The State of Digital Media Data Research, 2023

Dr. Josephine Lukito
Megan A. Brown
Ross Dahlke
Dr. Jiyoun Suk
Dr. Yunkang Yang
Dr. Yini Zhang
Bin Chen
Sang Jung Kim
Kaiya Soorholtz

MDDC
Media & Democracy Data Cooperative

The State of Digital Media Data Research, 2023

Dr. Josephine Lukito, Center for Media Engagement, University of Texas at Austin
Megan A. Brown, Center for Social Media and Politics, New York University
Ross Dahlke, Stanford University
Dr. Jiyoun Suk, University of Connecticut
Dr. Yunkang Yang, Texas A&M University
Dr. Yini Zhang, University at Buffalo
Bin Chen, Center for Media Engagement, University of Texas at Austin
Sang Jung Kim, Social Media and Democracy group, University of Wisconsin-Madison
Kaiya Soorholtz, Center for Media Engagement, University of Texas at Austin

February 2023

This work has been generously supported by the Social Science Research Council. All comments and recommendations presented here reflect the independent opinions of the authors.

Executive Summary

The purpose of this report is to provide an account of digital media data (DMD) research practices and to highlight its ongoing challenges. We define DMD as data that are collected, extracted, gathered, or scraped from a web-based platform such as a website, social networking site, mobile application, or another virtual space. We break these practices and their challenges into three stages—collection, analysis, and sharing.

We argue that continuing digital media data (DMD) research should be guided by four principles: collaboration, transparency, preparation, and consistency.

1. **COLLABORATION**: Working together on protocols for DMD research can occur across all stages of the research pipeline, including setting norms for data sharing, producing baseline research, and co-developing archives.
2. **TRANSPARENCY**: Make code and data accessible to other researchers, when possible. Open-source software development is especially helpful for advancing research in a transparent manner. Given the cost for collecting, storing, and analyzing DMD, we also encourage researchers to provide these details in their publications.
3. **PREPARATION**: Researchers should anticipate the risks or challenges to collecting, analyzing, and reporting on DMD. Emerging methods for data collection, such as donated data, provide a new mechanism for studying user-consented data.
4. **CONSISTENCY**: We end with this principle because it builds on the aforementioned three as, when researchers are collaborative, transparent and prepared, research approaches will be more consistent, allowing us to compare across studies and identify situational contexts that require nuanced protocol.

Acknowledgements & Disclaimers

This report was written up as a follow up to the 2022 Digital Data Conference, which took place April 14-15, 2022. Many things have changed between then and the time when this report was released, including U.S. President Biden's recently released Blueprint for an AI Bill of Rights, the adoption of the Digital Services Act, Elon Musk's acquisition of Twitter, and announcements of new API access by platforms such as TikTok and YouTube. These developments highlight the constantly changing nature of digital media research. Correspondingly, this report should be read as a snapshot in time—a reflection of practical and ethical challenges in 2022. We are grateful to those who have written on, and continue to write on these issues. A link to our references can be found in the appendix.

We would like to extend a special thank you to the Social Science Research Council (SSRC), who sponsored the Digital Data Conference and this report, as well as the Knight Foundation.

This report was a collaborative endeavor of researchers across multiple research centers and universities who are a part of the Media and Democracy Data Cooperative. In particular, we would like to thank the Center for Media Engagement at the University of Texas at Austin; the Center for Social Media and Politics at the New York University; the Institute for Data, Democracy, and Politics at the George Washington University; the Center for an Informed Public at the University of Washington; and the Center for Civic Renewal at the University of Wisconsin, Madison.

A disclaimer: in this report, we most heavily consider the use of DMD to conduct political communication research, though our findings are relevant to many disciplines and industries.

Media & Democracy Data Cooperative
Dr. Josephine Lukito, Megan A. Brown, Ross Dahlke, Dr. Jiyoun Suk, Dr. Yunkang Yang, Dr. Yini Zhang, Bin Chen, Sang Jung Kim, Kaiya Soorholtz

What is "digital media data" (DMD)

This research report is about digital media data (DMD), defined as **data that are collected, extracted, gathered, or scraped from a web-based platform such as a website, social networking site, mobile application, or another online space**.

Digital media data includes digital life data (e.g., text, image, and videos), digital trace data (e.g., timestamp, author information), and digitized data (e.g., books, radio recordings, and broadcasts shared online) (Lazer & Radford, 2017). Digital media data are valuable for industry, non-profit, citizen, and academic researchers seeking to understand how people communicate with one another.

As people rely more on digital media platforms to converse, digital media data becomes increasingly "big"--meaning that they often require computational techniques to understand and analyze (Parks, 2014). The large quantity of data collected digitally also needs different security and privacy considerations from traditional research data. This is especially true of social media data (one type of digital media data), which may contain information that can identify private citizens without their realization.

Because there are many types of digital media data, researchers must often make different decisions depending on the type of data they are working with. Despite this, researchers often follow a broad template to approach digital media data. We describe this process as the digital media data research pipeline.

The DMD Research Pipeline

Digital media data research can be broken down into three broad steps: first, the researcher collects the data. Then, they analyze the data using inductive and/or deductive approaches. Finally, once the researcher has applied their analytical techniques, the researcher then reports on the data.

The **data collection** stage has developed tremendously as researchers have found new ways to collect data from users and platforms. Some of these approaches require consent procedures, whether it is getting approval from a platform to collect from their application programming interface (API) or gathering consent from a user to get a data connection or capture their screen. Others are more independent or adversarial, such as the use of unofficial scraping techniques. Despite this variety, fiscal, ethical, and practical challenges remain.

Similarly, **data analysis** has improved with the rise of computational techniques. Advances in cloud storage have made it easier for researchers to handle large datasets and open source intelligence has made different analytical tools more accessible to researchers with limited resources. This explosion of research, however, is complicated when considering what is missing from these analyses.

Finally, when reporting on their findings, researchers must make decisions regarding how much data to share, and in what form. These **data sharing** practices require a difficult balance between transparent research practices, terms of service violations, and users' rights.

[1] Collecting Digital Media Data

Researchers have many ways to collect digital media data. *However, because of the piecemeal nature of these collections, the research community lacks financial resources, technical infrastructures, and standardized practices. In this section, we discuss current strategies for collecting digital media data, challenges to these strategies, and recommendations.*

[1.1] Current Collection Tactics

Below, in Table 1, we list seven types of collection strategies, as well as the advantages and disadvantages of each. It is worth noting that these tactics are not mutually exclusive—in one project, a research team may employ a combination of these tactics to collect their data.

Table 1: A typology of strategies for collecting digital media data

Strategy	Examples	Pros	Cons
Partnering with Social Media	Social Science One; Facebook 2020 election project	Consent of platform; potential access to otherwise inaccessible data	ToS limits to access and sharing; Nondisclosure agreements
Documented and Official APIs	Twitter API; YouTube API	Consent of platform; often free to access; easy to start data collection	ToS limits to access and sharing; replication crisis (Freelon, 2018); programming knowledge needed
Commercial Data Brokers	Salesforce; Synthesio; NewsWhip	Consistency of meta-data; accessible to non-coders	Cost; opaque/proprietary collection procedure
Third Party Research Archives	Internet Archive; Pushshift; MediaCloud	Often free to access; accessible to non-coders	Lack of control over collection procedure; data completeness may be unclear
Scraping	Selenium; rvest	Free; easy to start data collection	Often not ToS-compliant; requires substantial cleaning; programming knowledge needed
Undocumented API Usage	Third-party collections of Gab, Gettr	Free	Possibly not ToS-compliant; programming knowledge needed; replication difficulties
Hacked Data[1]	DDoS	Free; potential access to otherwise inaccessible data	ToS-violating; ethically questionable collection methods

[1] This is not an endorsement of using hacked data in research, but an acknowledgement that it is available and that scholars have previously considered using this data.

[1.2] Current Collection Challenges

While there are a variety of techniques that researchers can use to collect digital media data, three common issues emerge for most researchers. The first collection issue is that of cost: a high financial cost to accessing or using the data hinders access for researchers with limited resources. Even if a researcher uses a so-called "free" approach, there are often additional costs to storing and managing the data. The second collection issue is the broad legal and ethical risks to data collection, particularly for more independent data collection methods, such as undocumented API usage. The third collection issue is that of deleted data, particularly when studying content that is generally unwanted in a media ecosystem (e.g., violent content).

[1.2.1] Cost Challenges to Data Collection

Costs for accessing data can vary widely. While some platforms make their data accessible for free to researchers or journalists via APIs, researchers also have the option to pay for data from data vendors (e.g. NewsWhip or Brandwatch) or purchase data in bulk directly from the social media companies.[2] These contracts typically involve a researcher selecting parameters such as time range, keywords, account handles, or other fields (which vary by platform) to get a sample of posts. However, as we note above, it is not always clear or consistent what sampling strategy these brokers use to generate the datasets in the first place. Additionally, these contracts can be prohibitively expensive, often costing thousands of dollars for a single dataset.

Finally, regardless of the source of the data collection, collecting, storing, and analyzing digital media data can be incredibly expensive if researchers do not have the in-house computing resources for this type of data collection. Using cloud computing platforms (e.g. Amazon Web Services or Google Cloud Platform) can cost thousands of dollars per month. A single terabyte of storage can range anywhere from US $25-$1,000 per month, depending on the format and storage location of the data. This does not include the price that researchers incur when they move the data off of the cloud platform onto their local machine or elsewhere for analysis. On Amazon Web Services, moving a terabyte of data out of their cloud storage costs around US $92, as of October 2022. Using the most basic example, Twitter's public sample stream, a 1% random sample of all tweets, is approximately 25 gigabytes per day (uncompressed). First, collecting this data on a cloud computer would cost around US $5 per month depending on the platform. Storing this in volume storage (where it is easily accessible for research) would cost around $750 per month and would increase over time as more data is collected and added to storage. Finally, moving the data out of cloud storage and onto a local machine or university cluster for analysis would cost around US $100 per month of data. In addition, beyond storing text data, video and audio data can cost even more to collect and store due to its size. Overall, collecting digital media data can come with an exorbitant price tag.

[2] As far as we are aware and at the time of writing, Twitter is the only social media company that sells bulk data purchases to researchers.

[1.2.2] Legal and Ethical Challenges to Digital Media Data Collection

Conducting research involves important legal and ethical considerations generally, but the novelty of digital media data (relatively to traditionally collected data) means there are few frameworks and many potential challenges. Correspondingly, digital media researchers must carefully assess risks in every stage of a study, but especially in the early stages prior to data collection. Current collection tactics, as listed above, vary in their legal and ethical risks, both to internet users and to researchers.

Legal risks associated with digital media research include violation of "The Common Rule," ToS (Terms of Service), and NDA (Nondisclosure Agreements) (Mislove & Wilson, 2018). The Common Rule refers to the regulations for the protection of human subjects in research which are laid out in the US Department of Health and Human Services (HSS) 45 CFR 46. According to The Common Rule, for any research involving human subjects, researchers must 1) get the research reviewed by an institutional review board (IRB) before data collection, and 2) obtain informed consent from experiment participants. Even if a study may qualify for an IRB exemption, researchers should obtain a formal exemption from IRB. And yet, IRB expectations around digital media data research varies across universities, with little to no standardization.

Then, there are Terms of Service, which are written to protect companies and often contain terms that restrict use of services and data for users. Violation of ToS may cause legal consequences. Restrictions include how the site may be accessed, and whether data can be collected or shared. For example, crawlers and scrapers may be forbidden and cause a violation of ToS on some websites. NDAs are often introduced when researchers want to require data directly from online service providers. Most NDAs state that the data should not be shared publicly. Noticeably, NDAs are legal contracts and violating the terms risks legal consequences.

Ethical issues in digital research often speak to potential harm to participants. First, compared to studies that involve direct interactions with human subjects (e.g., surveys, experiments, focus-group interviews, etc.), obtaining informed consent becomes much more difficult in digital research. Second, in experiment studies that involve digital media data, ethical risks should be assessed if the study involves exposure to uncomfortable content, deception, or exposure to mis/disinformation that may influence one's attitudes on significant issues (e.g., vaccine effectiveness, presidential election). Third, researchers should be careful to publicly share the data they collected. Publicly available data does not mean users are willing to share their data to the public - *being in public* is not the same as *being public* (boyd & Crawford, 2012). More complicated and newer ethical challenges are emerging with the evolving techniques and tools used in DMD research. But our goal for doing ethical research holds - how to minimize potential harm to participants and to the digital space.

[1.2.3] Deleted Data Collection Challenges

DMD is often ephemeral: content that exists on a platform one day will not necessarily be there the next day. This is particularly true of unwanted digital media content, such as pro-violence or pro-harm content, encouragements of harmful practices (e.g., self-harm), medical misinformation, and hate speech. For unwanted digital content, studying this content can produce a dichotomy: researchers do not want deleted content to persist on these platforms, and yet they may want access to this data to understand its spread or to build system-wide

solutions. However, the ephemerality of the content extends to all digital media: content can be deleted when users are suspended or when they choose to remove their content. At a greater scale, content also disappears when platform services end, such as when Amazon Web Services suspended its web hosting services to Parler.

Despite these challenges, researchers and platforms alike has recognized the significance of studying this deleted content in certain contexts. For example, Twitter makes data from foreign information operations available through their Information Operations archive. Researchers have also sought to predict the likelihood that a post will be deleted in the future on platforms such as Instagram (Tinati, Madaan, & Hall, 2017) and Twitter (Volkova & Bell, 2017).

Challenges to research using deleted content emerge from the onset, in many ways. Practically, deleted content is difficult to find or collect. Ethically and from a security-perspective, researchers must also reflect on whether this data are reasonable to study. (For example, should the researcher study the data if it had been deleted, and does it matter if the platform or the user deleted the content?) Furthermore, collecting or retaining this content can constitute a Terms of Service violation risk.

[1.3] Recommendations for Advancing Data Collection

Drawing from these challenges, we recommend the following efforts:

First, it is essential to **collaborate on developing interdisciplinary ethical practices and guidelines**. For these principles to be effective, they should be widely applicable, regardless of discipline or analytical approach: general enough to have applicability across a range of fields, both inside and beyond the academy, and specific enough for researchers to apply it to their projects. Examples of potential communities suited to do this work include the Association of Internet Researchers (which has produced its third iteration of ethics guidelines), journals like *Big Data and Society* and *Social Media + Society*, as well as discipline-specific organizations and their affiliated conferences. More specific templates and resources will help researchers apply these principles to their context.

Such a collaboration can also help with **building data archives** that make data more accessible to researchers with fewer resources. Pioneering collaborative data collection efforts like SOMAR, MDDC, and the IRIE highlight the need for a variety of archives to address different challenges and collect data of varying levels of access. These collaborations, and smaller ones between groups of research centers, non-profits, and civic organizations, can also serve as an independent intermediary between informal data requests and fully public data sharing. Such data archives can also be helpful for storing and studying deleted content in a closed-group framework.

Finally, and this is always worth stating: if the data collection strategy poses risks to the quality of one's work or to the users being studied, **alternative research methods should be considered**. User volunteered data such as data donations, screenomics research, and other projects wherein individual users? give their data directly to researchers present new opportunities for digital media data research. Given the varied uses of digital media, it is important to recognize that a plurality of approaches, both independent and collaborative, will be necessary to understand the impact of digital media on society.

[2] Analyzing Digital Media Data

In this section, we discuss strategies for analyzing digital media data, often (though not always) using computational qualitative and quantitative methods. This variety of methods is both helpful and challenging, as there are few standards in this research, resulting in data representation issues and decontextualized analysis. Owing to the (understandably) outsized interest in harmful content such as pro-violence discourse, we also consider the consequences of studying traumatic digital content.

[2.1] Current Analytical Tactics

Current analytical tactics for quantitatively analyzing digital media data (DMD) broadly include natural language processing (NLP) for text data, computer vision for image and video data, and network analysis. For social scientists, NLP and CV are seldom used by themselves. Instead, features and variables are often extracted using these techniques which are fed into more traditional statistical methods. One trend in these tactics is that some researchers are quick to adopt the latest technical advances in analyzing text and image data (e.g. transformer models, CNNs). However, widespread adoption of these methods has not yet happened, and older methods remain the most popular (e.g. topic modeling, qualitative analysis of images).

Likely because of the ease of collecting text data, NLP remains the predominant method for quantitatively analyzing DMD. Some researchers have employed recent advancements in large language models (LLMs), particularly Google's BERT model for supervised classification of text (e.g. Cinelli et al., 2021; Moffit et al., 2021; Rains et al., 2021; Stormer-Galley et al., 2021). However, other methods continue to be widely used. Unsupervised machine learning, either using Latent Dirichlet Allocation (LDA) (e.g. Blei, Ng, & Jordan, 2003; Mahrenbach & Pfeffer, 2021) or Structural Topic Modeling (STM) (e.g. Roberts et al., 2019; Jiang et al., 2021; Yarchi et al., 2021), continues to feature in analysis for grouping texts together. Dictionary-based approaches also remain a key part of textual analysis, including the use of custom-made dictionaries (e.g. Borton et al., 2021; Markowitz et al., 2021) and more established dictionaries, such as LIWC (e.g. Argur & Gan, 2021; Evans et al., 2021; Rathje et al., 2021; Seraj et al. 2021) and VADER (e.g. Bathina et al, 2021; Dambanemuya et al., 2021; Matalon et al., 2021).

Although the amount of scholarship analyzing image and video data remains smaller than text, interest in these data has made considerable progress. This growing interest is seen in new books outlining how to use computational methods to study images (e.g. Casas et al., 2020) as well as special issues in academic journals, including *Computational Communication Research* (Casas & Webb-Williams, 2022) and *International Journal of Press and Politics* (Bucy & Joo, 2021). Although still nascent, images and video are proving useful for extracting unique features for analysis (e.g. Lu & Pan, 2022), including using off-the-shelves Open AI libraries such as Face++ to identify faces, low-level aesthetic features such as color and brightness, emotions, and objects (Peng & Jemmott, 2018),[3] supervised machine learning to classify images (e.g. Rossi et al., 2021; Steinert-Threlkeld, 2022), and unsupervised machine learning to reveal more advanced conceptual categories (Steinert-Threlkeld et al. 2022; Zhang & Peng, 2022).

[3] Despite the usefulness, we acknowledge ethical concerns associated with scientifically classifying gender, sex, and race based on facial recognition algorithms (Scheuerman et al., 2021). We hope that researchers continue discussions of ethical considerations as the field of visual machine learning advances.

Computer vision technologies can also be applied to videos. Communication researchers have already experimented with frame sampling methods to convert videos to frames first and then apply image-based methods (Lu & Pan, 2022). Future research on videos may also adopt "video as data" techniques from computer science, which has made significant progress in video retrieval (He et al., 2021) and video similarity learning (Kordopatis-Zilos et al., 2019).

Mixed-methods and qualitative analysis also continue to be important tactics for researching digital media data. Some scholars take a mixed-methods approach in answering research questions, producing novel insights that could possibly be missed when taking a single-method approach (e.g. Kligler-Vilenchek et al., 2021; Yan & Yang, 2021). Purely qualitative research also continues to contribute to scholarship on the digital sphere (e.g. Belotti et al., 2022; Cellard, 2022; Snelson, 2016). However, much of social science research using DMD continues to be quantitative, and more qualitative research could be beneficial in developing new research questions and analyzing existing questions in new ways.

[2.2] Current Analysis Challenges

While there is a cornucopia of research about digital media data, challenges persist: there are few baseline datasets, and computational approaches may over-reduce the data to the point of decontextualization. Below, we highlight three ongoing challenges to digital media data analysis: data representation issues, decontextualized analyses, and the focus on harmful or unwanted digital media content.

[2.2.1] Challenges with Data Representativeness

Owing to the uneven access to digital media data, research tends to skew towards digital spaces that make their data readily accessible. Thus, what we know about digital media comes from a narrow range of platforms. Such a challenge is most concretely seen in studies of social media. As illustrated in Table 2, for example, most published research in communication has focused on Twitter (n = 2,861 articles) and Facebook (n = 3642). Using the search box within each journal, we searched for mentions of social media platforms' names since the publication of the first issue of each journal (The count is the number of search returns).

Table 2: Most studied platforms in communication journals

	Twitter	*Facebook*	*YouTube*	*Instagram*	*WhatsApp*	*Telegram*	*WeChat*	*TikTok*
Social Media + Society	615	707	287	376	108	25	37	48
New Media & Society	887	1218	622	386	149	43	52	46
Information, Communication & Society	692	917	405	199	104	33	44	25
Political Communication	121	120	41	22	8	14	2	1
Journal of Communication	208	239	104	30	17	36	7	5

Communication Research	72	117	41	16	10	3	0	2
Journal of Computer Mediated Communication	94	160	70	10	3	5	2	0
International Journal of Press/Politics	172	164	73	41	29	9	3	12

As noted, the over-emphasis on some social media platforms is not just a matter of what researchers are interested in: it is also an issue of data accessibility and how we extrapolate our findings. The platforms most studies, including Twitter, YouTube, and Facebook, have mechanisms for researchers to collect data about these platforms. In contrast, platforms that are less studied—including WhatsApp and TikTok—tend to be more difficult to analyze, either because data collection tactics are limited or because the data are hard to analyze?.

A related data representation issue is that of modality: owing to the ease of storage and analysis, DMD research overemphasizes text data despite the popularity of audio-visual content on platforms such as YouTube, TikTok, and BitChute. Compared to natural language processing techniques, there are fewer resources for studying images, videos, and audio content. Similarly, multi-modal analyses (studies of multiple modalities of communication) are relatively new in the research (Xuanyuan et al., 2021; Singh & Sharma, 2021).

Lastly, existing DMD research heavily focuses on certain languages, particularly Romanized languages. Studying DMD originating in non-English speaking countries and conducting comparative analysis should be an important agenda.

[2.2.2] Challenges with Decontextualized Analyses

Computational approaches do not necessarily consider the technical and social contexts of how DMD is created. Often, analysis of DMD becomes decontextualized, treating the data as if it was created in a vacuum or generalizing too broadly beyond the specific circumstance in which the data was created.

First, literature on the algorithms that govern digital environments are seldom considered in much of the empirical research on DMD. Scholarship on algorithms and underlying metrics concerns Facebook (e.g. Meese & Hurcombe, 2021; Schwartz & Mahnke, 2021), Twitter (e.g. Bandy & Diakopolous, 2021; Russo & del Gobbo, 2021), Instagram (e.g. Fouquaert & Mechant, 2021), TikTok (e.g. Bhandari & Bimo, 2021; Klug et al., 2021; Peterson-Salahuddin, 2021; Schellewald, 2021; Zeng & Kaye, 2021), and news rooms (Christin, 2020). Yet, other studies using DMD from these platforms seldom consider this body of research. How can one truly understand behaviors in Facebook Groups without considering the algorithm which influences information exposure and encourages specific behaviors?

Relatedly, there has been a recent push to incorporate more descriptive analysis into digital media research. Namely, the creation of the *Journal of Quantitative Description: Digital Media*, which seeks to elevate research on "mere description" to better understand the world (Munger et al., 2021). This precise focus on description can possibly lead to a better understanding of the underlying platforms, systems, and user experiences.

In addition, the observational data that researchers acquire are likely not created for explicit scientific purposes. As a result, these data often do not easily translate into constructs that researchers want to study (Lazer et al. 2021). Furthermore, researchers should more thoroughly consider the generalizability of their conclusions (Lazer et al. 2021; Lovett & Munger, 2021), with the recognition that not all things need to be generalizable. Does the population in this study on this specific platform represent everyone on other platforms? Can the findings be generalized across different cultural, geographic, and national contexts (Matassi & Boczkowski, 2021; Correa & Valenzuela, 2021)? How time-specific are the findings? These considerations are especially important because variables, even basic demographics, are often not available to researchers. For example, age and digital literacy are key moderators of some effects of social media (Munger et al., 2021), but not very often are these variables accessible to researchers using DMD.

[2.2.3] Challenges with Analyzing Traumatic Digital Content

Social media platforms serve as a hub for user-generated content created by global contributors. Because everyone can create content on social media platforms, there is a broad spectrum of DMD, from socially beneficial content to social media posts most users find objectionable. Some social media users constantly upload harmful content, such as child pornography, gratuitous violence, and hate-filled messages (Arsht & Etcovich, 2018).

Although popular social media platforms provide functions to block or report content that negatively impacts one's psyche, some professionals are required to investigate undesirable digital media content. Content moderators should monitor extreme content that violates platform policies (Roberts, 2017). Journalists are under pressure to investigate misinformation or notorious content on social media to inform the public (Bessey, 2019). Researchers also examine digital media data containing violence or hate to inform the public and expand academic knowledge.

The emotional cost of content moderation work has been reported repeatedly (Steiger et al., 2021). Significant psychological damage is caused by repeated exposure to harmful content, including experiences of post-traumatic stress disorder (PTSD) and other mental issues (Ofcom, 2019). The psychological health of journalists working with user-generated content containing images of extreme violence is also at risk (Feinstein et al., 2014). Although there have been discussions about the psychological health of content moderators and journalists encountering extreme content, there still is a lack of discussion on how researchers are psychologically impacted by extreme content on digital platforms.

[2.3] Recommendations for Advancing Data Analysis

We recommend following efforts to address these problems, challenges, and risks in digital platform data analysis:

First, **researchers would benefit from more research comparing different datasets and establishing baselines**. Data validation techniques that help highlight data representation issues should not be relegated to appendices, but instead discussed systematically to synthesize challenges and propose uniform solutions.

Second, **research on a variety of social media platforms, mediums/modalities (e.g., image and visual data), and under-represented languages (e.g., non-Romanized languages) will help produce a fuller understanding of the digital media ecology**, supplementing the already rich literature using natural language processing to study text communication. As image-as-data and computer vision techniques become more readily accessible, fields studying digital media data are likely to experience a visual turn in the study of this content.

Third, **researchers studying harmful content should have access to mental health resources throughout the analytical process**. Previous scholarship has highlighted how occupations with exposure to trauma (e.g., content moderators, social workers) experienced greater degrees of secondary trauma stress (Gil & Weinberg, 2015; Ruckenstein & Turunen, 2020). Such concerns should be extended to researchers, particularly content labelers and those reading and consuming harmful digital content.

Finally, to overcome concerns of decontextualization, **researchers should continue to do mixed-methods and interdisciplinary research**, synthesizing literature from STEM, social science, and digital humanities disciplines. Qualitative and human-in-the-loop approaches will help ensure that data is not reduced to a view from nowhere. Given the increased effort involved with mixed-methods research, particularly synthesizing the literature across disciplines, publishers may also want to consider extended paper lengths or non-traditional formats to publish this work.

[3] Sharing Digital Media Data

In the third section, we turn our attention to the practice of sharing digital media data. Owing to pro-transparency and open science efforts, researchers have tried to make their data more accessible. However, with digital media data, these efforts must be balanced with a digital media users' right to privacy, as well as different platforms' Terms of Services. We explore this transparency-privacy dilemma and discuss approaches that researchers have taken to balance decisions about data sharing.

[3.1] Current Data Sharing Tactics

Researchers have a variety of strategies they can employ to share data with other researchers. This ranges from open science practices that make data publicly accessible to more informal means of sharing data between trusted collaborators. When making decisions about whether to share digital media data (and what variables or meta-data to share), researchers also balance data sharing decisions with concerns about TOS or privacy violations.[4]

The most public form of data sharing is to make the data fully available, or available with limited censorship of personally identifiable information (effectively distinguishing public figures from private individuals). For example, the Proceedings of the National Academy of Science has created several standard options for authors to share data, including sharing all original data or sharing anonymized data only.

While sharing data publicly is admirable at face value, some data collection strategies may limit this approach. For example, Twitter's terms of service limits data sharing for non-academic researchers in their Development Agreement Policy. In this paradigm, researchers can share identifiers (like the Tweet ID) that allow other researchers to "rehydrate" the dataset using platform APIs. However, as discussed, this presents challenges for data sharing and replicability because researchers cannot rehydrate deleted data.

Increasingly, social media platforms are also only making data accessible in more limited formats, such as Meta's Facebook FORT or access to the Twitter Consortium. Data sharing is often quite limited for partnership projects between academics and social media companies such as the Facebook 2020 election project.

Finally, researchers can sometimes share data with one another in an interpersonal form that includes (but is not limited to) sharing amongst research teams and informal sharing for the purposes of paper evaluations, replications, and person-to-person requests.

Regardless of the strategy used, researchers must grapple with a question of quantity: how much data should researchers provide? In the spirit of open science, researchers may opt to provide as accessible data as possible. However, raw data may provide identifying information that puts users at harm or violates terms of service. Thus, the transparency-privacy dilemma: how do researchers balance between making data accessible for replication while also maintaining the privacy of users?

[4] Many limits to data sharing stem from the data collection practices, as the use of APIs or a third-party tool may come with contractual obligations.

[3.2] Current Sharing Challenges

The range of options for sharing data aligns frustratingly with the balance between making data openly available and accounting for Terms of Service limits and expectations of privacy among users. In this section, we highlight the two poles of this balance: the first being the digital replication crisis and open science principles, the second being the importance of user privacy and informed consent.

[3.2.1] The Digital Replication Crisis

Replicability, defined as the verification of research results by independent peer researchers, is the bedrock of science because it undergirds the credibility of scientific claims. In the past decade, there has been heated discussion about the replication crisis, like in psychology (Shrout & Rodgers, 2018). Over the years, open science as a culture and a practice have advanced. To facilitate scientific replication, scientists should provide transparent information about the various stages of research: design, methodology, data, and analytic methods (Nosek et al, 2015).

So far, conscientious efforts from various social and natural science disciplines have been made to push toward the open science practices and conventions. For example, the Open Science Framework (OSF) created by the Center for Open Science (COS) (Foster & Deardorff, 2017) provides a tool for researchers to document various stages of the research process to promote "openness, integrity, and reproducibility" in scientific research. Besides creating the infrastructure in place to facilitate replication, journals should provide guidelines to incentivize open science practices, though journals across scientific disciplines have unevenly adopted the open science framework in their article submission guidelines (Nosek et al, 2015). Some journals like *American Economic Review* have mandated data sharing, and some journals like *Psychological Science* made it optional and provided incentives like "open data," "open materials," and "preregistration" badges.

Research leveraging DMD should without exception be replicable. In their seminal 2009 essay that articulates and establishes the then-nascent field of computational social science, Lazer and colleagues point out the importance of creating an open academic environment that lends itself to "critique and replication." Like in other fields, journals publishing digital media research using computational methods do not enforce uniform publication standards for replication. Variations in data access also creates challenges: even if a researcher shared all the identifiers of social media posts in the study, other researchers attempting to replicate the study might not be able to download the dataset from the social media platform due to lack of access.

While providing datasets is one critical step toward promoting replication, proper documentation of the research process is another key step that has been often overlooked. Gebru and colleagues (2021) put forth the practice of creating a datasheet for each machine learning dataset that "documents its motivation, composition, collection process, recommended uses, and so on." Such datasheets not only help the data producers reflect on their own data collection and generation processes, but also inform data consumers of the crucial background and features of the data. Similarly, Bender and Friedman (2018) advocate for including a data statement on the characteristics of data in all papers on natural language processing.

We would like to end with one caveat. Even in an ideal condition where journals encourage open science practices and researchers embrace those standards, scientific replication might not necessarily follow due to the equity issue. For instance, a researcher seeking to replicate a study might not be able to do so due to the high-performance computing resources required by the analysis yet unavailable to the researcher. In this respect, building open science is also dependent upon creating a more equitable research environment that provides access to the research process as well as distributes research resources more evenly.

[3.2.2] Privacy and Informed Consent

Gaining permission or informed consent to use digital media data for research has provided unique challenges and discussions among researchers. As one of the principles of research ethics, informed consent allows researchers to access full information about the research subjects. However, the nature of DMD has created difficulty in applying the traditional ethics code to online spaces.

The current challenges and debates include: first, DMD collection is often conducted before the purpose of data collection has been identified (e.g., collections via streaming archives). Second, given the collaborative and participatory nature of DMD production, content ownership can involve multiple parties and stakeholders. Furthermore, DMD often lacks clear (or verified) information about the users, which often involves minors and other vulnerable groups, making the informed consent process more difficult.

A fundamental question of these challenges is whether DMD is private or public, especially on social media (Fuchs, 2018). Social media users have all agreed to terms of service, including information about how their data can be used by third parties (Nunan & Yenicioglu, 2013). However, this does not indicate that research ethics are automatically granted: as scholars have noted, simply being able to access the data does not confer ethical use (boyd & Crawford, 2012; Proferes et al., 2021), especially given that users may not be aware their data can be used this way (Fiesler & Proferes, 2018) and, "the process of evaluating the research ethics cannot be ignored simply because the data are seemingly public" (boyd & Crawford, 2012). In other words, there can be numerous factors to consider: is content posted on a public forum, community or platform, or private network? Does the content aim for public visibility or who is the intended audience? Does the content include sensitive topics (e.g., political violence, trauma, etc.)?

In this regard, the research practices differ across methodologies, resulting in little agreement. For example, qualitative researchers who use in-depth interviews or online ethnography find potential research subjects on digital platforms, with opportunities to directly ask for permission for research participation. When digital media data research involves a massive quantity or computational method, such an approach is not feasible. In many studies, aggregate-level data is used and reported. Research has also suggested de-identifying ordinary users, by either not mentioning usernames or using pseudonyms, except for well-known public figures or organizations (Fuchs, 2018). Such approaches can minimize the privacy issue, but always need to be used with caution as identifying information can still be available.

[3.3] Recommendations for Advancing Data Sharing

Drawing from this dilemma, we propose the following recommendations:

First, **researchers should establish norms around de-identifying social media and requesting access to quoting individual posts**. One common threshold used by Twitter researchers, for example, includes anonymizing users who are not verified and have fewer than 5,000 followers (see Hua, Ristenpart, & Naaman, 2020; Yin et al., 2018) or are not public figures, including elected officials and self-described journalists (Kennedy et al., 2022).[5] Some archives, such as 4cat, embed this process in their collection strategy. It is worth noting that one threshold will not be sufficient to account for a range of platforms and communication dynamics. We therefore recommend that researchers make these decisions on a project-by-project basis.

Second, when possible, **researchers should make every best effort to make data and coding materials accessible**. In instances where sharing the data is not possible, researchers can still provide the code used to collect the data, for example. When sharing data, researchers should strive to adhere to the FAIR principles: data should be findable, accessible, interoperable, and reusable.

Finally, the digital media data research community should **provide academic or industry credit for producing open data ethically or software to assist with data collection**. Some conference proceedings do accept datasets as a contribution (e.g., the Pushshift dataset in Baumgartner et al., 2020). Alternatively, there may be a benefit to developing a publication akin to the Journal of Open Source Software, but for data. More formal mechanisms of sharing deleted data will also advance research on harmful digital content; however, additional collaborative infrastructure would be needed to protect researchers doing this data.

[5] Kennedy et al. anonymize accounts that "1) are not verified; 2) are not public figures, including elected officials and self-described journalists; and 3) had<250,000 followers" when the tweets were collected.

General Recommendations

The state of digital media data research is both hopeful and precarious: owing to the constantly changing nature of not one, but multiple digital media platforms (across different governmental and cultural contexts), researchers need to continually adapt their data collection, analysis, and sharing strategies. At the same time, our conference and ongoing research highlight the need for a set of standards and practices that can produce more broadly comparable and practically useful research.

In this report, we highlight challenges and potential solutions that are specific to each of the stages of digital media data research: collection, analysis, and sharing. Below, we summarize recommendations that can apply to digital media data research more broadly. While less specific than the recommendation of each section, they can also serve as guiding principles for evaluating one's own research practices.

Be collaborative.

Collective buy-in is essential (after all, what is the purpose of a standard if no one adheres to it). Collaborating on guideline development, data archives, and institutional support for DMD research will ensure a greater degree of consistency. In a related vein, researchers should be open to a plurality of DMD research, inclusive of both collaborative and independent work.

Be transparent.

While there may be restrictions to the data being specifically provided, researchers can still be transparent about their collection, analysis, and sharing approach, whether it is making their code accessible and reproducible or providing cost estimates regarding their research. Additionally, conferences and journals can promote ethics and transparency statements that require researchers to be self-reflexive about their approach (see Chancellor et al., 2016, for one such example).

Be prepared.

As we note in our executive summary, the internet is always changing. The data we study, and our approaches to studying them, must continually be adapted to respond to these changes. For this reason, it is essential that researchers be prepared with alternative plans for data collection and risk assessment plans for a study's design. Researchers should also be prepared when public engagement about one's research invites harassment.

Be consistent.

While we recognize that different projects and studies will require different protocols, standard conventions are essential for a young discipline to succeed. Many recommendations here– whether producing guidelines, data archives, baseline research, and norms for data sharing– focus on standardization at the research-team level. However, for consistency across a field to occur, their researchers must be prepared, collaborative, and transparent.