't anymore. You can get to it through a service called CrowdTangle, which is owned by Meta. But they've been... A lot of the people who worked on CrowdTangle have since been placed in other services within Meta, or some have even quit. So we never... We don't really know what's going to happen with CrowdTangle" (Interviewee 6)*

> *"what we've seen, Twitter opening up the API for researchers a few months ago, I think that was a very big thing. I'm not necessarily sure how long that stays, because, well, it's always at the whim of the company. [...] Facebook tried to, with the Social Science One thing. I think with that consortium, I think that is more or less a case for how you don't do it. Because that was basically – I mean, that was a transparent PR move by the company. They tried to show 'we take things seriously'. But then guaranteed to basically replicate academic inequalities with regard to access to the data" (Interviewee 14)*

Thirdly, **skill and time requirements to access data continue to be high thresholds for entry**, despite the popularity of these data sources. Developments such as the Academic Twitter API minimises the need for researchers to do their own programming, but can sometimes lead to lower quality data frames that create more pre-processing work before analysis can be conducted. For example, one interviewee compares the data collection with Twitter's previous API and the new Academic API:

> *"the new API had come out. [...] on the one hand, it was great because in a couple days I was done with the data collection, which was not at all the experience that I had with the other paper. But what I found is that the data was a lot more messy, so the data frame was messier and I had to do many like cleaning operations to have, you know, a dataset that looked more or less as clean and tidy as the data set that I obtained from the other API's and the other R packages. And at the same time, I had some issues: so sometimes the data would contain some metadata and sometimes it wouldn't, which was a problem." (Interviewee 19)*

Indeed, several interviewees express frustration that the data gathering process becomes more strenuous than it need be because of how companies structure their APIs, and the information they provide to guide researchers:

> *"Whenever you have gathered [data], everything is fine. [...] The problem is gathering. What are you allowed to gather? Which kind of data can you gather? Et cetera. So that that one is, yeah, basically it's the most important obstacle that we have to overcome" (Interviewee 9)*

3.1.2 Content removal

Content removal creates two primary issues among interviewees: it complicates the ability to replicate analyses and it limits the ability to study communication that is most likely to be deleted (hate speech, conspiracy theories, anti-government rhetoric, etc.). Interviewees using social media data experience that posts are removed, made private or edited by the person or page who created the post, or indeed because the user deleted their account. Or, content is removed when a page, user or group is blocked.

In certain cases, content relating to a particular topic is removed systematically. This can be the case because of states or other powerful actors choosing to censor online discussion of a topic. One interviewee shares that they and their colleagues have the suspicion that activist groups with large followings sometimes hand over their online spaces, either due to threats or for financial gain, to outsiders who then delete the accounts.

When asked about content removal or censorship, one interviewee mentioned the phenomenon of mob censorship (Waisbord 2020). While the concept refers to citizens mobbing journalists as a form of silencing, the interviewee explained how for certain research topics, citizens may anticipate potential responses before posting about that topic and therefore choose not to post publicly or at all:

> *"People turning older posts into private, or deleting them. You never know. Or people doing exactly*

like that, doing selective exposure so [the post] would remain among their trusted few." (Interviewee 2)

Regardless of why content is removed, there is a shared frustration with how this impacts the ability to study certain phenomena where deletion is highly likely, especially when working with data from platforms that are particularly likely to remove content.

> *"Facebook pulls everything that it finds offensive. Then you have the user review process that might pull the next thing and then people pull their own things or edit them later. And yes, I know that you can technically get to the pre-edit version. But let's face it, who does that? So this is a problem" (Interviewee 5)*

> *"I cannot possibly be on top of everything at all times. Things may get deleted, I cannot check many thousands of tweets whether or not they got deleted constantly." (Interviewee 4)*

There is a **lack of certainty of how to handle content removal**. In particular, interviewees' practices when it comes to checking for deleted posts differ. On one end of the scale, some believe that as long as they only analyse those posts that were available at time of data collection[3], they are in line with guidelines. Indeed, depending on the source of data this can sometimes be the only possible recourse when using large datasets, and interviewees feel that by engaging in proper anonymising protocols they are acting in line with the terms of service of platforms who state that deleted content should not be kept in records. In the middle of the scale is the practice of not checking for deleted content in the aggregate analysis, but taking care that any examples being quoted in the analysis have not been deleted at time of publishing. On the other end of the scale, interviewees explain the process of 're-hydrating' social media posts: regardless of what content is available at time of data collection, the version of data that is archived and analysed should go through the following process, in the example of Twitter:

> *"essentially, you have this list of tweet IDs and then you send that up to Twitter and back comes the tweets themselves in an orderly like CSV format or something. And then if a tweet ID has been removed, i.e., if the tweet itself has been removed, this is not coming back to you, [...] But the idea then is that you have some sort of token - like the tweet ID - that you then re-populate your data set based on what's available - what's still available." (Interviewee 6)*

Yet, even those interviewees who are familiar with the concept of re-hydrating data often state that they currently do not engage in this practice, or are not sure exactly how it works.

3.1.3 Finding text

Despite being the third most common issue among survey respondents, interviewees did not report issues identifying or capturing relevant text. Neither in trying to find the online or offline spaces where relevant political discussions are going on, nor excessive difficulty in specific problems such as dealing with issues of misspelt words, colloquial language or dialect. Instead, capturing relevant text data was primarily related to issues of non-English or multiple languages – see section 3.3.

3.1.4 Finding (and using) tools

Regarding tools for accessing CPPT data, a primary issue is time investment. For example, several interviewees outline the significant amount of coding that is needed when using existing tools. The time-consuming nature is also a result of the high threshold of knowledge needed to use tools, discussed already in this report. Moreover, because many of the techniques are rather new, it is difficult to find sufficient guidelines, leading to mistakes that prolong data collection. One interviewee shared an experience when conducting research on a website popular in their country:

> *"I was crawling their website and my crawler was getting that data at a very huge speed, [...] so when I was crawling this data the guys from this company reached out to [interviewee's university] and they blocked my IP for a week or so. And they were like 'OK. You're not going to do anymore crawling because they're very mad at us'. What were we doing? We're just jamming their servers!" (Interviewee*

[3] As multiple interviewees explain, it is not possible to retroactively access deleted posts using standard APIs for many platforms, however in certain cases such posts can be accessed through other means.

18)

Another recurring issue is that **tools developed for other purposes (such as marketing research) often do not result in exactly the type of data needed by CPPT researchers**, with one interviewee also stating that they felt stronger trust in tools that are developed by discipline-related researchers themselves, such as Netvizz[4]. Moreover, tools that are trained for use with other types of text data than the short posts common to social media and other CPPT sources often perform worse when applied to CPPT data.

> *"I think given the structure sometimes of this text, it's not that straightforward. Some of the models cannot be applied or perform very differently with these short texts. Others are quite OK with this. […] But often you see that the behaviour is quite different than to text with normal length" (Interviewee 21)*

A final issue is that, regardless of how well certain tools might suit researchers' needs, there is a **lack of knowledge of what tools are out there** – one interviewee expressed that it is primarily via word-of-mouth that one becomes aware of existing tools, rather than having a reliable place to browse alternatives to choose between. This results in sometimes choosing suboptimal research methods for the question at hand.

3.1.5 Ethical restrictions for access

In many cases, researchers are not certain whether they are ethically allowed to use CPPT data they have accessed or wish to access. In other cases, there are ethical restrictions in place in one country that differ from those in another country – one interviewee discussed the difficulty of finding out whether restrictions of the country where the researcher conducts their research, and restrictions of the county (or countries) where data is from, are in line with one another. As another interviewee expressed, **the uncertainty and extra work required to access CPPT data that may be restricted for either ethical or legal reasons makes it easier to simply choose other data sources**.

3.2 Ethical challenges

Survey respondents (218 of which answered ethics questions) overall indicate having awareness of the best practices relating to using text data, with 77% agreeing to that statement when asked. However, among PhD students, only 54% feel they have awareness of best practices, compared to 92% among senior researchers. Moreover, only 52% agree that they received sufficient training in data protection and privacy and 55% agree that they have not always been able to ask for informed consent to participate even when this would have been best practice.

3.2.1 Uncertainty or lack of awareness

Although a majority of survey respondents indicated knowing the best practices for using CPPT data, interviewees share their **uncertainty regarding what would be ethical in specific research contexts**. Several interviewees wish for regularly updated information that is available publicly, which takes into account ethical standards of specific disciplines and platforms. One interviewee with ethics committee experience highlights some of the main issues that vary greatly from project to project:

> *"Who counts as a public figure online, so that you can use their statements freely as you want, and where someone is not a public figure, and you need their informed consent if you want to present their statements or posts explicitly? And what degree of anonymization is enough - like what degree of anonymization each project needs based on what kind of data it handles?" (Interviewee 4)*

Moreover, one interviewee working primarily on issues of research integrity in the EU shared that perhaps researchers overestimate their understanding of best practices, leading them to resist a standardised approach to ethics with internet data in particular:

> *"each institution has its own culture. Not everything can be transplanted, and if it's forcefully being transplanted then it won't be implemented, it will be just an empty regulation or standard. So now we're just going into circles to be honest." (Interviewee 1)*

A potential lack of awareness of ethically sound practices may lead to engaging in research with good

[4] http://thepoliticsofsystems.net/permafiles/rieder_websci.pdf

intentions but bad practices, and therefore shed doubt on CPPT studies at large, even among those studying CPPT themselves:

> *"if I would have to just make a guess, I would say that half of the studies done on Facebook, if not more of them, are unethical." (Interviewee 5)*

3.2.2 Legal versus ethical

A recurring verse in the interviews was the **tension between what is legally allowed in CPPT research and what is ethical**. In particular, that access to certain text is not allowed due to legal restrictions of either GDPR or social media companies, but that in fact there is no ethical reason to restrict access. There is some concern that legal restrictions limit the studying of important topics. As expressed by one interviewee:

> *"I'm generally a big fan of this idea of controlling your own data and right to your own data and everything, but that must have limits. Because if the right to your own data leads to a situation where we start getting bubbles of Neo Nazis who can refuse scrutiny of their discourse because they say 'I do not consent'." (Interviewee 5)*

Another interviewee shared that despite knowing it was not legally allowed by Facebook, they joined private Facebook groups in order to collect data for analysis, explaining the extreme danger these particular groups can pose and therefore the necessity of studying them. Indeed, this type of behaviour may be commonplace, as multiple interviewees feel certain studies are not possible without potentially crossing some legal restrictions. For example, one interviewee recalls questioning the legality of data access in a journal article they were reviewing:

> *"I thought, 'how come they can have that data? And they do not account for that'. But there the answer that I gave to you earlier was that to me, there was a quiet voice that said 'ok. They did not harass anyone and the people that are there were not influenced at all because of the [researchers'] behaviour, so who cares?'" (Interviewee 9)*

The previous quote highlights a sentiment that many interviewees share, in that researchers have an innate sense of what is ethically correct from their training and familiarity with the literature, thereby prioritising this sense over legal barriers of companies. Indeed, several interviewees expressed that strictly following legal restrictions can make citizen data not useful enough to engage with:

> *"of course, there's, you know, more convoluted and possibly illegal ways that you could do this still, but that's not a good idea. [...] If the ethical hurdles weren't as difficult to handle in terms of - it takes time, and it's an administrative burden - I would probably do more research into [citizen text]" (Interviewee 6)*

Finally, the incongruity between certain legal and ethical standards can also create a problem when researchers feel that by complying with the legal standards and some minimal ethical guidelines, they don't need further reflection:

> *"when people feel they are caught up in these huge administrative processes, and maybe that their admin at the universities would say 'you cannot do this, you're not allowed to do this'. That is not about research ethics, it's about formal procedures and the fear of sort of doing something illegal. Where the ethical responsibility - and I think that's important - is mainly on the researchers themselves. And you cannot exempt yourself from research ethics by saying 'this was pre-approved by somebody' because it should be an integrated part of the whole process." (Interviewee 12)*

3.2.3 Different interpretations of guidelines and context-specific differences

Different interpretations of ethical guidelines are commonplace, according to interviewees. Interviewees experienced different interpretations by members of the same ethics committee, considering the same research proposal, but also more informally when at the research design phase, including vastly different views of ethics by country.

> *"It's not that there's no agreements, but there's some leeway still on how we interpret the limitations imposed by Twitter. And yeah, I've seen some practices that I don't necessarily agree with." (Interviewee 19)*

"my socialisation in [country in Europe] was the jungle. I never, never, never was asked to do any ethics checking. I did a PhD without any ethics checking." (Interviewee 10)

With such nuanced data sources, ethical grey areas are common and deep reflections over the ethical, legal and social implications of using the data are necessary (Larsen 2022). As one interviewee explains – in reference to a specific project with unclear ethical implications – researchers and committees can sometimes have opposing motivations:

"I don't have a clear position on whether we are OK to study that. I can see good reasons for why we absolutely must be OK, and I can see good reasons for why this is a problem. It's not like this is clear. I completely understand why IRBs[5] are confused and everybody does it a little bit differently, because in the end it is a trade-off between different, equally – or maybe not equally, but all of them relevant – protected rights. [...] The problem is not there because of under-regulation. The problem is there because there is a genuine norm conflict." (Interviewee 5)

The interviewee specialising in EU research integrity explained that ethical best practices were more easily self-regulated within academia when research communities were smaller and had greater possibilities to discuss openly with one another, compared to today where **many researchers fear legal or social repercussions if they showcase uncertainty about ethical practices**. Indeed, sharing best and worst practices is missing according to other interviewees:

"Maybe it's me, maybe it's my colleagues, I don't know, but we just talk about what we do and we didn't talk a lot about – or I've never talked a lot about ethics" (Interviewee 20)

"this is also what's needed in research ethics, that we discuss these dilemmas, and we have forums and fora to discuss them and reflect upon them and challenge each other. And also discuss worst practise in things that we've done" (Interviewee 12)

Discussions are also needed because of the **necessity of tailoring ethical procedures to contexts**. Not just because of regional or country differences in what is considered ethical, but because specific topics and communities require different ethical procedures. Without discussion and openness from an institutional point of view to adapt ethical procedures to the context, data collection and analysis of CPPT may be compromised. For example:

"the link that you sent me, the consent to do the interview. That would be very problematic to copy paste that practise into our context because people would be... How would you say, anxious, about where that would go. Maybe they would prefer paper for example, because paper can be stored offline and cannot be hacked" (Interviewee 2)

3.2.4 Lack of institutional knowledge

A common theme in interviews was the lack of institutional knowledge of relevant ethics guidelines.

"you know, internet ethics is already a bit of a blurry area, every institution has their own idea of what ethical research looks like online." (Interviewee 4)

With GDPR, researchers felt their institutions were slow in updating their understanding of CPPT ethics, leading to what one interviewee called a sunk cost: individual researchers had to invest their time in understanding the impact of GDPR, in order to move their research forward, instead of institutions engaging in this work. Beyond GDPR, generally many researchers feel that their **departments are unable to provide the necessary guidelines for the complexities of using CPPT data**. When researchers use guidelines available to them via their university, but from other disciplines, this creates a number of issues. For example, one interviewee who engaged in observational-style research in a project using data from a less studied app used guidelines from another field due to a lack of other alternatives:

"this was an issue raised, you know, adapting an ethic guidance from another area which has different reasons for performing undercover research. And it was, I would say, criticised when I presented my work" (Interviewee 11)

Institutional lack of knowledge is perceived as particularly strong within ethics committees. One

[5] Institutional Review Board (equivalent of ethics committee in other institutions)

interviewee on an ethics committee expressed that it was important for them to be there as a voice of qualitative internet research in their departments' ethical committee, since otherwise there would be insufficient understanding of proposed projects using internet data. For universities with one ethical committee across all departments, this does not necessarily solve the problem, because **often the committee is made up of individuals from legal, philosophical, medical, psychological or other backgrounds rather than those main disciplines using CPPT**.

> *"I think the main issue is that most of these ethical review boards come from psychological research. And for them, stuff like party preferences already counts as highly sensitive information. But for us it is kind of the baseline that we need to know" (Interviewee 13)*

3.2.5 Informed consent and public data

Just over half of the anonymous survey respondents agreed that they would have wished to ask for consent in certain projects but were unable to. Among the interviewees, not asking for informed consent is justified by one of three factors: 1) claiming that the data used is public; 2) the unfeasibility of gaining consent when using big data; 3) that asking for consent would lead to a change in behaviour and therefore risk the research project. For example:

> *"sometimes I have thought about sending a consent form or something like that, but I'm quite sure it would mean the end of my research. Because in these kinds of groups – they are not debate groups. They are propaganda groups. And if you have a contrary opinion or if you are seen as a spy or something like that, you are thrown out immediately" (Interviewee 15)*

> *"the reason I didn't seek informed consent is that it's already a difficult time [...] people couldn't interview people without seeking police approval first, that kind of thing. And during that time, I thought seeking informed consent creates a digital trace." (Interviewee 4)*

Given the public nature of posting in online spaces where anyone could see the post, many feel that the data is free to analyse without anonymisation. However, some interviewees are highly critical of claiming that informed consent is not needed in such cases, highlighting that publicly posting does not mean you anticipate having a spotlight shone on your heat-of-the-moment reaction to a political event, for example. They therefore advocate for strenuous anonymisation procedures and omitting identifiable quotes and screenshots of even publicly available posts:

> *"Most people are not informed of what they are saying 'ok' to online. [...] we can't say 'oh, that's their problem that they didn't read the 300 pages of terms of service' [laughs] because the thing is that, unfortunately, being on these platforms is a requirement of life in our current age. [...] I feel like in a system where people are technically coerced into selling their data, in order to be able to exist and be well informed and be a part of the times, we cannot get away with saying: 'Terms of service!'" (Interviewee 4)*

> *"the general main rule of asking for consent to participate in research also applies to the online world. This is how we uphold trust in research and research integrity. Of course there are exceptions, but in many cases there is no practical reason not to ask for consent, because people are easily available. You know who they are." (Interviewee 12)*

In other cases, interviewees express that they **avoid CPPT data where consent would be needed**, adapting their research questions according to the type of data that they can more easily get ethics approval to gather:

> *"I'm not looking for it right now. In the future I might be, but yeah, it would be more difficult to get kind of the right approval. And also ethically, you should inform the people you study and, well, some of the people I study would probably be very angry – more angry than they already are – so it might also be ethically very difficult." (Interviewee 3)*

3.2.6 Exploitation and vulnerable groups

A final ethical issue of note relates to **vulnerable groups and exploitative research practices**. One interviewee highlights that one can never be certain whether social media posts come from underage accounts, meaning that special care might need to be taken on a case-by-case basis depending on the type of research

questions being addressed; but also extra care in anonymising individuals when studying stigmatised groups is needed. Interviewees are concerned with the tendency of researchers to not make research results available to the communities they are studying, in particular when Western researchers study non-Western communities and do not provide any research summary in the languages spoken by those communities.

3.3 Language or regional differences

Table 3 Languages studied by survey respondents

	Non-interested (N = 20)	*Interested (N = 56)*	*Users (N = 146)*
Only English	3 (15.00%)	2 (3.64%)	22 (15.60%)
Multiple incl. English	16 (80.00%)	43 (78.18%)	105 (74.47%)
Only non-English	1 (5.00%)	5 (9.09%)	10 (7.09%)
Multiple not incl. English	0 (0.00%)	5 (9.09%)	4 (2.84%)

Note: cell entries are number of responses and column percentages.

Survey respondents were more likely to study text in multiple languages than in only one. The 216 respondents who indicated which language(s) they study are shown in Table 3 above. Among CPPT users, 75% report using English and at least one other language, while only 3% study multiple languages not including English. In interviews, when asked about language-related challenges, it became clear that many interviewees see regional differences as equally important factors.

3.3.1 Data availability

Access to text varies by language and region. In the survey, 42% of CPPT researchers indicated that a major reason they use the languages that they do, is because of which languages it is possible or easy to find textual data for. In interviews, text access mattered more due to region than language: either easy-to-study platforms have greater usership in some countries than others or because of specific regional challenges.

> *"I see that there are some papers coming out like using Twitter data and different kinds of other forums where they have analysed like millions of tweets [...] but I guess in the [small country in Europe] context, to find enough data to do it well? I don't think it's possible" (Interviewee 3)*

> *"the problem is that in the North African context, there are no archives. The Internet also, even the Internet deletes, the web pages are not stable. And that applies to citizen produced text, which generally have less durability then print media or audio-visual materials." (Interviewee 2)*

To complicate matters further, social media is used differently by different communities – **for certain topics, it may be more likely to find open discussion on social media in some populations than others**:

> *"Twitter does not stand for the globe. For example, one critical premise is that the majority of the Twitter users are in the US or in Western countries [...] It's easy data to collect honestly when compared to interviews, especially in an authoritarian context. But we also need to frame it in its relevance, and its scope" (Interviewee 2)*

> *"People from my own country, [country in Asia] and in my part of the world. They would not be that vocal on social media, so it's very hard to gather - or even if they are vocal, they won't talk much about these topics that we are interested in" (Interviewee 18)*

3.3.2 Methodological possibilities

Regional differences in methodological training, and therefore ability to engage in CPPT research, was also a theme discussed by interviewees:

> *"we still have a problem here in [country in South America], working especially in human sciences, in working with big data. We are not trained to do that" (Interviewee 11)*

Relating instead to the language of study, methodological possibilities are perceived as limited because of the difficulties of studying multiple languages. In the survey, 28% of CPPT researchers indicated that

availability of tools is a major reason for studying the languages that they do. Interviewees also highlight the uncertainty of what to do when special characters in languages are not handled properly by software. Generally, the perception is that **tools work best for English, and the more different a language is from English, the more likely that tools will be unavailable or require significant work to apply them**. Issues such as handling accents, languages using different symbols, or finding validated stop-word lists, make researchers either forgo certain languages or forgo certain methods. For example:

> *"I had quite some students that were using Arabic text, [...] there we had also some encoding things just to start with, you know, to get the texts properly into R and work with them when they have very different letters etc. [...] I mean, UTF 8[6] is great. But they often have different encodings on their computers so when they read even a proper Arabic text into R, for instance, they transfer to very weird systems that they are not aware of." (Interviewee 21)*

3.3.3 Translation

Unlike elite text, **citizen text is more difficult to translate in a way that retains relevant contextual information**. A concern almost universal among interviewees is that the short, unstructured and less strategic text from citizens is difficult to properly analyse or interpret, and this issue becomes much more entrenched when text must be translated. For example:

> *"are you translating in a way that really reflects the true meaning of those particular citizens, or social movements, posting that?[...] And that intersects with the second challenge, finding the keywords to find those materials traceable. I can give you an example, the term or the keyword 'Arab Spring'. Arab Spring has been standard in Western and English language bibliographies and library searches and even keywords in general and so on. But it is a colonial term that has been externally imposed on the people in the region, and people in the region never use the 'Arab Spring' as a term" (Interviewee 2)*

The **additional resources needed for translation, in terms of time, money and collaborators**, is seen as a significant hurdle. For survey respondents, the difficulty finding collaborators for certain languages is more likely to be a concern compared to the quality of machine translations (33% compared to 19% give it as a major reason for which languages are chosen). In the case of collaborators, interviewees explain that there can be greater difficulty for some languages than others, when knowledge of the topic is needed in order to code or when funding is limited.

> *"It's extremely easy to find people who will code specific languages, or content in specific languages. It's much more difficult to find multiple people in some languages in comparison." (Interviewee 17)*

3.4 Inequalities

3.4.1 Gender differences

Interviewees revealed certain **gendered barriers in the experiences of CPPT researchers**. Firstly, the time investment required can exacerbate gender differences in producing and publishing research, given that male researchers are less likely to experience care-taking responsibilities – a pattern that has only increased in the aftermath of Covid-19 (Minello et al. 2021):

> *"I have to admit that one of the reasons why I kind of sometimes seem productive is because I can dedicate evenings and weekends to some extent" (Interviewee 5)*

Secondly, in the case of fieldwork, administrations can have gendered (and racialised) risk perceptions when approving fieldwork applications, thereby creating more work for a female researcher going to a non-Western region than may be necessary:

> *"they classified my field request as high risk. And then I asked them: 'Tell me what makes you think this is high risk?' And they said: 'Oh, you know, you're taking a boat and then you are there in the middle of the jungle, exposed to so many risks' and I say 'well. Do you know what happened with*

[6] UTF 8 is an encoding system (needed for computers to store letters and symbols in text) that is able to produce a unique code for any symbol, in any language.

women in [European capital city] streets at night?'" (Interviewee 15)

Thirdly, women and minority groups are underrepresented within decision-making roles in academia, with one interviewee highlighting the differences in prevalence of female deans and university presidents within the EU. Though these barriers are not unique to CPPT research, they remain issues in this field also.

3.4.2 Collaboration opportunities

Unequal access to networks of potential CPPT collaborators was experienced by interviewees for many different reasons. Interviewees who were in departments with greater networking between disciplines – for example, linguistics and computer science – discuss benefitting from this collaboration in terms of what methods they were able to use in CPPT research. But access to data is also impacted by networking, according to interviewees. The type of networks CPPT researchers need range from contacts within communities of interest for those engaging in field work, to collaborators at universities where ethical approval can be gained in a more appropriate way for CPPT research.

A pattern in interviews was that more junior people were less likely to collaborate with others in their CPPT research, compared to respondents who were at the professor level – a few interviewees with established careers did indicate a tendency to work with others around the same career stage. However, CPPT often requires skillsets from different disciplines, multiple languages and methodological experience, not to mention that the often time consuming or expensive nature of the research can be alleviated when collaborating. Moreover, some interviewees experienced being outsiders to international networks because of the country in which their institution is located, or a lack of openness when moving to Europe:

> *"Sometimes I think if I would have had this network, maybe the problems that I faced in gathering data during my PhD would not have been so pronounced, because I was not able to gather a whole lot of data back then. But yeah, you learn as you go on. So with networks – again, I'm a migrant. When migrants are new to a place, they don't know their ways around things" (Interviewee 18)*

> *"opening science, decentralising science, is also about establishing connections with our peers in other places" (Interviewee 15)*

3.4.3 English bias

A bias in favour of English language is common throughout many aspects of CPPT research. Firstly, because non-English speaking countries are seen as case studies while English-speaking countries are seen as reflecting the world – an experience shared by interviewees studying countries from Norway to Spain, to Germany to Poland, but even more so when studying non-European contexts.

Secondly, it is seen as more prestigious to publish in English-language journals. Regardless of prestige, it is simply harder to find and cite relevant research for CPPT topics that are published in languages one does not speak.

> *"it's kind of a bet. 'Oh this I would save to publish in English', and then the chances of being refused by the publisher, by the journal, are higher." (Interviewee 11)*

> *"I do believe that it's better to publish in English. Because I've been to many conferences abroad where – especially I think for linguists, it's important to have comparative linguists. It's important to have access to the data about other languages that's conducted in the native language, but described in the English language." (Interviewee 7)*

Thirdly, non-native English speakers experience that their ability to work with CPPT materials in English can be questioned – both in international conferences and journals – in a way that is not commonplace for English-speaking researchers working with non-English material. Although not specific to CPPT research, interviewees also express frustration with the cost of paying for a language edit when they submit English articles, in particular when this seems to be for minor language issues.

3.4.4 Methods bias

In a field with such variety in methods that are used, many interviewees expressed a sense that **quantitative methods are perceived as more valuable or rigorous than qualitative**. Despite this, even those interviewees at the forefront of developing complex computational methods argued for the necessity of qualitatively analysing text in order to ensure that any big data analysis is not 'missing the mark' in interpreting

CPPT.

Relatedly, **given the diversity of disciplines engaging in CPPT research, interviewees expressed concern that methodological choices of peers are not easy to understand**. Although disciplines attempt to speak to each other, there is some scepticism of, e.g., topic models from CPPT scholars in one corner, and of correspondence analyses from another corner. One interviewee engaging in discourse analysis, a CPPT method common to other disciplines than their own, felt fear of not being accepted when disseminating their research. Other times, interviewees using methods from other disciplines than their own instead felt insecure about their ability to apply the methods.

3.4.5 Structural support

A final inequality relates to the **vastly differing levels of structural support for guidance in CPPT research**. As previously discussed, a large part of the problem is that some countries have well established institutions (that researchers are aware of) to turn to when researchers require support[7], or instead have a culture wherein it is common for individual universities to devote resources towards keeping up ethical guidelines for different types of CPPT research. In other countries, such practices are less common, there is a lack of national-level institutions to turn to, the institutions are not well known to researchers, or they are not accessible when researchers have ethical dilemmas to discuss. As the field grows and guidelines become stricter, inequalities may grow:

> *"to me it seems that it is fairly easy to incorporate and to like amplify inequalities between places that have resources, and places that do not have resources. Because if you want to be GDPR compliant, there will be more steps to carry out. If there is institutional support, mostly legal support [laughs], then that will be done in a much better way. But not all institutions can provide that, [...] and this obviously might lead to some people not embarking on specific research questions, because there are six more steps that need to be carried out." (Interviewee 17)*

Moreover, universities in certain contexts have greater funding available for researchers in social sciences and humanities, minimising those resource constraints that CPPT research often has, such as translation or data access costs. Finally, PhD students and early-career researchers were more likely to mention taking on additional sources of incomes during the interviews, as a way to support themselves and their research costs.

4 Summary and Outlook

To summarise the results, both survey and interview data suggest that access to CPPT data and following ethical procedures are the most challenging aspects of the research. Not all challenges may have been identified via the survey and interviews, but **the following patterns appear dominant: 1) CPPT researchers sometimes avoid research questions they are interested in, because access or ethics are perceived as too difficult to navigate; 2) additional resources are needed for certain types of CPPT projects; 3) inequalities in the ability to engage in CPPT research between institutions, regions, genders and scholars at different career levels remain**.

Jointly, the access issues mean that a large portion of CPPT studies suffer from the inability to gather (all) the data that is actually of interest for studying a certain phenomenon. As one interviewee summarises for the case of Facebook data:

> *"We're talking about Facebook, but what we really have is the public pages of Facebook. That's a tiny fragment, and that's not at all representative of what's going on on Facebook, right? [...] In a way, the main problem is that we pretend that we have studied Facebook when we haven't. But another part of the problem is that many of the things that we really would need to study are legally, technically or ethically unstudiable." (Interviewee 5)*

While ethical issues are of concern for both survey respondents and interviewees, several interviewees expressed a sense that greater consensus in research ethics of CPPT is forthcoming due to greater attention to the issues. Indeed, a 2022 special issue in *Journal of Information, Communication and Ethics in Society* (Zimmer 2022) highlights some of the challenges brought up by interviewees, such as extra care needed when

[7] For example, Norway's thorough guide to internet research ethics is publicly available: https://www.forskningsetikk.no/en/guidelines/social-sciences-humanities-law-and-theology/a-guide-to-internet-research-ethics/

studying vulnerable groups like youth (Mackinnon 2022); the risk and difficult emotional journey that researchers themselves may face when collecting certain CPPT data (Eneman 2022); potential exploitation of crowd-work which may be used in some CPPT research (Xia 2022); and taking stock of the current situation relating to how ethics in internet research is taught (Reeve et al. 2022). A challenge going forward is therefore to provide regularly updated information that is available publicly, that takes into account ethical standards of specific disciplines and platforms data comes from.

Our data suggest four concrete areas wherein the OPTED infrastructure could benefit CPPT researchers, and the institutions aiming to support CPPT research:

1. **Introduction to tools for accessing and analysing citizen text from social media**

OPTED's WP3 (see D3.3) has already created a platform where researchers can learn about tools for analysing media texts. Many of these tools can be used for CPPT research, and additional CPPT specific tools could also be inventoried. CPPT researchers who currently rely on tools used by their close colleagues or learned about via word-of-mouth, would benefit greatly from a closely monitored and updated list of tools and their possible usages. Likewise, departments may benefit from this resource to support in particular junior researchers who are unsure of what tools are available for CPPT research.

2. **Regularly updated ethical guidelines for different types of CPPT research, and a forum to discuss ethical best practices**

The lack of institutional knowledge, and individual researchers' uncertainty, about ethical use of CPPT research is made more complicated by the fact that different social media platforms, different countries and different disciplines do not share the same guidelines. Researchers and ethical review board members could greatly benefit from the OPTED infrastructure providing information about existing guidelines that are as specific as possible, and pointing to those organisations that have additional information such as the AoIR[8]. Moreover, researchers would benefit greatly from having a forum to discuss best practices relating to the context-specific applications of ethical guidelines when engaging in cutting-edge CPPT research, for example studying newly created social media platforms.

3. **Awareness of others researching the same language or region**

As a research hub, OPTED could contribute to closing the unequal access to networking experienced within CPPT research communities. As one interviewee highlighted, the ability to find others working on text in the language (or region) one is studying can reduce duplicating work (such as creation of stop-word lists) that has already been done as well as providing a network of potential future collaborators across the world.

4. **Long abstracts in English, for CPPT research published in other languages**

Much CPPT research is conducted on fundamental democratic issues that occur in one region, but mirror closely the themes or patterns discovered by research from other regions. By providing or linking to depositories of long-form abstracts in English, of research conducted in other languages, OPTED can improve the study of democracy by making it easier for researchers to learn from their peers in other parts of the world who publish in non-English languages.

A common infrastructure that provides these necessary CPPT resources, and/or provides information about where to find such resources, thereby alleviates challenges for both individual researchers and for the institutions working to support researchers. Studying CPPT amplifies the voices of citizens in political communication. By minimising inequalities in access to the resources CPPT researchers need, OPTED may contribute to advancing the study of citizens' voices in the realm of politics.

[8] See: https://aoir.org/ethics/

References

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, 16(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Boulianne, S., & Hoffmann, C.P. (2021, September). *Platforms for politics: A comparison of six social media platforms*. Paper accepted for presentation at the American Political Science Annual Meeting (Online)

Burgess, J., & Bruns, A. (2015). Easy data, hard data: The politics and pragmatics of Twitter research after the computational turn. In G. Langlois, J. Redden, & G. Elmer (Eds.), *Compromised data: From social media to Big data* (pp. 93–111). Bloomsbury Publishing.

Eneman, M. (2022). Ethical dilemmas when conducting sensitive research: Interviewing offenders convicted of child pornography. *Journal of Information, Communication and Ethics in Society*, 20(3), 362–373. https://doi.org/10.1108/JICES-03-2022-0028

Gelovani, S., Kalsnes, B., Koc-Michalska, K., & Theocharis, Y. (2021). *A review of citizen-produced political text (CPPT) across time and languages: Data, tools, methodologies and theories*. Project Report for OPTED. Available at: https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D2.2.pdf

Hargittai, E. (2020). Potential Biases in Big Data: Omitted Voices on Social Media. *Social Science Computer Review*, 38(1), 10–24. https://doi.org/10.1177/0894439318788322

Larsen, R. (2022). 'Information Pressures' and the Facebook Files: Navigating Questions around Leaked Platform Data. *Digital Journalism*, 0(0), 1–13. https://doi.org/10.1080/21670811.2022.2087099

Mackinnon, K. (2022). Critical care for the early web: Ethical digital methods for archived youth data. *Journal of Information, Communication and Ethics in Society*, 20(3), 349–361. https://doi.org/10.1108/JICES-12-2021-0125

Minello, A., Martucci, S., & Manzo, L. K. C. (2021). The pandemic and the academic mothers: Present hardships and future perspectives. *European Societies*, 23(sup1), S82–S94. https://doi.org/10.1080/14616696.2020.1809690

Niekrewicz, A. (2020). Funkcja politycznych słów kluczy w memach internetowych. *Forum Lingwistyczne*, 7, 105–116. https://doi.org/10.31261/FL.2020.07.08

Özkula, S. M., Reilly, P. J., & Hayes, J. (2022). Easy data, same old platforms? A systematic review of digital activism methodologies. *Information, Communication & Society*, 0(0), 1–20. https://doi.org/10.1080/1369118X.2021.2013918

Reeve, J., Zaugg, I., & Zheng, T. (2022). Mapping data ethics curricula. *Journal of Information, Communication and Ethics in Society*, 20(3), 388–399. https://doi.org/10.1108/JICES-12-2021-0124

Waisbord, S. (2020). Mob Censorship: Online Harassment of US Journalists in Times of Digital Hate and Populism. *Digital Journalism*, 8(8), 1030–1046. https://doi.org/10.1080/21670811.2020.1818111

Xia, H. (2022). The original sin of crowd work for human subjects research. *Journal of Information, Communication and Ethics in Society*, 20(3), 374–387. https://doi.org/10.1108/JICES-12-2021-0126

Zhu, J., & Wang, C. (2021). I Know What You Mean: Information Compensation in an Authoritarian Country. *The International Journal of Press/Politics*, 26(3), 587–608. https://doi.org/10.1177/1940161220963572

Zimmer, M. (2022). Guest editorial: Introduction to AoIR 2021 papers on emerging ethical practices and platform challenges. *Journal of Information, Communication and Ethics in Society*, 20(3), 345–348. https://doi.org/10.1108/JICES-08-2022-143

Appendix

Please find the survey questionnaire in the appendix of D9.3.

Survey respondents summary table

Figures relate only to respondents who chose to give demographic information in the survey.

	All respondents (N = 295)	*CPPT non-interested (N = 73)*	*CPPT interested (N = 59)*	*CPPT users (N = 163)*
Region (N=188)				
Africa	9 (4.79%)	1 (5.88%)	1 (2.13%)	7 (5.65%)
Americas	31 (16.49%)	1 (5.88%)	5 (10.64%)	25 (20.16%)
Asia	18 (9.57%)	1 (5.88%)	3 (6.38%)	14 (11.29%)
Europe	126 (67.02%)	14 (82.35%)	37 (78.72%)	75 (60.48%)
Oceania	4 (2.13%)	0 (0.00%)	1 (2.13%)	3 (2.42%)
Gender (N=208)				
Male	120 (57.69%)	10 (52.63%)	31 (59.62%)	79 (57.66%)
Female	84 (40.38%)	8 (42.11%)	21 (40.38%)	55 (40.15%)
Neither	1 (0.48%)	0 (0.00%)	0 (0.00%)	1 (0.73%)
Prefer not to say	3 (1.44%)	1 (5.26%)	0 (0.00%)	2 (1.46%)
Academic rank (N=209)				
PhD student	41 (19.62%)	5 (26.32%)	15 (28.30%)	21 (15.33%)
Junior	37 (17.70%)	4 (21.05%)	10 (18.87%)	23 (16.79%)
Mid	73 (34.93%)	5 (26.32%)	14 (26.42%)	54 (39.42%)
Senior	50 (23.92%)	4 (21.05%)	12 (22.64%)	34 (24.82%)
Other	8 (3.83%)	1 (5.26%)	2 (3.77%)	5 (3.65%)
Main academic field (first field listed if multiple) (N=209)				
Communications	97 (46.41%)	8 (42.11%)	15 (28.30%)	74 (54.01%)
Political Science	67 (32.06%)	10 (52.63%)	31 (58.49%)	26 (18.98%)
Psychology	5 (2.39%)	0 (0.00%)	0 (0.00%)	5 (3.65%)
Sociology	15 (7.18%)	0 (0.00%)	2 (3.77%)	13 (9.49%)
Other	19 (9.09%)	0 (0.00%)	4 (7.55%)	15 (10.95%)

Interview questionnaire and summary table

The following key questions were asked to all interviewees (sometimes phrased slightly differently depending on their expertise). Additional questions were asked to subsets of interviewees (e.g. questions about teaching only asked to teachers), or specific questions based on individual interviewees' expertise. Moreover, background questions were asked about the interviewee themselves as well as the type of CPPT research they engage with.

- Do you collaborate with others in your CPPT research, or work alone on this?
- What issues have you come across, when trying to access CPPT data?
- Could you describe any difficulties you have come across, when researching text from a hidden,

vulnerable or hard to reach population?

- Have you experienced text that you were planning or in the process of using for research being removed or censored from the platform it was on?
- In what ways have changing allowances of social media companies impacted your past, present or planned future research?
- How have you been impacted by GDPR (the EU's General Data Protection Regulation) and other changing requirements relating to using private citizen data in research?
- In your experience, how are ethical guidelines for CPPT research interpreted differently by different actors?
- What support is needed, for yourself and other researchers, to be aware of all the guidelines in place for data protection, consent and confidentiality for the type of CPPT research you do?
- Have there been times when you faced barriers due to the language you study or wanted to study?
- What would you say are the biggest challenges specifically for research involving text produced by citizens, which may not be a problem for research involving text produced by other actors (such as public figures, organisations, governments)?

The following table shows demographic characteristics of the interviewees.

Characteristic	*Breakdown*
Gender	13 women, 8 men
Primary discipline	5 communication/journalism, 1 computer science, 2 international relations, 1 law, 2 linguistics, 8 political science (2 interdisciplinary with no clear main discipline)
Seniority	8 junior (<5 years post PhD), 10 mid (5-15 years post PhD), 3 senior (>15 years post PhD)
Region of origin	5 outside Europe (Asia, MENA, South America); 3 Northern Europe, 3 Eastern Europe; 4 Southern Europe; 6 Western Europe
Region of current affiliation	4 outside Europe (MENA, North America, South America); 5 Northern Europe; 2 Eastern Europe; 3 Southern Europe; 7 Western Europe