RE-07-18-0008-18
University of Texas at Austin

Abstract: Investigating Platform Development for Mobile and Social Media Data Preservation

In this Early Career Development project, Dr. Amelia Acker (School of Information, University of Texas at Austin) requests $308,921 from the Laura Bush 21st Century Librarian program for a three-year empirical investigation into emerging preservation tools and new data stewardship practices to answer the following research questions: (1) How do platform developers working in non-library contexts design and construct systems for the creation, transmission and preservation of mobile and social media data? (2) How does the provision of networked information services in mobile and social media platforms, including preservation technologies for mobile and social media data differ from established preservation infrastructures and professional practices in libraries, archives, and museums? (3) Which emerging preservation tools and new data stewardship practices are potentially transferrable to libraries, archives, and museums? In answering these questions, the research project will contribute both theoretical and practical knowledge about digital preservation models, approaches, and techniques across a number of communities with concern for the stewardship and future access of mobile and social media data collections. Outcomes of the investigation have the potential to create a much-needed point of connectivity between two domains of information provision and digital preservation.

The project director (Acker), with support from a graduate research assistant, will gather qualitative, ethnographic data from a variety of different digital media organizations (industry, government, not-for-profit) that focus on social software and mobile communication technology development. In this project, we plan to investigate a range of platform development ecologies, by focusing on the platform developers (technologists, software designers, product managers, and engineers) who contribute, design, and build social media platforms and mobile applications. The research will contribute to the theoretical and practical knowledge in the domains of library and information science by examining (1) the role that platform developers play in the software ecology of digital preservation, (2) if platform developers' conceptions, models, or theories of digital memory contribute to or have significant impact on social media and mobile media's durability, vulnerability, or preservation possibilities; (3) how these conceptions of digital memory and long-term access influence future possibilities of digital preservation for creators or primary users, as well as secondary users such as researchers, librarians or archivists; yielding knowledge about (4) the practical, technical, and social requirements and considerations needed for the digital preservation of mobile and social media data.

The deliverables will include scholarly research publications and an Open Educational Report on the status of digital preservation of social and mobile media data outside of traditional digital stewardship contexts in libraries, archives, and museums in the United States. The design, conduct, and interpretation of findings will be informed by an Advisory Board made up of digital preservation and social computing experts, who will help assess and circulate the final Open Educational Report to educators, practitioners, and community stakeholders. Results of this research should be useful to several LIS communities and beyond, including policy makers and regulators interested in expanding information and communication technology regulation to address the vulnerabilities of social and mobile data as digital evidence; building out the social infrastructure of the IMLS National Digital Platform and the broader library and archives community to include platform developers as preservation stakeholders; community archivists and social activists interested in documenting social movements across social media platforms; digital preservation practitioners in the field and students in training; and scholars examining the relationships between digital media firms, social networking platforms, mobile information and communication technologies, and the future of digital cultural memory.

University of Texas at Austin

Investigating Platform Development for Mobile and Social Media Data Preservation

In this Early Career Development project, Dr. Amelia Acker (School of Information, University of Texas at Austin) requests $308,921 from the Laura Bush 21st Century Librarian program for a three-year empirical investigation into emerging preservation tools and new data stewardship practices to answer the following research questions: (1) How do platform developers working in non-library contexts design and construct systems for the creation, transmission and preservation of mobile and social media data? (2) How does the provision of networked information services, including preservation technologies for mobile and social media data differ from established preservation infrastructures and professional practices in libraries, archives, and museums? (3) Which emerging preservation tools and new data stewardship practices are potentially transferrable to libraries, archives, and museums? In answering these questions, the research project will contribute both theoretical and practical knowledge about digital preservation models, approaches, and techniques across a number of communities with concern for the stewardship and future access of mobile and social media data collections. Outcomes of the investigation have the potential to create a much-needed point of connectivity between two domains of information provision and digital preservation.

1. Statement of Need

Social and mobile media platforms are systems that bind together documents and digital traces (data), users (creators), and producers (content and platform providers). As new and unique forms of digital cultural heritage, these data are created and move across a diverse ecology of platforms ranging from enterprise platforms (e.g. Outlook, Google Apps), to consumer offerings (e.g., Android or Facebook), to platforms that provide the infrastructural underpinning of the internet and mobile networks (e.g., Amazon Web Services, Verizon). Currently, data created by users in social and mobile platforms represents the fastest form of data creation and collection in internet connected information infrastructures. 77% of the world's mobile social traffic is dominated by Facebook and its mobile social subsidiaries, Instagram, Messenger, and WhatsApp (Taplin, 2017). Even more, most users accessing social content, news, and entertainment websites on the web are using apps and mobile devices connected to cellular networks. Yet, these data traces that are created when people connect to the internet and communicate are varied, born networked, and vulnerable to loss (Acker, 2015). Moreover, they exist in a platform ecosphere that is heterogeneous in content and context, ranging from activity streams like Facebook posts, Tweets, Snaps, to mobile data uploads such as mobile video, text messages, or even telephony metadata about GPS location (Duggan, 2015; Poushter, 2016).

Social and Mobile Media Data Preservation

There are currently several high-profile efforts by the digital preservation community that collect and provide research access to social media data in the US. Libraries, research institutions and community archives such as the Library of Congress Twitter Archive, the George Washington University's Social Feed Manager, or DocNow's app and tool suite each offer user-centered collections that extract data from platforms ("Documenting the Now," 2018; "Social Feed Manager," 2012; Pass, 2010). Despite these important efforts, mobile and social media data growing in size and transforming in content, acquiring new layers of networked engagement data and metadata. For example, recently the Library of Congress announced that it would be scaling back its Twitter Archive collection strategy based on the sheer volume and the changes of tweets themselves—now mobile media with moving images, videos, polling capabilities, and emoji stickers (amongst many other features) (Library of Congress, 2017). Social media and mobile platforms such as Facebook, Instagram, and Twitter often frame digital preservation as user-centered and *outside* platform infrastructure, and as such, provide tools for account holders to extract data from platforms that are authored by the user (only).

As in traditional archives, each of these models begin with the creator or author and are user-centered in terms of ownership, data-extraction and secondary use cases. While these "archive tools" allow users to download their personal social media data and metadata, they prohibit the extraction of engagement metadata—or that information that makes social media *social*. They also prohibit users from 'reading' their individual collections through older versions of the platforms or apps as opposed to playing an old video game in an emulator for example. Each of these platforms' models of memory begin (and end) with the creator as author: user-centered in terms of ownership, data-extraction and secondary use cases. However, critics of user-centered data extraction approaches have pointed out that they fail to reflect the user experience of social infrastructures, platform development and updates through time, or the reality of social networks up-stream—where platforms, user theories, and data products are designed, rolled out, and used *in situ*. This points to a double bind with the current reality of divesting data from platforms and the limits of digital object reconstruction from mobile and social media platforms in digital archives and preservation contexts. Because this active and ongoing decontextualization strips layers of context and limits secondary future uses, the current user-centered paradigm of social media data extraction for archives has stakes for how digital stewards describe, theorize, and confront information control in platforms, but also future contexts of accessing our digital cultural memory created in and embedded in social media platforms.

The techniques and tools of collecting social media data and their connection to digital preservation, reproducibility and comparability have been described extensively by social media researchers, computer scientists, and computational social scientists. In the past decade, several researchers and digital preservation practitioners have offered techniques for scraping, naming, labeling, and collecting these data collections in number of meaningful ways. For example, the developers of Social Feed Manager have discussed the provenance of a tweet and its powerful metadata (Kerchner et al., 2016). Axel Bruns and Katrin Weller have written widely about the impact of changes to the terms of service for the Twitter API and its impacts on researchers, and how it will impede future use cases (Bruns, 2013; Bruns & Weller, 2014, 2016). Light, Burgess, and Duguay have discussed an extensive and beguiling ethnographic walkthrough method for 'reading' mobile apps and describing the creation and transmission because accurate storage and backup are not possible (Ben Light, Jean Burgess, & Stefanie Duguay, 2016). And for many years, archival scholars, such as Michelle Caswell to Anne Gilliland have noted the power of cell phone records, networked metadata, and the power and vulnerabilities of mobile records (Caswell, 2009; Gilliland, 2014). While many LIS scholars and digital preservation practitioners have engaged with the power of web archives, few have begun to examine the impact of (and dearth of techniques for) preserving the mobile and social web in all its forms and dynamism. Clifford Lynch's recent call to document algorithms is a welcome innovation explicitly in this direction, by problematizing software updates, platform feeds, and AI filters (Lynch, 2017). Recent research on social media collecting and preservation have pointed to problems with user-centered models of ownership over data ownership and terms of service, problems with API rollbacks, the blackboxing of algorithms, frequent feature updates (or their deprecation), and forgotten platform features (Driscoll & Walker, 2014; Weller, 2016). Moreover, problems of access, retrieval and description continue to be an issue with regard to users' privacy and the ethics of data collection for social science and technology regulation (Weller & Kinder-Kurlanda, 2016).

Many legal thinkers and policy makers have commented on the inability to opt-out of enrollment into these social infrastructures as data subjects who must submit (intentionally but more often unintentionally) to creating swaths of data that are then collected and re-purposed by third party companies and data brokers in exchange for the use of the tool or service (Gandy Jr., 2006; Lyon, 2014; Pasquale, 2016). And thus creating data to be collected by corporations and data brokers is increasingly compulsory to participate in society, public life, and

even state programs (Braman, 2006; Gandy Jr., 1993). In the United States for example, enrolling in Medicaid involves e-mail and thus setting up (and accessing) an email account. Paying for a street parking and parking tickets in local municipalities with cashless meters, such as the city of Austin, now involve the ParkX mobile app (2017). In both cases, having access to a mobile phone and a personal e-mail address are not meaningfully voluntary for most people to participate in these kinds of public services. As a result, the ethics of collecting personal information from networked platforms remains open to consideration, ongoing in scope and impact.

Beyond the compulsory creation of social and mobile media data, researchers have raised ethics concerns about reusing personal data of users who may be unaware of long-term preservation verses the competing risks of not preserving and collecting data (Thomson & Kilbride, 2015). Collecting social media and networked mobile media in archives or for scholarly research raises a number of secondary-use risks for stewards that involve safeguarding for accidental disclosure, de-anonymization, and the ability to opt-out for creators. Scholarly associations, such as the Association of Internet Researchers has provided a list of guidelines and questions to identify vulnerable and risks of personal and ethical data ("Ethics – AoIR," 2018). Increasingly digital methods textbooks tackle these issues and provide strategies for scholars. However, little is known about the perspective of reusing user data from the experts who build these platforms. What do platform developers think of these issues? Their perspective, completely falls outside current conceptual framing of mobile social media data collection for scholarship and programmatic boundaries of digital preservation models deployed in libraries, archives, and museums (LAMs).

Research Gap

Librarians, archivists, and preservation practitioners have been developing tools, standards and best practices for bitstream preservation, metadata standards, research data management, and documentation of digital media for several decades (Galloway, 2004, 2010). However, existing tools, standards, techniques and best practices are a poor fit for mobile platforms and social media data due to proprietary secondary-use restrictions, dynamic cross-platform conditions of content creation, the ascendency of cloud storage solutions for users, and limited capacities to emulate platform environments. Outside of law enforcement and records compliance laws, most mobile and social media platforms do not ensure the preservation of long-term user data for creators, researchers, or journalists.

Despite the lack of preservation resources for users and scholars, we increasingly see some of the most innovative, rapid, iterative, and impactful digital preservation technologies coming from industry, proprietary technology developers, and social network platforms themselves, such as the Facebook *Legacy Contact* product, which provides users an in-platform digital stewardship feature (Brubaker & Callison-Burch, 2016). Yet, social media platforms such as Facebook, Twitter, and LinkedIn do not have contractual obligations to preserve user's data for long-term access, while users may delete their personal data, no platforms agree to act as a digital preservation repository—this may change. Presently, there is little knowledge of how platform developers themselves make sense or conceive of future access, secondary use-cases, ethics of data collection, or digital preservation of mobile and social media data. Indeed, no empirical work exists in the theory or practice of digital preservation that specifically focuses on the platform developers who create social and mobile media platforms, tools, and experiences for creators. While computational social scientists and social media researchers have interrogated the longitudinal persistence of social media data sets and future use cases (Zubiaga, 2017), archival scholarship and digital preservation initiatives have largely overlooked the impact that platform developers have and continue to have on the long-term preservation of digital culture that is created and often 'locked-in' to mobile and social media platforms. More work is needed to understand the conceptual, technical, and moral impact of platform development practices on digital cultural memory and the possibilities

for future preservation mandates (Acker, 2015). More research is also needed to assess existing proficiencies, such as skill sets and tools that already exist for users (creators), and stewards, such as librarians and archivists, even researchers who use social and mobile media data in their work.

Overall there is a clear national need for expanded preservation models focused on social media across platforms and mobile-based content. While special collections of social media exist, they largely focus on mainstream platforms such as Twitter and Facebook. Relatively few, high profile libraries, archives, and museums are attempting to preserve and provide access to mobile social media for reproducible research or cultural posterity, even fewer resources exist for early career information professionals or LIS students of digital preservation to steward vulnerable and ephemeral digital culture created and accessed in contemporary platforms. Further, there is a need for educational materials that support this knowledge acquisition. This project will serve as a bridge between corporate platform creators, industry technologists, and public information professionals by providing empirical data about emerging and experimental approaches to long-term access to mobile and social media data in platform development cycles. The project director (PD) has already conducted preliminary research and completed proof-of-concept projects on mobile digital forensics, ethnographic work studying personal digital archives and Facebook, and used emerging and experimental preservation technologies in her metadata and mobile ICT classes.

Current scans of the digital preservation community indicate that leading approaches for social media preservation focus on artifact-based extraction techniques, such as data scraped from the web or from platform extraction tools (such as APIs), instead of engaging with the platform development cycle or developers directly. The ongoing omission of platform developers' perspectives from social media preservation discussions is problematic: It means that current and future claims by LIS researchers about platforms and preservation will need to be substantiated by a deeper body of evidence of platform developers, including their conceptualization, theories, and approaches to platform development for information service provision. There is an urgent need to develop comprehensive resources describing existing approaches and to establish known standards specific to long-term preservation of social media and mobile media data that comes out of the development cycle and practices of developers. This includes a conception of long-term access and a general vocabulary for understanding digital preservation from a platform development perspective.

This research project will collect empirical data about the current state of preservation infrastructure at a range of digital media organizations engaged in platform development and information service provision. Designed as a comparative project across a number of different field sites, Acker the PD, will observe and interpret data stewardship and digital preservation challenges in contemporary organizations that embrace diverse ownership models of social and mobile media data through platforms, mobile applications, and platform feature development. The PD has established contacts at each site, which is vital for securing access and effectively studying guarded organizations. Acker has already conducted research on digital preservation strategies for social media and mobile platforms (Acker, 2017; Acker & Kriesberg, 2017), personal digital archiving with social media (Acker & Brubaker, 2014), personal identity, device ownership and mobile apps (Acker, 2015; Acker & Beaton, 2017). Building on her existing research agenda and using digital and qualitative methods, Acker will employ a multidimensional approach to inquiry, from ethnographic observation, interviews, case study, and trace data analysis. A graduate researcher will be recruited to assist in the cleaning, coding, and synthesis of the data, as well as the circulation of results. The research project aims to cultivate cross-domain expertise, social infrastructure, and digital preservation knowledge for information practitioners working in LAMs and beyond. The proposed research will have an impact on professional education, scholarly research,

digital preservation theories and models, electronic evidence policy, information privacy perspectives, and professional practice across US cultural heritage organizations.

2. Project Design

Using embedded research techniques including fieldwork observations, interviews, content analysis of documentation and workflows, and ethnography the PD will generate empirical data and gather technical documentation from platform developers (such as software engineers, product managers, and data workers) at different organizations as they build tools, programs, and new products addressing contemporary digital preservation of mobile and social experiences as part of their internal operations and information service provision. At each site, the long-term preservation, authenticity, versioning, and access to mobile and social media is a concern, yet, the values, ethos, and community traits that motivate their business models and development perspectives of long-term stewardship for platforms will be varied and are as yet unknown. What can we learn from these sites? How do technologists such as platform developers in different preservation ecospheres with different information supply chain logics conceptualize long-term access to social and mobile media data? What can libraries, archives, museums, and LIS educators learn from these contexts? Findings from the project will prepare the ground to develop new digital preservation curriculum, professional development opportunities, and further current scholarship on digital preservation and information infrastructure.

In software and platform development, large portions of developers' time are spent iterating models, analyzing use and performance, incorporating suggestions for changes, and creating updates into the design or the experience of the system. Similar to platform developers and software engineers, the ethnographic method involves observing (Spradley, 2016), documenting (Maanen, 2011), and analyzing the routines and actions of those under study (Lofland, 2006). For scholars, social and mobile platform development poses new challenges for sociotechnical research because the development of features is largely hidden and guarded. However, a serious outcome of organizations that guard social or mobile platform development processes from scholarly observation is that social and mobile computing products may influence a great number of people (or users) but the processes of development and standardization are unknown and may impact future use cases and preservation mandates of cultural heritage institutions. The investigation will test the hypothesis that a better understanding of such modes of production and provision of information services from social and mobile media data platform providers can inform, impact, and potentially transform the current capacities of public sector LAMs to preserve mobile and social media.

Goals and Objectives

With graduate research assistance, Acker will gather qualitative, ethnographic data from a variety of different digital media organizations (industry, government, not-for-profit) that focus on social software and mobile communication technology development. In this project, we plan to investigate a range of platform development ecologies, by focusing on the platform developers (technologists, software designers, product managers, and engineers) who contribute, design, and build social media platforms and mobile applications. The goals and objectives of this research is to contribute to the theoretical and practical knowledge in the domains of library and information science by examining (1) the role that platform developers play in the software ecology of digital preservation, (2) if platform developers' conceptions, models, or theories of digital memory contribute to or have significant impact on social media and mobile media's durability, vulnerability, or preservation possibilities; (3) how these conceptions of digital memory and long-term access influence future possibilities of digital preservation for creators or primary users, as well as secondary users such as researchers, librarians or archivists; yielding knowledge about (4) the practical, technical, and social requirements and considerations needed for the digital preservation of mobile and social media data.

Research Questions

The project's core research questions (RQs) reflect these goals and objectives, as well as larger issues described in the statement of need.

1. How do platform developers working in non-library contexts design and construct systems for the creation, transmission and preservation of mobile and social media data?
2. How does the provision of networked information services in mobile and social platforms, including preservation technologies for mobile and social media data differ from established preservation infrastructures and professional practices in libraries, archives, and museums?
3. Which emerging preservation tools and new data stewardship practices are potentially transferrable to libraries, archives, and museums?

Sample: The research investigation will examine technology development organizations engaged in platform development work for mobile and social media data applications. These include digital media firms, internet entertainment companies, not-for-profit arts organizations, and cybersecurity training outfits. For the purposes of this investigation "platform developers" represents a range of technologists, including software engineers, product managers, forensic investigators, product managers, and data workers employed in the development, testing, maintenance, and administration of mobile and social media platforms.

Implementation: The PD plans to interview informants and then visit each site for 2-6 weeks for in-depth observation, gathering data from developers over Years 2-3 as they work on projects over time. Having two points of data collection, interviews and then observational field work, will provide an opportunity to examine the development process over time, particularly across cycles and lifetimes of product rollouts. This is important since one potential outcome is to understand preservation concerns during design, implementation, and evaluation stages of platform development. Our qualitative data will include interviews with developers engaged in platform iteration, thus contributing RQ1 and RQ2. Investigators may observe advanced technological software, tools, and data that may be confidential and proprietary but potentially available in the future to LAMs once features have been released. As a result, and in order to respond to RQ3, we will anonymize data and synthesize broadly to preserve confidentiality and incorporate available secondary sources, such as data from W3C or ITU that is pertinent to our research findings in order to develop more detailed understandings of these platform development ecosystems.

Expected Outcomes

This is a three-year project, projected to begin in September 2018 and ending August 2021, activities in the first two years involve the development of the study's interview and recruitment protocols, data collection in the field, and data processing. In years 2-3 we will conduct analysis and interpretation of findings, prepare and present research products, disseminate an Open Educational Report, and submit the final report to the IMLS. Outcomes of the project will include: 1) a conceptual model of preservation infrastructures for mobile and social media data from a platform development perspective; 2) a needs assessment in terms of the training requirements of mobile social digital preservation for LIS students, practitioners in the field, and social media researchers; 3) an Open Educational Report for the education of digital preservationists in LIS programs, practitioners in the field, community archivists, and social media researchers.

Project Activities and Communication Plan

- Year 1 Activities (September 2018-August 2019): In the first six months of the project we will secure IRB approval for the project; prepare protocols for subject recruitment, interview protocols and data collection instruments; followed by recruitment of participants; the Advisory board will convene at the start of 2019 to plan ongoing consultations; and the project director will embed for the first rounds of ethnographic fieldwork during the summer of 2019. We will interview and hire the project's Graduate Research Assistant in spring 2019 to begin in Year 2. Preliminary findings and research design will be reported on at scholarly and professional conferences (such as ASIS&T, iConference, ALISE, PASIG, ALISE, SAA) throughout the first year.

- Year 2 Activities (September 2019 to August 2020): Data analysis and transcription of data collected from Year 1 will be performed by the project director and the GRA from September through February, this will include the inductive and deductive analysis of observational data, interviews, journals, visual data sources, and relevant secondary source materials; a coding schema based on this corpus will be developed in the spring. From September through December we will continue to interview participants from field sites. The Advisory board will convene in January and plan ongoing consultations. For the first six months of 2020 we will synthesize findings, draft preliminary results and circulate to partners, stakeholders, and the Advisory board. The project director will embed for the final ethnographic fieldwork observations during the summer of 2019. Preliminary results from Years 1-2 will be published in scholarly venues (JASIST, JDOC, JELIS) and reported on at conferences throughout the year.

- Year 3 Activities (September 2020 to August 2021): From September to December the team will analyze and process data collected, conduct inductive and deductive analysis, complete analysis of qualitative and quantitative data by February. We will share our synthesis with the Advisory board and meet in March for feedback. In the final year, the PD and GSR will produce an Open Educational Report which will be made publicly accessible through the Texas Digital Library. The OER will be publicized on professional listservs, as well community networks represented and influenced by the Advisory board members and research team. From March to August we will prepare drafts of the final products of the study, drafting conference papers, journal articles, press releases; we will develop OER report, prepare and present research products, anonymize data and deposit in appropriate repositories; then submit a final report to IMLS.

Roles and commitments of field sites and informants

Acker, the project director, has secured and continues to build access to field sites and recruit potential participants. Negotiating access and participation is an ongoing process in ethnographic research (Bryman, Bryman, & Burgess, 2002), and involves building relationships with "known sponsors" or orienting figures to each environment (Patton, 2001). Paul Leonardi has described these types of sponsors as a "project champion" or someone who can see the benefits of the proposed research and can ensure that it can occur at the proposed site (Leonardi, 2015). Having identified project champions at five sites, the PD has secured access to through a long-term process known as a reciprocity model of engagement with field sites which involves giving presentations, circulating research questions, teaching with tools developed by these organizations, and communicating how this research will benefit the organization, amongst other activities. While deep insight can be achieved by individual relationships with informants and project champions, articulated affiliations with field sites is most desirable, and if funded, the PD will negotiate organizational legitimacy through formal affiliations such as fellowships and visiting research status. Designed as a comparative project, each field site is marked by information service provision where experimental preservation technologies are being researched, designed,

tested, and implemented. Project champions at each of the field sites have agreed to partner in this research, providing the PD with access to personnel (engineers, system administrators, and product teams); organizational meetings; and in some cases, access to technical documentation, tools and products, including workflow data.

On confidentiality and the transferability of results

Many information scholars have discussed the nature of negotiating access to ethnographic research sites engaged in engineering, software development, and technical innovation (Kelty, 2008; Leonardi, 2015). While some technoscientific domains, such as a publicly funded research lab or a university research library may be easier for researchers to access, high-tech organizations, particularly in product development stages, are known to be notoriously difficult to negotiate access for observation. While the project director has been negotiating access, and building ongoing relationships to these guarded sites since 2014, challenges continually arise particularly with regard to confidentiality. Part of the challenge in obtaining access to such places is that practitioners, for any number of reasons, would prefer not to be observed or questions by nosy ethnographers. This concern is valid: negative representations of work ongoing at field site could lead to public scrutiny, legal actions, even the dissolution or failure of innovation. A central problem with gaining entrée to organizations undertaking the development and implementation of information technology is that they may be corporate, for-profit environments or operating under government contracts. Organizations engaged in creating new technologies must guard trade secrets and protect their proprietary information as part of on-going operations; ideally, these development projects will result in useful and profitable products for their company, start-up, or organization. Thus, negotiating entrée to sites where individuals work intensively with proprietary information technologies, can be difficult because researchers must balance rights to privacy, confidentiality, and proprietary information at multiple registers, from field sites, to individual platform products or processes, to informants' personal identities. Another practical difficulty is observing or witnessing the platform development cycle without betraying the concerns of organizational representatives who may not be the informants or developers themselves, but may represent leadership, marketing, or legal interests of the firm. In response to these concerns, we will provide confidentially to each field site to minimize their concerns according participants, sites themselves, organizational profiles, or products.[1] To this end, we will ensure that descriptions of participants, products, and sites are generic so that outsiders cannot identify the product, technology, or development process until it has been released publicly (or, possibly even deprecated or shuttered) or previously agreed upon with organizations. We will only collect digital documents, work products, and artifacts that do not personally identify informants or field sites, and these artifacts will be stored in a locked drawer within the PD's office or on a password protected encrypted external hard drive. For more on the project's data security protocols, please see the Digital Product form.

Finally, the confidentiality concerns of each field site organization must be balanced with concerns about supporting scholarly research often rest on the transferability of results to broader audiences, public-private, or scholarly and occupational communities. While it is necessary to protect the privacy and confidentiality of high-tech corporations and organizations involved in platform development, it is vital to study secretive, closed, or

[1] In each case this will be negotiated with a Memorandum of Understanding (MOU) that describes what confidentiality means in practice during observation, synthesis and reporting. IMLS full proposals and supporting documents can be subject to FOIA requests, as such the project director has decided (in consultation with informants, project champions, and Advisory board leadership) not to name or provide details of the five field sites and pre-emptively violate any confidentiality concerns. For brief descriptions of proposed field sites please reference the original Preliminary Proposal.

private organizations to expand the empirical record, to create public awareness, and to create new conditions of social change and possibility. In this case, following developers through the processes of platform development, in its conception stages, design cycles, implementation and use, will give early verification to the hypothesis and research questions outlined above, which, until now, have lacked description and theorization in terms of digital preservation for LAMs and in LIS research. In order for digital stewards to create the communicable value of digital preservation to mobile and social media data contexts, we need empirical data about the beginnings of platform development.

Research Team

Project Director: Amelia Acker, PhD, is an Assistant Professor at the UT iSchool where she teaches digital preservation, metadata, literacy and memory technologies. Acker's current research agenda is concerned with the emergence, standardization, and preservation of data in mobile and social media platforms. Her impact on the field of LIS can be seen in four areas of focus: (1) data literacy, (2) the preservation of mobile media, (3) social media metadata, and (4) critical data studies. In 2017, Acker received early access to the Obama White House social media data archive for her research on social media metadata.

Graduate Research Assistant: To be recruited for assistance in Years 2-3 of the project**.** The roles and responsibilities of the GRA will include active participation in data collection, processing the data (transcribing interviews; sorting; labeling; organizing files; depositing data to secure storage); active participation in data analysis, including coding and working with the PD to develop conceptual framing; collaborative writing, presenting the projects findings at conferences. For job description, see Key Project Staff document.

Advisory board

Three experts have agreed to serve as advisors: Jed Brubaker (CU Boulder), Ed Summers (UMD), and Jessica Meyerson (Software Preservation Network); each of which have expertise in areas of software preservation, digital identity and stewardship, ethnographic methods, digital curation, and web archives. The advisers will provide advice, ongoing assessment, evaluation, and contribute a brief report at the end of the three-year project. The Advisory board and project director will meet for three conference call meetings annually throughout the project as a method of evaluating data collection and progress towards sharing findings. In their capacity as an advisory committee, these scholars will evaluate the design and progress of the project, assuring the quality and dissemination of conclusions. As needed, the investigators will consult with each advisor individually based on their knowledge and expertise. For more on advisory board members, please consult the Key Project Staff and Consultants document.

3. Diversity Plan

The research project intends to study a population (platform developers) and technology development organizations who have heretofore not been explicitly engaged with as an essential part of the digital preservation ecosphere. Because each of the sites are often guarded, the subjects and sites constitute a relatively understudied set of organizational profiles, and meet the criterion of diversity. In addition to bringing a plurality of new perspectives to the theory and application of digital preservation, this research has the potential to serve future users, as well as community archivists and social groups who may wish to document their organizations or communities in alternative ways not afforded by traditional libraries, archives, and museums. The communication plan is to circulate the Open Educational Report to different stakeholders, the Advisory board will offer guidance on distribution for best possible reach and uptake in engaging diverse communities as well. The project will also create inclusive social infrastructure—or bridges across communities of practice, and thus it will benefit participating organizations and platform technologists, developers, and social media designers by

improving relationships, and providing knowledge about the needs of researchers, user communities, and cultural heritage institutions.

4. National Impact

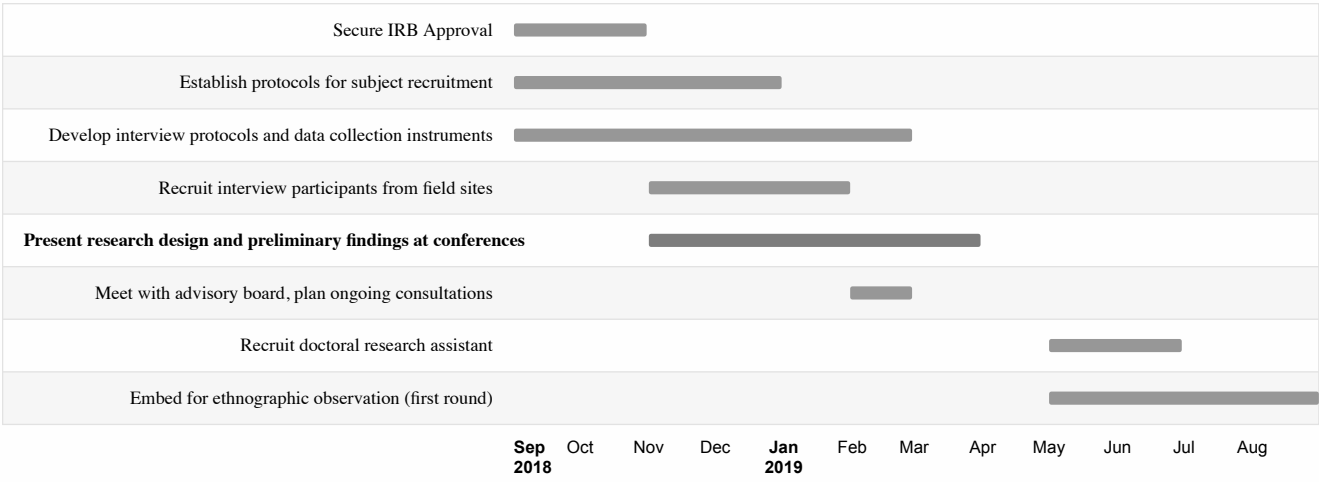
This project on preservation models and platform development cultures in the US will aid archives, libraries, and museums as they further develop best practices for digital preservation, particularly with regard to social media collections and emerging infrastructures, a key priority of focus the IMLS's National Digital Platform (Institute for Museum and Library Services, 2015). The goals of this research are (1) to gather information about how platforms are conceived in order to be able to inform the library and archives community about practical and technical matters regarding working with platforms; (2) to address conceptual gaps of social and mobile media preservation in platform development between technology organizations and current LAM approaches; (3) identify how these conceptions development and long-term access influence future possibilities of digital preservation for creators and secondary users such as researchers, librarians or archivists; yielding knowledge about (4) the practical, technical, and social requirements and considerations needed for the digital preservation of mobile and social media data.

As the LIS field continues to grapple with preserving and providing access to new forms of digital culture and data collections, studies of non-library preservation settings can expand current approaches and identify future solutions. The project will investigate a diverse set of technology development cultures where long-term data stewardship and digital preservation strategies are being re-envisioned in exciting ways with the potential to greatly benefit American archives, libraries, and museums. Understanding preservation contexts in platform development organizations can inform and potentially transform the current capacities and shared-service approaches of public sector libraries, archives, and museums, which are in the early stages of collecting and preserving mobile and social media. In addition to its potential impacts on LIS theory, research, and education in the US, the project aims to cultivate cross-domain expertise, social infrastructure, and digital preservation knowledge for information practitioners working in the public sector. The proposed research will have an impact on professional education, scholarly research, digital preservation theories and models, electronic evidence policy, and professional practice across U.S. cultural heritage organizations. These benefits will be sustained beyond the conclusion of the award through a communication plan that involves scholarly publications, presenting results at professional conferences, and circulating an Open Educational Report for a diverse group of stakeholders. The research project extends Dr. Acker's long-term research agenda on social media metadata and mobile media data preservation, focusing on the development, testing, and outcomes of preservation infrastructures in different development and innovation communities. Finally, the research will have an immediate pedagogical impact on early-career LIS professionals when MSIS students take Dr. Acker's courses on metadata and digital preservation at the UT iSchool, and apply their knowledge and skills in the field shortly thereafter.

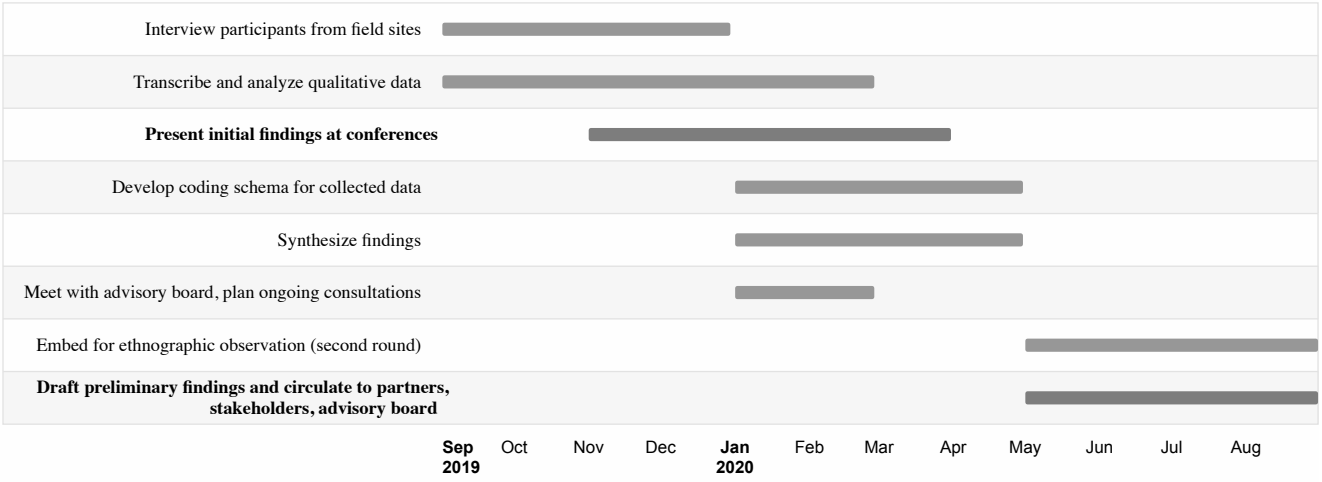
University of Texas at Austin

Schedule of Completion
(Deliverables in bold)

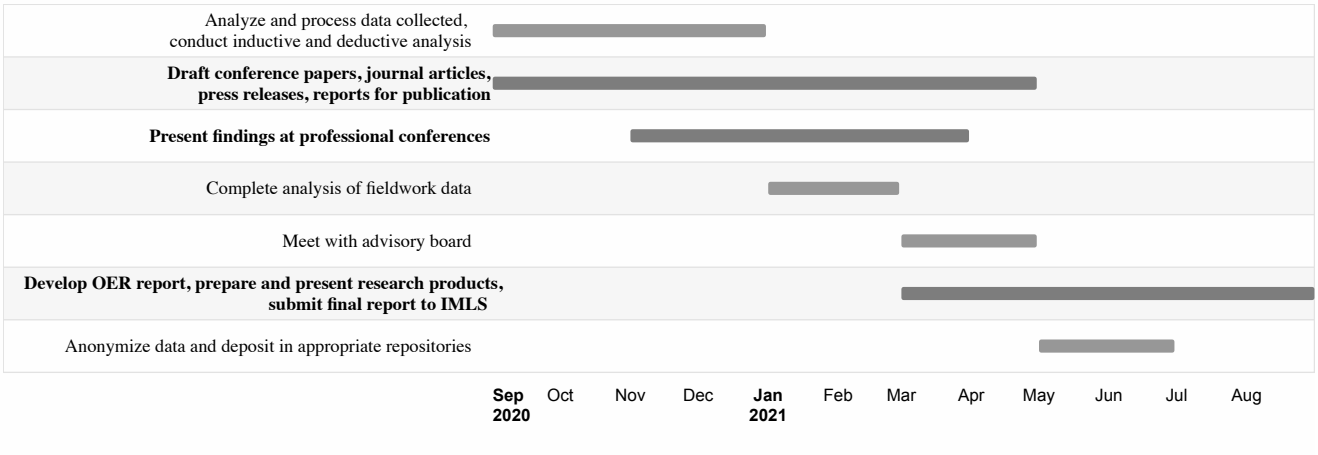
Year 1: Sept 2018-August 2019



Year 2: Sept 2019-August 2020



Year 3: Sept 2020-August 2021



DIGITAL PRODUCT FORM

Introduction
The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., digital content, resources, assets, software, and datasets). The products you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products can be challenging. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and practices that could become quickly outdated. Instead, we ask that you answer questions that address specific aspects of creating and managing digital products. Like all components of your IMLS application, your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

- ☐ Please check here if you have reviewed Parts I, II, III, and IV below and you have determined that your proposal does NOT involve the creation of digital products (i.e., digital content, resources, assets, software, or datasets). You must still submit this Digital Product Form with your proposal even if you check this box, because this Digital Product Form is a Required Document.

If you ARE creating digital products, you must provide answers to the questions in Part I. In addition, you must also complete at least one of the subsequent sections. If you intend to create or collect digital content, resources, or assets, complete Part II. If you intend to develop software, complete Part III. If you intend to create a dataset, complete Part IV.

Part I: Intellectual Property Rights and Permissions

A.1 What will be the intellectual property status of the digital products (content, resources, assets, software, or datasets) you intend to create? Who will hold the copyright(s)? How will you explain property rights and permissions to potential users (for example, by assigning a non-restrictive license such as BSD, GNU, MIT, or Creative Commons to the product)? Explain and justify your licensing selections.

The main research products will appear as text and images in peer-reviewed journal articles, conference proceedings, book chapters, and a final educational report. Primary source data will include transcripts from interviews, observational notes, sketches in order to facilitate re-use. The investigators will own the intellectual property of the dataset and resulting publications. The final dataset will be submitted to Texas Digital Library repository in accordance to current best practice guidelines for preservation, future access, and re-use. Investigators will also deposit preprint copies of all published research in the Texas Digital Library repository, which is open access. The open educational report (OER) and materials produced in Year 3 will be deposited and remain open access. The OER will be published under a Creative Commons Attribution 4.0 license, which will allow libraries, archives, and LIS educators to access and build off of the published materials. The CC license allows users to share and adapt the report for any purpose. All products will include attribution of the investigators and the funder, IMLS for supporting the research collecting and generating the data.

A.2 What ownership rights will your organization assert over the new digital products and what conditions will you impose on access and use? Explain and justify any terms of access and conditions of use and detail how you will notify potential users about relevant terms or conditions.

The University of Texas will impose no condition on access and use of the published research made available through the Texas Digital Library. Primary data and other supporting materials created or gathered in the course of the research project will be shared with other researchers and LIS practitioners upon reasonable request. Access to research data will follow a set of access protocols guided by the standards of the University of Texas

at Austin's Institutional Review Board in order to the protect privacy, confidentiality, and to respect any proprietary, intellectual property, or trade secrets of the project's field sites or interview subjects as described in the narrative of the grant. Terms of use will include proper attribution to the investigators along with disclaimers of liability in connection with any use of distribution of the research data. We support open access and will ask for attribution of IMLS support and in terms of possible publications.

A.3 If you will create any products that may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities, describe the issues and how you plan to address them.

Investigators from the project will publish the results of their work, which may include a selection of the data. Papers will primarily be published in peer-reviewed journals and conference proceedings. These data will be generated from ethnographic observation and semi-structured interviews with developers, technologists, and platform engineers. Qualitative data (interviews, transcripts, field notes, photographs, sketches) may involve privacy concerns and trade secrets. Personal information about participants will by anonymized and in some cases not collected to maintain research subjects' privacy. As such, transcripts will be de-identified and transcripts will not be published to reduce the possibility of disclosure or identification of persons or field sites. The PI will publish coding schema resulting from inductive analysis making anonymized data available to researchers if requested.

Part II: Projects Creating or Collecting Digital Content, Resources, or Assets

A. Creating or Collecting New Digital Content, Resources, or Assets

A.1 Describe the digital content, resources, or assets you will create or collect, the quantities of each type, and format you will use.

In addition to published journal articles, conference proceedings and book chapters, the project will publish an open-access Open Educational Report which will summarize the state of the art of digital preservation of social and mobile media data, and a survey of community assessment and educational resources and tools. Each of these resources will be published as PDFs and downloadable documents. Transcripts and codebooks will be formatted as PDFs or .docx word documents, photographs will be formatted as .jpg, audio recordings of interview will be .wav or .mp3.

A.2 List the equipment, software, and supplies that you will use to create the content, resources, or assets, or the name of the service provider that will perform the work.

The PI and research assistant will use machines and software provided by the University of Texas, in addition to encrypted cloud storage provided by the university. Transcription services will be contracted later at a competitive market rate.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to use, along with the relevant information about the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

File formats will include .docx, .pdf, .mp3, .wav, .pptx, and .jpg.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

The PI will manage the proposed deliverables, review all data and data project products before publication. The graduate research assistant will also review products as they are synthesized, and participate in quality assurance

of the coding, transcription, and anonymization of collected materials.

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance. Your plan may address storage systems, shared repositories, technical documentation, migration planning, and commitment of organizational funding for these purposes. Please note: You may charge the federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the federal award (see 2 C.F.R. § 200.461).

C. Metadata

C.1 Describe how you will produce any and all technical, descriptive, administrative, or preservation metadata. Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content, resources, or assets created during your project (e.g., an API [Application Programming Interface], contributions to a digital platform, or other ways you might enable batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content, resources, or assets available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name(s) and URL(s) (Uniform Resource Locator) for any examples of previous digital content, resources, or assets your organization has created.

Part III. Projects Developing Software

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) it will serve.

A.2 List other existing software that wholly or partially performs the same functions, and explain how the software you intend to create is different, and justify why those differences are significant and necessary.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software and explain why you chose them.

B.2 Describe how the software you intend to create will extend or interoperate with relevant existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the software you intend to create.

B.4 Describe the processes you will use for development, documentation, and for maintaining and updating documentation for users of the software.

B.5 Provide the name(s) and URL(s) for examples of any previous software your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under open-source licenses to maximize access and promote reuse. What ownership rights will your organization assert over the software you intend to create, and what conditions will you impose on its access and use? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain and justify any prohibitive terms or conditions of use or access and detail how you will notify potential users about relevant terms and conditions.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will deposit the source code for the software you intend to develop:

Name of publicly accessible source code repository:

URL:

Part IV: Projects Creating Datasets

A.1 Identify the type of data you plan to collect or generate, and the purpose or intended use to which you expect it to be put. Describe the method(s) you will use and the approximate dates or intervals at which you will collect or generate it.

The intended purpose of this ethnographic, observational, and interview data is to support qualitative inquiry and theoretical development of the digital preservation of social media and mobile media data. This qualitative data will result from field site observations, fieldwork activities, and semi-structured interviews. Interview and

observational data collected will be memos, audio transcripts, interview transcriptions, photographs, field notes, technical documentation, and workbooks. In Years 1-2 interview data will be transcribed and memos will be synthesized to create a codebook. Transcripts will then be coded using the codebook in Years 3-4. The data will be used and interpreted in scholarly publications in Years 1-3 and one freely accessible open educational report placed in the Texas Digital Library under a CC 4.0 license for access in Year 3.

A.2 Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

Yes, IRB approval is required and all data collection practices will be reviewed by the University of Texas at Austin's IRB. An IRB application will be submitted after grant approval, as outlined in the schedule of completion we expect to secure approval in the beginning of Year 1, with reapplications for approval as directed by the IRB.

A.3 Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

The fieldwork will not collect personally identifiable information from interview subjects, but may in the course of observation collect confidential information or proprietary information such as new technologies in development. Every effort will be made in compliance with UT's IRB research protocols to preserve confidential and proprietary information of field sites. Final datasets submitted for public release will be anonymized and contain no personally identifiable information.

A.4 If you will collect additional documentation, such as consent agreements, along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

Researchers will obtain verbal consent according to a human subjects' protocol and such verbal consent for participation will be recorded before the interviews. Once interviews have been transcribed, the audio recordings will be deleted to preserve identifiable voice information. During the implementation of the research project, associated research data will be backed up on a password-protected secure server maintained by the University of Texas's computing infrastructure, in order to protect from loss of data from hardware failures, fire, theft, etc.

A.5 What methods will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Investigators will use notebooks, audio recorders, cameras, word processing software, and qualitative transcription software to generate and capture data from the project. These data will be converted to open formats (e.g., ODF, PDF, JPEG) for preservation and future re-use. Interviews will be transcribed and subsequent coding schema, reports, articles, and conference proceedings will be generated as .pdf, .docx, and .csv. In Year 3 anonymized interview data and accompanying codebook will be included in the transcription data deposited in the Texas Digital Library repository.

A.6 What documentation (e.g., data documentation, codebooks) will you capture or create along with the dataset(s)? Where will the documentation be stored and in what format(s)? How will you permanently associate and manage the

documentation with the dataset(s) it describes?

The codebook and interview transcripts will be stored as .docx documents.

A.7 What is your plan for archiving, managing, and disseminating data after the completion of the award-funded project?

During the implementation of the project, investigators will adhere to the data creation and management best practices outlined by the University of Texas at Austin's Human Research Protection Office. All members of the research team will familiarize themselves with good practices around data management by attending data management workshops at the university library and reviewing library resources related to data management. In implementing the data management plan for this project we will follow the policies set out in the University of Texas's guidelines on research data management. The investigators will be responsible for ensuring the data management plan is adhered to. Upon completion of the project, final dataset (including documentation, codebooks, and proper citation) will be submitted to the Texas Digital Library, a trusted, public and authenticated data repository. We will use our time in Years 1-2 assembling the data, we address dissemination in Years 1-3 in our narrative section.

A.8 Identify where you will deposit the dataset(s):

Name of repository: Texas Digital Library

URL: https://www.tdl.org/

A.9 When and how frequently will you review this data management plan? How will the implementation be monitored?

A long-term strategy for the maintenance, curation, and archiving of data will be implemented. Our advisory board will monitor the implementation, reviewing the data management plan three times during annual meetings. Investigators will review the data management plan every 6 months throughout the course of the project. Datasets will be curated and preserved according to the practices of the Texas Digital Library.