Yale MACMILLAN CENTER
Genocide Studies Program

Mass Atrocities in the Digital Era Initiative (MADE) *Working Paper No. 3*

March 2022

Digital Archive of Memorialization of Mass Atrocities (DAMMA) Workshop Whitepaper[1]

Authors:[2]

Daniel Bultmann[i], Mykola Makhortykh[ii], David Simon[iii], Roberto Ulloa[iv], Eve M. Zucker[iii]

Abstract:

This whitepaper addresses the prospect of creating a digital archive for the memorialization of mass atrocities (abbreviated herein as DAMMA). It is based on the proceedings of a virtual workshop held in October 2021 that addressed questions regarding the scope, form, usages, and development of such an archive.

To cite:

Bultmann, Daniel, Mykola Makhortykh, David Simon, Roberto Ulloa, & Eve M. Zucker. (2022). Digital Archive of Memorialization of Mass Atrocities (DAMMA) Workshop Whitepaper. Yale University Genocide Studies Program: Mass Atrocities in the Digital Era Working Paper #3.

---

[1] We thank each and every one of the participants of the workshop for sharing their fascinating projects and their expert perspectives on the project. The full list of names is provided within the whitepaper itself.

[2] The authors' names appear in alphabetical order to connote the equal contribution of each to this whitepaper. Their respective affiliations at the time of publication are as follows:
[i] Humboldt-Universität zu Berlin, Germany
[ii] Universität Bern, Switzerland
[iii] Yale University, USA
[iv] GESIS - Leibniz Institute for the Social Science, Germany

Table of Contents

Introduction

The growing availability of information about mass atrocities both past and current enables new possibilities for individuals and societies to access the memories and histories of these tragic events and to form new meanings and narratives that shape how we remember those events. The ease of digitizing documents and other archival material has created an explosion of resources for researching the traumatic past independent of institutional support in a manner unimaginable three decades ago. At the same time, there is an increased ability to share information, such as through open access digital archiving and platform-based grassroots memory practices, which has resulted in webs of relationships between various actors and the cross fertilization of ideas and understandings of different instances of mass atrocities. This collaborative effort in remediating and reinterpreting the past can lead to new narrative constructions based on the information and documents found. Because of the sheer scale of information available, there are a myriad of narratives that may be employed by individuals, states, groups, or societies for varying purposes. Of course, a wide range of narratives was possible prior to the abundance of digital resources but a key difference is the wider access to various anchors for narratives and possibilities for wider circulation of the narratives themselves.

These changes present conceptual challenges for traditional ways of understanding the notions of *memory* and *archive*: compared with analogue data, content on the internet can be ephemeral and unreliable. Emergent narratives, which quickly spread across the platforms, are not always based on fact: they may represent only portions of the truth or may be blatantly false. Simultaneously, the relative ease with which one can collect data on the web allows for individuals or groups to assemble portions or versions of the past that suit their interests or purpose. These assemblages may come to serve as historical archives, memorials, or both. Finally, the prospect of an expanding digital "metaverse" adds potentially a new virtual space for memorial activities.

In this context, there is a strong need for a *digital archive for the memorialization of mass atrocities (DAMMA)*.

The components of that concept can be defined as follows: An *archive*, a "place for the storage of documents and records"[3], is *digital* when its content is stored in "the form of especially binary numbers" [4], or, in practical terms, when it can be stored in a local or cloud-based virtual storage. The concept of archiving is then in close connection to that of *memorialization*, "the act of marking a physical or conceptual space for posterity in remembrance of a person or an event"[5], as an archive could serve as such space. Finally, the preservation and remembrance of *mass atrocities*, that is, the "large-scale, systematic violence against civilian populations"[6], can be achieved through the archive and the memorial expressions within it. The archive of mass atrocities would thus provide digital drawers of memory, within which are stored the lives of those who were lost and the experiences of those who suffered. In the scope of this conceptualization, this paper summarizes a virtual workshop held in October 2021 on the topic of how to conceptualize a DAMMA, as well as how to prepare for its construction.

I. Background of Project

The initiative to build an archive of mass atrocity memorialization on the web grew out of a need that Eve Zucker and David Simon identified in the course of a project they were working on concerning digital memorialization of genocide and mass violence that began in 2018. David is a political scientist and expert on the genocide against the Tutsi in Rwanda and Eve is a sociocultural anthropologist with expertise on memory and social healing in the aftermath of the Cambodian genocide and more recently the Holocaust. As their research progressed, it became apparent that the memorialization of mass violence was increasingly prevalent on websites, blogs, and on social media. Web search engine results (e.g., Google) rendered limited results, and it was clear that these results were not necessarily representative of the memorial activities taking place on the web. It is not surprising given that web search is often driven by commercial interests and relies on user-tailored search

[3] Featherstone, M. (2016). Archive: Theory, Culture & Society.
[4] "Digital." 2022. In Merriam-Webster.com. Retrieved Jan 4, 2022, from https://www.merriam-webster.com/dictionary/digital
[5] Moncur, W., & Kirk, D. (2014, June). An emergent framework for digital memorials. In Proceedings of the 2014 conference on Designing interactive systems (pp. 965-974).
[6] Straus, Scott. *Fundamentals of Genocide and Mass Atrocity Prevention*. United States Holocaust Memorial Museum, 2016.

algorithms which do not necessarily take into consideration ethical considerations relevant in the case of retrieving information about mass atrocities.

David and Eve thus sought another way in which sites and content could be located and collected. A database or archive provided a solution to this problem if they could find the means of collecting the data in a manner that would overcome the limitations of commercial search engines. However, neither of them had experience writing or using programs that would scrape this type of data from the web. Seeking assistance with the programming challenges associated with this emergent task, they reached out to their colleague, Daniel Bultmann, also a scholar of genocide who was working on a project to collect data from social media relevant to the Cambodian genocide. After several discussions at Daniel's suggestion, they invited Mykola Makhortykh and Roberto Ulloa to provide the technical expertise necessary for such a project. Mykola is an Alfred Landecker lecturer who combines computer science, humanities, and communication science to study how the adoption of online platforms and algorithmic systems affects information-seeking behaviors, including the ones dealing with mass atrocities. Roberto has experience in collecting online data using diverse methodologies (web tracking, automatic browsing, and APIs), as well as facilitating access to the collections.

The now-five scholars soon realized the complexities that a digital archive of mass atrocities entails and sought to gain insight and decided that they would benefit from learning from the experiences of those already working on digital archive projects. To this end they organized a workshop with several members from this community of digital archivists and invited them to share their work and address a number of specific questions that were prepared relating to their projects. The response was overwhelmingly positive resulting in a total of twelve participants (beyond the five organizers) contributing to the workshop. (See Table 1 for a list of scholars who contributed to the workshop). This whitepaper reflects the conglomerate of ideas that emerged during this workshop through the presentations, discussions, and activities that ensued.

Table 1. Participants of the workshop. Name, affiliation, and project of the participants of the Digital Archive or Mass Atrocities workshop.

Participant	Institution	Project
Paul Salmons	Paul Salmons Associates	Holocaust Denial and Distortion, UNESCO
Ed Summers	University of Maryland	Document the Now
Niels Brügger	Aarhus University	Web Archives Studies network
Alexandra Drakakis	Madison Square Garden, Archives	9/11 Museum Digital Archive
Daniel Gomes	FCCN - Scientific Computing Unit of the Foundation for Science and Technology	arquivo.pt
Elisabeth Fondren	St. John's University	Archiving and Preserving Social Media at the Library of Congress
Alex Thurman	Columbia University Libraries	Web Resources Collection Coordinator
Pamela Graham	Center for Human Rights Document and Research, Columbia University	Human Rights Web Archive-Archived Index
Jonathan Bright	Oxford Internet Institute	Holocaust Denial and Distortion, UNESCO
Katrin Weller	GESIS - Leibniz Institute for the Social Sciences	Archiving the German web
Roland Moerland	Maastricht University	Webs of denial: A discourse network analytical approach to genocide denialism
Victoria Walden	University of Sussex	Digital Holocaust Memory
Heather Mann	University of Oxford	Holocaust Denial and Distortion, UNESCO

Despite the benefits of the extensive volume of memorialization-related data and the connectivity between various actors that have resulted from the digital turn, there remain a number of challenges for understanding this turn's long-term effect, in particular in the context of memory about genocide and mass atrocity. Several of these concerns are the management of information considering the ease with which records in digital form may be deleted or lost, the sheer abundance of information making it difficult to organize or interpret, and the use and prevalence of commercial algorithms for the purpose of information curation that might interfere with their use in particularly sensitive domains. Outside of the management of digital information there are additional challenges stemming from the often-siloed conversations and knowledge clusters by those accessing and working with mass atrocity data. Because of these challenges there is a necessity for the dialogue

concerning digital memorialization to address a number of questions that are summarized in Table 2.

Table 2. Questions of a Digital Archive of Memorialization of Mass Atrocities. List of questions regarding the implementation of a digital archive of memorialization of mass atrocities organized in categories.

Category	Questions
Scope	● Which online sites, activities, or collections qualify as something that might be called a memorial to a mass atrocity or a memorialization practice about a mass atrocity? ● Who owns a digital memorial? Do they belong to individuals who pay hosting? To the community? To the historians studying them? ● Are there any cases that do not constitute a memorial or memorializing activity based on the identity of the creator of the site or the practitioner of memorialization? ● Are there individuals or groups who create or engage with online memorial activities who because of their political agendas or belief sets are disqualified from having their content considered to be memorialization? ● Are denialism and other counter-narratives a form of memorialization? ● Does this depend on who created them? ● Should newer forms of digital content such as games and second life / metaverse be included?
Form / Implementation	● What would a digital archive of mass atrocities look like? ● Would it be a collection of social media posts on a mass atrocity over a given time, or an online historical archive of a genocide, or a series of blogs or some portion of the above? ● If online games constitute or incorporate memorials to genocide or mass atrocities (for example within “massively multiplayer online role-playing games” - or MMORPG), how can the data on the platform be archived? Does it bear equal weight as an institution creating an online commemoration to a genocide for example?
Implications	● How might DAMMA influence memorial practices now and in the future? ● How might the selection process inadvertently legitimize some memorials and forms of memorialization and sideline others? ● Would the structure of the archive influence how its component parts were perceived and used? ● How does our collection change the understanding of the atrocities by placing events or information in connection with one another either directly or through findings resulting from investigations on the archive? ● How can we enable better collaboration and knowledge sharing among scholars and researchers and at the same time build strong links to the public so that the archive would serve the needs of not only academics, but also victims and society at large?

The rest of the whitepaper is structured as follows. First, it discusses three different groups of motivations behind the development of DAMMA: the academia-, the victim- and the society-based ones. It then looks at different forms of (digital) archives which can be used for the implementation of DAMMA. This is followed by the discussion of different sources of

data which can be included in DAMMA as well as the aspects which are to be taken into account when collecting them. Then, the paper examines methods for collecting and archiving data together with different forms of public interfaces for using DAMMA. It ends with the discussion of challenges for organizing DAMMA, followed by the summary of the whitepaper and the discussion of the next step of the project.

II. Motivations

Different types of archives serve different purposes. Three clusters of motivations emerged from the workshop: 1) academia-centered motivations (material for future scholars / possibilities for research / possibilities for teaching); 2) victim-centered motivations (memorials / therapeutic benefits); and 3) society-centered motivations (preservation of materials which might disappear / fulfilling of ethical responsibilities for the victims / preventing societal fragmentation and radicalization). Each of these clusters has a distinct set of implications which will be discussed below. As discussed later these motivations link up to some or all iterations of the archive as a tool for scholars, a historical resource for victims, and as a resource for educators.

i. Academia-centered motivations

One of the central aims of an archive (and the ultimate motivation that started this project) is to provide scholars and researchers the means with which to observe, analyze, and to some degree understand what types of memorial activities concerning mass atrocities occur in online spaces, by whom and to what end. Recognizing the limitations of ad hoc searches, often over commercial search engines, which produce only limited (and sometimes questionable) data, an archive or database of online mass atrocity memorialization offers a superior method of capturing and cataloging the data for research purposes. It might be of particular use for scholars from several academic fields including (but are not limited to) historians, anthropologists, political scientists, sociologists, computational social scientists, media and communications scholars, cultural studies scholars, and cultural geographers. Additionally, DAMMA could serve the interests and needs of professionals and practitioners who work on topics connected to mass atrocities and trauma such as museum curators,

librarians, policy makers, and psychologists studying and treating trauma. Members of the public seeking information about a mass atrocity event and its representation online would also benefit from the archive.

A second motivation is to create a resource for future historians and other scholars who desire to study the way in which past atrocities were remembered and framed at a particular moment in time. Such a resource can help answer a number of questions, such as what narratives are employed to remember past atrocities? What symbolism is employed to frame past suffering and what local or global cultural dimensions does it engage with? What references are used to interpret the traumatic past? With this in mind, the point was raised at the workshop that the emphasis should be on collecting more, not less, of the web. Having a collection of forms of online memorialization of particular mass atrocity events taken in whole or in part offers a brilliant opportunity for future scholars to travel back in time by getting something close to a holistic sense of what people were thinking and what they were doing in regard to managing and remembering mass violence. The capacity to fulfill this motivation presumes enduring future accessibility to the results of the archiving process.

Third, a digital repository of memorialization of mass atrocity could help support efforts to understand how perpetrators/deniers of past atrocities use (and abuse) digital memory for their own respective aims. This was the motivation in fact of the UNESCO project, which collects instances of antisemitism in the forms of distortion and denial on the web and allows the researchers to see what tropes and memes are employed by various groups and individuals over the platforms included in the study.

A final motivation for academics is that such an archive could open new possibilities on how we teach about mass atrocity events and their aftermaths (particularly those aftermaths focused on memory and trauma). An archive of archives and memorials could open pedagogical possibilities for educators seeking to incorporate knowledge of (and lessons from) historical episodes of mass atrocities into their curricula.

ii. Victim-centered motivations

Archives of mass atrocities may also serve as memorials for victims and a form of remembrance of the individual suffering by sharing victims' identities and experiences during the mass atrocity events through the records held in the archive. For example, the Arolsen Archive recently made the decision to call itself an archive *and memorial*. To this end they created both the #everynamecounts crowdsourcing initiative to digitize all records of individuals from their analogue archive and the meta-memorial "A Paper Monument," which memorializes the Arolsen archive itself. Behind the shift from the traditional archive to the archive-memorial was a recognition that in some cases the records were the only memorials available to the victims. In those cases, the archive represents an entity of collective memory of the atrocity. These two dimensions of the archive-memorial – that is the individual records (e.g., photos and documents) within the archive and the collection as a separate mnemonic entity - both serve as a mechanism for remembering those who suffered and perished in genocides and mass atrocity events.

Archives as historical repositories provide victims and their families with a resource where they can search for information about the past including in some cases the plight of their own family members. In addition, such archives may facilitate contact between survivors and other individuals or entities connected to that time period. Connecting to the past through its history and/or people who share in that past may also convey therapeutic benefits to victims and generations that follow by opening dialogue and forming the narrative that meets the needs of survivors by perhaps granting them the sense that the past suffering is not forgotten, and the future generations can learn from its painful lessons.

Survivors and their descendants may also find meaning in participating in the selection of what is included in the archive as was discussed during the workshop. Here the archive also becomes a shared initiative where members of the community choose how they and their past will be represented. In this manner, those groups represented in the archive have some control over what is shared and therefore to some extent control the narrative.

iii. Society-centered motivations

In addition to the benefits for scholars as well as victims and their families, DAMMA can also convey benefits upon the societies in which the atrocities occurred and other societies where survivors have settled. Through the archives, societies may find ways of articulating the traumatic events and helping those who lived through them by acknowledging what they experienced and taking measures to prevent such conditions that allowed the mass atrocity to occur in the first place. By creating and expanding initiatives like DAMMA, the sometimes-atomized gestures of remembering have the potential to come together to make a larger collective statement about the ways in which a given society approaches past suffering.

An additional societal benefit raised in the workshop relates to the preservation of material that might otherwise disappear. Regarding the former, forms of digital material do not always last both online and offline due to changes in technologies, lack of maintenance, the disappearance of online websites over time, and the easy deletion of content whether purposefully or inadvertently. Moreover, the dynamic and subjective nature of the web contributes to the ephemerality of content on the web (as noted at the workshop). The creation of a digital archive of mass atrocities would mean that certain data can be captured and preserved (provided that it is properly maintained on a server and/or using an archiving service). To do so would provide a social benefit, serving an ethical imperative to remember the victims of past atrocities in perpetuity to keep the digital embodiments of their memory and to facilitate future efforts to understand the roots, process, and impact of those atrocities.

Furthermore, the web itself in many ways is a manifestation of who we are as a species and as a culture in this moment. As we capture a portion of the web, we preserve not only what we find, but equally we preserve our choices and methods of remembering. In such a manner, the archive may allow for the continuity of dynamic engagement with the representations of the past found in the archive. Moreover, the atrocities themselves as they are represented in the archive are ensured some degree of continuity in that the archive preserves them. For some of the projects shared at the workshop the idea of preserving who we are through the archiving process manifested the *raison d'etre* of the project.

III. Forms of archives

Throughout the workshop and in response to our discussions, we identified three forms of archives that cater to building one large digital archive of mass atrocities, as discussed above. While the underlying database remains the same, the three forms differ in purpose and audience and hence in modes of access, levels of involvement of the affected communities and institutions, as well as degree of data contextualization.

The first database form serves as a research tool for varieties of atrocity remembrance and denial, as exemplified by the UNESCO Holocaust Denial and Distortion project headed by Paul Salmons, Jonathan Bright, and Heather Mann. As a lasting tool open to research only, it is less contextualized and features greater access restriction than the other two forms. The emphasis in this data provision lies on the source that is archived (its subject, date, location, type, etc.) so that researchers who use the archive can easily find and organize information during searches on given subjects. To some extent, this first database places additional emphasis on a source's connection to other sources of similar material. Although this form of archive is more comprehensive and less contextualized than the others discussed below, it still falls under the category of an archive, as information is selected, stored, and provided according to set principles and methodologies. Accordingly, it thus does not differ from other archives that are part of digital memorialization in general and those envisioned for this project. This also means that this database form is an online tool and platform that shapes the digital memorialization of mass atrocities; to a certain degree, it even co-creates what it studies.

This aspect of co-creating online memorialization is most obvious in the second form of database, which serves as a historical archive repository with the purpose of recording and preserving collections of a subject, case, or event-specific material online. This historical archive is less restricted in access and provides a certain degree of contextualization for the pre-defined organization of materials. Notably, one of the themes that emerged in the workshop was the desire to encourage the participation of the communities included in the archive creation, as what is being archived is a community experience and thus part of collective memory. The participation of affected communities and of already existing archives, as recommended by the workshop participants, may involve deciding what to include in the new archive and/or providing context for the materials in the collection. This

addition of participation within the historical archive not only has an ethical dimension in that it allows the subjects to determine their own representation, but it also features a moral dimension in that the archive takes a certain perspective. At the same time, the archive is – at least more explicitly – part of a political space in mass atrocity memorialization.

While a historical archive allows for a more disparate collection, the third form, an archive as a tool and source for education, must be more tightly organized, potentially in different dimensions given its place in the public realm and socio-political space of online memorialization. An education-focused archive would certainly be similar in many ways to the historical archive, but with addition of pedagogy-specific metadata: i.e., supplementary material to explain and raise questions about the items in the collection. It would make sense for this version of the archive to arise after the first two forms not only in terms of organizational evolution and technical demands, but also because it might be informed by and hence benefit from studies and experiences derived from the first two.

Individually and together, these three alternative visions of the proposed archive offer innovative ways to study the memorialization of mass atrocities. They make it possible to address issues within the realm of digital memorialization of mass atrocity while also presenting novel theoretical and empirical opportunities for the study of trauma, how we teach it in a digital age, and supporting the research of psychologists, sociologists, historians, internet researchers, computational social scientists, media and communication scholars, and many others. The archive – as many participants highlighted throughout the workshop – also preserves “at risk” data that might otherwise disappear, all while providing a record potentially for use in human rights cases, and – especially in the connections between sources – capturing single items as well as their spheres of influence, patterns of diffusion, and network communities linked through subjects, themes, and discourses.

IV. Data sources

The decision on what primary purpose an archive of digital memorialization should serve informs the question of what material (or types of materials) should be archived. A non-exhaustive list of data sources that could be collected is presented in Table 3.

Table 3. Categories of data sources. Categories of data sources that could be collected for a digital archive of mass atrocities. The left column indicates the category, and the right column describes the category including some examples.

Category	**Description**
Digitized non-digital sources	Digital forms of the non-digital artifacts typically found in traditional archives (e.g., digitized books, articles, photographs)
Institutional/ official websites	Websites maintained by: ● Government agencies ● Government affiliates ● Officially sanction memorialization bodies
Non-institutional / non-official websites	Websites maintained by ● Businesses, ● Civil society organizations (including groups, clubs, and non-profit organizations devoted specifically to memorials (such as *Together We Remember*) ● Individuals
News	Journalist coverage (across various mediums) including: ● News articles, ● Interviews, ● Eyewitness accounts (local and international) The material included could be accounts of the episode in questions or coverage of efforts to frame and remember past episodes, including those by institutions and organizations engaged in memorialization,
Social media	Including: ● Platforms that tend to feature verbal posts (e.g., Twitter) ● Platforms that tend to feature photos or video (e.g., Instagram or TikTok) ● Platforms that feature both (e.g., Facebook)
Dark web forums	There are typically communal discussion forums featuring non-modal and/or extreme narratives about the past. May or may not be accessible to the public.

Each of the forms of data sources can be characterized by a variety of dimensions, some of which we identify and highlight in Table 4.

Table 4. Dimensions of data sources. The different data sources can be characterized by the dimensions listed in the left column. The right column describes the dimension and presents some examples on how they can be applied.

Dimension	**Key Issues**
Manageability of collection	Some material, like ordinary websites, are relatively straightforward to collect once parameters of inclusion and frequency of collection are established. Others – especially social media – entail enormous logistical challenges.
"Reliability" of content	Two important issues: ● The underlying veracity of the narrative being communicated through a digitally memorialized narrative, which would help researchers, educators, and social users understand how digital memorialization is used in the service of the production of history or the construction of narratives; ● The credibility of the purveyor of the content on a given site, which might vary especially between (and among) official versus non-official websites (and even so-called dark web forums, Telegram, 4Chan, 8kun). Awareness of these issues might help address the subjective nature of narratives and could enable a deeper investigation of how narrative construction develops.
Legal accessibility	The rights of a third party – the proposed archive – to store, display, and disseminate digital material not originally produced, collected, or collated by the archive itself likely vary: ● Social media, for example, may be governed by end user licensing agreements that limit access to posts, even if they were originally posted publicly. ● Government-affiliated websites may be required to archive all posts and information that pass through them, creating at least a linkable archive that could be brought into the service of the archiving project.
Time/context dependence of content	The experience of engaging with digital content is different depending on the contextual parameter such as ● Date ● Region ● User profile While variability in a digital interface across these dimensions is likely to be difficult to capture, they nonetheless represent a key element of how digital memorialization is experienced.
Vulnerability of content	Much content is at risk of changing or disappearing. For example: ● The Snapchat app is built to make posts disappear once they've been viewed; ● Some social media companies may remove offensive (and possibly illegal) posts in an effort to not become party to the instigation of violence – which may be a positive development but poses a challenge for the enterprise of archiving. ● Non-social media content may be lost to site updates, planned obsolescence, domain expiration, or curatorial neglect.

Two of the issues listed in the table – "manageability" and "legal accessibility" – function as constraints on the other three ("reliability," "context-specificity," and "vulnerability"). Decisions regarding the ultimate focus of the archive should be made based on what is most needed or desired, but with the knowledge that the constraints may substantially determine

the final scope – and perhaps the shape – of the project. A scope-limiting strategy would have to establish firm limits on what social media would be included (if any) or rely on strong definitions of what sources (official or unofficial) of archive and memorial production would be subject to collection.

V. Methodologies

Another recurring theme discussed during the workshop concerned the methodology of collecting and archiving data as well as the tools available for this purpose. The discussion primarily evolved around the following two components of the process: *discovery* of content to be archived and *maintenance* (updating) of archived content. Interestingly, the aspect of processing of the archived content to achieve certain research aims was rarely noted in this context.

The first component - discovery of the content - was the one that attracted most of the discussion. Several options were suggested by the participants ranging from manual retrieval of content to be archived to the large-scale web crawling to relying on social media APIs. The selection principles suggested also varied from the use of random samples of content to be archived to the items manually selected (e.g., via social media monitoring; Docnow) to the content prioritized by the search engines to theme- or institution-based collections (e.g., to crawl all content related to a commemorative event and/or produced by a specific heritage institution).

Also noted was the importance of archiving not only the content (e.g., the website devoted to mass atrocities), but also the experiences using these websites. Referred to as "walkthroughs", these experiences involve looking at "how the producers may explore it; how educators who use it might explore it; how different user groups might explore it; how a researcher might [explore it]. This idea is informed by the notion that digital experiences are mostly informed by symbiosis of human and computational agents".

The second component - maintenance - was primarily discussed in the context of longitudinal data collections (i.e., the reiterative process of archiving which goes beyond a single capture of data). As one participant noted, "The web-crawling is more involved than snapshots – it's a very iterative process of trial and error to get full captures. There's also the

question of whether [one's] goals include longitudinal documentation of the evolution of the resources being archived. If so, then there are overlapping phases of new crawls, repeat crawls". The procedure, however, was noted to be a complicated one, considering the "networked and iterative" nature of the online environment.

In terms of the toolbox which can be used for these aims, the participants primarily noted existing large-scale archival projects such as Archive-It and Internet Wayback Machine. Conifer was also proposed, cited for having the "look-and-feel" of using an archive. Certain live search engines[7] might also be feasible. One of the participants also noted Webrecorder as a tool of selective archiving.

VI. Public interface

Beyond the archiving process itself, an archive should have a public interface. This looks different for digital archives, and specific projects have opted for alternative interfaces depending on their aim and scope. Commoncrawl, for example, provides the data in the form of monthly Internet dumps, that the user would then mine to find what is needed. Services like Webrecorder or Conifer focus on fidelity, so that the user would obtain an exact picture of how a page looked in a particular point of time. We identified three types of public interfaces according to the input of our participants: *documentation*, *navigation,* and *aggregation*.

The *documentation* interfaces involve any services that record the process of archiving itself. This form of interface not only helps understanding the data in terms of the scope and possible limitations, but it is valuable for those who want to build upon the methodology to create similar projects. In its simplest form, it would involve the methodological details of the construction of the mapping of the collected data, e.g., manually curated lists of websites or links posted in specific social media channels, and the technological details behind the process of accumulating the data, e.g., the frequency of the snapshots or the methodologies used social media APIs or web crawlers. In the case of incremental archives, a monitoring dashboard can be built to track the volume and current state of the archive;

[7] For some examples of the projects using them, see collections of Arquivo.pt for Afghan sites preservation or preservation of websites of research and development projects.

keeping track of an incremental mapping could be time consuming but extremely important. It is also relevant to start building an archive of archives referring to the need of documenting these processes and accumulating lessons and recommendations for future archivists. This whitepaper is part of that effort.

The *navigation* interfaces refer to the ways in which the user can find the data that they need. There are multiple forms in which this could be achieved. The data mapping unit (e.g., URLs, social media accounts) and archiving timestamp are the most basic forms of navigation elements that should exist in every archive. Even such distinct forms of archiving such as Common Crawl or Webrecorder[8] have these two features for accessing data: URL and timestamps. Additionally, the data can be manually or automatically categorized, indexed so that quick searches can be performed over the collection, or further transformed into sophisticated representation of knowledge such as semantic networks. Any of these forms of processing would then require an intuitive interface to allow the users to navigate the archive.

The *aggregation* interfaces are those related to results obtained through analysis of the archive. Although it is possible that the archive remained present for a long time before resources are allocated to explore the patterns that it hides (e.g. sometimes the motivation for archiving is the hope that future researchers with access to more powerful methodologies and technologies will be able to process it), analyzing the data so that it can be presented in a meaningful way is the ultimate goal of archiving it. Note that the categories mentioned in the navigation interfaces could be part of the aggregation interface, but the focus of the functionality would be to display information related to the categories, e.g., histograms. Most of the forms of aggregation come directly from Natural Language Processing (NLP) techniques such as sentiment analysis, entity recognition, or topic modeling, however an additional opportunity is the construction of networks between the archived items (or between entities recognized by the NLP techniques).

In the case of mass atrocities, a comprehensive archive should integrate the three types of interfaces, but the documentation and navigation interfaces are fundamental to address the different motivations that sparked this initiative. The documentation interfaces will help

[8] See ReplayWeb.page; Webcrawler link.

scholars understand the structure of the data to perform large analyses, and replicate our methodological approach for further collection, whereas the navigation will allow all members of society to explore the content and answer specific questions regarding the events. More importantly, the navigation interface represents the tangible part of the archive, a cultural object that memorializes the events. The aggregation interface can further contribute to realization of the society-centered motivations by enriching the archive with elements that help interpreting the content, for example, uncovering hidden patterns that would contribute to the better understanding of the mass atrocities.

VII. Challenges

As one participant highlighted during the workshop, web archiving might always deal with sensitive data, but in the case of mass atrocities, the obligation to protect the memory of the victims is paramount. There needs to be an awareness and clear understanding that an archive is a social practice and set of methodologies that might lead to the co-existence of the memorial and denialist/distortionist narratives in the same space that it aims to study and represent. Each decision to include cases, events, communities, individuals, or organizations, or to draft methodologies for webscraping, webcrawling, and storage, has ethical and political implications. There is a need to make not only robust and ethics-guided choices about what to collect, but also for those choices to be followed consistently.

When dealing with ethical obligations, a digital archive memorializing mass atrocities – and particularly the one that contains denialist or distortionist materials – must 1) incorporate ethics from a human rights perspective, 2) deal with difficult questions about the reliability of sources, and 3) consider a proper involvement of affected communities and institutions due to an increasing emphasis on participatory culture and the positionality of researchers and archivists (both in heritage/archive practice and in media studies). Furthermore, the scope of data collection and case selection further feeds into ethical challenges, such as the cases to include and exclude, thereby defining what “counts” as a mass atrocity. Finally, depending on how the archive was structured – i.e., if archived material were to be publicly accessible via Google (or any other search engine) – the potentially sensitive and distressing nature of the materials being stored and represented in the archive might render it appropriate to use content warnings.

Any archive – and especially one concerning mass atrocities – needs to consider how legal practices differ across countries and continents. How these regulations vary poses a particular challenge. One critical area of regulation to consider, for instance, is the General Data Protection Regulation (GDPR) of the European Union. In the U.S., copyright law provides libraries with exceptions for collecting "published works" (§108 of the Copyright Act), yet this does not address digital preservation and web archiving. Different approaches to regulation will need to be considered particularly when archiving social media content and dealing with the question of whether people can apply to have their material removed. Under EU jurisdiction, for example, the "Right to be Forgotten" imbues individuals with control over whether their images and other representations are publicly accessible on the web. Thus, several ethical and legal questions that may arise include:

- Will subjects have a right to opt out of the repository?
- How will the archive handle copyright matters that differ across countries and platforms?
- Who – whether subject, content creator, or user – will have the right to make decisions about access to archived material?
- Will there be a need to use certain technologies – like permission-based access or image protection – to control how archived material is accessed, used, and circulated?

Amongst others, these examples point to how the ethics of memorialization ethics may be doubly challenging in view of the regulatory transformations often brought about by changes in governments or political regimes. At the same time, memorialization ethics may also be more stable than expected due to centuries-old regulations that profoundly influence the question whether or not users can opt out.[9]

[9] A participant framed it in the padlet this way: "In many European countries, web archives have continued the centuries-long tradition of so-called legal deposit, a framework going back to the 1660s. In this way, to be allowed to print public material, the printer had to deposit a copy of it in the king's library. Many national web archives operate under such conditions, namely that a copy of anything made publicly available on the web must be handed over to the library/web archive. If this is not done, the library could fetch it itself. Therefore, 'what to collect' is easier to answer in such cases: everything related to the nation state in question, and very often using the country-code top-level domain as a

The scope of the archive, the reasons for it, and its visibility to the public present another set of challenges. As a means to define scope, cases of mass atrocity should be properly collected for archiving, and the reasons for their selection justified transparently. The archive needs to define parameters for involving its community of users (researchers, general public, and victims, as well as students and perpetrators), including the involvement of, and its coexistence with, traditional forms of preservation and representation (museums in particular), participation by language, region, and connection to the subject matter also merit attention. Feeding into this is a need for projects not to over-promise what they can deliver, as highlighted by a participant in view of the failure of the Library of Congress Twitter Archive to manage public expectations. Public communications should be drafted in a way that they do not raise unrealistic expectations in view of a timeline, and what DAMMA will consist of in terms of data, expected results, funding, and overall scope.

Considerations of scope lead to the challenge of scale, and to the inevitable tradeoff between the comprehensiveness of the data collection and the quality of its curation. While the recent advancements in the field of information retrieval allow facilitating the work of curators with the help of algorithmic systems, such a facilitation still requires fine-tuning which becomes problematic as the size of the collection grows and the variety of materials included (as well as available formats) increases. Under these circumstances, it might be important to consider the balance between the volume of the collection and the quality of what is collected as well as its usability.

Furthermore, there is a risk to DAMMA by underestimating the scale of the data that will be retrieved, even for small-scale projects (especially when data in the "heavier" formats, such as videos, are being added). The project needs to decide on what types of assets and content it will collect and preserve, and this will in turn inform the choice of tools and platforms. Will it include social media? If so, one must also consider whether dark web forums, Telegram, 4chan, and 8kun will be incorporated alongside mainstream platforms

---

minimum – in the Danish case everything that has the .dk suffix. This also implies that there is no means for opting out. Once something has been made public, it is archived, whereas opting out would be like deleting something that has been said in public, which would never happen. However, in many countries (but not in Portugal) access is restricted to researcher access."

such as Facebook, Instagram, TikTok, Reddit, and Twitter. How will the project deal with regional differences in media usage that influence the platform that users turn to (for instance, when trying to represent Holocaust memorialization platforms compared to Cambodian genocide memorialization). The question here is also whether to limit it to being online or to include offline digital assets which might play a role in atrocities' memorialization.

One possibility raised at the workshop was to focus on memorialization events such as commemorative anniversaries. This would not only provide a solid rationale for longitudinal web crawling and collecting but also for systematic comparisons across different atrocities. The cultural significance and value of memorialization events also serve as moments which may feature surges in denialism. The archiving effort should consider the possibility of capturing such distortions as instances of harmful memory culture, corrosive yet worth preserving for the sake of study and confrontation.

Questions of scope and size inevitably raise more technical questions about the sustainability of data collection beyond project cycles. While setting up a basic storage architecture per se is a quite trivial task, requiring only a decent server that would be running without interruptions, making sure that the process is not interfered on the platform side is a different story. For example, Google began to deploy a cookie agreement as a means of disrupting crawlers as they accessed the search page, thus resulting in the interruption of their work.

Projects face a challenge in whether they will be able to produce deliverables that are not bound to their life expectancy. Sustainability and accompanying technological problems were some core challenges identified by the participants. How will projects make sure that cooperation, data collection, and maintenance of their archive continue beyond their funding? This would also rest upon robust selection criteria in a dynamic environment. As highlighted by one participant, data are highly context-specific (for instance the delineation between memorialization and distortion). This means that data collection is not just about gathering single items but also about archiving context. The dynamism of data poses a problem not only at the time of implementing and running a project but also beyond it due to evolving user practices and changes to platform regulations, access, and web architecture, especially with respect to social media.

Beyond context- and platform-related barriers, the ephemerality of constantly disappearing content poses the need for proper documentation and also for extra explanations to users so that the logic behind the collected sample is made clear. At some level, this ephemerality is impossible to deal with (indeed, it serves as a motivation for the project in the first place), so transparency will be the key when setting up crawlers. The identification of vulnerable content and context may be a crucial step towards this. Another concern that was raised involved the complexity of identifying the authenticity and the original source of the content, since simply collecting time, date, and geolocation would not be sufficient. Another key technological challenge involves interpreting content and embedding it in digital and non-digital contexts and into a web universe of references. Promoting awareness and supporting the use of the archive will be an additional task.

The ability to perform all these tasks sustainably invokes questions of resources and funding. As highlighted by the participants at the workshop, a project needs to assess the resources necessary for its intended scope and for creating a sustainable archive. Quality of access is always a problem, but, as emphasized by the participants, the bar is especially high due to the parameters imposed by private companies. An in-house alternative would afford more flexibility but would add to the resource needs of the project. How do you quantify the best access that they can provide? How do you finance it? Will you use algorithmic affordances to potentially increase the usability of the archive (e.g., by customizing information delivery) while risking the potential transparency tradeoffs? As suggested by a participant, it might even be possible to promote a web archive as a working environment free of customization and non-transparent algorithms. Although this would come with less functionality and ease of use, it would provide equal access to data that would otherwise be hidden.

VIII. Conclusion

This whitepaper summarized the prospect of creating a digital archive for the memorialization of mass atrocities (DAMMA) based on the proceedings of a virtual workshop held in October 2021. Specifically, it discussed a selection of questions regarding the scope, form, usages, and development of such an archive.

By doing so, this whitepaper has laid out several of the themes of digital archiving related to mass atrocities, from the range of scopes that such projects might have, to the various forms the resulting archive might take, and to the ways in which the archive might serve researchers and the public. Many challenges – indeed, layers of challenges from the conceptual in nature, to technical ones, legal ones, and logistical ones – that were identified at the October 2021 workshop have also been described. This whitepaper has helped to identify the tradeoffs with respect to each of these that will be central to the decisions to be made on the DAMMA project.

The October 2021 workshop proved invaluable both as an opportunity to bring an esteemed and experienced group of scholars together for the exchange of insights and advice. Fittingly, the workshop provided many more questions than it did answers about the conceptual framework of the digital archive for the memorialization of mass atrocities (DAMMA) and its practical implementation. In doing so, it illuminated many of the decisions about dimensions and parameters that face the DAMMA project in the short, medium, and long runs. The bottom line, as summarized in this whitepaper, is that there is both considerable enthusiasm for the idea of the project and its goals – and recognition of the challenges that must be overcome to get the project off the ground.

As a result, the core team (Bultmann, Makhortykh, Simon, Ulloa, and Zucker) have concluded that the next step of the project, before moving to the first stages of constructing the actual archive, is to hold a floor-up workshop, focusing on more technical issues. Doing so will help delineate the realm of what is feasible, from which we expect to be able to work backwards more effectively in determining the parameters of the DAMMA.

The technical and process issues to be addressed in the follow-up workshop include:

I. Data collection, invoking questions like what tools can be used to identify data to be collected, and what tools could be used to collect the data itself;
II. Data storing, considering issues of how much data are likely to be collected, how it might be held securely, and how it might be organized for the joint objectives of security and access;
III. Legal and institutional aspects, addressing the relationships between the collectors of the data, the users of these collections, the original creators of the data, and the

subjects of the data – understanding that all of these parameters may vary considerably across digital memorial collections and across jurisdictions in which they legally exist;

IV. Funding parameters, involving the effort to identify sources for initial start-up costs (including the pilot project, described below) as well longer-term, more sustainable support for the effort.

Following that workshop, the core team hopes to be able to propose (and solicit funding for) a pilot study. This study would be much smaller than the envisioned DAMMA, involving a limited scope across several parameters. For example, it might pre-identify a set of archive targets related to a specific historical episode, and then engage in the basic exercise of constructing a linked archive of those sites, as well as thematically adjacent sites identified through a scraping exercise.

The goals of the pilot study will be to understand, in real-life terms, the choices that the abstract issues laid out in this paper seem to implicate, to develop a technical/technological blueprint for larger efforts, and to being to understand the parameters of engagement with the project, both as creators behind it and for potential users.