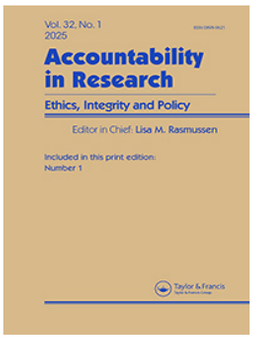

Taylor & Francis
Taylor & Francis Group

Accountability in Research
Ethics, Integrity and Policy

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/gacr20

Web archives for data collection: An ethics case study

Deanna Zarrillo, Mat Kelly, Erjia Yan & Chaoqun Ni

To cite this article: Deanna Zarrillo, Mat Kelly, Erjia Yan & Chaoqun Ni (08 Sep 2024): Web archives for data collection: An ethics case study, Accountability in Research, DOI: 10.1080/08989621.2024.2396940

To link to this article: https://doi.org/10.1080/08989621.2024.2396940

ACCOUNTABILITY IN RESEARCH, 2024
https://doi.org/10.1080/08989621.2024.2396940

Taylor & Francis
Taylor & Francis Group

Web archives for data collection: An ethics case study

Deanna Zarrillo[a], Mat Kelly[a], Erjia Yan[a], and Chaoqun Ni[b]

[a]College of Computing & Informatics, Drexel University, Philadelphia, USA; [b]Information School, University of Wisconsin, Madison, USA

ABSTRACT

Background: Web archives offer researchers a promising source for large-scale longitudinal data collection; however, their complex social and technical infrastructures create an array of ethical concerns. In addition, there is a notable lack of guidance available for researchers hoping to conduct0 ethical research using web archives.

Methods: We present an ethical decision-making case study based on an ongoing research project using the Internet Archive's Wayback Machine to study faculty appointments and mobility at Historically Black Colleges and Universities (HBCUs).

Results: This paper contributes to information ethics discourse by expanding on the Association of Internet Researchers' recommendations for ethical decision-making, and mapping ethical considerations for each stage of the project within existing conceptual frameworks for research using web archives.

Conclusions: By utilizing internet research guidance and web archive research frameworks in a case study approach, we hope to aid future researchers conducting internet research of a similar nature by serving as a useful reference.

ARTICLE HISTORY
Received 11 June 2024
Accepted 22 August 2024

KEYWORDS
Case study; data ethics; web archives; privacy; job mobility

1. Introduction

Big data and rapidly emerging Internet-based technologies have created a complex socio-technical environment that presents a unique ethical challenge for researchers. The development of new ways of finding, accessing, aggregating, and analyzing information about people far outpaces the creation of guidelines and best practice policies for using this data. New types of data, the growing availability of personally identifiable information (PII), and the ease with which this data can be accessed exaggerate the resulting policy gap. In turn, path-setting through ethical decision-making becomes a herculean effort (Carusi and Jirotka 2009). For researchers, this introduces a variety of concerns in the way risks are mitigated for study populations, especially when their data is Internet-

CONTACT Deanna Zarrillo dz364@drexel.edu College of Computing & Informatics, Drexel University, 3675 Market St., Philadelphia, PA 19104, USA

based and publicly accessible. The public vs. private nature of online data is a core issue in research ethics, and there is ongoing discourse on where Internet-based research overlaps with human subjects research. Internet-based research creates risks surrounding the privacy of individuals and unintended consequences for the use of PII (Zimmer 2010). However, few comprehensive guidelines exist for how researchers can ethically interact with online data.

In 2019, the Association of Internet Researchers (AIR) approved and released their Internet Research: Ethical Guidelines 3.0 document (IRE 3.0). This living document outlines ways to ethically orient a research project and guides the various considerations a scholar should address at each stage of research. The AIR makes clear that the ambiguity of Internet research and the wide range of ethical decision-making procedures, ideologies, and cultural lenses through which data can be understood means that readers must consider documents such as these to be "guidelines, not recipes" (Franzke et al. 2020). These guidelines advocate for a pluralistic approach to ethical decision-making where each research question invites its own diverse set of social, legal, technical, or other particularities that may be resolved through dialogic reflection and contextually aware judgment calls. Dialogical approaches to ethical decision-making such as the dual use of ethics in research as both a methodological resource and a topic of study itself have been advocated for and described as "ethics in action" (Mondada 2014). In this vein, many researchers have made great efforts to thoroughly describe their ethical decision-making using case studies that may serve as references for others conducting similar work (see Lomborg 2013; Ogden and Maemura 2021; Tiidenberg 2018).

The case study approach is a useful resource for researchers when navigating projects with new and complex sources of data. Web archives are one such source of complexity as they contain vast quantities of diverse data. Sources like these, which are gaining popularity in research, complicate the development of standard guidelines. Web archives introduce particularly idiosyncratic ethical concerns, as they are an attractive source for collecting historical online data about people, places, things, and events in a technically complex information access venue. Maemura (2018) distills issues in conducting web archive research down to three common challenges; namely, how to select and organize data from web archives, how to approach critical examination of sources in web archives, and how to approach ethics and consent. All of these issues are further confounded by infrastructural imbalances within archives created by the inherent socio-political power dynamics between the Internet and society (Maemura 2023).

Web archival collections attempt to capture and preserve the fluidity and evolution of the Internet through facsimile-style reproductions of the live web. However, the immense scale and dynamic nature of the Internet means

that it is impossible to collect and preserve it in its entirety. Hegarty (2022) describes web archives as a "sliver of a sliver of the Internet." By this, Hegarty means archives like the Internet Archive aim to be representative in their collections, rather than exhaustive. The precise socio-technical processes behind the development of archive infrastructure are a "black box" to users, which means research using data that is a byproduct of these unknown processes risks the reaffirmation of biases that may already exist in online sources (Milligan 2016). Web archive collections are often made up of snapshots of the live web, collected by crawls, appraised with human and computer-mediated methods, stratified by numerous levels of governance, and wrought with missing, incorrect, and de- or re-contextualized information (Summers and Punzalan 2017). Put simply, web archives are "actively created and subjectively reconstructed" (Brügger 2011). Ultimately, using web archives as a source for research data collection, especially when the collected data is about people, compounds the already present ethical issues of internet research.

Following the dialogical precedent set by previous scholars and in alignment with the concepts laid out in IRE 3.0, we present a case study of ethical considerations and decision-making in an ongoing research project in which the Internet Archive's (IA) Wayback Machine is used as the primary source for data collection about faculty appointments and mobility at Historically Black Colleges and Universities (HBCUs). This paper contributes to information ethics discourse by providing a rich description of our data collection decision-making in case study form. We interrogate the ethical use of specific web archive data through the lens of our guiding frameworks. We expand on the guiding questions found in AIR's Ethical Decision-Making recommendations (Markham and Buchanan 2012), and map these considerations for each stage of the project within Maemura & Ogden's 3-dimension conceptual framework for research using web archives (2018). By utilizing both broad and niche subject area guidance and expanding on them through the combination of their recommendations, we hope to address many of the concerns laid out above.

2. Methods

2.1. The case

The research project this case study is based on takes advantage of modern information technologies, including the Internet Archive's Wayback Machine, data from Academic Analytics Research Center (AARC), and the Web of Science to collect a large-scale, heterogenous, longitudinal data set of HBCU faculty appointments and mobility from 2005 onward. Data collected from IA include name, rank, department, college, and e-mail address as

Table 1. Sample data collection worksheet. Names and emails have been anonymized to maintain the privacy of individuals.

Year	Home capture URI	Faculty page URI	School or college name	Department name	Faculty name	Faculty rank	Faculty email
2020	web archive link	web archive link	School of Business	Accounting	Fred Harris	Professor	fred.harris@howard.edu

identified from department-level faculty directories (Table 1). Additionally, we conducted a survey and interviews to contextualize and capture the reasons for faculty mobility. The survey aimed to reach scholars who have been associated with an HBCU from 2020 onward and yielded 361 valid responses from former or current HBCU faculty members. Follow-up interviews were conducted with respondents who indicated their willingness to participate in further discussions.

This project examines the effect of academic mobility on the productivity, impact, and career paths of professors employed at HBCUs, as well as what institutional factors impact faculty attrition and retention. The emphasis on the mobility of HBCU professors aims to focus analysis on concerns surrounding "brain drain" of talented individuals to predominantly White institutions (PWIs) (Seymore 2005). Despite the impact social mobility and civil rights movements have had on academic mobility (Sugimoto et al. 2017; Van Noorden 2012), there is little contemporary research on academic mobility at HBCUs. Results from this large-scale, longitudinal analysis will provide important evidence regarding the career paths of professors moving to and from HBCUs.

Here, we present the guiding frameworks used to make sound ethical decisions during the phase of data collection from IA. Ethical decisions made during interviews and surveys are not discussed here as these data are not linked directly to the manually collected faculty information found in the Internet Archive. Decisions in those steps were guided by appropriate frameworks and contemporary ethical discourse for each methodology.

2.2. Guiding frameworks

We use existing frameworks to help guide the ethical discussion in this paper. The longitudinal nature of this project, in combination with the use of multiple sources of job history data, brings a few ethical considerations to the forefront, in particular, the use of the IA as our primary data source for the collection of faculty information. During the data collection portion of this work, we frame our ethical decision-making with a conceptual understanding of the immense efforts taken by a researcher to begin using web archives for research (Ogden and

Maemura 2021) and guided by a series of foundational considerations for conducting Internet-based research using AIR's decision-making guidance. Ogden & Maemura's three dimensions for researching using web archives include the often iterative acts of *orienting*, *auditing*, and *constructing*, all of which they claim are necessary for a researcher to have a comprehensive understanding of the scope of the archive in which they are working given its nature and complexity. The overall goal of their framework and analysis was to highlight the procedural and epistemological entanglements researchers contend with when developing research methods for digital and archival sources. In contrast, Markham and Buchanan's (2012) Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0) addresses more generalized concerns surrounding Internet research. This document sets the stage for ethical decision-making by presenting relevant questions about data use in each respective linear stage of research.

For our analysis, we broadly categorized these high-level questions into the following five larger ethical considerations: 1) legalities; 2) privacy; 3) study population; 4) data considerations; and 5) dissemination. Throughout the following case discussion, additional questions that we deliberated as part of the HBCU faculty mobility project are adapted into translatable considerations that may serve as web-archive-specific addendums to AIR's decision-making guidance (refer to Table 2).

Table 2. Crosswalk between the three dimensions of web archive research and AIR recommended guiding questions categorized into 5 larger ethical considerations. Included in the table are integrated and expanded considerations suggested specifically for web archive research.

3-Dimensions	Considerations	AIR Decision-Making & Internet Research guiding questions [*Expanded*]
Orienting	Venue	How is the context defined and conceptualized? [*Have you considered historical, local, international, and/or emergent policies, standards, and guidelines (i.e., the right to be forgotten)?*]
Auditing	Privacy	How is the context (venue/participants/data) being accessed? If access to an online context is publicly available, do members/participants/authors perceive the context to be public? What particular issues might arise around the issue of minors or vulnerable persons? What are the potential harms or risks associated with this study? What are potential benefits associated with this study?
	[*Source Criticism*]	[*Have the chosen sources been critically examined? (i.e., Brügger 2011)*]
Constructing	Study Population	Who is involved in the study? What is the primary object of study? How are we recognizing the autonomy of others and acknowledging that they are of equal worth to ourselves and should be treated so?
	Data Considerations	How are data being managed, stored, and represented? How are texts/persons/data being studied?
	Dissemination	How are findings presented? [*How will privacy be protected post-publication?*]

3. Results and decision-making

The following sections present our ethical decision-making processes throughout each phase of research and include discussions on the particular considerations for this case study. Relevant literature and possible expanded ethical considerations for web archives are also discussed.

3.1. Orienting

The following consideration is part of the orienting phase of research that involves "finding one's position in relation to unfamiliar surroundings; tailoring or adapting to specified circumstances" (Ogden and Maemura 2021). Understanding the environment of web archives is paramount to coordinating a research project that utilizes them. In the following, we explore AIR questions and recommendations related to the venue of data collection.

3.1.1. Venue

The AIR recommends researchers initially ask themselves, "How is the context defined and conceptualized?" (2012). Due to the complexities of web archives, we determined the context of the IA's venue through detailed consideration of legalities. Legalities can refer to regional, national, institutional, and platform-level policies, rules, and regulations regarding the collection and use of data and are essential for researchers to familiarize themselves with before starting an Internet-based research project (Cilliers and Viljoen 2020). During this phase of research, we consulted a variety of regulatory documents to ethically assess the use of the Internet Archive as our primary source of HBCU faculty data. Because personal information is included in our data, we had to make sure it was appropriate for use in the first place and orient ourselves to possible obstacles that regulatory policies might cause in later stages of research. The documents included the Internet Archive's Terms of Use, Privacy Policy, and Copyright Policy (Mondada 2014), crawler policies of HBCU websites, institutional disclosure requirements in the Higher Education Act of 1965 (NPEC 2009), and guidance from the Association of Internet Researchers (AIR) for the provision of faculty and staff data (King 2023). The Internet Archive's Terms of Use, Privacy Policy, and Copyright Policy (2014) clearly outline that researchers hold sole responsibility for appropriate data collection and handling, cementing the need to consult multiple regulatory documents. In consulting these documents, we learned about several key factors that support our use case. First, academic institutions in the United States are required to disclose certain information to students (current and prospective) and

the public. Under the HEA §485 Academic Programs disclosure requirement, universities must publish faculty information such as name, rank, and department (NPEC 2009). Additionally, the AIR published guidance on institutional data sharing and classified directory listings as "public" requiring the lowest levels of governance and stating they are *purposefully* made public (King 2023). Finally, in an audit conducted by the research team, it was concluded that it was typical for university websites to allow crawlers to collect data, which serves as the primary accession technique for web archives.

These documents helped *orient* us to what standards and accepted practices already exist and provided us with the necessary confidence to move forward with data collection. However, we are limited to legalities on the live web since guidance for historical data collection about people is scarce. Crossen-White (2015) begins to address this by calling the protection of the "personal reputations of individuals from the past" part of a researcher's duty of care. Web archives and modern Internet technology have given us the ability to resurface information about individuals who may have otherwise been forgotten. In doing so, a researcher must approach historical data collection and analysis about these individuals that is reflexive of original contexts and safeguards autonomy of the persons involved. This idea is a central tenant that appears in policies outside of the US context such as the EU's General Data Protection Regulation's (GDPR) right to erasure or "right to be forgotten." Further, certain documents or regulatory bodies like institutional ethics committees may require different and sometimes contradictory ethical procedures for handling different types of data, adding breadth to the considerations researchers must utilize to orient themselves to a research question. As part of our approach, we utilize both quantitative and qualitative data collection in this project to build relationships with study participants and enhance our responsible decision-making. From our experience, we suggest additional considerations for Internet-based research, which may include historical policies, international standards, specific institutional ethics board advice, and contextual information about data ethics and standards at the time of capture in a web archive. Additionally, because preferences around personal protections like privacy are constantly evolving, we add that considering emergent standards and best practices would benefit researchers' decision-making.

3.2. Auditing

The auditing phase is described as "inspecting, reviewing or assessing systematically; seeing the collection in the round, understanding boundaries"

(Ogden and Maemura 2021). In this stage, we assess the various idiosyncrasies that exist in the data, documents, platform, and live web sources. We utilize the recommendations from AIR related to privacy and source criticism to guide decisions.

3.2.1. Privacy expectations

Once the Internet Archive was determined to be a valid source for data collection, we needed to examine the archival documents we intended to use not only for data completeness and accuracy, but for possible privacy issues as well. The AIR decision-making document poses four privacy-related questions to internet researchers for consideration (Markham and Buchanan's 2012):

- How is the context (venue/participants/data) being accessed – especially when considering the perceived context of public vs. private?
- What particular issues might arise around the issue of minors or vulnerable persons?
- What are the harms or risks associated with the study?
- What are the potential benefits associated with this study?

Understanding the data privacy and terms of use policies addressed in the auditing phase helped us to begin evaluating the context of the venue being accessed, which further informed the context of the data we collected. In addressing the first question, we determined that as evident from the policies discussed above, not only was it reasonable for faculty to be observed on institutional websites, but their information is published on those websites for the express reason of being discoverable by prospective students and the public. We also agreed that both platforms we collected the data from, the Internet Archive and university websites, were generally understood to be public-facing platforms. Our determination here is grounded in Nissenbaum's contextual approach to privacy online (Nissenbaum 2011). This approach serves as a general rule of thumb when determining a website's location on the spectrum of private to public (Kelly, Nelson, and Weigle 2018). Nissenbaum suggests that researchers take into account the privacy norms of an equivalent physical space paying special attention to how the Internet may mediate interactions. With this in mind, universities and archives are placed closer to the public side of the spectrum.

The second two questions were of particular interest to us. Our study includes data about HBCU faculty. Because faculty demographics at HBCUs are more diverse than at PWIs (Gasman 2021), there may be a variety of intersectional groups that require particular attention to privacy. Because of this, our study participants could be classified as a vulnerable niche community. Additionally, the subject matter is primarily

PII collected over multiple years. Faculty members may recognize their information as publicly accessible during employment at one university, but may not necessarily predict that their historical employment data could be collected and interrogated. Despite this, archiving such data may indeed be essential to preserving and understanding of marginalized communities (Baker 2011). Through examining the archive, we may come to identify larger trends in the social factors affecting these communities and produce more effective interventions. However, when studying specific demographic profiles or communities with strong identity ties using descriptive or predictive analysis, there is a great risk of creating stereotypes that may result in overly simplistic representations of a population in a research study (Edwards and Edwards 2016). Just because an analysis *can* be done with available historical data, does not then necessitate that it *should* be done (Lin et al. 2020). Researchers should take note when any recognizable demographic pattern is exposed and avoid opening up communities to exploitation.

The positionality of our project is rooted in advocacy for HBCUs. We recognize the social and cultural significance of HBCUs as historical institutions dedicated to the education and enrichment of Black students (Gasman 2013). Thus, we hoped to address some concerns about privacy expectations and risks by utilizing both quantitative and qualitative data in our analysis in addition to traditional data anonymization techniques. This allows us to balance potential biases a researcher may introduce into their analysis with contextual anecdata that characterize the lived experience of the study population and generally enhances responsible decision-making (Crossen-White 2015). In this auditing phase, we identified which aspects of the data presented risks to the study population and consciously developed research protocols and strategies to mitigate said risks.

Finally, the AIR recommendations advise researchers to consider the benefits of their study. In particular, the decision-making document further asks who specifically benefits from the study and whether potential benefits outweigh potential risks (Markham and Buchanan's 2012). In our case, our work could help positively influence the future of HBCUs, including current and future faculty that make up a portion of our study participants. Making data available could help administrators, policy-makers, and other actors produce better data-backed decisions regarding HBCUs. However, because these benefits are not guaranteed, we cannot conclude that they outweigh the privacy needs of individuals. While privacy does not always have to outweigh all other social values (Baker 2011), the anonymization of data can be determined by its purpose. Our efforts are aimed at maintaining a balance of privacy for individuals while still sharing impactful and

actionable insights. Specific details on our anonymization techniques are discussed later (3.3.2. Data Considerations).

3.2.2. Source criticism

While not part of the original framework, the addition of Source Criticism and ethical questions surrounding the completeness and accuracy of data from web archives would benefit scholars utilizing frameworks like the AIR recommendations for similar research projects. We argue that source criticism is an essential part of the auditing phase – it ensures researchers take into account the infrastructural nature of web archives when developing data collection guidelines and protocols. The number of archival captures or records a web archive holds for a particular URI over time, in addition to the quality, completeness, and accuracy of different captures, can vary widely for many technical and sociopolitical reasons (for further reading see Maemura 2023). For instance the coverage of archival captures across 35 HBCU homepages had an average of 5,322 per institution, whereas PWIs like Drexel University and the University of Tennessee, Knoxville have well over 10,000 archival captures (Zarrillo et al. 2022). Further, some archival captures may differ substantially from what was present on the live web at the time of accession to the archive (Brügger 2011). Because our study's data requirements meant annual faculty information was needed from department-level pages, we had to determine specific criteria when selecting which archival capture to ultimately collect data from and ensure our data set was as complete and accurate as possible.

Brügger (2011) defines a set of rules for critically examining web archive sources that are utilized throughout the data collection process. Rule one says that the least deficient version is deemed closest to the original (deficiencies being elements which are lost during the process of archiving a live web source). In identifying institutional pages with faculty information at IA, many captures had missing and incomplete information. Foraging for the least deficient pages within our set criteria was integrated into the data collection process. Brügger's (2011) second rule says that comparing versions of an archived page by proximity in time and space is an effective way to increase the accuracy of data. One criterion for selecting captures for our study was the specific timeframe in which the capture occurred each year. Our study uses data collected from IA captures within a given monthly time frame within a collection year. Rule three considers the speed at which information may change on a live webpage and how that differs from changes in web archive captures. Faculty appointments, attrition, and retention rates are inconsistent from university to university. Additionally, there is no guarantee that the administrators responsible for updating faculty listings are accurately or consistently performing these updates, even if an archival capture can be determined identical to the historical live web's version. This

increases the chances that we miss essential mobility data through our chosen collection method. Given that new job postings become available by late ummer (Bohn 2022), we limited our time frame to captures from October to December. By collecting data from captures during the typical fall term at U.S. universities, and before the end of the typical calendar year, we hope to capture the maximum amount of faculty turnover as possible. The remaining rules suggest critical analysis based on the types of texts, the genre characteristics of the live web source, and taking into account web characteristics that were typical of the period under examination. During faculty data collection we assessed the types of faculty listings available each year and maintained consistency within each university. In collecting a large period's worth of data, we witnessed how academic webpages, university infrastructure, and information sharing have changed over time. For example, some universities approach faculty listings using individual profile pages, whereas others create central directories. Sometimes universities interchange these methods, or even utilize both. Critical examination of these factors reveals the "texture" of the archive: what, how, and why documents are included, what is missing, and what are the inconsistencies (Hegarty 2022). Revealing this texture helps uncover bias within web archive content that may inadvertently affect insights.

Addressing these areas of source criticism provided confidence in the data collected from IA and allowed us to identify artifacts in the data that may have otherwise been hidden or ignored. For these reasons, we have expanded Table 2 to include this step, as it is an imperative consideration when conducting research using web archives.

3.3. *Construction*

Ogden and Maemura describe the construction phase as "building or making something, to form an idea or theory by bringing together conceptual elements" (2021). In our case, construction included the development of a data collection protocol, and other data management checks and balances while navigating the Internet Archive.

3.3.1. Study population

The first requirement in the constructing phase was to take into consideration different aspects of the project's study population. Specifically, the AIR document asks who is involved in the study, what is the primary object of study, and how we recognize the autonomy of others and acknowledge that they are of equal worth to ourselves and should be treated so (Markham and Buchanan's 2012). As discussed earlier, the participants of the study are those whose data was manually collected. This population is defined as faculty members employed at 11 HBCUs with appointments in the years of 2012 to

2021. Table 1 provides an anonymized example of the information collected in this phase. Ethical considerations around study population recruitment, selection, and provision of informed consent have received a lot of attention in internet research literature. Conducting research with an ethic of care requires not only responsible decision-making but responsible action (Cowan and Rault 2018). In academic research, even with careful decision-making there still exists an inherent power imbalance between the researcher and the study population. Cowan & Rault describe researchers themselves as a primary privacy issue because it is impossible to fully mitigate potential harm once a group of people is chosen for a study and opened up to academic scrutiny (2018). One major method of reducing harm involves the study population directly by acquiring their express consent to collect and use data about the individual for research purposes.

Informed consent is a common struggle for internet researchers. For our project, there are several reasons already discussed that led to the decision not to proceed with acquiring informed consent from the individuals included in our manually collected data from IA. First, our chosen study population for mobility analysis is made up of longitudinal affiliation data of faculty members found on a web archive. It is important to note that these faculty members are collected from a small subset of doctoral and master's degree-granting HBCUs with higher levels of research intensity. This limitation influences the inferences we draw from the analysis. By excluding non-research-intensive HBCUs, our analysis favors institutions with more access to resources within an environment where resources are already limited. Despite this limitation, the number of individual faculty members that would be required to contact would have been unmanageable for the team.

The methods in this study are subject to the practical limitations of acquiring informed consent, yet we still wanted to ensure the data in our project would not open the population to undue attention while at the same time maintaining our participants' sense of autonomy. We utilized a concept called the "distance principle" to scrutinize the type of data collected about the individual and their likely expectation of privacy about that specific data with respect to the source. In an ethics case study, Lomberg explains the "distance principle" as "the conceptual or experiential distance between the object of research and the person who produced it" (Lomborg 2013). For example, collecting social network data can facilitate a vivid picture of an individual's personal and behavioral patterns without knowing specific identifiers. This type of rich descriptive data risks reidentification of research participants if anonymity is not handled effectively. In Lomberg's case study, the data studied were personal artifacts like tweets, blog posts, etc., which creates a small distance between individuals and their data. Ultimately, because of this and the fact that the users' expectation of privacy was unclear in her population, informed

consent was pursued in this study. Because of information gathered from prior decisions in our project, such as the various policies consulted, we felt that the type of data required about our population (name, rank, e-mail, and department) could be sufficiently distanced from individuals' identities through other data considerations. This demonstrates the usefulness of the distance principle in ethical decision-making regarding data and the study participants from which said data is sourced.

3.3.2. Data considerations

After deciding to proceed with data collection, we had to consider the ethical questions surrounding how it was to be managed, stored, represented, and studied (refer to Data Considerations in Table 2). Early on in the project, we began managing our data by using an aggregator to holistically collect historical URLs of HBCU homepages and department pages to form the foundation of a crawler for the Internet Archive. The intention was for this crawler to identify the data we required on each page it was fed to prioritize the speed of collection. This method resulted in a variety of blockers inherent to the infrastructure of the Internet Archive, such as evolving department-level pages over time and the inconsistency of longitudinal archive captures (Kelly et al. 2022). While we did not proceed with the crawl in favor of manual collection methods, the development stages were crucial in the teams *orienting* to the intricacies of the Internet Archive and *constructing* the best process to collect and analyze the data it contains. The decision to manually collect data was not solely derived from practicality. Manual collection has ethical implications insofar as the labor associated may hinder replication efforts, which may be argued to slow the progression of science, but also may act as a disincentive for bad actors to find and use individuals' data. Our finalized manual data collection protocol largely considers the rules developed by Brügger (2011) when conducting in-the-moment source criticism of various department-level pages on IA. By carefully following the protocol, our team of research assistants carried out the data collection of professor affiliation data from 2012 to 2021 for 11 doctoral-level HBCUs on the Internet Archive.

The data storage for our project is facilitated through our affiliated institutions. PIs from each institution are working with their respective information communications and technology (ICT) staff to store encrypted data and troubleshoot the XML-utilized databases and created datasets. The distribution of data storage for each phase of our research adds an additional layer of protection from data privacy risks. We intend to study the data to identify mobility patterns. In this case, anonymization and aggregation of individuals' data does not skew the ultimate goal, as visualizing the to-and-from movement between HBCUs and other institutions does not require the department-level or individual-level data.

3.3.3. Dissemination

The final consideration relates to the dissemination of data and how findings are presented. NSF funding policies require open access to publications and research data. Before beginning the data collection for this study, the PIs created a data management plan that listed key deliverables: (1) master lists of professors at the HBCUs; (2) classified professors in four categories (non-mobile, mobile within HBCUs, HBCU to non-HBCU, non-HBCU to HBCU); (3) aggregated domain-level yearly publication and citation numbers for each HBCU; and (4) anonymized researcher-level publication and citation data. Data anonymization is often an integral part of ethically conducting research that involves personal data (Carusi and Jirotka 2009). In line with best practice, our data will be anonymized using unique identifiers for names, institutions, and departments. The e-mail addresses we originally collected will not be included in the final dataset of the project. Since the start of data collection, we have since determined the best way of moving forward would be to deposit anonymized data to NSF repositories while storing raw researcher data and identification keys through institutional ICTs. These keys can then be made available to peer scholars for reproducibility via an approval process conducted by the original team. As of this writing, the larger project is not yet closed, therefore datasets are not currently available.

In conducting internet research using web archives, we suggest adding a consideration for the protection of privacy in the post-publication stages of a research project. We remain uncertain how exactly the data in our study will be used in the future and by whom. We know that guidelines are slow to develop and new technologies develop fast; so researchers are often paving their own way through ethical decision-making (Carusi and Jirotka 2009). Part of this project's dissemination plan is the development of an open access dashboard of mobility patterns. This dashboard will be limited to anonymized and aggregated data to reduce possible re-identification and other privacy risks, especially as we intend to update the live dashboard annually for three years after the project's close. Possible versions of the dashboard with the ability to drill down into more detailed views of the overall mobility patterns may be made accessible to individuals based on certain criteria determined by the team.

Finally, another researcher-led intervention called "un-Googling" could be explored as presented by Shklovski and Vertesi (2013). In addition to expanding traditional anonymization techniques to include data like places and environmental contexts, this technique involves carefully choosing keywords to enhance discoverability in certain communities and reduce it in others. With this in mind, keywords were selected for this paper that were broad enough to reduce the introduction of bias in search results, but

relevant enough to certain fields of research that discoverability is not hindered.

4. Conclusion

The case study described here uses the Internet Archive's Wayback Machine to collect personally identifiable information about past and present faculty members at Historically Black Colleges and Universities to study academic mobility patterns. Longitudinal affiliation data is then contextualized with interview and survey data to examine the experiences of faculty at HBCUs and the factors that lead to retention or "brain drain" from these institutions. The dynamic creation, maintenance, and infrastructure of web archives and the ease with which users can access them opens analysis up to a significant number of ethical obstacles. There is not a wealth of guidance on using web archives as a source of people data. By utilizing a conceptual framework developed by Ogden and Maemura (2021) in tandem with the AIR Ethical Decision-Making and Internet Research recommendations from Markham and Buchanan (Markham and Buchanan's 2012) in our ethical considerations, we pursued our best effort to navigate data collection from the Internet Archive ethically.

4.1. *Future work and FAIR data principles*

One important consideration for the larger project not discussed at length here is to verify that our data collection and management procedures align with and support FAIR (Findable, Accessible, Interoperable, Reusable) data principles (Wilkinson et al. 2016). Ensuring all of our data is FAIR is an important overall process for ethical research writ large as it encourages research communities to build trustworthy, sustainable, and open data sharing infrastructures (Rauch et al. 2022).

The main goal of FAIR principles is to make data more easily reusable while retaining ethical safeguards. Appropriate management of Internet Archive data is complicated by the dynamic facets of this work as previously discussed. Making data "Findable" refers to the availability and transparency of the dataset. Data and metadata should be richly described, indexed in a searchable resource, and have global and persistent unique identifiers (*FAIR Principles* n.d.). Our intended anonymization and data deposit procedure should facilitate the findability of project data; however findable data must also be "Accessible." For our data to be accessible our data management plan does not include any third party or commercial software required to access the deposited data. The procedure is simple, open, and free, but will require access authorization from a team member to ensure the intention of data reuse does not contradict the ethical considerations determined through the collection process (refer to earlier section 3.3.3. Dissemination).

The third principle requires data to be "Interoperable," meaning it is able to be integrated, understood, and processed by both humans and machines for broad applications (*FAIR Principles)*. Effectively, this principle advises researchers to avoid data language or metadata terminology that is inaccessible or non-standardized without properly shared specifications. The data in our deposited set will be easily interoperable and will provide sufficient documentation for such use if required. Finally, the FAIR principles state that data should be "Reusable." Guidance defines reuse requirements like rich metadata descriptions including provenance, alignment with local or community data standards, and clear usage rights. These requirements allow others to become intimately familiar with the data for further interrogation. FAIR principles are foundational considerations throughout the course of a research project, not just during data collection and dissemination. By doing so, we maximize the potential impact and return on investment of our work.

In situating our data collection approach within multiple frameworks, we also help to bridge them by expanding on AIR's recommendations and offering addendums specific to internet research that takes place on web archives. Underscoring our data collection work by aligning our project with FAIR principles will aid in future research or policy implementations of our findings. We hope this case study will be of use to future web archive researchers and extend the discourse on practical uses of internet research ethics across the research lifecycle.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Science Foundation under awards no. [2122525], [2121861], and [2122691].

References

Baker, A. 2011. "Ethical Considerations in Web 2.0 Archives." *School of Information Student Research Journal* 1 (1). https://doi.org/10.31979/2575-2499.010104 .

Bohn, M. 2022. "Let's Talk—A Look at the Academic Job Search Timeline | NIAID: National Institute of Allergy and Infectious Diseases." *NIAID*. March 15. https://www.niaid.nih.gov/about/lets-talk-look-academic-job-search-timeline .

Brügger, N. 2011. "Web Archiving – Between Past, Present, and Future." In *The Handbook of Internet Studies*, edited byM. Consalvo and C. Ess, 24–42. John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444314861.ch2 .

Carusi, A., and M. Jirotka. 2009. "From Data Archive to Ethical Labyrinth." *Qualitative Research* 9 (3): 285–298. https://doi.org/10.1177/1468794109105032 .

Cilliers, L., and K. Viljoen. 2020. "A Framework of Ethical Issues to Consider When Conducting Internet-Based Research." *SA Journal of Information Management* 22 (1): 22. https://doi.org/10.4102/sajim.v22i1.1215 .

Cowan, T. L., and J. Rault. 2018. "Onlining Queer Acts: Digital Research Ethics and Caring for Risky Archives." *Women & Performance: A Journal of Feminist Theory* 28 (2): 121–142. https://doi.org/10.1080/0740770X.2018.1473985 .

Crossen-White, H. L. 2015. "Using Digital Archives in Historical Research: What are the Ethical Concerns for a 'Forgotten' Individual?" *Research Ethics* 11 (2): 108–119. https://doi.org/10.1177/1747016115581724 .

Edwards, M. R., and K. Edwards. 2016. "Chapter 12: Reflection on HR Analytics—Usage, Ethics and Limitations." In *Predictive HR Analytics: Mastering the HR Metric*, Kogan Page. https://drexel.skillport.com/skillportfe/assetSummaryPage.action?assetid=RW$3433:_ss_book:112616#summary/BOOKS/RW$3433:_ss_book:112616 .

FAIR Principles. (n.d.). GO FAIR. Accessed July 24, 2024 https://www.go-fair.org/fair-principles/ .

Franzke, A. S., A. Bechmann, M. Zimmer, and C. Ess. 2020. Association of Internet Researchers. *Internet Research: Ethical Guidelines 3.0*. https://aoir.org/reports/ethics3.pdf .

Gasman, M. 2013. "The Changing Face of Historically Black Colleges and Universities." https://repository.upenn.edu/handle/20.500.14332/35096 .

Gasman, M. 2021. "The Talent and Diversity of HBCU Faculty." Forbes. July 19. https://www.forbes.com/sites/marybethgasman/2021/07/19/the-talent-and-diversity-of-hbcu-faculty/?sh=7f207b344d90 .

Hegarty, K. 2022. "Representing Biases, Inequalities and Silences in National Web Archives: Social, Material and Technical Dimensions." *Archives and Manuscripts* 50 (1): 31–45. https://doi.org/10.37683/asa.v50.10209 .

Internet Archive's Terms of Use, Privacy Policy, and Copyright Policy. Internet Archive Terms of Use. (2014, December 31). https://archive.org/about/terms.php .

Kelly, M., M. L. Nelson, and M. C. Weigle. 2018. "A Framework for Aggregating Private and Public Web Archives." *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 273–282. Texas. Fort Worth.

Kelly, M., D. Zarrillo, C. Jackson, and E. Yan. 2022. "First Steps in Identifying Academic Migration Using Memento and Quasi-Canonicalization." *Presented at the ACM/IEEE JCDL 2022 Workshop on Web Archiving and Digital Libraries (WADL 2022)*, Cologne, Germany.

King, B. R. 2023. "The Provision of PII Data on Faculty and Staff by Institutional Researchers." AIR. January 27. https://www.airweb.org/article/2023/01/27/the-provision-of-pii-data-on-faculty-and-staff-by-institutional-researchers .

Lin, J., I. Milligan, D. W. Oard, N. Ruest, and K. Shilton. 2020. "We Could, but Should We? Ethical Considerations for Providing Access to GeoCities and Other Historical Digital Collections." Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, 135–144. https://doi.org/10.1145/3343413.3377980 .

Lomborg, S. 2013. "Personal Internet Archives and Ethics." *Research Ethics* 9 (1): 20–31. https://doi.org/10.1177/1747016112459450 .

Maemura, E. 2018. "What's Cached is Prologue: Reviewing Recent Web Archives Research Towards Supporting Scholarly Use." *Proceedings of the Association for Information Science and Technology* 55 (1): 327–336. https://doi.org/10.1002/pra2.2018.14505501036 .

Maemura, E. 2023. "Sorting URLs Out: Seeing the Web Through Infrastructural Inversion of Archival Crawling." *Internet Histories* 7 (4): 386–401. https://doi.org/10.1080/24701475.2023.2258697 .

Markham, A., and E. Buchanan, 2012. "Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)." https://aoir.org/reports/ethics2.pdf .

Milligan, I. 2016. "Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives." *International Journal of Humanities and Arts Computing* 10 (1): 78–94. https://doi.org/10.3366/ijhac.2016.0161 .

Mondada, L. 2014. "Ethics in Action: Anonymization as a Participant's Concern and a Participant's Practice." *Human Studies* 37 (2): 179–209. https://doi.org/10.1007/s10746-013-9286-9 .

Nissenbaum, H. 2011. "A Contextual Approach to Privacy Online." American Academy of Arts & Sciences. October 1. https://www.amacad.org/publication/contextual-approach-privacy-online .

NPEC (National Postsecondary Education Cooperative). 2009. *Information Required to Be Disclosed Under the Higher Education Act of 1965: Suggestions for Dissemination (Updated)*. NPEC 2010831v2. Washington, DC: National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2010831rev .

Ogden, J., and E. Maemura. 2021. "'Go fish': Conceptualising the Challenges of Engaging National Web Archives for Digital Research." *International Journal of Digital Humanities* 2 (1–3): 43–63. https://doi.org/10.1007/s42803-021-00032-5 .

Rauch, C. B., M. Kelly, J. A. Kunze, and J. Greenberg. 2022. "FAIR Metadata: A Community-Driven Vocabulary Application." In *Metadata and Semantic Research*, edited by E. Garoufallou, M.-A. Ovalle-Perandones, and A. Vlachidis. 187–198. Springer International Publishing. https://doi.org/10.1007/978-3-030-98876-0_16 .

Seymore, S. B. 2005. "I'm Confused: How Can the Federal Government Promote Diversity in Higher Education Yet Continue to Strengthen Historically Black Colleges." *Wash & Lee Journal Civil & Social Just* 12 (2): 287–319. https://scholarlycommons.law.wlu.edu/crsj/vol12/iss2/9 .

Shklovski, I., and J. Vertesi. 2013. "'Un-Googling' Publications: The Ethics and Problems of Anonymization." *CHI '13 Extended Abstracts on Human Factors in Computing Systems*: 2169–2178. https://doi.org/10.1145/2468356.2468737 .

Sugimoto, C. R., N. Robinson-Garcia, D. S. Murray, A. Yegros-Yegros, R. Costas, and V. Larivière. 2017. "Scientists Have Most Impact When they're Free to Move." *Nature* 550:29–31. https://doi.org/10.1038/550029a .

Summers, E., and R. Punzalan. 2017. "Bots, Seeds and People: Web Archives as Infrastructure." *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 821–834. https://doi.org/10.1145/2998181.2998345 .

Tiidenberg, K. 2018. "Ethics in Digital Research." *The SAGE Handbook of Qualitative Data Collection*. https://doi.org/10.4135/9781526416070 .

Van Noorden, R. 2012. "Global Mobility: Science on the Move." *Nature News* 490 (7420): 326. https://doi.org/10.1038/490326a .

Wilkinson, M. D., M. Dumontier, J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3 (1): 160018. https://doi.org/10.1038/sdata.2016.18 .

Zarrillo, D., M. Kelly, C. Jackson, and E. Yan. 2022. "Collecting Diachronic Affiliation Data for Faculty at HBCUs Using Memento." *Proceedings of the Association for Information Science and Technology* 59 (1): 527–532. https://doi.org/10.1002/pra2.664 .

Zimmer, M. 2010. "'But the Data is Already Public': On the Ethics of Research in Facebook." *Ethics and Information Technology* 12 (4): 313–325. https://doi.org/10.1007/s10676-010-9227-5.