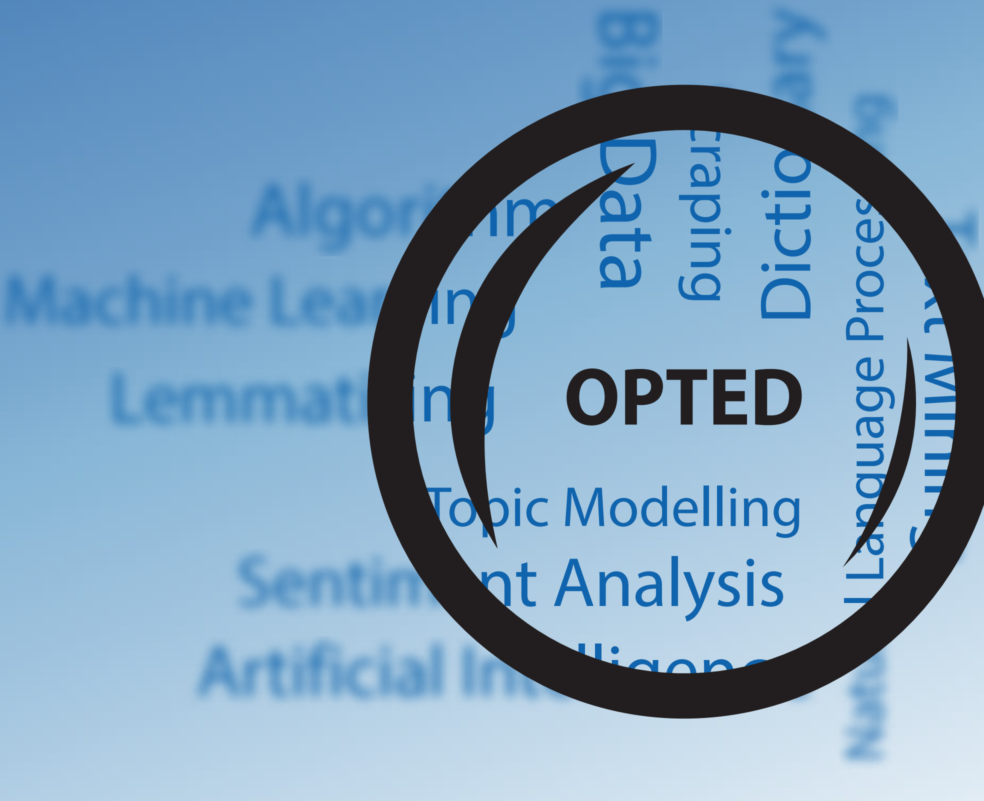
OPTED

Sharing *is* Caring (about Research): Three Avenues for Sharing (Copyrighted) Text Collections and the Need for Non-Consumptive Research

Deliverable 7.6

Johannes B. Gruber, Wouter van Atteveldt, and Kasper Welbers

Vrije Universiteit Amsterdam



D7.6: Sharing *is* Caring (about Research): Three Avenues for Sharing (Copyrighted) Text Collections and the Need for Non-Consumptive Research

Disclaimer

This project has received funding from the European Union’s Horizon 2020 research & innovation programme 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

Sharing *is* Caring (about Research): Three Avenues for Sharing (Copyrighted) Text Collections and the Need for Non-Consumptive Research

Deliverable 7.6

Authors: Johannes B. Gruber, Wouter van Atteveldt, and Kasper Welbers

Vrije Universiteit Amsterdam

Due date: 30. September 2023

Contents

Abstract

A decade ago, the computational turn in communication science was heralded by promises of unseen treasure troves of available digital text containing communication of journalists and social media users. As these treasures become more closly guarded today, the field needs to think of new strategies to continue to enable researchers who want to engage in computational communication research. One of these strategies is to make text data sharing a more common practice in the field. In this article, we outline three avenues to share as much as your data as possible, while still honouring ehtical and legal restictions. Given the relative lack of infrastructure for some of these avenues, we also highlight the capacities of the Amsterdam Content Analysis Toolkit (AmCAT) to enable non standard sharing strategies. We especially highlight the functions for *non-consumptive research* – which means analyses methods that can be performed without access to the full data set.

1 Introduction

Text-as-data has arrived as a prominent method for content analysis in the social sciences. The twin promises of making large scale analysis of text corpora feasible while keeping the costs for manual annotation to a minimum convinced institutions and individual researchers to make considerable investments into methods training and embrace the computational revolution of communication science in the last decade (e.g. Grimmer and Stewart 2013; see overview articles of Brady 2019; Hilbert et al. 2019; Lazer and Radford 2017; Van Atteveldt and Peng 2018; Atteveldt et al. 2019). However, since the early days, a lot of the hopes and enthusiasm has evaporated as access to the data gold mines of social media databases and digital news archives are progressively limited by their owners. And more importantly, pressing questions about political communication stay untackled, as researchers are shut out.

Just this year, Twitter, which was renamed to 𝕏 recently, eliminated the free academic access to their API[1] and Reddit has shut down access to its API[2] for Pushshift, a service on which most academic research of that social media site relied. Meanwhile Meta's tool to gather data from Facebook and Instagram, Crowdtangle, has been publicly rumored to be closing down for several years[3], making it difficult to plan, or request funding for, any projects relying on this data. Likewise, digital news archives, like LexisNexis, increasingly guard their sites against large scale data collection or make it out of reach expensive. Meanwhile, newly developed or strengthened privacy and ethical standards and laws – while both important and welcome – further limit access to data for analysis. The Post-API Age, as predicted by Freelon (2018), seems to have reached a new stage and the promised

[1] https://web.archive.org/web/20230831000123/https://www.theverge.com/2023/3/30/23662832/twitter-api-tiers-free-bot-novelty-accounts-basic-enterprice-monthly-price

[2] https://web.archive.org/web/20230829045754/https://www.reddit.com/r/modnews/comments/134tjpe/reddit_data_api_update_changes_to_pushshift_access/

[3] https://www.bloomberg.com/news/articles/2022-06-23/meta-pulls-support-for-tool-used-to-keep-misinformation-in-check; https://web.archive.org/web/20230816171112/https://www.theverge.com/2022/6/23/23180357/meta-crowdtangle-shut-down-facebook-misinformation-viral-news-tracker; https://web.archive.org/web/20230827165945/https://www.abc.net.au/news/science/2022-08-16/facebook-crowdtangle-meta-disinformation-transparency/101325544

gold mines of digital communication traces are no longer accessible.

In sum, this limited data availability also broadens the gap between academics once more: While computational methods have decreased the funding needs for analysis, limited availability of data leads to new divides in the field between those who can afford to negotiate access, or have industry ties that enable them to collaborate with social media companies, and those who cannot – also limiting who can reproduce or check such research. Likewise, web scraping content, as Freelon (2018) suggests, requires know-how and resources. As more platforms, outlets and individuals try to block their content from being scraped by AI companies – and de facto everyone else[4] – the investment in know-how and resources might soon surge and put it out of reach for many. These problems have been recognized in the field and are discussed as important challenges at methodological and issue specific conferences[5] as well as in specialized working groups like the EDMO Working Group on Access to Platform Data[6].

In this contribution, we thus want to highlight a mitigation strategy that grows in importance given the narrowing ways to access communication data: sharing already gathered corpora for secondary analysis. For data collected through, e.g., surveys, this strategy has been established for decades. For text corpora, there are often legal or ethical barriers to what can be shared: A large portion of textual data, especially media content, is copyrighted, and moreover for some data (such as interpersonal communication) there are privacy and other ethical concerns (Van Atteveldt, Althaus, and Wessler 2021).

Specifically, we present three avenues to share data sets, structured by what is legally and morally justifiable in light of the interests of data owners: making pre-processed versions of the corpus available; making metadata of a corpus available; making *non-consumptive* research capabilities available – that is enabling analyses of a corpus without sharing the text data for consumption (i.e., reading). We put a particular focus on the last avenue of enabling analyses through *non-consumptive research*, in cases where other strategies appear impossible. Some of the options we discuss rely on infrastructure that currently does not exist, like servers to host tools for *non-consumptive research.* We thus additionally offer an open source software toolkit which allows universities, other institutions, or even individual researchers to share at least a portion of their collected data online, while respecting legal and ethical limitations to what can and cannot be shared. In particular, we developed this toolkit based on a modular design and open source frameworks, focused on the needs we identified for ourselves and the community.

2 The case for sharing textual data

A core argument of this paper is that that textual (and other) data sets should be made publicly available for secondary research where possible. This can enable others to quickly build on work after the original data set has been gathered. It saves researchers from duplicating work and enables them to focus their energy on answering important questions

[4] https://www.nytimes.com/2023/07/15/technology/artificial-intelligence-models-chat-data.html

[5] Like the keynote roundtable at Comptext 2023 or the *The Post-API Conference* at the Annenberg Public Policy Center in 2023.

[6] https://web.archive.org/web/20230606203119/https://edmo.eu/2021/08/30/launch-of-the-edmo-working-group-on-access-to-platform-data/

for society that might otherwise stay uncovered. What might be less obvious is why those who gathered the data in the first place might want to do that. We believe that there are at least three convincing arguments why one should consider the step (see e.g. Dienlin et al. 2021; Klein et al. 2018; Miguel et al. 2014; Nosek et al. 2015 for similar arguments):

1. **Lend your research credibility by making it reproducible:** One of the cornerstones of robust scientific research is reproducibility, that is that others than the original researcher can take the original data and methods to reproduce the findings of a study (Barba 2018). When the original data is not available, reproducibility is impossible by default. By making data sets available, original researchers expose themselves to more rigor, which lends transparency and credibility to the original research, but also strengthens the foundation of scientific inquiry as a whole.

2. **Gather citations (beyond the field of the researcher):** data sets that are available and are used for secondary analysis are usually also cited in new studies. Given the importance for citation counts in most academic systems worldwide, this is a big incentive for those who spent the resources to gather the original data. What researchers should also keep in mind here is that the impact of the data set might well end up surpassing that of the original research, given that data collections might attract researchers from other disciplines than their own. As computer and data scientists become more interested in social science questions, gathered and annotated data has become a valuable resource in these fields, for example.

3. **Promote the original research:** Sharing data also acts as a promotional tool. When other researchers access and use this data, they will almost always become aware of the original research and build on it. It then often makes sense for them not to build upon the original findings and further promote the findings through their own contribution.

When researchers can access and build upon one another’s data, it accelerates scientific progress and helps to avoid duplicative efforts. Useful text collections will also be referenced repeatedly, providing additional recognition for those who collected the data. Of course, even with those incentives and with good support, there is considerable effort in preparing material for publication, which is not always sufficiently rewarded under the current academic norms, at least not on the short term. For this reason, we echo the call on journals that don’t do so yet to require materials to be published to the extent possible, even for sensitive or copyrighted materials; and on hiring and evaluation committees to appreciate the effort and worth of data contributions (Atteveldt et al. 2019).

3 Three Avenues for Sharing Text Collections without Sharing the Texts

As argued above, it is often desirable and even beneficial to share parts of textual data sets, even if the data set is proprietary, copyrighted, or sensitive. In this section, we describe three strategies for allowing other researchers access to textual data in situations where the texts cannot be shared publicly.

3.1 Making pre-processed versions of texts available

Many researchers might be aware of the advantages that data sharing have for the research community and themselves, but refrain from it anyway, given obstacles and dangers involved with sharing them. However, the legal and ethical limitations to free access to information where they infringe on copyrights or on the privacy of data owners might not in all cases extend to their pre-processed versions. Pre-processing of textual data can be roughly divided into *destructive* and *non-destructive* steps. In the best case scenario, data owners would give permission to make a tokenized version of their text – that is, a text split into its words and punctuation characters in their original order – available. Other researchers can then rebuild the data for all the same analyses that can be performed on the original set. The caveat of such non-destructive pre-processing, however, is that the same restrictions that apply on the original data likely also apply to the pre-processed data.

Where such restrictions are a concern, *destructive* pre-processing steps could instead be performed before sharing textual data. This could involve turning text into a document-feature-matrix – that is a table which contains counts of how often each word, punctuation character or symbol occurs in a text. In these cases bag-of-word approaches can still be employed to analyze the data, but the original texts can usually not be recreated[7]. For most limitations regarding copyright laws, making data available in such a format constitutes fair use, as "the effect of the use upon the potential market for or value of the copyrighted work" (Hennesy and Samberg 2019) is essentially non-existent.[8] Essentially, someone who wanted to read (i.e., consume) an article or book would gain little from staring at a document-feature-matrix, leaving the data owner's chances to make money from selling the work untouched. In fact, the platform Media Cloud has practiced this avenue for a while to share web-scraped content from hundreds of online news sites without known issues. Replicability of analysis suffers only for computational approaches where the word order is important – which is a little unfortunate given that the field currently departs from the bag-of-words paradigm. In cases where it is crucial to protect someone's privacy, destroying the word order might still not be enough, however. Additional steps could be applied to create a privatized version of the data, for example, by replacing each feature description of a document-feature-matrix with a synonym or an entirely random symbol. This decreases the usefulness of the data somewhat, but it can still be used to reproduce findings or improve upon original models.

A second destructive pre-processing step that has gained popularity in recent years is to encode texts into embeddings. Word or text embeddings produced through models like BERT (Devlin et al. 2019) are a common transformation process in computational text analysis now and are usually employed before a classification model is trained (e.g., Theocharis et al. 2020; Rodriguez, Spirling, and Stewart 2023; Laurer et al. 2023; Simon et al. 2022). This means that sharing the encoded data, along with the original training and test data, would allow other researchers to improve upon the original model and use it in similar contexts. In these cases as well, the original texts can not be recovered, evading most concerns for data owners. It also makes interpreting results harder and removes the possibility to add genuinely new annotations. Nevertheless, many analysis methods remain

[7] If original texts are very short, it might still be possible to some degree.

[8] Whether this commonly accepted legal argument fully satisfies data holders is a different question, which one should keep in mind.

possible, meaning that the presence of these datasets might still turn out to be invaluable for the research community.

3.2 Make Metadata Available for Reconstruction of the Data

Whether sharing pre-processed versions of texts is possible or not, making metadata available is another important avenue to sharing corpora, either in addition or as an alternative. The most important use of this metadata is that it can often be used to reconstruct the original data with far fewer resources or make the possibility available where the original data could otherwise not be gathered. Referencing URLs, status IDs or, where not available, the dates of publication, headlines and authors of text, can enable others to collect the same data once more. This might not seem important at first, but consider how time-consuming and difficult it often is to identify the relevant portion of a population text. At the minimum, a database must be queried using carefully selected keywords, before validating the obtained corpus to make sure false positives are eliminated (King, Lam, and Roberts 2017). Skipping this step in secondary analysis of data means that it can proceed more quickly, freeing up resources to focus on important substantive research questions.

In other cases, obtaining a selection of relevant texts might be impossible. As many platforms and outlets do not offer a way to query the content or demand additional compensation, using keywords to select relevant portions of a database might be impossible. However, gathering all content to filter afterward is often infeasible as well as it might conflict with rate limits and/or require enormous resources for download and storage. In sum, this can make a curated list of items the most valuable part of a shared dataset.

So while obtaining the full text still involves some extra efforts and challenges, if metadata of the original data is available, this enables researchers to reproduce (often in a limited fashion as not all items might still be available) and extend the original analysis. In the case of Twitter data collections, the practice of sharing status IDs had become common and helped researchers to get around API limits as they were able to 'rehydrate' databases easily.[9] A practice that has unfortunately also become punishingly expensive under the new API rules. A more encouraging example is again the platform Media Cloud from which researchers can obtain curated lists of URLs to news stories about specific topics. Using tools such as `news-please`[10] or `paperboy`[11], one can then quickly extract the full text, if the websites are still available.

3.3 Make Non-consumptive Research Capabilities Available

A final avenue for making datasets available for secondary analysis that of *non-consumptive research*. Non-consumptive research means that a researcher analyzes texts without consuming, that is reading, them. The term was popularized by two court cases which also exemplify its meaning: that of Authors Guild v. HathiTrust, 755 F.3d 87 (2d Cir. 2014) and Authors Guild v. Google 721 F.3d 132 (2nd Cir. 2015). In essence, both cases were about sharing tools to search copyrighted texts en masse to make counts, trends in usage and page numbers available for users. In the case of *Google Books* the service's function to

[9] http://dfreelon.org/2012/02/11/arab-spring-twitter-data-now-available-sort-of/
[10] https://github.com/fhamborg/news-please
[11] https://github.com/JBGruber/paperboy

display text snippets surrounding a keyword hit was also reviewed by the court and found to be covered by fair use exceptions to copyright:

> "*The creation of a full-text searchable database is a quintessentially transformative use [. . .] [T]he result of a word search is different in purpose, character, expression, meaning, and message from the page (and the book) from which it is drawn.*" (quoted from Hennesy and Samberg 2019, 294).

Transformative use here refers to the first factor that determines fair use: if a work that uses copyrighted material adds something new, "with a further purpose or different character, and do not substitute for the original use of the work" (17 U.S.C. § 107).

While *non-consumptive research* was made popular through these mostly legal arguments to receive exceptions from copyright claims, the principles of *non-consumptive research* apply for most other texts that have limitations to free access, for example where they infringe on the privacy of data owners. If the original text is not available to third parties, harm to those whose data is being analyzed is unlikely. And to go even a step further: if the original content is not consumed by the primary researchers, potential harm to participants and the researchers is minimized.

The model of *non-consumptive research* can thus be employed and extended for most data collections where concerns or limitations against publications exist. In Addition to allowing users to search the database and show which documents fit a query, aggregates (how many documents fit the query) and comparisons (e.g., do more speeches held by members of one party fit a query than those of another) are possible. In this way, researchers might assess the relevance of a data collection for their own research questions before engaging with data owners to ask for permission or collecting the data again themselves.

Beyond the simple functions most people know from Google Books though, the definition of *non-consumptive research* is broader and comprises all computational analysis in which a researcher does not display substantial portions of the original text or image (Center 2023). Based on this, HathiTrust and others have experimented with "data capsules". HathiTrust describes a capsule as a "system that grants a user access to a virtual machine which is a dedicated, secure desktop environment [. . .] through which a user can carry out non-consumptive research", yet which does not give researchers the ability to display or download the data (Center 2023). Given such a system, a wide range of unsupervised methods and destructive pre-processing steps become available to

4 Sharing Corpora with AmCAT

The Amsterdam Content Analysis Toolkit (AmCAT) has been in development in various guises since about 2001 as a text research / content analysis platform. The core of the project has always been a database of documents, combined with graphical user interface and API to make it possible for casual and power users respectively to quickly gather the texts they need. In the current fourth iteration of the toolkit, a particular focus was put on sharing data with an audience as broad as possible, but as narrow as legal and ethical circumstances allow. It comes at the right time for researchers who want to take one of the avenues described above to share their corpora, but so far miss the infrastructure for it.

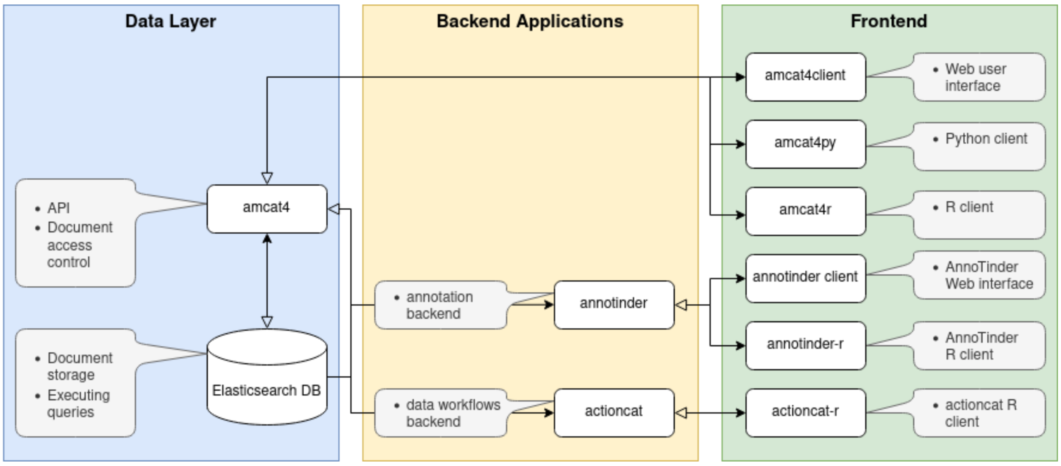


Figure 1: Design diagram of the Amsterdam Content Analysis Toolkit (AmCAT)

Besides providing a new dashboard and API wrappers for `R` and `Python` – which make them great for working with collected corpora, or selecting a relevant subset – one of the biggest improvements of the current 4.0+ versions of AmCAT is that it enables and facilitates the access and processing of data that cannot be shared openly by keeping data owners in direct control of the data, and by employing trusted connections and role-based access control methods. Below, we describe some of the functions of the flexible framework access control and guest access and the (pre-processing) actions which extend AmCAT to be a toolkit for *non-consumptive research.* Besides querying a database to get counts and trends of about the appearance of certain features, we developed a new framework, which we called `actioncat`, that can be used, for example, for pre-processing actions that allow users without access to the original data to decide which pre-processed versions of texts is then made available. The toolkit can be used for free and deployed by most users with a basic to advanced technical background. Because of the modular design, shown in Figure 1, deployments of AmCAT can be integrated in many existing infrastructures. In the future, we hope to encourage universities and other organizations to make deployed versions available for usage by the research community.

4.1 AmCAT Dashboard to present corpora

For the first avenue we discussed above – sharing an entire dataset including the full text – there are good options already available. The most commonly used is the open source Dataverse Project[12], which many might know from the most widely used distributed version at Harvard. Using this route is popular among researcher, with one reason likely being that data sets receive a DOI (Digital Object Identifier), which makes them easily citable, and the other being that the service is free.

One downside that we identified specifically for text and even more so for image and video data is that data sets can only be downloaded in full or in the parts the data set creators have pre-determined, for example, one file per country or period. For survey

[12] https://dataverse.org/

data, for example, this is not an issue as files are usually small and using a subset of the data is often less relevant. For corpora with typical sizes of tens or hundreds of thousands of documents, the files can take substantial amounts of time to be downloaded from a repository (depending on the speed of the connection and the bandwidth of the repository host).

In AmCAT, we provide graphical user interface and API packages in `R` or `Python` to enable users to query the database using the powerful query string syntax developed by Elasticsearch[13]. This can help researchers to evaluate whether a corpus contains the relevant data for their research or and select only the required parts for download. As an example, we downloaded the *ParlEE* plenary speeches data set, which contains speeches legislative chambers in from eight EU states (Sylvester, Greene, and Ebing 2022), and made it available through AmCAT. The data is provided as one csv file per country on Harvard Dataverse, with file sizes ranging from 281MB to 1.8GB. In Figure 2 we show an example query, which selects documents that feature at least one word related to the migration debate. Someone trying to decide if this was a good data set to study this debate would quickly see the shard peak in frequency in 2015, without needing to download anything. Give that this is a data set that comes with no restrictions, they could then start to use it for their analysis immediately.

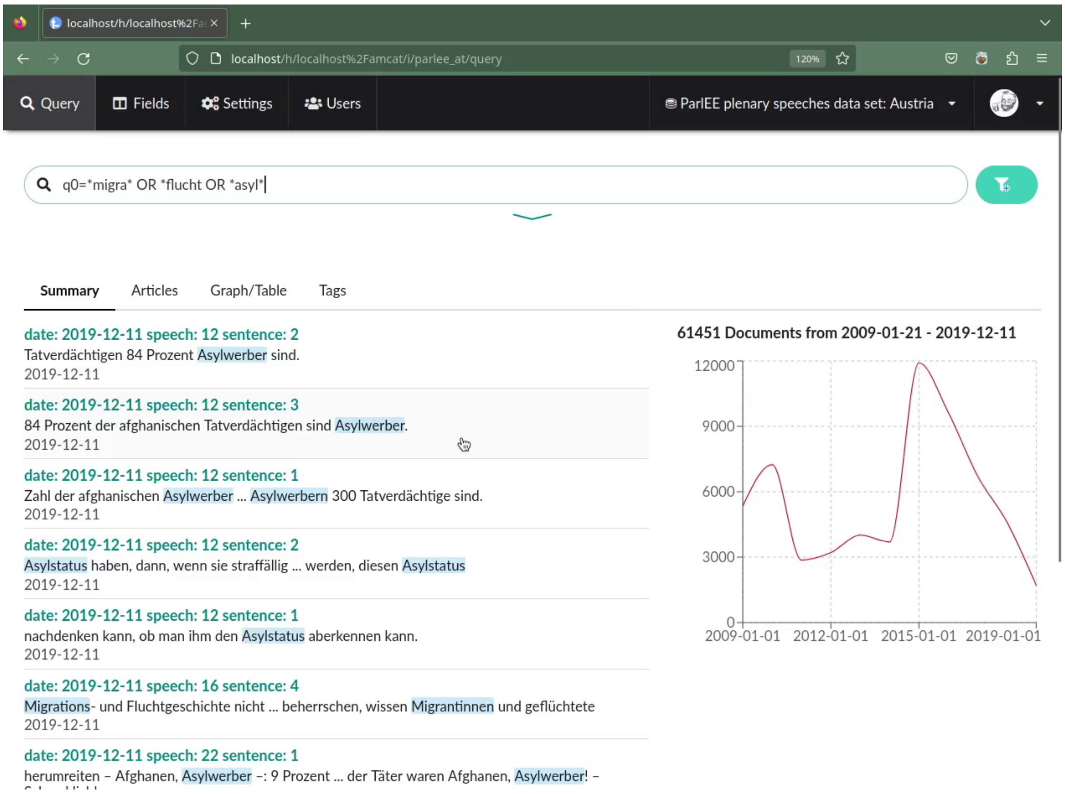
While AmCAT is not at a stage yet where a flagship distributed node could offer DOIs or free permanent storage, like the Harvard Dataverse does, there are no technical limitations that prevent that. For the ability to present ones collected data to other researchers, or even the general public, to reproduce, extend and build on original analysis AmCAT offers attractive options that would otherwise need to be built from ground up.

4.2 Access Control in Amcat4 Explained

For the remaining three avenues – sharing pre-processed text, sharing only metadata or enabling *non-consumptive research* – the access control features of AmCAT are key. `amcat4`, the database module of AmCAT, provides fine-grained access control, which enables administrators of an instance to share data in exactly the way they want and control what members of their team and outsiders can do on an index. This section explains how access control works in `amcat4` and highlights which features and settings provide infrastructure for the above described avenues.

We define two sets of roles that can be set by administrators or users: one that is configured globally per user on an `amcat4` instance and one set that can be used per index (i.e., a corpus. In combination the three global (reader, writer and admin) and index roles (none, metareader, reader, writer and admin) offer twelve different role levels that are shown in Table 1.

[13] http://web.archive.org/web/20230908093502/https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html


localhost/h/localhost%2Fa
localhost/h/localhost%2Famcat/i/parlee_at/query
120%
Query
Fields
Settings
Users
ParlEE plenary speeches data set: Austria
Summary
Articles
Graph/Table
Tags
61451 Documents from 2009-01-21 - 2019-12-11
12000
9000
6000
3000
0
2009-01-01
2012-01-01
2015-01-01
2019-01-01

q0=*migra* OR *flucht OR *asyl*

date: 2019-12-11 speech: 12 sentence: 2
Tatverdächtigen 84 Prozent Asylwerber sind.
2019-12-11

date: 2019-12-11 speech: 12 sentence: 3
84 Prozent der afghanischen Tatverdächtigen sind Asylwerber.
2019-12-11

date: 2019-12-11 speech: 12 sentence: 1
Zahl der afghanischen Asylwerber ... Asylwerbern 300 Tatverdächtige sind.
2019-12-11

date: 2019-12-11 speech: 12 sentence: 2
Asylstatus haben, dann, wenn sie straffällig ... werden, diesen Asylstatus
2019-12-11

date: 2019-12-11 speech: 12 sentence: 1
nachdenken kann, ob man ihm den Asylstatus aberkennen kann.
2019-12-11

date: 2019-12-11 speech: 16 sentence: 4
Migrations- und Fluchtgeschichte nicht ... beherrschen, wissen Migrantinnen und geflüchtete
2019-12-11

date: 2019-12-11 speech: 22 sentence: 1
herumreiten – Afghanen, Asylwerber –: 9 Prozent ... der Täter waren Afghanen, Asylwerber! –

Figure 2: AmCAT Interface for the Austrian corpus in ParlEE, queried for texts on migration

Table 1: Access roles in amcat4

Role Level	Global Role	Index Role In Index A	Description
0	-	Guest	If a user does not have an explicit role on an index, the guest role (if any) is used. An unauthorized user can still get guest roles, so they can see any indices with a guest role. This is not a specific level, but it serves as a fallback for undefined index roles.
1	Reader	None	Can see which indices exist, but cannot access them. Index A is invisible to the user.
2*	Reader	Metareader	Can see which indices exist. Can read all properties, do queries, etc. in Index A, but cannot read the 'text' attribute.
3	Reader	Reader	Can see which indices exist. Can read all properties, do queries, etc. in Index A, but cannot make changes.
4	Reader	Writer	Can see which indices exist. Can add/delete documents, add/delete users (up to their own level), and make other changes (but not delete) in Index A.
5	Reader	Admin	Can see which indices exist, can add/update/delete documents and users (up to their own level) in Index A and delete Index A itself.
6	Writer	None	Can create new indexes and users (with at most their own global role). Index A is invisible to the user.
7*	Writer	Metareader	Can create new indexes and users (with at most their own global role). Can read all properties, do queries, etc. in Index A, but cannot read the 'text' attribute.
8	Writer	Reader	Can create new projects and users (with at most their own global role). Can read all properties, do queries, etc. in Index A, but cannot make changes.
9	Writer	Writer	Can create new projects and users (with at most their own global role). Can add/delete documents, add/delete users (up to their own level), and make other changes in Index A, but can't delete Index A.
10	Writer	Admin	Can create new projects and users (with at most their own global role). Can add/update/delete documents and users (up to their own level) in Index A and delete Index A itself.
11	Admin	Admin	Can delete projects and assign themselves a role on any index role. Can do whatever they want, including deleting the index.

* Relevant for non-consumptive research

For the three avenues we described above for researchers who can not share their full data, role levels 2 and 7 in Table 1 are crucial. What is important to note is that query functionalities stay enabled even in levels where the user has no access to the full text, as shown in Figure 6a. For avenues two and three – metadata access and access to pre-processed data – this can be a valuable addition to just making the data available: what we described in the last section, the additional effort needed to obtain a data set from the Dataverse before it is clear if the data is relevant, is exponentially larger if the goal is to rehydrate a corpus give the URLs of pre-selected documents; in case of data that has undergone destructive pre-processing steps, the query function can provide invaluable validation capabilities as researchers can test if their assumption about the original data are correct. The simple counts, frequency plots and query functionality therefore fill an important need for researchers who want to use a dataset, as they allow for initial non-consumptive research on a corpus. The functions available through the dashboard are additionally mirrored in the API.

Additionally, to these twelve roles, we can control the default role users have on a given index (see level 0 in the table). We call this the guest role of an index. This way, it is possible, for example, to have one index where everyone can see the metadata while other indexes on the AmCAT instance are hidden. Users' global and index role and the guest roles of indexes can be controlled via the dashboard or the API. We show this illustratively in the example below.

However, first we need to highlight another setting that can modify access to data in `amcat4`: the authentication mode. The authentication mode of an `amcat4` instance controls who has access to the data in the first place. Unlike the roles, this setting can only be accessed by administrators through the command line on the machine where the instance is hosted. If you follow our recommended way of installation (see the online manual), `amcat4` will run in a Docker container. When invoked with the command `docker exec -it amcat4 amcat4 config`, an interactive configuration menu will guide the user through various settings as shown in Listing 1.

Listing 1 Authentication modes in amcat4

```
auth: Do we require authorization?
  Possible choices:
  - no_auth: everyone (that can reach the server) can do anything they want
  - allow_guests: everyone can use the server, dependent on index-level
      guest_role authorization settings
  - allow_authenticated_guests: everyone can use the server, if they have
      a valid middlecat login, and dependent on index-level guest_role
      authorization settings
  - authorized_users_only: only people with a valid middlecat login and
      an explicit server role can use the server

The current value for auth is AuthOptions.no_auth.
Enter a new value, press [enter] to leave unchanged, or press [control+c]
  to abort:
```

The explanations of the different levels should hopefully be clear, except the term **middlecat login**. For authentication modes beyond allow_guests, people whom you want to grant some kind of access need to log in. To make this more secure, we wrote

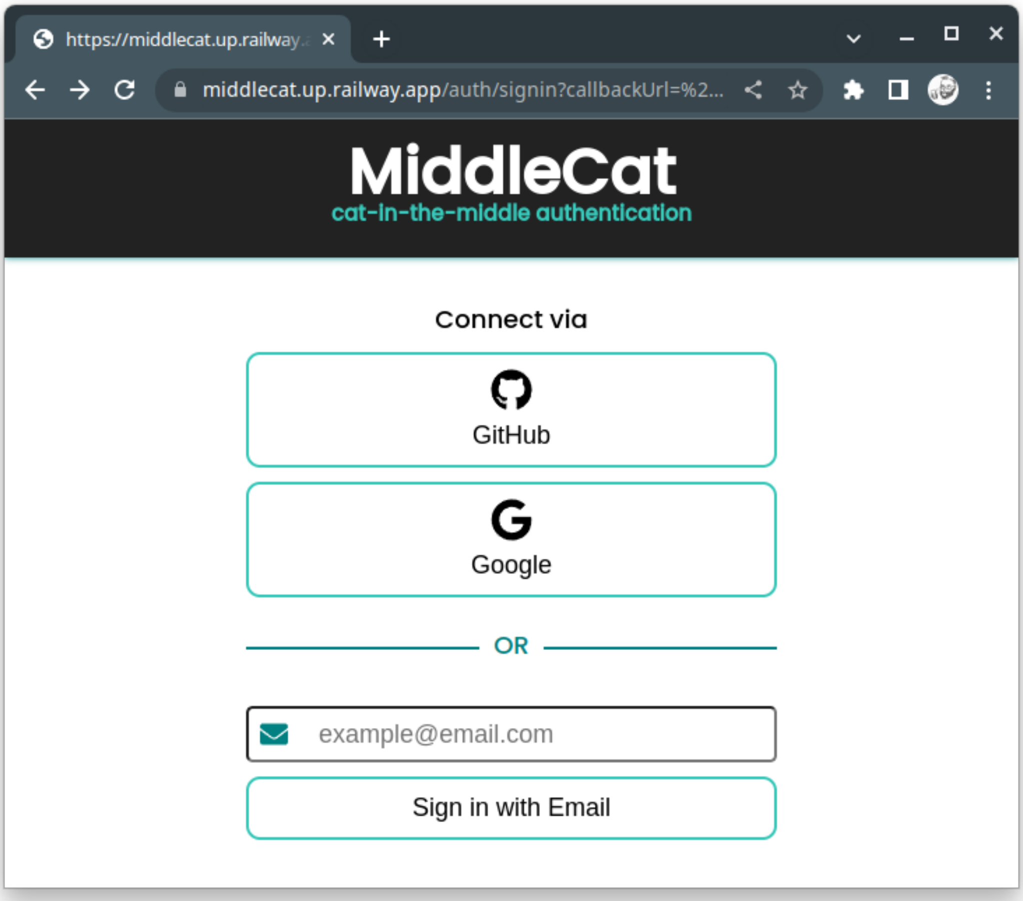

https://middlecat.up.railway.
middlecat.up.railway.app/auth/signin?callbackUrl=%2...
MiddleCat
cat-in-the-middle authentication
Connect via
GitHub
Google
OR
example@email.com
Sign in with Email

Figure 3: Middlecat login screen

our own authentication provider called **middlecat**. It enables the administrator to set up authentication via different identity providers like Google or GitHub, with a fallback solution to let users log in via a one-time email link (if they want to log in again, they need to request a new link). We host a middlecat instance at https://middlecat.up.railway.app, but as the software is open source, everyone can set up their own instance and negotiate with different identity providers to make other authentication options available (e.g., through their university or company).

In cases where non-consumptive research should be made possible, all modes except *no_auth* can be used. The indexes can then be configured to let users have a metareader role explicitly or via the guest role of an index, as explained above.

4.3 Non-Consumptive Research via the AmCAT Dashboard: an Example

As Illustration of the access control, we can use an instance where the authentication mode is set to *allow_guests* and where we added a test index. To reproduce this setup, one could spin up an instance of the AmCAT suite using Docker and Docker Compose as shown in

Listing 2 (you can find a more detailed explanation in the AmCAT manual)

Listing 2 Creating an AmCAT instance through Docker

```
# download our docker compose file with curl or manually
curl -O https://raw.githubusercontent.com/ccs-amsterdam/actioncat/main/
  actions/dfm/docker-compose.yml
# run docker compose to download and start the AmCAT applications
docker-compose up --pull="missing" -d
# create a test index to use in this example
docker exec -it amcat4 amcat4 create-test-index
# configure the instance to run in allow_guests (make sure to also set
  your email address as admin email, or you are locked out)
docker exec -it amcat4 amcat4 config
```

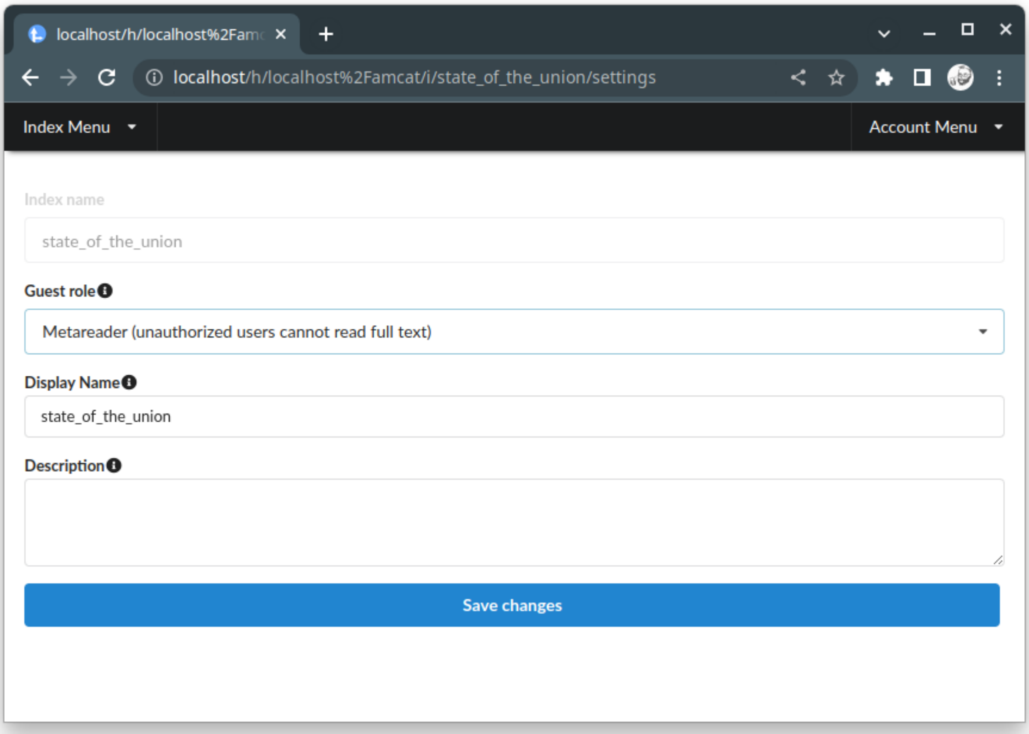
We can then change the guest role of the test index through the web dashboard at http://localhost/ (if you hosted the instance locally on your computer): First, users need to log into the dashboard as explained above. If their mail address, which is supplied either by the identity provider or directly, matches an admin account or has another role that can change the roles of other users, a setting page then becomes available to them, which is shown in Figure 4.

Switching perspectives for a second, we can look at an index, in this case the example data included in `amcat4`, from the perspective of a user. After selecting the index to which the user has access – *state_of_the_union* – they can perform a search as shown in Figure 5a. They can also switch from *Summary* to *Graph/Table* tab to compare document groups based on meta fields. In Figure 5b we show a frequency comparison of documents fitting the search "america and europa", which retrieves documents containing both terms. Party is a meta information of the texts, but can be used to group documents. For these simple frequency analyses and text query functions there is no difference for users with and without access to the text data. Only when a user attempts to access the full text – beyond the short snippet shown in the summary – is there a difference. Figure 6b and Figure 6a show the comparison between the view of a *reader* and *metareader* respectively.

In sum, users with no specific rights can still explore the corpus to a large degree. Through the fine-grained access control and the dashboard, AmCAT offers ways to query, visualize and analyze text data without giving users access to the underlying data. If they then decide that the data would be relevant for their research, they can try to obtain usage right from the data owners somehow. The administrator of the AmCAT instance can then grant them a *reader* role or above, so they could access the full text. Alternatively, they could use the meta information of the data set or a filtered subset of it, to 'rehydrate' the corpus from its original source. On the other side of things, AmCAT allows researchers to make data sets available for non-consumptive research where it is not possible to publicly share the full data due to copyright, privacy, or other concerns.

4.4 Packaged Non-Consumptive Analysis Workflows

While `amcat4` offers rudimentary analysis methods for users with *metareader* access like counting and cross-tabulating documents using queries, we also developed and addon framework to AmCAT that makes it possible to run entire analysis workflows in a non-


localhost/h/localhost%2Famc
localhost/h/localhost%2Famcat/i/state_of_the_union/settings
Index Menu
Account Menu
Save changes

Index name

state_of_the_union

Guest role

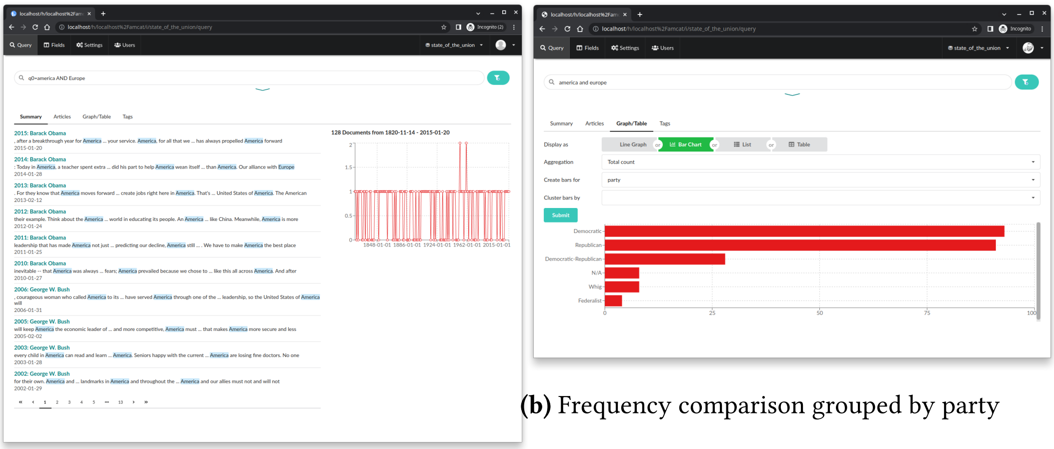
Metareader (unauthorized users cannot read full text)

Display Name

state_of_the_union

Description

Figure 4: Change guest role of index state_of_the_union



q0=america AND Europe

128 Documents from 1820-11-14 - 2015-01-20

2015: Barack Obama
, after a breakthrough year for America ... your service. America, for all that we ... has always propelled America forward
2015-01-20

2014: Barack Obama
: Today in America, a teacher spent extra ... did his part to help America wean itself ... than America. Our alliance with Europe
2014-01-28

2013: Barack Obama
. For they know that America moves forward ... create jobs right here in America. That's ... United States of America. The American
2013-02-12

2012: Barack Obama
their example. Think about the America ... world in educating its people. An America ... like China. Meanwhile, America is more
2012-01-24

2011: Barack Obama
leadership that has made America not just ... predicting our decline, America still ... . We have to make America the best place
2011-01-25

2010: Barack Obama
inevitable -- that America was always ... fears; America prevailed because we chose to ... like this all across America. And after
2010-01-27

2006: George W. Bush
, courageous woman who called America to its ... have served America through one of the ... leadership, so the United States of America will
2006-01-31

2005: George W. Bush
will keep America the economic leader of ... and more competitive, America must ... that makes America more secure and less
2005-02-02

2003: George W. Bush
every child in America can read and learn ... America. Seniors happy with the current ... America are losing fine doctors. No one
2003-01-28

2002: George W. Bush
for their own. America and ... landmarks in America and throughout the ... America and our allies must not and will not
2002-01-29

america and europe

Display as: Bar Chart

Aggregation: Total count

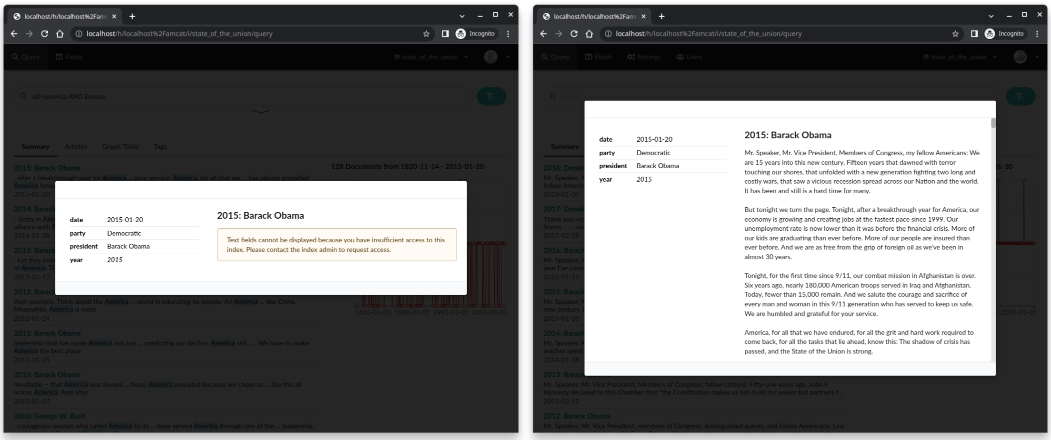
Create bars for: party

Cluster bars by:

(b) Frequency comparison grouped by party

(a) Frequency analysis

Figure 5: Analyses in the AmCAT dashboard



localhost/h/localhost%2Famcat/i/state_of_the_union/query

q0=america AND Europe

date	2015-01-20
party	Democratic
president	Barack Obama
year	*2015*

2015: Barack Obama

Text fields cannot be displayed because you have insufficient access to this index. Please contact the index admin to request access.

(a) User with no role

localhost/h/localhost%2Famcat/i/state_of_the_union/query

date	2015-01-20
party	Democratic
president	Barack Obama
year	*2015*

2015: Barack Obama

Mr. Speaker, Mr. Vice President, Members of Congress, my fellow Americans: We are 15 years into this new century. Fifteen years that dawned with terror touching our shores, that unfolded with a new generation fighting two long and costly wars, that saw a vicious recession spread across our Nation and the world. It has been and still is a hard time for many.

But tonight we turn the page. Tonight, after a breakthrough year for America, our economy is growing and creating jobs at the fastest pace since 1999. Our unemployment rate is now lower than it was before the financial crisis. More of our kids are graduating than ever before. More of our people are insured than ever before. And we are as free from the grip of foreign oil as we've been in almost 30 years.

Tonight, for the first time since 9/11, our combat mission in Afghanistan is over. Six years ago, nearly 180,000 American troops served in Iraq and Afghanistan. Today, fewer than 15,000 remain. And we salute the courage and sacrifice of every man and woman in this 9/11 generation who has served to keep us safe. We are humbled and grateful for your service.

America, for all that we have endured, for all the grit and hard work required to come back, for all the tasks that lie ahead, know this: The shadow of crisis has passed, and the State of the Union is strong.

(b) User with reader role

Figure 6: Comparison: *reader* and *metareader* text preview

consuptive way. We call this framework `actioncat` The basic idea we implemented is to leverage the open source Docker infrastructure, which we also employ to make the AmCAT suite of packages available, to let users create specialized workflows in order to perform analyses on data they do not have access to.

We offer two example *actions* (which is what we call predefined workflows that are packaged in a Docker image/container), one in `R`, one in `Python`:

- The `R` action adds a tidy document-feature representation field to an index
- The `Python` action adds a document embeddings field to an index

Both of these actions are destructive preprocessing workflows in the sense that the original text cannot be reconstructed from the new field. This makes these actions well suited for indexes where the full text can not be shared because of copyright, privacy or other concerns. Using AmCAT's finegrained access control features, the full text can be hidden from users without specific permissions, but the preprocessed data can still be shared with a wider audience. Users can imitate these examples to create their own workflows and send them to administrators of an `amcat4` instance. After approval, the administrator can then run a workflow using just two commands: one to download and one to run the action. Compared to sending just `R` or `Python` files for processing, this approach has the advantage that the action will have all the right dependencies already and perform the action exactly as on the user's machine (thereby standardizing the process to a certain degree and making the admins life a little easier).

4.4.1 Non-Consumptive Research via `actioncat`: an Example

To illustrate the basic steps from the administrator's perspective, we use one of the example workflows in the repository here: https://github.com/ccs-amsterdam/actioncat. First, we spin up an instance of the AmCAT suite using Docker and Docker Compose as in Listing 2. Administrators can then use actions with the same basic approach as shown in Listing 3.

In case of a user defined action, the only difference for the administrator will be the link to the Docker Compose file – which could also be sent via email, for example. The action will download the container image, start a process and this specific one will run until it has added a tidy document-feature representation to all texts in the test index. The user can access the results immediately as they are processed via the Web interface as shown in Figure 7a or using the code in Listing 4 to access the API. By changing the `docker-compose.yml` (shown in Listing 5), it is possible to control which index the action is applied on, the name of the text field and the name of the new dfm field by changing the environment variables.

Listing 3 Running the predefined dfm action

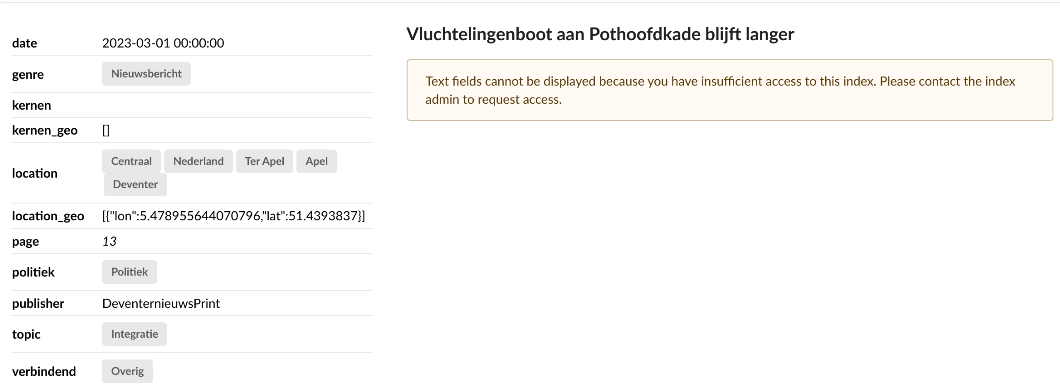
```
# download the action file with curl or manually
curl -O https://raw.githubusercontent.com/ccs-amsterdam/actioncat/main/
  actions/dfm/docker-compose.yml
# run the action
docker-compose up --pull="missing" -d
```

4.4.2 Non-Consumptive Research via `actioncat`: creating and analyzing annotations

This is just one of the possibilities with `actioncat` for non-consumptive research. You can also run custom analysis scripts that use the textual content to annotate a document, for example with topic classification, sentiment analysis, or geotagging. These annotations can then be viewed and queried using AmCAT, even without access to the texts that were used to create the annotations. As an example, Figure 7a shows a ‘metareader’ view of a document that has been preprocessed with location and topic information. The logged-in user has no access to the underlying text, but can still view and query the generated annotations.

5 Conclusion

As the gold mines of digital freely available text data are being more closely gated by the companies who own them, the field needs to rethink the value of data sets. When social media data was cheaply available, it might not have mattered much that corpora collected digital dust on the hard drives of most researchers after their research on it was published. Today, however, these corpora are more valuable than ever, as they can fuel the research efforts of scientists who have lost access due to limited funding and a lack of industry connections. In this contribution, we presented three avenues to sharing a corpus for secondary analysis even when it is not possible for legal or ethical reasons to share the full text of a corpus: versions of the text that have undergone destructive pre-processing, meaning that full texts cannot be reconstructed and hence can’t be consumed (Avenue 1) metadata, including identifiers of the documents, so other researchers can rebuild the same pre-filtered corpus about a topic (Avenue 2); or making the data set accessible for

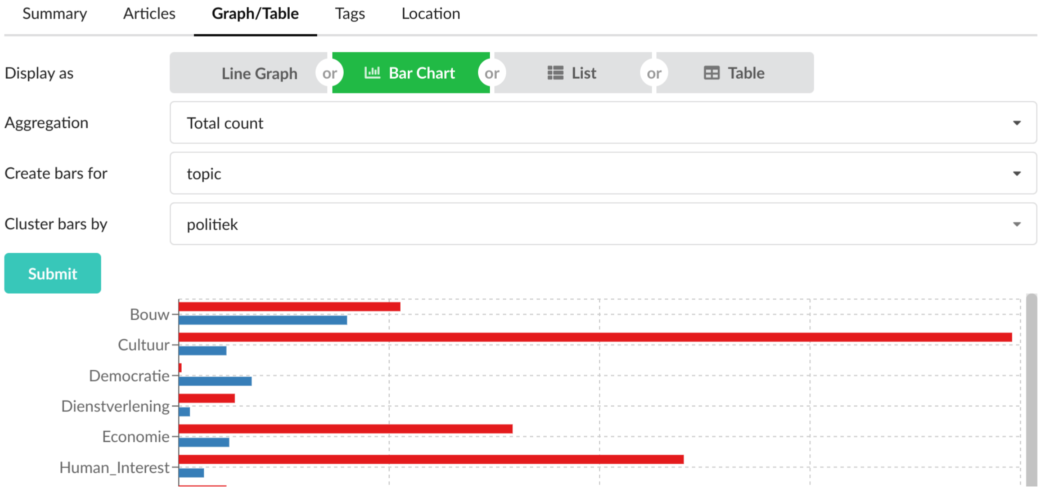


date	2023-03-01 00:00:00
genre	Nieuwsbericht
kernen	
kernen_geo	[]
location	Centraal Nederland Ter Apel Apel Deventer
location_geo	[{"lon":5.478955644070796,"lat":51.4393837}]
page	13
politiek	Politiek
publisher	DeventernieuwsPrint
topic	Integratie
verbindend	Overig

Vluchtelingenboot aan Pothoofdkade blijft langer

Text fields cannot be displayed because you have insufficient access to this index. Please contact the index admin to request access.

(a) Example metareader view of processed document



Display as: Bar Chart

Aggregation: Total count

Create bars for: topic

Cluster bars by: politiek

(b) The metareader can still query the text and analyse annotations

Figure 7: Comparison: An unauthorized user can see and analyse the annotations, even without access to the underlying text

Listing 4 Querying the new field through the API

```
if (!requireNamespace("amcat4r", quietly = TRUE))
  remotes::install_github("ccs-amsterdam/amcat4r")
library(amcat4r)
amcat_login("http://localhost/amcat")
sotu_dfm <- query_documents(index = "state_of_the_union",
                            queries = NULL, fields = c(".id", "dfm"))
sotu_dfm


    # A tibble: 232 × 2
       .id      dfm
       <id_col> <list>
     1 9d8...0d0  <list [3,370]>
     2 846...068  <list [2,176]>
     3 2b6...aa5  <list [2,895]>
     4 4f3...8bf  <list [3,172]>
     5 c36...4b0  <list [3,739]>
     6 5a2...8ba  <list [3,745]>
     7 8b0...5f2  <list [3,927]>
     8 484...893  <list [3,308]>
     9 a3a...a70  <list [2,554]>
    10 57c...840  <list [1,729]>
    # ℹ 222 more rows
```

non-consumptive research, meaning texts can't be consumed (read) by still analyzed via computational methods (Avenue 3).

We additionally discussed the fourth iteration of the Amsterdam Content Analysis Toolkit (AmCAT), which can be used to realize all of the discussed avenues. This means that researchers are not hindered by a lack of technical options to work around data sharing limitations. AmCAT enables researchers to share their data in a way that it is possible to query the database using (a combination) of keywords, allowing those who consider using a corpus for secondary analysis to check whether it contains relevant documents. As the query functions constitute *non-consumptive research*, they are enabled for data set with open as well as limited access (i.e., where users can't see the full text).

Additionally, we provide a framework that enables arbitrary workflows to be performed on data with limited access. Given approval by an administrator who has to check that a workflow does not return protected data, users can run any pre-processing or analysis scripts they wish, without ever having to see the actual data. They can then access a document-feature-matrix, embedding or whatever other version of the text they might want. This does not allow quite the interactive exploration that, e.g., HathiTrust's "data capsules" do, as workflows need to be pre-approved, but the framework is a lot more flexible in that any workflow can be packaged and sent to the admin of an AmCAT instance for quick and seamless review and authorization.

In sum, a crucial criterion for data storing infrastructure going forward should be scientific transparency to facilitate replication, collaboration, and efficient re-use of academic contributions in terms of data collection and tool creation. The features in the AmCAT enable researchers to make data available in ways that conform with copyright, privacy, or other concerns. Following the ideas developed within the concept of *non-consumptive research*, we can enable new analyses or replication effectively without sharing access to

Listing 5 Querying the new field through the API

```
version: "3.8"
services:
  action-dfm:
    image: ccsamsterdam/amcat-action-dfm:4.0.13
    build: .
    network_mode: "host"
    environment: # behaviour of the R script is controlled through these variables
      - amcat4_host=http://localhost/amcat
      - index=state_of_the_union
      - queries=NULL
      - text_field=text
      - dfm_field=dfm
    # for authentication, this container needs access to the httr2 cache directory.
    # You can find it with `rappdirs::user_cache_dir("httr2")`
    # volumes:
    #   - ~/.cache/httr2:/root/.cache/httr2 # [local path]:[container path]
```

the full material.

References

Atteveldt, Wouter van, Joanna Strycharz, Damian Trilling, and Kasper Welbers. 2019. “Computational Communication Science| Toward Open Computational Communication Science: A Practical Road Map for Reusable Data and Code.” *International Journal of Communication* 13 (0). https://ijoc.org/index.php/ijoc/article/view/10631.

Barba, Lorena A. 2018. “Terminologies for Reproducible Research.” https://arxiv.org/abs/1802.03311.

Brady, Henry E. 2019. “The Challenge of Big Data and Data Science.” *Annual Review of Political Science* 22: 297–323.

Center, HathiTrust Research. 2023. “Non-Consumptive Use Policy.” https://web.archive.org/web/20230906154601/https://www.hathitrust.org/the-collection/terms-conditions/non-consumptive-use-policy/.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

Dienlin, Tobias, Niklas Johannes, Nicholas David Bowman, Philipp K Masur, Sven Engesser, Anna Sophie Kümpel, Josephine Lukito, et al. 2021. “An Agenda for Open Science in Communication.” *Journal of Communication* 71 (1): 1–26.

Freelon, Deen. 2018. “Computational Research in the Post-API Age.” *Political Communication* 35 (4): 665–68. https://doi.org/10.1080/10584609.2018.1477506.

Grimmer, J., and B. M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts.” *Political Analysis* 21 (3): 267–97. https://doi.org/10.1093/pan/mps028.

Hennesy, Cody, and Rachael Samberg. 2019. “Law and Literacy in Non-Consumptive Text Mining: Guiding Researchers Through the Landscape of Computational Text Analysis.” *Copyright Conversations: Rights Literacy in a Digital World*. https://escholarship.org/uc/item/55j0h74g.

Hilbert, Martin, George Barnett, Joshua Blumenstock, Noshir Contractor, Jana Diesner, Seth Frey, Sandra González-Bailón, et al. 2019. “Computational Communication Science: A Methodological Catalyzer for a Maturing Discipline.” *International Journal of Communication* 13: 3912--3934. https://ijoc.org/index.php/ijoc/article/download/10675/2764.

King, Gary, Patrick Lam, and Margaret E. Roberts. 2017. “Computer-Assisted Keyword and Document Set Discovery from Unstructured Text.” *American Journal of Political Science* 61 (4): 971–88. https://doi.org/10.1111/ajps.12291.

Klein, Olivier, Tom E. Hardwicke, Frederik Aust, Johannes Breuer, Henrik Danielsson, Alicia Hofelich Mohr, Hans IJzerman, Gustav Nilsonne, Wolf Vanpaemel, and Michael C. Frank. 2018. “A Practical Guide for Transparency in Psychological Science.” Edited by Michéle Nuijten and Simine Vazire. *Collabra: Psychology* 4 (1). https://doi.org/10.1525/collabra.158.

Laurer, Moritz, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. “Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI." *Political Analysis*, 1–17. https://doi.org/10.1017/pan.2023.20.

Lazer, David, and Jason Radford. 2017. "Data Ex Machina: Introduction to Big Data." *Annual Review of Sociology* 43: 19–39.

Miguel, E., C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, et al. 2014. "Promoting Transparency in Social Science Research." *Science* 343 (6166): 30–31. https://doi.org/10.1126/science.1245317.

Nosek, B. A., G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, et al. 2015. "Promoting an Open Research Culture." *Science* 348 (6242): 1422–25. https://doi.org/10.1126/science.aab2374.

Rodriguez, Pedro L., Arthur Spirling, and Brandon M. Stewart. 2023. "Embedding Regression: Models for Context-Specific Description and Inference." *American Political Science Review*, 1–20. https://doi.org/10.1017/S0003055422001228.

Simon, Mónika, Kasper Welbers, Anne C. Kroon, and Damian Trilling. 2022. "Linked in the Dark: A Network Approach to Understanding Information Flows Within the Dutch Telegramsphere." *Information, Communication & Society* 0 (0): 1–25. https://doi.org/10.1080/1369118X.2022.2133549.

Sylvester, Christine, Zachary Greene, and Benedikt Ebing. 2022. "ParlEE plenary speeches data set: Annotated full-text of 21.6 million sentence-level plenary speeches of eight EU states." Harvard Dataverse. https://doi.org/10.7910/DVN/ZY3RV7.

Theocharis, Yannis, Pablo Barberá, Zoltán Fazekas, and Sebastian Adrian Popa. 2020. "The Dynamics of Political Incivility on Twitter." *SAGE Open* 10 (2): 215824402091944. https://doi.org/10.1177/2158244020919447.

Van Atteveldt, Wouter, Scott Althaus, and Hartmut Wessler. 2021. "The Trouble with Sharing Your Privates: Pursuing Ethical Open Science and Collaborative Research Across National Jurisdictions Using Sensitive Data." *Political Communication* 38 (1-2): 192–98. https://doi.org/10.1080/10584609.2020.1744780.

Van Atteveldt, Wouter, and Tai-Quan Peng. 2018. "When Communication Meets Computation: Opportunities, Challenges, and Pitfalls in Computational Communication Science." *Communication Methods and Measures* 12 (2-3): 81–92. https://doi.org/10.1080/19312458.2018.1458084.