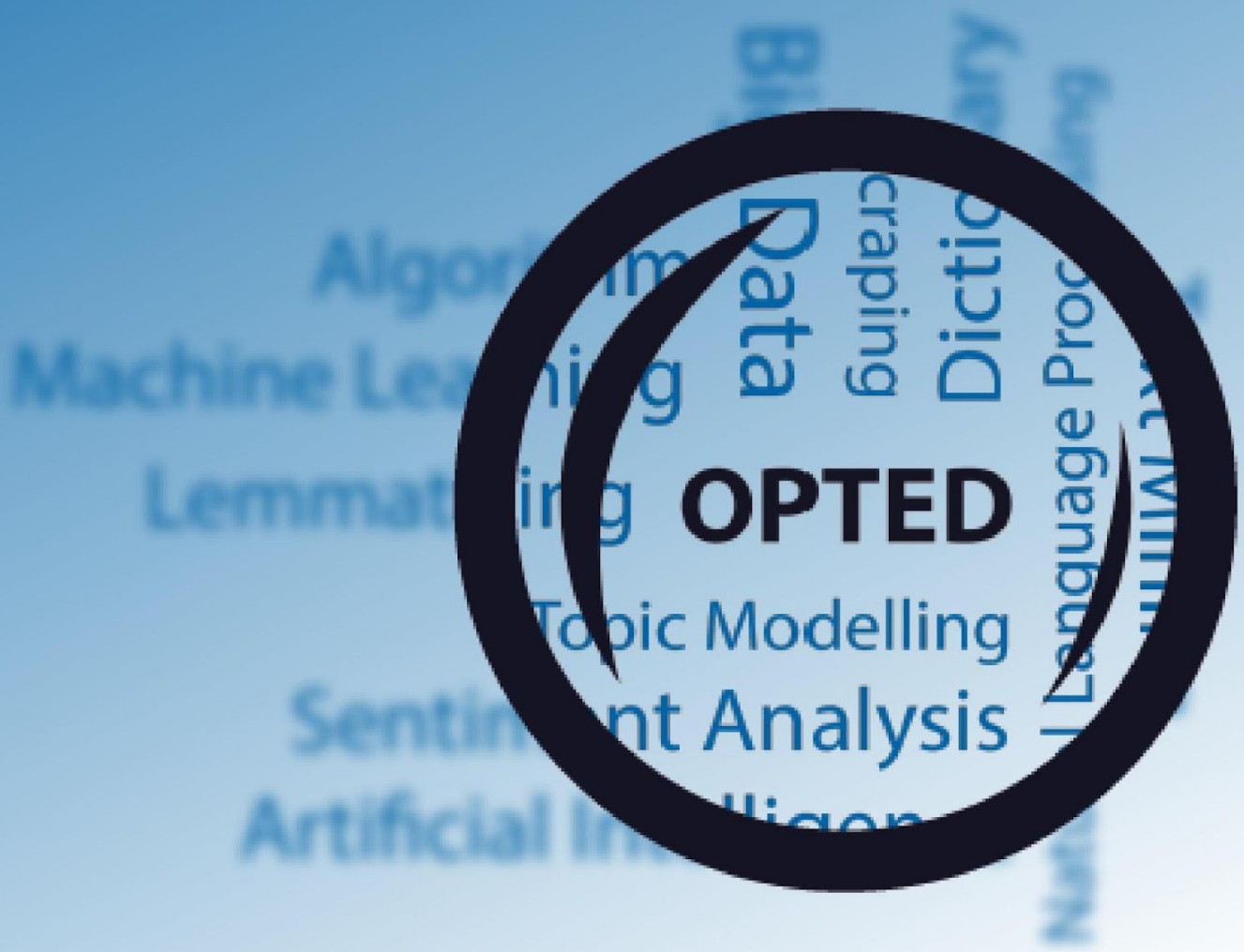
OPTED

Update of the CPPT inventory

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



Disclaimer

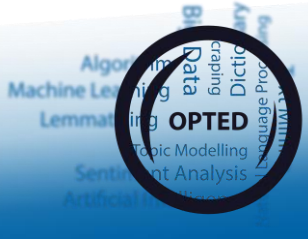
This project has received funding from the European Union's Horizon 2020 Research & Innovation Action under Grant Agreement no. 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report



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

Update of the CPPT inventory

Deliverable D2.4

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: March 2023

1 Executive Summary

In our first report, D2.1, WP2 created an inventory of research using citizen produced political text (CPPT) in academic fields relating to political communication. The inventory provides a database of the rich field of research using citizens' own voices as a way to understand political questions within democracies and beyond. In the inventory, research from 2014-2020 was included, giving insight into how citizens' political text is researched. In D2.2, selected analyses from the inventory provided a summary of themes and patterns in CPPT research. In particular, regions of the world where CPPT originates, the methods of attaining CPPT data and the source of data employed by researchers. Beyond this, the analytical methods of CPPT scholars and the languages of the text were also summarised.

To update the inventory, research from 2021-2022 is now added. The same methodological approach is employed, however with the limitation that only English-language publications were added (as explained below).

In this report, we explain the focus and limitations of the update to the inventory, and briefly remind the reader about our data collection procedure which can be found in full in D2.1. We then make some descriptive presentation of the data and compare findings from 2021-2022 with the original findings of the 2014-2020 CPPT research.

2 Updating the CPPT inventory

2.1 Selection criteria for CPPT research in 2021-2022

The selection criteria for the update was largely the same as for the creation of the inventory: target literature was peer-reviewed journal articles which use CPPT data. However, instead of collecting scholarship written in different European languages, the update only includes English-language journal articles. This choice is caused by three main factors. Firstly, the financial constraint given that the inventory requires detailed coding by coders. Secondly and relatedly, the time constraint to not only scrape and clean the results of our literature search but for the coders to code. Thirdly, the publication cycle and online accessibility of some non-English language journals where 'online first' options do not exist. Moreover, the majority of CPPT literature in 2014-2020 was written in English – the original database scraped publications in eight additional languages (French, German, Italian, Portuguese, Spanish, Polish, Norwegian, Swedish), yet 2,109 of the 3,260 CPPT publications identified for the original inventory were written in English. Although these 1,151 non-English language publications provide valuable information, the resource constraints made focusing on English language publications more feasible for the inventory update. In this report, we use only the English language publications from the original inventory when providing figures and data for comparison purposes.

2.2 Brief summary of coding procedure

The procedure for updating the inventory followed the original procedure used for creating the inventory. Please see the report D2.1 for a more detailed description of the abbreviated method description herein. The primary difference in method is that only English language articles were included in the update of the inventory.

Using the software Publish or Perish[1], 37 search terms identifying CPPT literature were used to scrape publications on Google Scholar that include these search terms. The search terms include for example: "comments section" political; instagram "political text"; user-generated AND "political text" OR "political comments". We also exchanged the word 'political' in relevant search terms with 'civil society' and 'democracy' to get an idea of how search terms including substantive, specific topic areas might have impacted the search results. Comparing search results, we found a great deal of overlap. E.g. the search term '"facebook comments" political' retrieved 500+ results, while the term '"facebook comments" civil society' retrieved 222 and '"facebook comments" democracy' 260 results, wherein 69 and 43 search results were unique to the 'civil society' and 'democracy' search terms respectively. We therefore accept that our inventory may exclude some relevant CPPT literature that does not explicitly mention politics. However, we choose not to include search

[1] Harzing, A.W. (2007). *Publish or Perish*, available from https://harzing.com/resources/publish-or-perish

terms relating to specific substantive areas of scholarship, given that this would bias our inventory to one field of research over another - we could not feasibly scrape an exhaustive list of all substantive areas that CPPT researchers are interested in. By only including the word 'political', we improve the chances of search results actually relating to CPPT rather than non-political topics, without skewing our results in favour of literature in one subfield over another.

Up to 500 results for each search term were scraped, leading to a total of 20,457 publications – compared to 16,595 for the time period 2014-2020 of the original inventory. Each search result from Publish or Perish includes information about authors, number of citations and more, which are included in the dataset.

The search results were then cleaned: Only peer-reviewed journal articles were kept, meaning a large number of results were removed due to being pre-print, books or chapters, repository entries, blog posts or similar. Duplicates were removed, and a careful cleaning of the scraped publications removed those publications clearly not relevant for CPPT or published in journals completely unrelated to the academic fields relating to CPPT.

This cleaning procedure led to 1,900 likely relevant publications being part of the dataset that was then coded, compared to 6,040 for the original dataset from 2014-2020. Clearly, a much larger share of scraped articles were identified as not relevant before being sent to coders in the update. One possible explanation is that more fields outside of those tangential to political communication are using social media (text) data. Examples of journals that were not considered directly relevant for CPPT (and therefore any search result from these journals was removed during this cleaning stage) are Journal of Medical Internet Research and European Journal of International Law. The fact that so many journals from fields not relating to politics or communication (or relevant fields) appeared in our search results may indicate the growing interest from other academic disciplines to engage with social media data in various ways, and therefore the potential for even greater cross-disciplinarity in future.

Another reason for a larger number of scraped search results being coded as not relevant could relate to covid-19, given the understandable reliance on social media data (the primary source of CPPT data) in covid-19 related studies. However, the majority of search results relating to covid-19 were considered irrelevant to CPPT and therefore did not get included in the final data sent to coders, for example because the journals (e.g. The Gerontologist or Health Education Research) were not directly related to CPPT or because it is not text data but some other social media metadata being analysed. In the final data set of 1,900 articles, 146 articles included the word 'covid' in the title, and 14 'corona'. Therefore, about 8% of the updated database is primarily investigating covid-19 related questions. On the one hand it is quite interesting to see such an impressive publication speed concerning an important exogenous event, however, on the other hand 8% of the thematic clustering suggests that deviations in trends between the original inventory and the update are only minimally driven by covid-19 related scholarship.

Next, the dataset was sent to coders for detailed coding. The codebook was the same in both coding years (i.e., for creating the original inventory and the update) and it allowed for additional variables to be added to the dataset which relate to the data used in each publication, alongside those variables provided by Publish or Perish which relate to the publication and the authors (e.g. number of citations for the article, all co-authors' names). The variables coded include how much text data is being analysed, the language and origin of the text being analysed, the methods used for collecting and analysing text. The coders were the same individuals who were used to code articles for the original CPPT inventory. An inter-coder reliability test was conducted for the four coders used to conduct the coding for the update, resulting in a Krippendorf's alpha of 0.61. This score is somewhat low, which can be attributed to the fact that much of the coding uses open-response answers. Although they had already been trained and were experienced in the specific coding task, coders were provided with another set of written training materials and an exercise to complete before starting to code again. Coders could also mark publications in the dataset as irrelevant.

The final dataset consists of 684 publications for 2021-2022 – compared to 2,109 English-language articles in the original inventory for 2014-2020.

A comparison of why articles were coded as not relevant by the coders, in the original and in the update of the data, can be seen in Table 1. In this table, as well as this report in general, only English-language publications from the original are included in analysis. The main change is that a larger share of scraped publications were marked by coders as not analysing CPPT data (i.e., analysing data that is not political text produced by citizens), while a smaller share of scraped publications were marked as irrelevant due to not analysing any data or due to analysing data that is not relevant to the political domain.

Table 1 Reasons for irrelevance coding

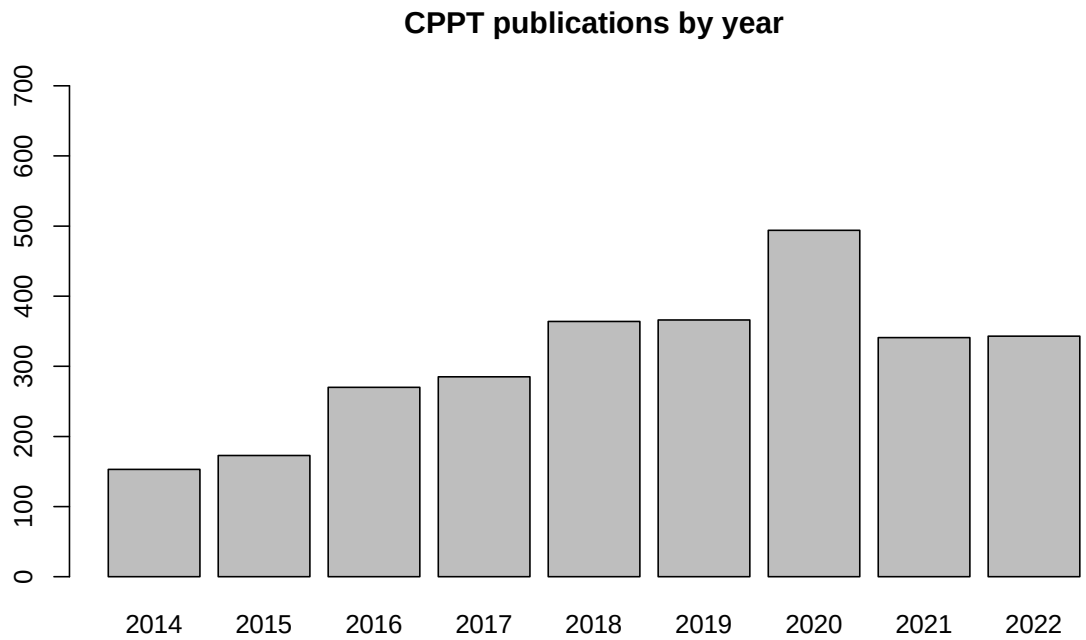
Reason for irrelevance	Original (out of 6,040 articles)		Update (out of 1,900 articles)		Change
Data analysed not CPPT	2209	56.2%	895	73.5%	17.3%
No data analysed	695	17.7%	130	10.7%	-7.0%
Data analysed not political	634	16.1%	86	7.1%	-9.1%
No access	220	5.6%	86	7.1%	1.5%
Book	66	1.7%	9	0.7%	-0.9%
Not written in English	51	1.3%	7	0.6%	-0.7%
Thesis of dissertation	24	0.6%	0	0.0%	-0.6%
Not academic	18	0.5%	3	0.2%	-0.2%
News article	10	0.3%	2	0.2%	-0.1%
Only published on a university website	3	0.1%	0	0.0%	-0.1%
Patent	1	0.0%	0	0.0%	0.0%

3 Data presentation

In this data presentation, key tables and figures from D2.2, using the English-speaking subset of the original data (for comparison purposes) of CPPT publications from 2014-2020, are compared with data from the update (CPPT publications from 2021-2022). The data presentation therefore aims to highlight the ways in which trends from 2014-2020 continue or potentially are disrupted in 2021-2022.

3.1 Volume of CPPT use in research

Figure 1 Number of publications by year



In 2014-2020, a clear trend showed an increase in CPPT publications every year. Considering the number of published CPPT studies by year, this trend of an increasing volume appears to taper off in 2021 and 2022, as shown in Figure 1. Despite the increased technological possibilities, for both CPPT access and analysis, there is some indication that less CPPT-based scholarship was published in 2021 and 2022 compared to previous years, perhaps due to the growing challenges faced by CPPT researchers (see D2.3). In particular,

covid-19 may have decreased the academic productivity of many researchers or slowed journals' publication speeds[2]. Another possibility is that our search terms are not catching keywords relevant to the very latest interests of CPPT researchers. As can be seen in the figure, it is also possible that 2020 was an outlier year in terms of CPPT publications, and that we can expect a slower, gradual increase in CPPT publications by year going forward.

3.2 Region of origin of CPPT data

Table 2 shows the regions being studied in CPPT publications, where multiple regions could be coded for a single publication if it studied multiple regions. In the original inventory, Europe and North America accounted for the region of study in about two thirds of studies. 2021-2022 saw somewhat fewer European and North American based studies on citizens' political communication, but overall very little variation in region of interest.

Table 2 Region of origin of CPPT data

Region	Original (2014-2020)		Update (2021-2022)		Change
North America	702	33.3%	203	29.7%	-3.7%
Europe	658	31.3%	189	27.6%	-3.6%
Asia	447	21.2%	135	19.7%	-1.5%
Middle East and North Africa	132	6.3%	44	6.4%	0.2%
Does not specify	121	5.7%	74	10.8%	5.1%
Sub Saharan Africa	111	5.3%	44	6.4%	1.2%
Australia and Oceania	96	4.6%	29	4.2%	-0.3%
South America	53	2.5%	23	3.4%	0.8%
Central America	13	0.6%	7	1.0%	0.4%

3.3 Languages of CPPT text studied

In D2.2, the top 10 languages of text studied were identified, and Table 3 shows changes in the popularity of these languages in the original inventory and the update. Publications studying text in multiple languages could be coded for each language studied. Somewhat fewer studies use English text (a decrease from 65% to 58%), with other languages remaining at about the same level of popularity. Still, the small decrease in over-representing English text in CPPT research hints at a potentially more diverse and globally representative scholarship.

Table 3 Top 10 languages studied

Language	Original (2014-2020)		Update (2021-2022)		Change
English	1371	65.1%	398	58.2%	-6.9%
Chinese	116	5.5%	53	7.7%	2.2%
German	105	5.0%	38	5.6%	0.6%
Spanish	86	4.1%	24	3.5%	-0.6%
Arabic	69	3.3%	19	2.8%	-0.5%
French	69	3.3%	17	2.5%	-0.8%
Italian	50	2.4%	16	2.3%	0.0%
Russian	50	2.4%	16	2.3%	0.0%
Korean	46	2.2%	9	1.3%	-0.9%
Swedish	39	1.9%	13	1.9%	0.0%

[2] See, for example: https://gdc.unicef.org/resource/effect-covid-19-pandemic-academic-productivity

3.4 Source of CPPT data

The top 10 sources of CPPT data in the original dataset have seen only small changes in popularity in 2021-2022. Table 4 shows how the top 10 sources in the original dataset have changed in the update. The most notable change is that blogs appear less popular among CPPT scholars in the last two years, suggesting perhaps that CPPT scholars are finding blogs to be less interesting to focus on. Newer social media platforms are not competing with the mammoths of Facebook and Twitter in CPPT studies from 2021-2022 either – with only 34 studies using Reddit, 34 using posts or comments from Instagram, 22 from Weibo, 21 from TikTok, and less than 10 each for Telegram, 4chan, Whatsapp and Twitch. This may partially be accounted for by the difficulty for individual researchers to come across research guidelines or navigate terms of service of platforms that are newer (see D2.3). Moreover, by the systematic data access challenges when platforms change or restrict previously accessible data – which could in fact be a growing issue, in particular as Twitter's previously free API becomes more restrictive for academic use[3].

Table 4 Top 10 sources of CPPT data

Source	Original (2014-2020)		Update (2021-2022)		Change
Facebook posts	448	21.3%	155	22.7%	1.4%
Facebook comments	446	21.2%	173	25.3%	4.1%
Original tweets	417	19.8%	166	24.3%	4.5%
Newspapers online	383	18.2%	86	12.6%	-5.6%
Forums	274	13.0%	80	11.7%	-1.3%
Retweets or replies	252	12.0%	94	13.7%	1.8%
Blogs	226	10.7%	26	3.8%	-6.9%
YouTube comments	100	4.8%	39	5.7%	1.0%
YouTube original videos	95	4.5%	14	2.0%	-2.5%
Political/deliberation websites	67	3.2%	13	1.9%	-1.3%

3.5 Method of collecting CPPT data

How did CPPT researchers collect data in 2021-2022? The most popular methods in the original dataset, shown in Table 5, have undergone considerable change in the update to the inventory. In particular, access via a company or via an API has become significantly more likely. The category 'company bought/API access' combining a typically free collection method (although Twitter's API access is, as mentioned above, changing) with paying companies for access makes it unclear how much the change can be attributed to better awareness of APIs as a data collection method across researchers, as opposed to other, more costly means of access through companies. Generally, it seems computational methods for accessing CPPT data may have increased, given that dictionaries/keyword searches and self-written programming to access text have also increased in popularity. In comparison, the more qualitative-research focused method of data collection via interviews has decreased in popularity.

Table 5 Collection methods

Collection method	Original (2014-2020)		Update (2021-2022)		Change
Self-copy-paste	861	40.9%	268	39.2%	-1.7%
Company bought/API access	363	17.2%	311	45.5%	28.2%
Dictionaries/keyword searches	310	14.7%	219	32.0%	17.3%
Interviews	298	14.2%	33	4.8%	-9.3%
Self-written program	190	9.0%	82	12.0%	3.0%

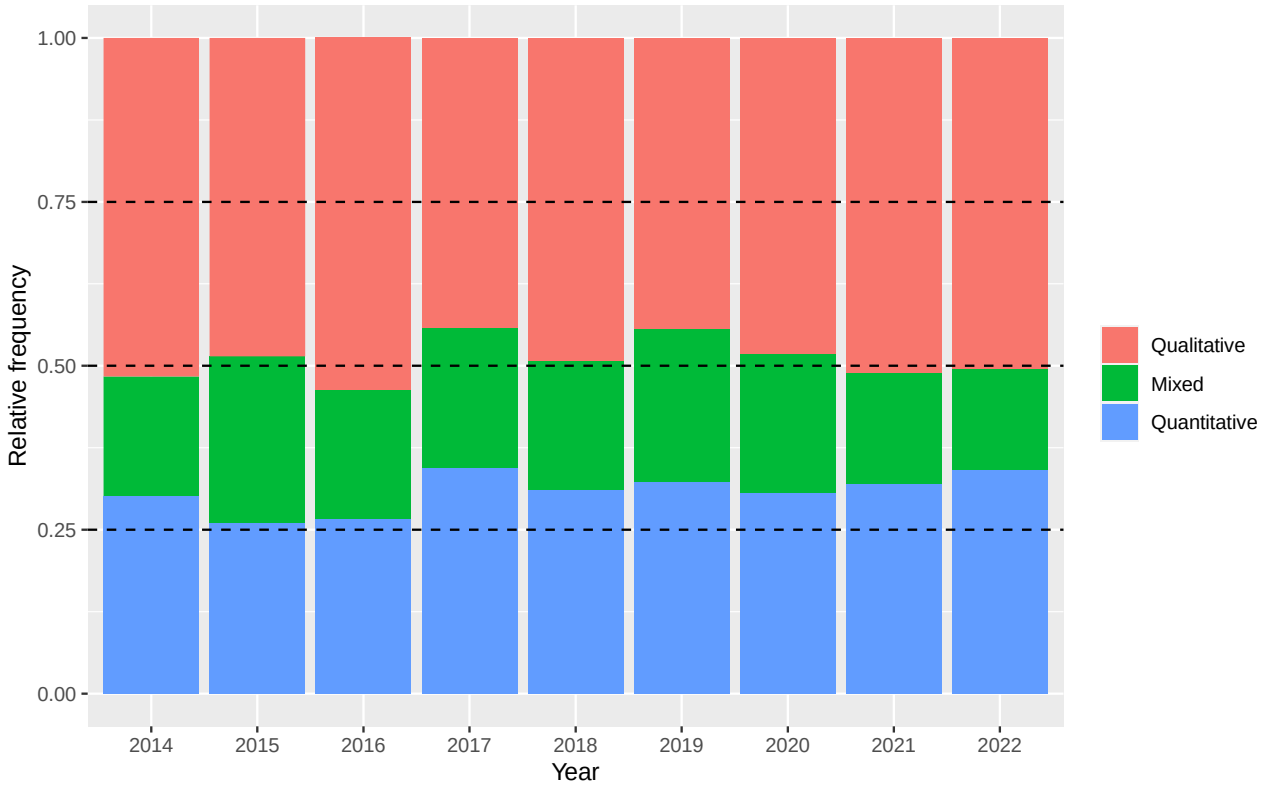
[3] See https://www.cip.uw.edu/2023/02/02/twitters-api-access-changes-academic-research/

Related to data collection method, is the practice of making data available to other researchers for either replication studies or the possibility of using the data to address new questions. In the original inventory, 39 out of the 2,109 English-language publications provided a URL link to the data they used. In the update, 30 out of the 684 publications provided a data link. While only around 4% of CPPT authors in the update to the inventory therefore transparently provide a link to their data, it is encouraging that the share of authors doing so has increased from less than 2% in the original inventory. In our next deliverable, we will examine the ethical and legal difficulties of sharing CPPT data.

3.6 Analysis methods in CPPT research

As noted in D2.2, about half of all CPPT publications in the original dataset used qualitative methods of analysis. By year, Figure 2 shows the share of CPPT studies using mixed methods, qualitative methods or quantitative methods. In 2021-2022, the primary change in method of analysis is a slight decrease in use of mixed methods, and slight increase in use of quantitative (including computational) methods. Qualitative methods continue to account for the methods used in about half of all studies. Perhaps the decrease in collaboration opportunities caused by covid-19 is partially responsible for decreasing mixed methods studies, given that these may be more likely to rely on joint efforts by researchers across academic fields that use different methods.

Figure 2 Analysis type by year



Considering specific types of methods, D2.2 identified the top 10 quantitative and top 10 qualitative methods of analysis. In the updated dataset, there is little change in popularity for methodological choices of CPPT researchers, as seen in Table 6. The main changes of note is an increase in quantitative manual coding, and decrease in qualitative interview as method. Given that interviews were also less likely to be noted as a method of data collection, it is hardly surprising that interviews are less likely to be analysed in CPPT publications. Once more, the impact of covid-19 on research possibilities may have contributed to the decrease in interviews as a method: the additional safety precautions needed for in-person interviews in 2020 and after may have delayed projects. Additionally, interviews can be a resource intensive method given the time and potential travel needed, and therefore it is possible researchers choose other methods when funding becomes more competitive. Overall, there is further support in Table 6 for a small increase in CPPT researchers' tendency to rely on quantitative or computational methods, despite the larger share of studies still using qualitative methods of analysis.

Table 6 Top 10 quantitative and qualitative analysis methods

Quantitative method	**Original (2014-2020)**		**Update (2021-2022)**		**Change**
Quantitative content analysis	518	24.6%	170	24.9%	0.2%
Text statistics	472	22.4%	119	17.4%	-5.0%
Sentiment scoring	200	9.5%	80	11.7%	2.2%
Hand coding	191	9.1%	148	21.6%	12.6%
Dictionaries keyword searches	113	5.4%	67	9.8%	4.4%
Topic models or text clustering tools	89	4.2%	57	8.3%	4.1%
Supervised machine learning	81	3.8%	44	6.4%	2.6%
Natural language processing tools	79	3.8%	53	7.7%	4.0%
Semantic network tools	72	3.4%	21	3.1%	-0.4%
Automated extraction	63	3.0%	14	2.0%	-0.9%
Qualitative method	**Original (2014-2020)**		**Update (2021-2022)**		**Change**
Qualitative content analysis	671	31.9%	228	33.3%	1.5%
Discourse analysis	482	22.9%	202	29.5%	6.6%
Thematic qualitative text analysis	369	17.5%	127	18.6%	1.0%
Interview	292	13.9%	33	4.8%	-9.0%
Observation	215	10.2%	37	5.4%	-4.8%
Evaluative qualitative text analysis	102	4.8%	5	0.7%	-4.1%
Grounded theory	73	3.5%	13	1.9%	-1.6%
Focus group	42	2.0%	5	0.7%	-1.3%
Type building text analysis	41	1.9%	2	0.3%	-1.7%
Survey	13	0.6%	0	0.0%	-0.6%

3.7 Journals publishing CPPT research

Finally, the top 10 journals publishing CPPT in 2014-2020 are compared with the top 10 journals in 2021-2022 in Table 7. As can be seen, the journals' likelihood of publishing CPPT in the original inventory and the update is relatively stable.

Table 7 Top 10 CPPT journals

	Original (2014-2020)	**Update (2021-2022)**
1	New Media & Society	New Media & Society
2	International Journal of Communication	Information, Communication & Society
3	Information, Communication & Society	International Journal of Communication
4	Social Media+ Society	Social Media + Society
5	Computers in Human Behavior	Social Science Computer Review
6	Social Science Computer Review	Journalism
7	Journalism	Discourse &Society
8	Journal of Information Technology & Politics	PloS one
9	Telematics and Informatics	Media and Communication
10	Discourse, Context & Media	Social Network Analysis and Mining

4 Conclusion

Adding two years of CPPT scholarship to the inventory of publications has shown some changes in trends, while also providing a richer and larger dataset to be referenced by interested parties. Firstly, there is some indication that the number of published articles (in English) has decreased in the last two years. However, this could be due to the search terms used to identify CPPT literature in this project. Although the search terms include many different social media platform names and text types (e.g. comments, tweets, posts), it is possible that some shifts in CPPT researchers' interests are not accounted for and therefore certain publications not identified or scraped. It could also be caused by covid-19 impacted academic productivity and slowing publication cycles, or the increased difficulty and uncertainty surrounding data access from platforms in the past years - despite the Digital Services Act which is intended to protect platform users and force greater transparency on the part of platforms[4].

Secondly, publications in 2021 and 2022 were less likely to use English language text to analyse, potentially signifying researchers' interest in studying groups of people speaking other, less well studied languages. In connection with the small decrease in the number of studies using Europe or North America as region of study, there is some indication that more global diversity is present in CPPT scholarship.

Thirdly, quantitative/computational methods for accessing CPPT data seem to be on the rise, with qualitative or manual techniques on the decline. This change speaks of an increased interest on the part of CPPT scholars to experiment with more advanced, potentially lower-resource, means of accessing data compared to e.g. interviews. Moreover, CPPT research is more and more using digital data (in particular from social media platforms), rather than more analogue, traditional data types (such as interviews or observational data) where computational methods may be less likely to be appropriate. However, there is still about half of CPPT research being published that uses only qualitative methods of analysis, and therefore the perspectives of qualitative researchers must not be overlooked.

Despite these interesting changes, overall there is stability in the CPPT scholarship considering sources of data, which continue to be predominantly social media based texts, and methods used for analysing CPPT data. The dataset, now covering the years 2014-2022, should prove a useful guide for researchers or others who hope to navigate CPPT scholarship.

[4] See: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package