The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research

Peeters, S.; Hagen, S.

DOI
10.2139/ssrn.3914892
10.5117/CCR2022.2.007.HAGE

Publication date
2022

Document Version
Final published version

Published in
Computational Communication Research

Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 4.2 (2022) 571-589
HTTPS://DOI.ORG/10.5117/CCR2022.2.007.HAGE

The 4CAT Capture and Analysis Toolkit: A Modular Tool for Transparent and Traceable Social Media Research

Stijn Peeters
Department of Media Studies, University of Amsterdam
stijn.peeters@uva.nl

Sal Hagen
Department of Media Studies, University of Amsterdam
s.h.hagen@uva.nl

Abstract

This paper introduces the 4CAT Capture and Analysis Toolkit (4CAT), an open-source Web-based research tool. 4CAT can capture data from a variety of online sources (including Twitter, Telegram, Reddit, 4chan, 8kun, BitChute, Douban and Parler) and analyze them through analytical processors. 4CAT seeks to make robust data capture and analysis available to researchers not familiar with computer programming, without 'black-boxing' the implemented research methods. Before outlining the practical use of 4CAT, we discuss three 'affordances' that inform its design: modularity, transparency, and traceability. 4CAT is modular because new data sources and analytical processors can be easily added and changed; transparent because it aims to render legible its inner workings; and traceable because of automatic and shareable documentation of intermediate analysis steps. We then show how 4CAT operationalizes these features through a description of its general setup and a short walkthrough. Finally, we discuss how 4CAT strives for an 'ethics by design' development philosophy that enables ethically sound data-driven research. 4CAT is then positioned as both an answer to and a further call for 'tool criticism' in computational social research.

Keywords: research tools, computational humanities, web scraping, 4cat, digital methods

Introduction

This paper introduces the 4CAT Capture and Analysis Toolkit (4CAT). 4CAT is an open-source[1] Web-based research toolkit designed to capture, manipulate, analyze, and visualize thread-like data from a heterogeneous set of online sources. After querying a specific data source, the created dataset can be processed further through a variety of modular, analytical processors. These processors comprise simple analyses such as frequency counts as well as more advanced operations like network visualization and Natural Language Processing (NLP). Since its inception in 2018, 4CAT has been used by hundreds of students and journalists and has been cited in multiple academic papers (see e.g. Al-Rawi, 2020; De Zeeuw et al., 2020; Jokubauskaitė & Peeters, 2020; Tuters & Hagen, 2019; Zelenkauskaite, 2021).

The aim of this text is to introduce 4CAT as a general-purpose software suite for capturing and analysing social media data. Our broader goal is to discuss some epistemological concerns within the increasingly tool-driven social sciences and humanities (SSH) and how we seek to practically address these through 4CAT's design. 4CAT and similar research tools can be seen as both a driver and an outcome of a 'computational turn' in SSH research (Berry, 2011), a concept denoting the increasing use of digitized data and quantitative software tools for socio-cultural research. This turn was at the root of the flourishing of new subfields like digital humanities (Berry, 2011; Kirschenbaum, 2012) and computational social science (Lazer et al., 2009). The study of digital culture itself has since the 90s also expanded from mostly ethnographic approaches (e.g. Baym, 2000) to include data-driven or 'Big Data' studies (boyd & Crawford, 2011), along with the introduction of 'digital methods' that study 'natively digital objects' like hyperlinks and likes (Rogers, 2013).

Methodological concerns have however been articulated regarding the growing prominence of software tools in SSH research. Since the humanities in particular have traditionally not prioritized technical skills (such as handling large quantities of data, computer programming, or data visualization) students and researchers often rely on tools that automate such tasks (Van Es. et al., 2018). Consequently, such software unavoidably co-structures research outputs. This calls for scrutiny of both the software's affordances as well as its position within a broader methodological and institutional context. In a humanities context, this also raises concerns on how to responsibly integrate computational tools and quantitative methods in traditionally qualitative research.

With this text and 4CAT itself we therefore seek to respond to the call for 'involvement' with methodological challenges emerging from quantitative

and tool-driven research, ranging from issues on objectivity, rigor, agency, black-boxing, and the varying 'epistemic cultures' of academic institutions (Knorr-Cetina, 1999; Rieder & Röhle, 2017). 4CAT allows practical engagement with such issues. In this text, after discussing the context in which 4CAT was developed, we therefore formulate three affordances that both touch on these issues and inform 4CAT's design: modularity, transparency, and traceability. We outline the tool's set-up in dialogue with these affordances. Finally, we end with a discussion on ethical and legal concerns regarding data capture. This paper is intended to describe 4CAT's core functionality and design considerations in a general sense; more detailed technical information and guides are available online via the tool's GitHub page.[2]

Context & grounding

4CAT is not the first tool to offer capture and analysis of online data. Most influential to 4CAT's design is the DMI Twitter Capture & Analysis Toolkit (DMI-TCAT; Borra & Rieder, 2014), which similarly separates the 'capture' and 'analysis' aspects of social media research, but focuses on one specific platform (Twitter) rather than offering a generic framework which supports multiple platforms. Other attempts at generalised tools include VINCA (Li et al., 2007) and Calico (Giguet & Lucas, 2013), both academic research tools that offer a generic capture and analysis framework for Web forums. Both single- and multi-platform tools have proliferated more recently, with academic (FacePager, Media Cloud, SMAT, CLARIAH Media Suite) and commercial (Dedoose, CrowdTangle, BuzzSumo) tools offering interfaces through which data from one or several social media platforms may be captured, analysed, or both.

These tools offer a wide variety of features, affordances, advantages, and disadvantages, and it is beyond the scope of this article to discuss these in detail. 4CAT seeks to add to this landscape with a particular focus on three affordances we believe are essential for sound computational socio-cultural research: modularity, transparency, and traceability. These core principles arose in dialogue with prior computational research projects and in the course of 4CAT's development. The term 'affordance' is used here in reference to Hutchby's 'communicative affordances', i.e. 'functional and relational aspects which frame, while not determining, the possibilities for agentic action in relation to an object' (2001, p. 5). This definition allows to emphasise the 'multi-directionality of agency and connectivity at work in approaching questions of affordances' (Bucher & Helmond, 2018, p. 242), in

this case touching on not only the relation between 4CAT and its users, but also its developers, the APIs it interfaces with, the data it stores, and so forth. Hutchby's definition is moreover useful to denote how these affordances are not *features* that strictly determine how the object is used. Rather, they often merely *frame* the agentic relation, or, in behavioural economic terms, they 'nudge' users towards a desired use (Thaler & Sunstein, 2011).

With this in mind, we next outline the three affordances centralised in 4CAT's design to facilitate methodological 'rigour', i.e. research practices that meet established academic standards like reproducibility, clarity, and objectivity.

Modularity

Many research tools focus either on a single input data source (e.g. Twitter data for DMI-TCAT) or a single form of data output (e.g. network visualizations for Gephi). While such focus is often a strength, it complicates broader applicability – something better afforded by a more agnostic and modular approach. As Manovich (2001) notes, modularity is one of the core principles of computer-based media. Research tools can leverage modularity by allowing smaller, incremental, or stand-alone contributions, which encourages collaboration and the reuse of research code. A modular approach additionally makes maintenance easier, as discrete parts of the software can be updated, redeveloped, or deprecated, without the need to re-work their integration with the larger whole. Moreover, since data-driven social media research is volatile, with code often suddenly becoming obsolete (e.g. when API access is terminated; see Freelon, 2018), a modular approach can more flexibly cope with this volatility, since the tool's functionality is not dependent on one access point and can be reconfigured quickly.

Transparency

Many tools and social media APIs are built for commercial purposes instead of for an academic audience. As a result, they often hide their inner workings, which potentially 'black-boxes' research. If commercial and closed tools become cornerstones of research, the academic pillar of an 'open process of scrutiny' is threatened, including the researcher's 'ability to understand the method, to see how it works, which assumptions it is built on, to reproduce it, and to criticise it' (Rieder & Röhle, 2012). However, concerns on methodological transparency go beyond the simple dichotomy of open or closed source, and this criticism can also be articulated towards academic software like 4CAT. Indeed, research tools in general are often attributed an 'inherent 'authority'' (Van Es et al. 2017, p. 172) or a 'lure of objectivity'

(Rieder & Röhle, 2012) that can camouflage 'hidden biases in both the data collection and analysis stages [that] present considerable risks' (Crawford, 2013). Availability of the tool's source code by itself does not sufficiently address this, since it does not *explain* how its outputs are in fact 'cultural entities [...] co-produced' by both the user and the software (Van Es et al., 2017, p.172). In other words, transparency should also be ensured by actively *explaining* and *rendering legible* how the data is 'cooked' (Bowker, 2013). Though some methods may be too complicated to explain with simple descriptions, tool interfaces can nevertheless make an active effort to nudge users into familiarising themselves with the underlying operations, e.g. with links to relevant external sources that explain the operationalized method.

Traceability

Closely related to transparency, methodological rigor benefits from the ability to retrospectively return to intermediate research outputs and the underlying qualitative decisions – e.g. regarding what and why certain parameters were used. This way, intermediate steps in the research process can be revisited, (peer-)reviewed, and adjusted. More abstractly, Latour has advocated for such 'traceability' since it allows 'going back and forth' within 'navigational datascapes', allowing one to move between aggregated results and individual data points without losing sight of the whole (2011, p. 804). Such circulation has the benefit of stimulating 'second-degree objectivity': objectivity that is derived from the 'multiplication of different viewpoints' (Venturini, 2011). However, as Latour et al. (2012) also acknowledge, these navigational practices require availability of compatible data and tools. As we will discuss, we aim to enable the 'possibilities of action' for such navigational practices with 4CAT.

4CAT's set-up

Having outlined the affordances that inform 4CAT's design, we now discuss the tool's concrete design. 4CAT is a Python-based application with which one can capture and process data from a variety of online sources. It consists of a back-end daemon that handles the retrieval and processing of data via a task queue, and a front-end based on the popular Flask library, which offers both a Web interface accessible with a browser, and an HTTP API. The back-end and front-end operate independently and communicate via a socket-based API and a shared PostgreSQL database. This means the back-end (which typically runs on a remote server) may be restarted or updated while the Web interface

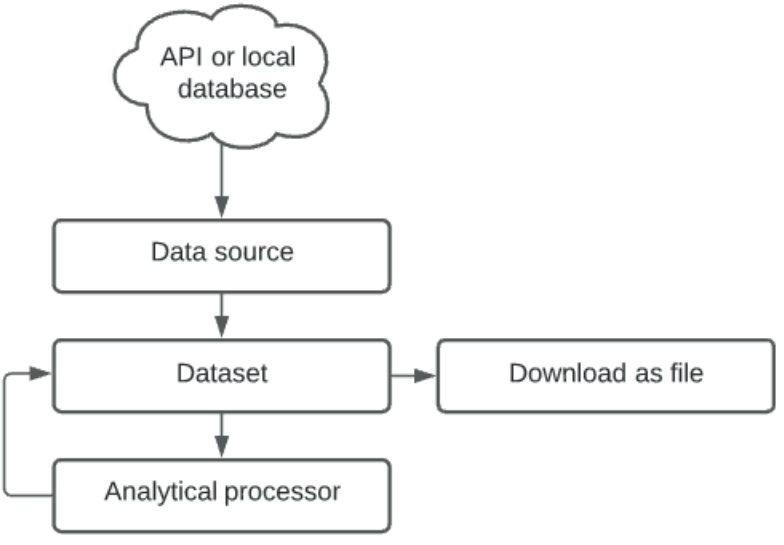


Figure 1. A schematic representation of 4CAT's architecture.

remains available. 4CAT primarily works with textual data with a 'thread-like' structure, that is, consisting of collections of chronologically ordered items that may further be grouped in separate conversational threads. This data is handled by 4CAT's three main components: code that handles the initial capture of data (*data sources*), the extracted and analysed results from these data sources (*datasets*), and scripts that can manipulate or analyse datasets (*analytical processors*). Figure 1 schematically shows how these interact.

4CAT's *data sources* are collections of workers, processors, and interface elements that handle the capture of data from a given platform. 4CAT has a Python API[3] that does most of the scaffolding around this, so the data sources *per se* can be relatively simple, focusing on the interaction with the external source the data is retrieved from. This will usually be an external Web API (e.g. the Pushshift API for Reddit data[4]) but data sources can also query e.g. a locally hosted database. At its core, each data source captures collections of discrete objects such as forum posts or messages, which are expected to have at least the following attributes:

- A unique identifier (*id*)
- An identifier for the thread or broader collection it is part of (*thread_id*)
- A time of creation (*timestamp*)
- An author (*author*)
- Textual content (*body*)

All items must have these attributes, though the latter two may be empty. Additional attributes may also be added, for example to expose platform-specific data such as the number of likes or retweets for an item. While these items directly correspond to threads of posts on forum-like platforms such as Reddit or 4chan, this way of handling data is 'naive' enough to also be compatible with differently structured platforms such as Tumblr, where a 'thread' can be conceived of as a Tumblr post and all comments or 'text reblogs' it received.

If a platform changes or limits access to its data (as Instagram did in 2019 when it severely limited automated access) or offers new APIs (as Twitter did in 2021 when it introduced a new 'Academic Track' API), data sources can easily be edited to work with these changes. 4CAT offers a logging system that data source authors can use to inform those managing a 4CAT server or instance of errors in data retrieval, signalling the need for such updates. A link to a page where users can report issues is additionally displayed prominently in the interface so tool maintainers can be notified of emerging bugs.

At the time of writing, the platforms 4CAT's data sources expose include but are not limited to Twitter, Telegram, Reddit, 4chan, 8kun, BitChute, Douban, Tumblr, and Parler. Additionally, .csv files that conform to the aforementioned structure can be uploaded to use 4CAT's analytical processors. Finally, there is the possibility to import data from other tools, such as Facebook and Instagram data via CrowdTangle, or TikTok data exported from an external scraper. We maintain a list of available data sources and instructions on how to add new ones on 4CAT's GitHub wiki.[5]

Having selected a data source in 4CAT, users can create a *dataset* of items matching specific parameters. Minding traceability and transparency, each dataset is accompanied by metadata that detail how it was produced, including the dataset's query parameters (such as a query string or date range), its file size, the time of execution, and the specific version of the code (identified by a git commit hash) that produced it. These metadata are prominently displayed in 4CAT's Web interface. After a query is finished, datasets can be downloaded to explore their contents or process them with other tools.

However, 4CAT itself can also analyse datasets with *analytical processors*. These are self-contained scripts that use 4CAT's Python API. Processors can range from simple scripts (that e.g. count how many items occur per month) to more advanced ones (e.g. that visualise network graphs). After running a processor, the result constitutes a new dataset, which can again be downloaded directly or, when available, manipulated

Table 1.

Examples of available processors	
Category	**Examples of processors**
Filtering	Filter by lexicon, Expand shortened URLs, Filter by unique posts
Networks	Co-link network, Co-word network, Sigma js network
Text analysis	Tokenise, Extract named entities, Word collocations, Word embeddings, Tf-idf
Post metrics	Count posts, Download images, YouTube URL metadata
Visualisation	Image walls, Histogram, RankFlow, Word tree

even further with subsequent processors (see Figure 1). Processors can specify different types of input: any of 4CAT's data sources, a specific data source, or the output of another processor. This way, it becomes possible to build 'pipelines' of processors that each perform a self-contained operation. For processors that commonly follow each other, it is possible to merge these into a single 'preset' (e.g. a 'Monthly histogram' preset that first executes the 'Count posts' processor and then the 'Histogram' processor).

However, with traceability in mind, the individual processors composing the pipeline can still be examined and on their own, and intermediate results can be shared and downloaded as well. As such, 4CAT allows navigation around different points in the 'paper trail' of a given result (see Figure 7). Our hope is that such traceability affords methodological reflexivity and navigational practices as discussed above. While outlining all processors is beyond the scope of this text, Table 1 provides a sample of what is available. A complete and updated list of processors can be found on the tool's GitHub wiki, which also contains instructions on how to create new processors.[6]

Using 4CAT

This section offers a brief walkthrough of the tool. Since 4CAT is actively developed at the time of writing, what follows is a specific snapshot of one of our own 4CAT instances[7] and other instances or later versions may look different. As 4CAT administrators may enable and disable modules as they see fit, not all of the modules shown below will be available in all 4CAT instances. The walkthrough below will nonetheless provide a general overview of the tool's use.

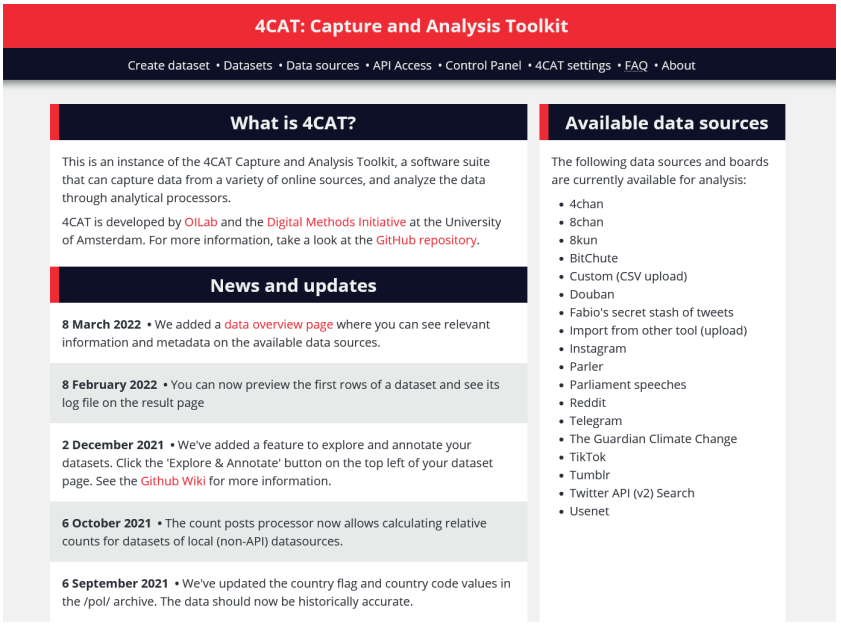


Figure 2. 4CAT's landing page.

Anyone can download 4CAT from GitHub and install it locally or on a server via Docker; our wiki provides a guide on how to install the tool.[8] Access to a 4CAT instance can be restricted through account registration and IP address matching (e.g. to allow access via a specific university's VPN). After installation, specific data sources and processors can be enabled or added via a configuration file. 4CAT's landing page (Figure 2) provides a general description of the tool, as well as a configurable welcome message. On the right side, a dynamic list of available data sources is shown. With constant changes in data access and the emergence of new or 'alt-tech' platforms (Donovan et al., 2019), this selection may fluctuate and draw from different stores of data. In the case below, the Twitter and Reddit data sources work with 'live' data through official and unofficial APIs (Twitter's v2 API and Pushshift), while 4chan data is retrieved from a static, locally stored database of historical posts.

Data capture

The 'Create dataset' (Figure 3) page allows querying a *data source* to produce a *dataset*. Data sources can use a range of predefined input

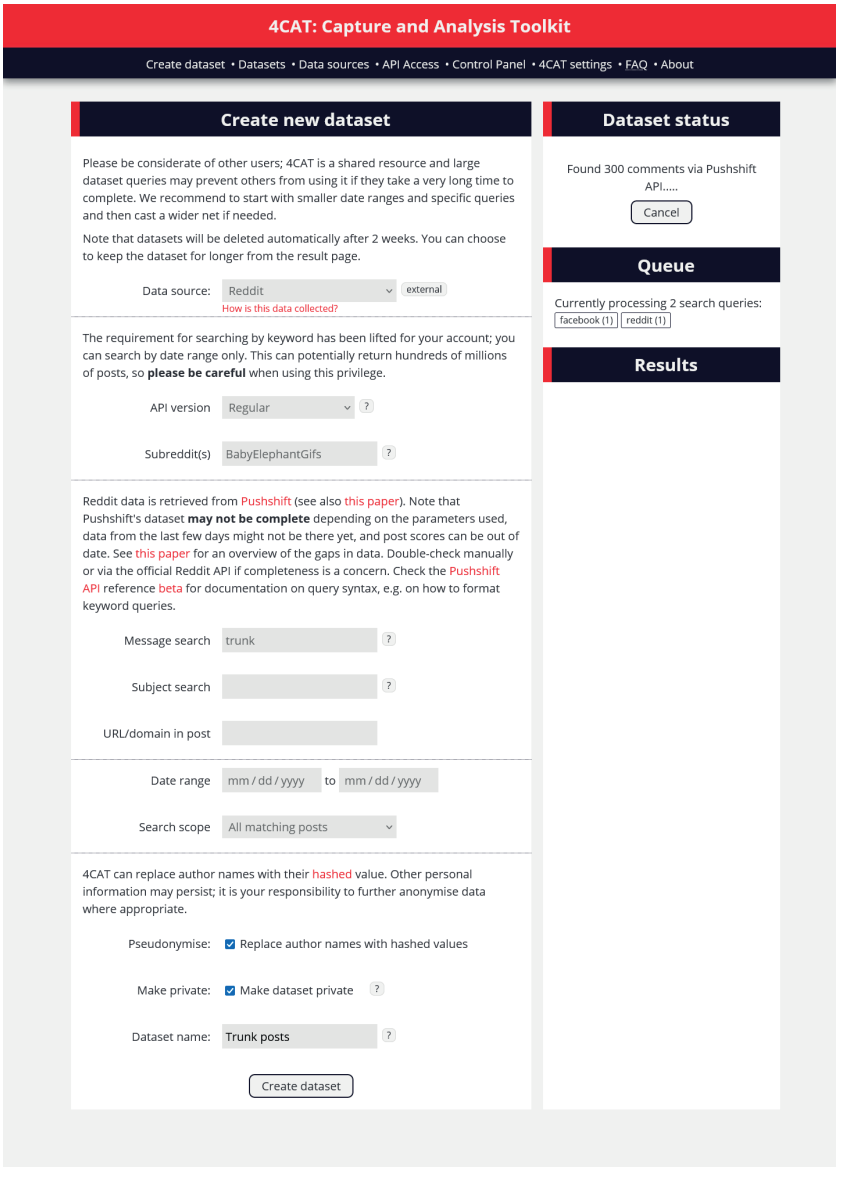


Figure 3. The dataset creation page, in this case showing an in-progress Reddit search.

fields for their interface, like a text field or a dropdown.[9] This allows the interface of this page to be customized to the logics of a specific data source, i.e. the parameters it exposes and the information a user needs to know to configure these properly. As an example of how 4CAT

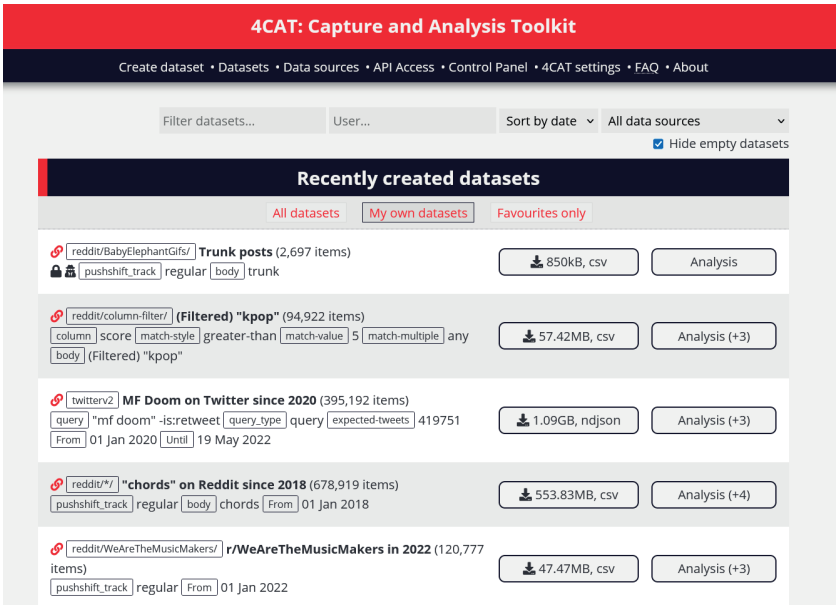


Figure 4. A results overview page.

aims to afford transparency through explanation, the interface for Reddit search in Figure 3 includes various references and warnings, e.g. on how the external data source it relies on, Pushshift, may not always provide accurate data (see Gaffney & Matias, 2018). This exemplifies the drawbacks of modularity since shortcomings of external systems may be 'imported' into 4CAT. However, with volatility and technical issues being an unavoidable part of Internet research, we at least try to render transparent possible limitations.

After submitting dataset parameters, 4CAT's Web interface signals the back-end task queue to start capturing data. On the 'Past results' page (Figure 4), users can retrieve datasets or check the status of a dataset being created. From here, the datasets can be downloaded directly, or one can view a dataset's page from which processors can be run. These dataset pages all have a unique, public URL which facilitates sharing datasets; though a dataset's creator can also make the dataset 'private' which prevents others from viewing it.

Figure 5 shows a dataset's page, in this case a Reddit dataset of individual comments containing 'rutte' on the subreddit *r/thenetherlands*. The top of the page shows the dataset's metadata, like the time of creation and query parameters. Above this are options to delete the dataset, re-run it, or retrieve

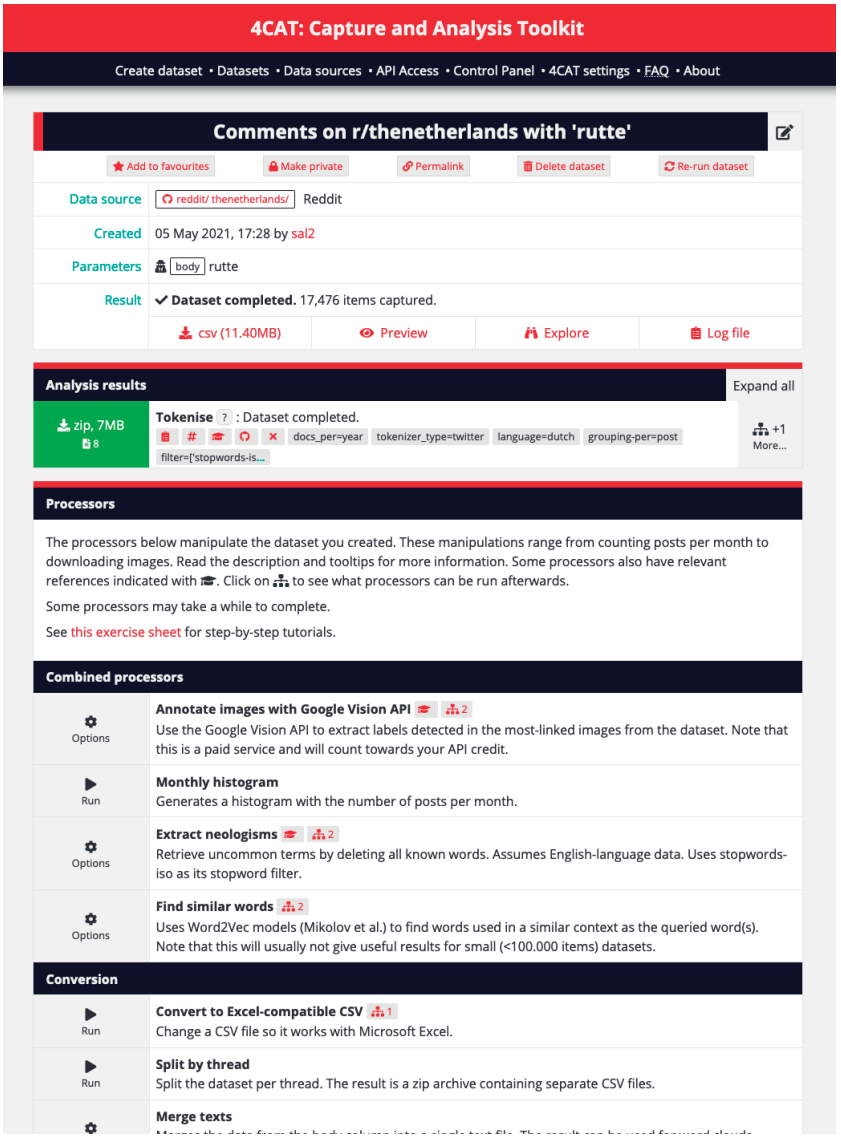


Figure 5. A dataset's page.

a permalink for sharing. Below this are the options to run processors. As a result of 4CAT's modularity, some of these processors may appear for every data source – like those retrieving simple metrics or text analysis processors – while others can be designated as data source-dependent, like 'Update Reddit post scores' for Reddit datasets.

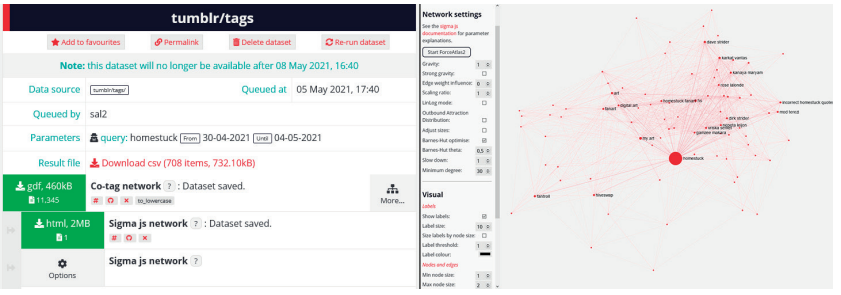


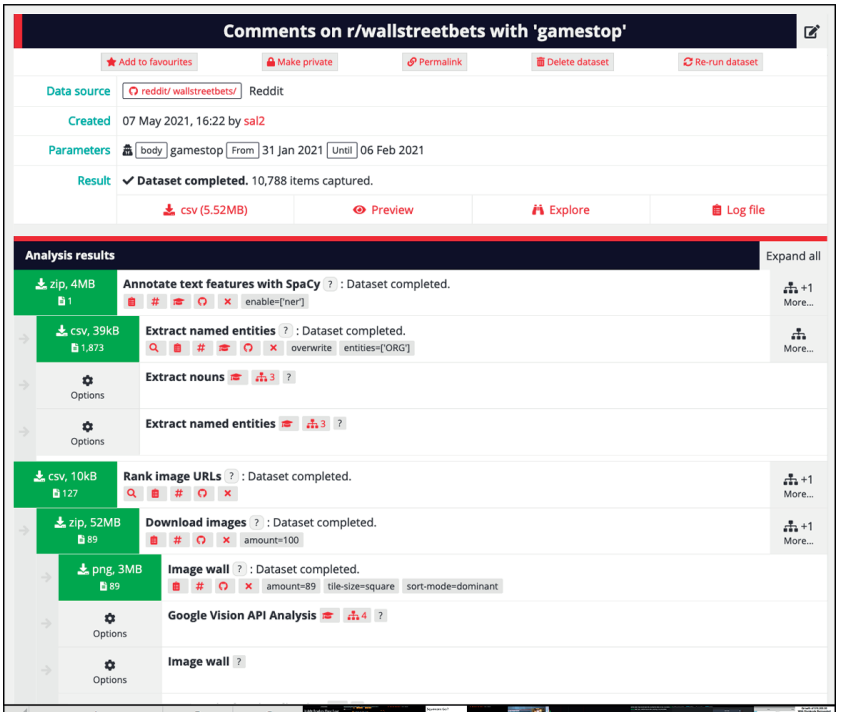
Figure 6. Running processors on a Tumblr dataset to generate a co-tag network graph.

Data analysis

Processors may require input by the user before being run; we encourage contributors to make processor parameters editable to afford transparency and decrease black-boxing. Furthermore, (academic) references may be added to a processor's description; clicking the academic cap icon (e.g. next to the 'Annotate images with Google Vision API' module in Figure 5) reveals a list of relevant websites and research papers that can help to understand the processor's functionality and (qualitative) epistemic assumptions.

When a processor is run, its progress is displayed and updated at the top of the page (Figure 6; here both processors have been finished, their status reading 'Dataset saved'). When finished, the output of the processor can be downloaded directly, or, if available, subsequent processors can be run. In Figure 6, two processors have been run on a Tumblr dataset: a 'Co-tag network' processor, which takes the 'tags' column of the parent dataset to create a network file compatible with Gephi (Bastian et al., 2009), and a subsequent 'Sigma.js network' processor, which produces a Web page through which this network can be visualized and manipulated in the browser. When the 4CAT instance is configured properly, each result will additionally display a versioned link to the GitHub page containing the exact version of the code that procured the result.

To provide one specific example of 4CAT's workflow, Figure 7 shows the results of two different pipelines using a dataset of *r/wallstreetbets* Reddit comments mentioning 'gamestop' in the first week of February 2021. Using *SpaCy*, a popular NLP library, the first processor extracts linguistic features like named entities from the comments ('Linguistic features'). The subsequent processor uses this output to extract and rank the most-mentioned named entities ('Extract named entities'). In this



	A	B	C
1	word	entity	count
2	gamestop	ORG	2730
3	gme	ORG	2530
4	amc	ORG	1211
5	wsb	ORG	305
6	drivewealth	ORG	298
7	nyse	ORG	292
8	amazon	ORG	212
9	sec	ORG	191
10	dtc	ORG	172
11	amc entertainment	ORG	163
12	revolut	ORG	146

Figure 7. A Reddit dataset processed through two analytical pipelines (top) and their results (bottom).

case, it shows how *GameStop* was often mentioned alongside companies like *AMC* and *DriveWealth*. The second pipeline first extracts image URLs from the text and ranks them by occurrence ('Top images'), followed by a processor that downloads and zips these images ('Download images'). Finally, the 'Image wall' processor sorts and combines these images – in this case showing r/wallstreetbets posters mostly posted screenshots of stock charts.

Ethical considerations

While methodologically productive, a tool like 4CAT can raise ethical questions. Notably, the fact that online data might be publicly available for capture through APIs does not mean it eludes privacy concerns (Zimmer, 2010). As Malin Sveningsson Elm (2009) notes, 'even if users are aware of being observed by others, they do not consider the possibility that their actions and interactions may be documented and analyzed in detail at a later occasion' (p. 77). Increasingly, how to handle such concerns is formalized in legislation, for instance through the EU's General Data Protection Regulation. While these give general guidance, reflecting on privacy concerns remains a continuous process for a modular tool like 4CAT, since each data source presents new ethical considerations. For example, anonymization may not be urgent for anonymous forums like 4chan, but becomes a more pressing issue in the case of Twitter data, where each item is linked to a specific user account.

This underlines how no tool is 'neutral' since normative decisions are always embedded in its design. It also emphasizes the responsibility of the tool's author and the value of an 'ethics by design' approach (Niemelä et al., 2014). While anyone can set up and control their own 4CAT instance, meaning we cannot *enforce* an ethically sound approach, we nevertheless try to *encourage* users to adopt one.

For example, 4CAT by default pseudonymizes (via a salted hash) any fields containing data about the author of a dataset item, e.g. their username. Here we leverage the 'effects of the default' (which is often left unchanged) within 4CAT's 'choice architecture' (Thaler & Sunstein, 2011). Furthermore, one can configure datasets from a given data source to be deleted automatically. For instance, Tumblr datasets may be configured to be deleted within three days, as mandated by Tumblr's terms & conditions for API usage (see the blue text in Figure 6). 4CAT can additionally be configured to delete *all* datasets after a set amount of time unless the dataset's creator opts out. Finally, one may add processors that remove particular types of information, and indeed 4CAT contains one by default to remove (rather than pseudonymize) all author information from a dataset. While ethical data capture remains an interplay between the agency of the tool, the researchers using the tool, and the platforms providing data access, through such features we hope to encourage and facilitate an ethically sensitive approach.

Conclusion

This paper situated and discussed 4CAT, a modular, Web-based tool that allows the capture and analysis of thread-like Web data in a transparent and traceable manner. Instead of merely offering a practical walkthrough of the tool itself, we sought to position 4CAT in dialogue with challenges emerging from the use of data-driven tools and computational methods within the humanities and social sciences. Notably, we outlined how the design of 4CAT engages with methodological concerns on modularity, transparency, and traceability. This short paper only briefly touches on these issues, however, and we openly invite further 'tool criticism' (Van Es et al., 2018; Koolen et al., 2019) of 4CAT, including its practical functions, methodological use, position within academic environments, and ethical dimensions.

Ultimately, we hope 4CAT can be used in the context of a 'digital *Bildung*' as coined by Berry (2011) and further developed by Rieder & Röhle (2017): 'a rolling process of reflexive thinking and collaborative rethinking' (Berry 2011, p. 22), one attentive to the complications and nuances of data-driven and computational techniques. While this eclipses what tools and their developers can single-handedly achieve, we hope to have shown practical design choices in applications like 4CAT can at least nudge (Thaler & Sunstein, 2011) towards such a digital Bildung, directing users to understand functions on a deeper level, render implicit computational techniques explicit, and make its shortcomings legible.

Acknowledgements

We would like to thank Bernhard Rieder and CCR's peer reviewers for their insightful comments, Dale Wahl for his significant contributions to 4CAT, the developers of the libraries, APIs, and archives used by 4CAT, and all those who have used and suggested improvements to the tool.

Funding

Both authors received funding from the ERC Horizon 2020 project ODYCCEUS, grant agreement number 732942. Stijn Peeters additionally received funding from the Dutch PDI-SSH foundation through the CAT4SMR project.

Notes

1. Dually published on GitHub and Zenodo; see github.com/digitalmethodsinitiative/4cat and doi.org/10.5281/zenodo.4742622.
2. See github.com/digitalmethodsinitiative/4cat/wiki.
3. This Python API is documented in the 4CAT's Github repository's wiki, at github.com/digitalmethodsinitiative/4cat/wiki/Developer-guide.
4. See pushshift.io (Baumgartner et al. 2020).
5. See github.com/digitalmethodsinitiative/4cat/wiki/Available-data-sources and github.com/digitalmethodsinitiative/4cat/wiki/How-to-make-a-data-source.
6. See github.com/digitalmethodsinitiative/4cat/wiki/Available-processors and github.com/digitalmethodsinitiative/4cat/wiki/How-to-make-a-processor.
7. To be precise, the following commit: github.com/digitalmethodsinitiative/4cat/commit/cebdfoc5ed54f4fad496aaefcoce3b1ecd3fd11d.
8. See github.com/digitalmethodsinitiative/4cat/wiki/Installing-4CAT. Generally speaking, 4CAT will run on relatively modest hardware; a four-core CPU with 8 to 16GB of RAM will typically suffice.
9. See github.com/digitalmethodsinitiative/4cat/wiki/Input-fields-for-data-sources-and-processors.

References

Al-Rawi, A. (2020). The convergence of social media and other communication technologies in the promotion of illicit and controlled drugs. *Journal of Public Health*, *fdaa210*. https://doi.org/10.1093/pubmed/fdaa210

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third International AAAI Conference on Weblogs and Social Media*, 361–362.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., & Blackburn, J. (2020). The Pushshift Reddit Dataset. *ArXiv:2001.08435* [*Cs*]. http://arxiv.org/abs/2001.08435

Baym, N. K. (2000). *Tune in, log on: Soaps, fandom, and online community*. Sage Publications.

Berry, D. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, *12*.

Borra, E., & Rieder, B. (2014). Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*. https://doi.org/10.1108/AJIM-09-2013-0094

Bowker, G. (2013). Data flakes: An afterword to "Raw Data" is an oxymoron. In L. Gitelman (Ed.), *"Raw Data" is an Oxymoron* (pp. 167–171). MIT Press.

Boyd, D., & Crawford, K. (2011). Six Provocations for Big Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1926431

Bucher, T., & Helmond, A. (2017). *The Affordances of Social Media Platforms*. Sage Publications. https://dare.uva.nl/search?identifier=149a9089-49a4-454c-b935-a6ea7f2d8986

Crawford, K. (2013, April 1). The Hidden Biases in Big Data. *Harvard Business Review*. https://hbr.org/2013/04/the-hidden-biases-in-big-data

de Zeeuw, D., Hagen, S., Peeters, S., & Jokubauskaite, E. (2020). Tracing normiefication: A cross-platform analysis of the QAnon conspiracy theory. *First Monday*. https://doi.org/10.5210/fm.v25i11.10643

Donovan, J., Lewis, B., & Friedberg, B. (2019). Parallel Ports: Sociotechnical Change from the Alt-Right to Alt-Tech. In M. Fielitz & N. Thurston (Eds.), *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US* (pp. 49–63). transcript Verlag.

Gaffney, D., & Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE*, *13*(7), e0200162. https://doi.org/10.1371/journal.pone.0200162

Giguet, E., & Lucas, N. (2009). Share and Explore Discussion Forum Objects on the Calico Website. *Computer Supported Collaborative Learning Practices*, 174–176.

Hutchby, I. (2001). Technologies, Texts and Affordances. *Sociology*, *35*(2), 441–456. https://doi.org/10.1177/S0038038501000219

Jokubauskaitė, E., & Peeters, S. (2020). Generally Curious: Thematically Distinct Datasets of General Threads on 4chan/pol/. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 863–867.

Kirschenbaum, M. (2012). What Is Digital Humanities and What's It Doing in English Departments? In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 3–11). University of Minnesota Press.

Knorr-Cetina, K. (1999). *Epistemic cultures: How the sciences make knowledge*. Harvard University Press.

Koolen, M., van Gorp, J., & van Ossenbruggen, J. (2019). Toward a model for digital tool criticism: Reflection as integrative practice. *Digital Scholarship in the Humanities*, *34*(2), 368–385. https://doi.org/10.1093/llc/fqy048

Latour, B. (2011). Networks, Societies, Spheres: Reflections of an Actor-network Theorist. *International Journal of Communication*, *5*(0), 15.

Latour, B., Jensen, P., Venturini, T., Grauwin, S., & Boullier, D. (2012). 'The Whole Is Always Smaller Than Its Parts' – A Digital Test of Gabriel Tardes' Monads. *The British Journal of Sociology*, *63*(4), 590–615. https://doi.org/10.1111/j.1468-4446.2012.01428.x

Manovich, L. (2001). *The Language of New Media*. MIT Press.

Niemelä, M., Kaasinen, E., & Ikonen, V. (2014). *Ethics by design—An experience-based proposal for introducing ethics to R&D of emerging ICTs*. ETHICOMP 2014 – Liberty and Security in an Age of ICTs. https://cris.vtt.fi/en/publications/ethics-by-design-an-experience-based-proposal-for-introducing-eth

Rieder, B., & Röhle, T. (2012). Digital Methods: Five Challenges. In D. M. Berry (Ed.), *Understanding Digital Humanities* (pp. 67–84). Palgrave Macmillan UK. https://doi.org/10.1057/9780230371934_4

Rieder, B., & Röhle, T. (2017). Digital Methods: From Challenges to Bildung. In M. T. Schäfer & K. Van Es (Eds.), *The Datafied Society: Studying Culture through Data* (pp. 109–124). Amsterdam University Press.

Sveningsson Elm, M. (2009). How do various notions of privacy influence decisions in qualitative internet research? In A. Markham & N. Baym (Eds.), *Internet Inquiry: Conversations About Method* (pp. 69–97). SAGE Publications, Inc. https://doi.org/10.4135/9781483329086

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness* (Rev. and expanded ed). Penguin Books.

Tuters, M., & Hagen, S. (2019). (((They))) rule: Memetic antagonism and nebulous othering on 4chan. *New Media & Society*, 1461444819888874. https://doi.org/10.1177/1461444819888746

van Es, K., López Coombs, N., & Boeschoten, T. (2017). Towards a Reflexive Digital Data Analysis. In *The Datafied Society: Studying Culture through Data* (pp. 171–180). Amsterdam University Press.

van Es, K., Wieringa, M., & Schäfer, M. T. (2018). Tool Criticism: From Digital Methods to Digital Methodology. *Proceedings of the 2nd International Conference on Web Studies – WS.2 2018*, 24–27. https://doi.org/10.1145/3240431.3240436

Venturini, T. (2011). *What is second-degree objectivity and how could it be represented*. http://www.medialab.sciences-po.fr/publications/ Venturini-Second_Degree_Objectivi-ty_draft1.pdf

Zelenkauskaite, A., Toivanen, P., Huhtamäki, J., & Valaskivi, K. (2020). Shades of hatred online: 4chan duplicate circulation surge during hybrid media events. *First Monday*. https://doi.org/10.5210/fm.v26i1.11075

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology*, *12*(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5