(1) Name of project

LGBTQ+ Audio Archive Mining project

(2) List of team members, titles, and roles on the project

- ***Project Leads***:
  - **Ann Hanlon**, Head, Digital Collections & Initiatives and Digital Humanities Lab, UWM Libraries
  - **Dan Siercks**, Interim Director, Web and Data Services, UWM College of Letters and Science
  -
- ***Senior Administrator***:
  - **Marcy Bidney**, Assistant Director of Libraries and Curator of the American Geographical Society Library, UWM Libraries
  -
- ***Disciplinary Scholar***:
  - **Cary Costello**, Associate Professor, Department of Sociology & Director, LGBT Studies Program, UWM College of Letters and Science
  -
- ***Team***:
  - **Shiraz Bhathena**, Digital Archivist, Archives, UWM Libraries
  - **Jie Chen**, Application Specialist, Digital Collections & Initiatives, UWM Libraries
  - **Karl Holten**, Information Systems Specialist, joint appointment: Digital Collections & Initiatives, UWM Libraries, and Web and Data Services, College of Letters & Science
  - **Ling Meng**, Digital Collections Librarian, Digital Collections & Initiatives, UWM Libraries

(3) Investigator Bios

Project Leads

Ann Hanlon: Ann Hanlon is Head of Digital Collections and Initiatives at the University of Wisconsin-Milwaukee. She also co-founded and leads the UWM Libraries Digital Humanities Lab. Ms. Hanlon has nearly twenty years of experience working with digital collections, including positions at the University of Maryland, Marquette University, and since 2012 at UWM. She has led projects to build digital archival collections of all shapes and sizes. She has also led initiatives in the areas of digital preservation and digital scholarship. She currently serves as Chair of the Network Council for the Digital Public Library of America (DPLA). She has published and presented in the fields of digital collections and scholarship, digital preservation, and digital humanities. Ann has an MA in History from the University of Maryland and an MSLIS from the University of Illinois at Urbana-Champaign.

Dan Siercks: For the past ten years Dan Siercks has led research computing efforts across the College of Letters and Science at University of Wisconsin-Milwaukee. In April 2018 he assumed the role of Interim Director of IT for the College of Letters and Science and currently leads its Web and Data office. Dan is an XSede (Extreme Science and Engineering Discovery Environment) campus champion, providing expert support for advanced digital services, including high-performance computing. He has presented on applications of high-performance computing in non-traditional disciplines at Educause Midwest, and in the UWM Libraries Digital Humanities Lab. Dan holds a B.S. in Computer Science from UW-Milwaukee and a M.S. in Data Science from UW-Eau Claire.

Senior Administrator

Marcy Bidney: Marcy Bidney is the Assistant Director for Distinctive Collections at the University of Wisconsin-Milwaukee Libraries. She also serves as the Curator of the American Geographical Society Library. Marcy's main interest and research focus in libraries includes utilizing evolving technologies to provide increased access to geographic information collections in libraries. Other interests include the digital and spatial humanities, geographic education, the history of cartography and the history of Geography.

Marcy's professional affiliations include the American Library Association where she has chaired several committees and discussion groups and served as Chair of the Map and Geospatial Information Roundtable in 2011. She currently serves as the Chair of the Map and Geospatial Information Curators group of the International Cartographic Association's Commission on Cartographic Heritage into the Digital and as a member of the Board of Review for the Osher Map Library and Smith Center for Cartographic Education at the University of Southern Maine. In 2018 she became a co-editor of the Journal of Map and Geography Libraries.

Disciplinary Scholar

Cary Costello: Professor Costello engages in research in embodied experience and interventions into embodied identity. In one line of research, he studies the regulation of sex and gender through medical interventions into the bodies of intersex and transgender people. In another, he examines embodiment in virtual settings. In a longitudinal study of avatar embodiment in the virtual world of Second Life, he studies how identification with the avatar body facilitates experiences of sensation in virtual flesh, and the therapeutic uses to which people put their avatars. He holds a PhD from University of California, Berkeley and a JD from Harvard Law School.

(4) Summary of Project

The UWM Libraries are asking for $49,790 to develop and make openly available models for extracting text, building usable text datasets, and developing public-facing data visualizations based on audiovisual (AV) materials in the LGBTQ+ collections in the UWM Archives. The *LGBTQ+ Audio Archive Mining project* will use machine learning tools and data analysis and visualization to build and process text datasets extracted from the AV materials in the UWM Archives LBGTQ+ collections. The UWM Libraries house one of the largest collections of historical and contemporary LGBTQ+ materials in Wisconsin, including a rich record of Milwaukee's LGBTQ+ communities. Researchers will be able to use the text dataset for analysis as well discovery, and we anticipate use of this "mined data" will not only help researchers discover people and relationships that might have remained hidden in bounded and difficult-to-use formats, but will also help researchers generate new questions about the collections and make connections with related collections outside our own repository. The project will provide a model for mining archival LGBTQ+ AV materials using open source tools and infrastructure in an ethical manner and with an eye toward deliverables that enhance the discovery and use of these collections, and the possibility of better understanding the sometimes-hidden history of the LGBTQ+ community in Milwaukee.

(5) Project Rationale and Statement of Significance

Audiovisual (AV) materials make up a significant portion of the formats found in mid-to-late twentieth century archival collections. However, their format can make it impossible to discern an object's topical content with a quick visual browse; understanding AV content is necessarily time intensive. However, the emergence of machine learning technologies that can identify patterns and extract text from digital AV files that are primarily spoken word presents opportunities for both archival processing and for research. This is especially important in the case of communities whose histories have largely been hidden, such as the LGBTQ+ community. AV materials in collections that are not identified as LGBTQ+ archival collections, especially broadcast media collections, often contain materials, some of it quite rich, pertaining to LGBTQ+ history. These materials can reveal changing public attitudes toward LGBTQ+ communities, the development of advocacy trajectories over time, and intersections with social justice movements.

The UWM Libraries house one of the largest collections of historical and contemporary LGBTQ+ materials in Wisconsin, including a rich record of Milwaukee's LGBTQ+ communities. Milwaukee is the largest city in Wisconsin, and in the early part of the 20th century, was considered home to one of the most vibrant and visible gay communities in America. While digitized textual materials in the collection are regularly utilized, AV materials in the LGBTQ+ collections, also important for research, are underutilized. Often, they are minimally described – if they are described at the item-level at all. Minimal descriptive access for AV materials can make it

difficult for researchers to identify items relating to more specific areas of research. For example, The Libraries' ACT-UP Milwaukee collection includes brief descriptions of the content, e.g. "Excerpt on statements made by Milwaukee Health Department's Medical Director Thomas Schlenker on HIV/AIDS." However, the brevity of that description does not convey to the researcher anything about the content of Schlenker's comments, which framed AIDS as a "gay disease," homosexuality as a mental illness, and gay men as promiscuous and deserving their fate. This video is of immense value to a researcher studying institutional homophobia, the medicalization of same-gender sexuality as mental illness, or sex panics. But researchers cannot locate this relevant archival item by searching for any of those terms. Content mining, theme-identification and tagging are necessary to surface important content like this.

The UWM Archives also have significant AV holdings in collections that aren't identified as LGBTQ+ collections, but nevertheless have important content documenting LGBTQ+ history in Milwaukee. These collections include the records of the campus radio station, WUWM, and the newsreel collection for local television station WTMJ. Converting the spoken word in these AV files to text and using tools to identify a wide range of themes that can be tagged can enable researchers to find these pertinent materials to use as data. We know, for example, that the collection of Milwaukee Journal Stations Records, 1922-1997, contains content that relates to LGBTQ+ topics. But the archive is vast, and which films and radio recordings are relevant, and the specific topics in sexuality or gender identity or expression to which they pertain, are not identified. And there are a rich host of historical events that took place in Milwaukee that are very likely documented in as-yet-unidentified AV sources in the collection. These include the activism of groups such as the Gay People's Organization and National Lesbian Feminist Organization in the 1970s; Milwaukee's passage of a bill protecting against discrimination on the basis of sexual orientation in 1980; the serial murders of men, mostly of color, by Jeffrey Dahmer between 1978 and; the activities of ACT-UP Milwaukee; and the founding of Alliance High School, a safe environment for LGBTA+ students in 2005.

What needs to be done

The possibilities for mining and describing archival AV has been noted by at least two recent and ongoing projects – the High-Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Humanities, and the Audiovisual Metadata Platform (AMP) being developed at Indiana University. While the HiPSTAS project articulated preliminary solutions that match the problem we have laid out ("…using machine learning tools to improve discovery with unprocessed audio collections"[1]), research for that project focused on the prosodic elements of sound files and not on speech extraction or topic modeling. The AMP project shares a PI with HiPSTAS and aims to develop a "metadata-generation interface made openly available to collection managers around the world."[2]

1 Clement, Tanya et al (2014). High Performance Sound Technologies for Access and Scholarship (HiPSTAS) in the Digital Humanities. ASIST 2014, November 1-4, 2014, Seattle, Washington

2 Press release, November 1, 2018. IU Libraries receives $1.2 million grant to develop ability to search digitized audiovisual files.

Our project does not seek to replicate the efforts of projects like HiPSTAS or AMP. Rather, the workflow we outline is one that provides a replicable model for creating digital audio files to enhance our understanding of the content and advance their discoverability and use both as historical evidence and as a database for analysis, to surface information and patterns not detectable from item level description. Using pre-trained machine learning models, we will develop methods for working with audio and text files at scale from our AV collections. We will organize the resulting text-based dataset for data analysis processes into a publicly accessible dashboard that will allow researchers to visualize relationships, topic clusters, and word frequencies. Our aim is to aid understanding of the contents of these collections, and discover previously unrecognized topics, relationships, and patterns that shed light on the history of the LGBTQ+ community in Milwaukee. The tools and methods we employ will be open source and we will endeavor to develop a workflow with a low barrier to entry in order to encourage replication at a variety of institutions, including as a potential case study for AMP or HiPSTAS.

The UWM Libraries are especially well-suited to hosting a pilot project that investigates AV formats, AV and text-mining technologies, and how those methods might be impactful for documenting the history of the LGBTQ+ community, especially in the upper Midwest. The UWM Archives works closely with the Department of Digital Collections & Initiatives (DC&I) to identify best practices and workflows for digitizing and managing digitized AV materials. DC&I was established in 2002, and as of October 2019, had developed sixty-five publicly accessible digital collections comprised of over 170,000 items[3]. While most publicly accessible digital objects in the UWM Digital Collections are images, a significant and growing portion include AV materials – an area of growing emphasis in both Archives and DC&I. Oral history projects have been an area of growth, with thirteen oral history collections available online and more in development, including several LGBTQ+ projects. Another major AV collection, the WTMJ-TV News Archives[4], is the largest surviving body of television news footage in Wisconsin. It dates from 1950 to 1980 and consists of approximately two million feet of 16mm film. A very small percentage of that footage is digitized, though a segment-level catalog is available for online search. With a mature digitization and digital collections program in place, workflows have been implemented to facilitate the reformatting of most AV materials either during initial processing or as a byproduct of patron request. This means that the availability of digitally reformatted AV materials in our archival collections is growing and presents an opportunity to better understand those materials through computational analysis.

The project will bring together three units from across campus in order to make this project happen: The UWM Archives, and the Digital Collections and Initiatives (DC&I) departments in the UWM Libraries (including the Digital Humanities Lab); and the Web and Data Services department in the UWM College of Letters of Science. The Archives and DC&I have a long history of working together to create community-engaged digital collections, including the

[3] https://uwm.edu/lib-collections/
[4] https://uwm.edu/wtmjsearch/

award-winning March on Milwaukee[5] and Milwaukee Polonia[6] collections, and oral history collections that include the Wisconsin HIV/AIDS History Project[7], and the Milwaukee Transgender Oral History Project[8]. Alongside that ongoing and well-established workflow, the Digital Humanities Lab was launched in 2013 to provide an interdisciplinary, collaborative space for experiments and inquiry into digital methods in the humanities to benefit both teaching and research.

In order to make this project successful and implement these models in a long-term and sustainable manner across collections – and as a model application for other cultural heritage collections, staff need four essential things:

1. Training in R and other methods for text analysis, and RShiny for visualization and delivery, as well as other data and text analysis tools and methods;
2. Training to understand and effectively use machine learning technology and applications;
3. Training in the ethical management and delivery of large datasets comprised of historical, sensitive, and personal data;
4. Support for cross-divisional coordination of time and resources; including strategic planning for staffing models in the DH Lab that support data-driven research and facilitate future project development for both scholarship and teaching.

While staff in L&S Web Services are well-versed in R and other data analysis methods, Library staff will benefit from building this skill set as an application for other data-driven research supported by the Libraries, especially as researchers begin to approach our cultural heritage collections computationally. All team members will benefit from jointly approaching the ethical questions posed by working with large datasets that document historically underrepresented communities. For the Libraries, we anticipate building a strong model for developing collections-as-data initiatives, as well as initiating further investigations into machine learning, with this model project as a testbed for understanding these processes in a concrete manner.

There are several communities who will benefit from the *LGBTQ+ Audio Archive Mining* project deliverables, as described in more detail in the Draft Use Model. Importantly, the project will provide a more well-marked path to our LGBTQ+ collections, and aiding students and community researchers who might not otherwise discover these materials. The project will be immediately useful to Sociology, History, Health Sciences, and School of Information Studies faculty, graduate students, and undergraduate students interested in the history of the LGBTQ+ community in Milwaukee, as well as in using data-driven research techniques for working with digital and digitized primary source material. Researchers will be able to use the text dataset for analysis as well discovery, and we anticipate use of this "mined data" will not only help

[5] https://uwm.edu/marchonmilwaukee/
[6] https://uwm.edu/mkepolonia/
[7] https://uwm.edu/lib-collections/wisconsin-hiv-aids-history-project/
[8] https://collections.lib.uwm.edu/digital/collection/transhist/search

researchers discover people and relationships that might have remained hidden in bounded and difficult-to-use formats, but will also help researchers generate new questions about the collections and make connections with related collections outside our own repository.

Statement of Significance

The UWM Libraries house one of the largest collections of LGBTQ+ materials in the Midwest, providing a rich record of Milwaukee's LGBTQ+ communities. The UWM Archives and Special Collections include important LGBTQ+ manuscript collections that tell the story of the gay rights movement in Milwaukee, lesbian organizing, and the local history of HIV/AIDS advocacy. Additionally, the collections include local LGBTQ+ newspapers, magazines, and other print publications.

An area of strength includes documentation of the local gay rights movement in the 1970s, including the records of the Gay Peoples Union (GPU), which produced a half-hour radio program called Gay Perspective that aired locally from 1971 to 1972. As far as we know, this was the first regularly scheduled, scripted gay and lesbian radio program in the nation, and is available as part of a digital collection. Like the GPU collection, many of the most highly used LGBTQ+ collections include AV materials and many are available online. These include the collections of ACT-UP Milwaukee, the AIDS Resource Center of Wisconsin (ARCW), Tri-Cable Tonight, a program produced by the Milwaukee Gay Lesbian Cable Network, and oral history collections, including the Milwaukee Transgender Oral History Project, the Milwaukee LGBT Oral History Project, and the Wisconsin HIV/AIDS Oral History Project.

Our LGBTQ+ collections include far more content that is not online, including numerous collections with AV materials. Collections might remain offline for reasons ranging from privacy issues to copyright, but most may be in the process of being made publicly accessible in our digital collections or undergoing additional processing as new accessions arrive.

The research value of these collections is significant, especially as they shed light on Wisconsin's unique contributions to LGBTQ+ history – a history that is often told disproportionately from the perspective of the major coastal cities. Most of our local holdings date from the period following the Stonewall Riots in 1969, although some provide glimpses of the lives of gay men and lesbians following World War II through the 1960s. Collections from the 1970s provide evidence of the local Gay and Lesbian Liberation Movement and the gradual emergence of groups within business, sports, theater, health care, and the media. Documentation from the 1980s shows the community's development in the face of the AIDS crisis and the mounting opposition from the New Right. In the 1990s and early years of the twenty-first century, debates about military service and marriage equity became more prominent. The breadth of coverage of topics makes these materials especially rich candidates for computational approaches that can reveal insights about language usage over time, sentiments regarding marriage equity across different populations, or patterns of negative or positive contexts for LGBTQ+ -specific terms and phrases.

UWM is home to the nation's second oldest independent LGBT Studies Program, which teaches nine courses per year, and seeks to involve students in original research using archival materials housed in the Libraries. Classes that have used the collections – and will benefit from additional access, description, and insight, include LGBT 200 (Intro to LGBT Studies), WGS 192 (Intro to Womens and Gender Studies), WGS 301 (Lesbian Feminism), HIST 294 (Historical Research Methods), and HIST 900 (Sports, Race, & Gender). Researchers have also benefited from the collections and recently, several books based on these materials have been published, including R. Richard Wagner's two-volume history of gay life in Wisconsin, *We've Been Here All Along: Wisconsin's Early Gay History,* published April 2019 by the Wisconsin Historical Society Press, and the forthcoming second volume, *Coming Out, Moving Forward*. And in 2017, Brice Smith published *Lou Sullivan: Daring to be a Man Among Men* (Trangress Press), a historical biography of Milwaukee-born transgender activist Lou Sullivan, drawing on materials in the UWM Archives collections. There is great hidden potential in the LGBTQ+ collections at UWM for far more teaching and research, especially utilizing the AV materials that lack sufficient descriptive information or the context that both students and researchers need. Surfacing the important content in these collections through better descriptive access and discovery mechanisms will undoubtedly benefit these communities not only in Wisconsin but throughout the LGBTQ+ research spectrum.

(6) Project Plan

Intellectual property

The collections we propose to include in our project are all 20th and 21st century collections and therefore include copyrighted materials. In all cases, the UWM Archives has explicitly obtained permissions to publish and disseminate those materials online, such as in the case of oral histories conducted for the UWM Libraries. In those cases where copyright may be an issue – for example, in the case of AV materials from the collections of WUWM or WTMJ – the secondary text data sets will be made available through an unmediated website, while AV files themselves may be downloaded with permission for research and educational purposes. The UWM Libraries has a strong working relationship with the copyright holders in these instances, including a copyright license agreement with the Journal Broadcast Group that grants to the UWM Libraries "a royalty free, non-transferable, non-exclusive, perpetual, world-wide license to digitize the Materials *[WTMJ news film]* and to use and make the Materials publically available in any and all online formats for educational and research purposes only." This agreement ensures that the selected WTMJ AV files may be shared with researchers as well as secondary text data sets.

Sensitive content and privacy

The LGBTQ+ materials in our collections include correspondence and other materials that may present privacy issues, either due to the nature of the content or because they include correspondence with third-party individuals. Protected information collected without the knowledge and consent of third parties will have been redacted as part of archival processing

procedures (social security numbers, credit card numbers, private medical information). Because the content of much of the AV materials is not fully known, we will review our text data set with the same criteria applied to our analog collections and redact any third-party or protected information that may appear. Because much of this material was originally broadcast or published in some form, we do not anticipate that this will be an undue burden, but we do want to be sensitive to the inadvertent existence of these materials, particularly on secondhand media that may include non-broadcast materials.

In cases where the content of the materials is sensitive due to disclosure of potential legal or other sensitive issues, or if protected information is discovered during digitization or analysis, those relevant selections from those collections will not be included in the data set or data analysis that is made available to the public. The UWM Archives has secured release forms for all of the oral history interviews included in the project and subjects were given the opportunity to designate portions of the recording for embargo or ongoing restricted access. All embargoes and restrictions will be honored. To that end, we have already identified one collection as out of scope due to privacy issues (We will not include the AV materials from the *Gay Farm Boys Oral History Project* due to the sensitive nature of many of the interviews.)

We will scrupulously review for any previously unidentified sensitive content during the workflow process and will use all pre-existing information at our disposal to ensure that sensitive data is not exposed to the public. Discovery and identification of this manner of content will be one of the first steps in our data analysis process, and we will work closely with UWM Archives to properly identify and redact any data that meets these criteria. We will document our processes and share our procedures as part of our final deliverables.

At the same time, we want to be careful not to err so far on the side of privacy and protection that our work has the effect of keeping this history in the closet. Because the UWM Archives has a continuing close relationship with many of the record creators for the AV materials included in this project, we will invite those record creators to review our work as we make progress with our speech-to-text dataset, and to incorporate their own research questions and input into our discovery and documentation. This includes the Papers of Miriam Ben-Shalom, the Milwaukee Gay/Lesbian Network Records, the AIDS Resource Center of Wisconsin, and Milwaukee Pridefest. We anticipate that doing so will not only ensure that our data set is as comprehensive, accurate, and ethically sound as possible, but will also generate new questions or knowledge about the materials in our LGBTQ+ AV materials in the process.

Draft Use Model

The *LGBTQ+ Audio Archive Mining* project will provide an intensive opportunity for key staff to develop and hone their latent data analysis skills, and to develop expertise in machine learning techniques and applications. Perhaps the most important outcome of this project is the opportunity it provides to dedicate time, identify a demonstration project, and give staff the opportunity to devote themselves to building expertise in a suite of skills that are adjacent to

the work already underway, but in need of further attention to rise to a level where the Libraries can effectively support additional services. Building new support services for data-driven research, especially in support of our digital and archival collections, aligns with our current Strategic Plan – especially "Research Excellence: Support the university's research goals by providing access to resources, stewarding unique digital content and historical collections, participating in activities designed to create change in scholarly communication and track the university's research outputs, and positioning the Libraries as a hub for research support services."[9] But that Strategic Plan "expires" in 2020 and so this project additionally serves as an opportunity to build new strategies and priorities around data-driven research services and collections-as-data initiatives in a new Strategic Plan that places added emphasis on this area of research and collection support.

The concept of collections-as-data aligns with the current goals of the DH Lab, including, "Engage faculty and graduate students in arts, humanities, and performing arts digital research; Encourage experimentation with and application of technology to arts, humanities, and performing arts research and teaching; and pool resources and expertise in order to develop broader expertise among faculty, staff, and graduate students, and develop shared and sustainable scholarship."[10] These goals were established at the outset of the Lab and inform our programming and partnerships. Going forward, we plan to use the time, training, and resources afforded by a *Collections-as-Data: Part to Whole* award to establish stronger and more formalized cross-departmental services for data-driven scholarship, and as a platform to develop models for how our own collections might be mined and queried using computational tools, and demonstrating how new research might emerge, as well as how these methods can benefit our students in the classroom.

The *LGBTQ+ Audio Archive Mining* project will engage existing roles from across the Libraries as well as partners in other divisions on campus. The Libraries intend to adopt this *Use Model* in order to continue to implement collections-as-data access for other formats and collections, and especially, to advance and sustain work initiated by the UWM Libraries Digital Humanities Lab. The DH Lab has been a site for discussion and collaboration since its launch in September 2013. Activities have included works-in-progress talks by faculty; external speakers to discuss high-profile DH projects or work with faculty and staff on emerging areas of interest, such as "sonic pedagogy" and collections-as-data; workshops on obtaining public data, GIS techniques, using the High Performance Computing cluster for humanities research, creating sound narratives, and numerous other topics; and programs such as our DH Teaching Fellows, aimed at developing models and peer support for integrating DH tools and methods in the classroom.

The **Digital Collections and Initiatives (DC&I)** Department in the UWM Libraries manages development and preservation of the Libraries' unique digital collections, and directs the activities of the DH Lab. It includes three full time staff (Department Head, Digital Collections

[9] https://guides.library.uwm.edu/c.php?g=700750&p=4972602
[10] https://uwm.edu/libraries/dhlab/

Librarian, and Application Specialist). Additionally, the Metadata Librarian has a 40% report to DC&I; the LAMP Specialist is also 50% DC&I and 50% L&S Web and Data Services.

Staff in DC&I will all make contributions to the *LGBTQ+ Audio Archive Mining* project. Their work on this project is an extension of work already underway in DC&I. Our Digital Collections Librarian, Ling Meng, will engage and expand his expertise in reformatting for AV materials, file management, and digital preservation; the Application Specialist, Jie Chen, has expertise in application development for digital collections, including working with APIs, PHP, Javascript, and other scripting languages, and adapting and customizing out-of-the-box applications for specific uses; Chen will work closely with the Linux system administrator, Karl Holten, to build public-facing data delivery mechanisms and to explore alternative and extended platforms for delivering data sets as well as dashboards that provide access to data analysis outcomes, such as topic clusters and relationship networks.

DC&I will continue its close working relationship with the **UWM Archives** department throughout this project. An important position in Archives is the Digital Archivist. This is a relatively new position (January 2019) focused on digital archives management, including born-digital and AV reformatting. The incumbent, Shiraz Bhathena, brings with him a deep bench of experience with AV materials in particular. This project will further develop his skill sets regarding digitized and born-digital AV materials, and enhance the Archives' workflows for processing, describing, and making those materials accessible for teaching and research.

Web and Data Services, in the College of Letters and Science (L&S), provides support for research computing, data reporting, and web applications. The Director of Web and Data Services, Dan Siercks, is the co-Project Lead for this proposal, as well as the primary support for research computing in L&S. Karl Holten is a member of the L&S Web Applications Team as well as a part-time report to DC&I in the Libraries. The DH Lab has worked with Web and Data Services since 2013, primarily to provide workshops to promote research computing services to the Humanities. This project provides a path to building a stronger and more formal working relationship that will increase our ability to support digital humanities and data-driven research that engage cultural heritage materials.

New services will include:

- **Classroom support**: LGBT 599 (Capstone Course in LGBT Studies): In this course, offered annually, LGBT Studies juniors and seniors culminate their work in the program by producing their own contribution to that scholarship. The form these final projects take varies by instructor, topic, and student, but often primary-source research is involved. Students often focus on the local Midwestern context for these projects, making the archival material in our collections potentially of great value to the students as data for their projects. However, students have traditionally underutilized the audio and audiovisual materials in the archives because of the difficulty in identifying which of the materials are relevant to their specific research focus. The project deliverables themselves will support deeper research in the Archives by these students. There is also

the potential to support further audio mining and text analysis work with the LGBTQ+ materials by the students themselves, to build their archival research skills with an emphasis on a data-driven approach to primary sources.

- **Workshops**: The DH Lab hosts regular workshops on discrete skill sets and tools. Building on the model developed here, staff will develop a new workshop series focused on the tools and methods used in this project, including (a) use of speech-to-text algorithms for mining archival audio and assessing accuracy, (b) methods for text analysis, including R and NLP approaches, (c) methods for data visualization using R, (d) and ethical approaches to working with archival data, using this project as a model and testbed. Workshops, like all DH Lab workshops, are open to anyone on campus or in the community. We make a special effort to identify classes that might benefit and schedule additional workshops accordingly, in order to accommodate class schedules. Instructors for the workshops will include the two Project Leads, Ann Hanlon and Dan Siercks, and may also include additional staff or faculty as appropriate. This is a model already in place in the Lab as we draw on expertise from across the Libraries and across the campus. Because we will foreground project development and training during the grant period, we anticipate putting the workshop series in place immediately post-award as a series in Fall 2021 and continuing into Spring 2022. See *Section (7) Timeline of completion* as well. For more on the DH Lab's programming and workshop schedules, both past and present, see the DH Lab events page: https://uwm.edu/libraries/dhlab/events/
- **Project support**: We will use the project outcomes to promote our digital collections as a resource for data-driven research projects and solicit project proposals that further explore our existing digital collections and identify additional archival collection candidates for reformatting and analysis. These faculty and student-proposed projects should ideally lead to grant proposals and cross-disciplinary collaboration that furthers the work initiated during this project, especially work that invests additional resources in uncovering and telling the story of the LGBTQ+ community in Milwaukee and the Midwest.

There is a considerably broad swath of communities potentially interested in the *LGBTQ+ Audio Archive Mining* project deliverables. The project will be immediately useful to Sociology, History, and School of Information Studies faculty, graduate students, and undergraduate students interested in the history of the LGBTQ+ community in Milwaukee, as well as data-driven research techniques for working with digital and digitized primary source material. For graduate student and faculty researchers, the most exciting element of this proposed project is that the outcome will not just be a set of tagged materials, but the analytic capabilities of the proposed end product. One of these is the simple but powerful tool of being able to search word frequencies within the data mined from an audio or audiovisual source. For example, a researcher studying the emergence of lesbian social recognition could search throughout a wide range of holdings to compare the relative frequency of the terms "gay community," versus "gay and lesbian" or "lesbian and gay" community over the second half of the 20th century. Similarly, a researcher could track the social recognition of bisexual and transgender people through word frequency counts. A much more robust tool that is planned would be of even

greater use in data analysis. That anticipated tool will enable researchers to examine and visualize relationships between terms within the archival sources. So, a researcher could examine the phrase "coming out," and see if it is surrounded by negative contextual terms (e.g. suicide, rejection, isolation, discrimination) and/or positive ones (liberation, freedom, love, honesty). Further, the tool will permit the researcher to visualize how closely these positive and negative terms appear when “coming out” is mentioned in an audiovisual source.

These capacities to visualize concepts appearing in audiovisual sources make the analytic process into one that generates new directions for research. More than just finding videos that mention a term, or counting how often it is used, visualizing relationships between concepts in the audiovisual data reveals patterns that the researcher may not have anticipated. This makes it of great use to the qualitative researcher, as these are not merely data that can be used to fit or falsify a hypothesis--the mined data are transformed into theory-generating tools.

Additionally, Health Sciences faculty and students are a key constituency for the project, especially as it expands access to AV materials that document broad coverage of the AIDS epidemic in Milwaukee and the upper Midwest. Outside academia, members of the LGBTQ+ community in Milwaukee, especially members of organizations whose collections are housed in the UWM Archives and activists, community historians, and allies who want to explore and learn more about this aspect of Milwaukee’s history, will be afforded greater access to the contents of our LGBTQ+ archival collections, especially to AV materials that had been under-described and underutilized.

Draft Implementation Model

Overview: Considering digital reformatting of analog audio and AV materials is already integrated into the archival processing workflow, this workflow addresses tasks post-digitization. Because the project requires multiple people across Library departments and across campus to collaborate effectively, the Planning phase is key to identifying clear roles and expectations. The Pilot phase will clarify methodology and set the threshold for accuracy for a useable text data set that will be extracted from the audio collections.

Phase 1: *Planning*: The first phase of the project will establish roles for team members, affiliated staff, and administration; identify communication tools for project participants; and describe deliverables in more specific detail for each phase of the project. This will include:

- Project leads will set up an email reflector and shared virtual project resources for team members to facilitate group communication and share documentation.
- Library Director will charge a *Collections as Data Steering Committee* to identify opportunities for integration with other units, areas for additional professional development, and to encourage broader discussion within the Libraries about core skills and competencies, and impact on services and collection development.

- Project leads, Administrative lead, and Disciplinary scholar will identify specific deliverables (beyond the data sets and project outcomes) and outline a plan for sharing those deliverables. Examples may include white papers, conference proposals, organizational charts, and updated position descriptions.

Phase 2: ***Pilot***: The pilot phase includes two testing phases aimed at developing and refining workflows for extracting and analyzing text from digital audio files. The pilot phase is key as we will identify the best approach to extracting textual data, including possibly co-mingling multiple modes of extraction and output, and verify that approach for implementation.

- ***Unsupervised learning pilot***: Our first phase of testing will focus on extracting text from batches of digital audio files using several unsupervised speech-to-text solutions to create a variety of "transcripts" to varying extents of correctness. These tests will be run using a collection that already has accurate human generated text transcripts, *The Milwaukee Transgender Oral History Project:* https://collections.lib.uwm.edu/digital/collection/transhist/search. This will enable the team to assess the accuracy of the speech-to-text solutions and define a workflow to extract the most accurate text data set as well as identify approaches to topic modeling and other pattern identification methods.
- ***Supervised learning pilot***: The second phase of testing will focus on supervised learning, based on topic modeling and other data mining approaches informed by a known corpus. We will use the Queer Zine Archive Project, http://archive.qzap.org/, as our known and highly curated text corpus to help align the analysis of the UWM Archives audio with a more human readable text-mining classification model. Supervised classification will generate models that facilitate discovery and surface topic clusters within and across the collections.

Phase 3*: Data preparation*: During phase 3, we will narrow down the specific collections to draw our audio data sets from, and outline steps to prepare audio files to meet uniform specifications (if necessary); and document methods and specifications for the creation and sharing of text files generated, including the degree to which text files should be "cleaned up," and methods for sharing and dissemination.

- The disciplinary scholar, Cary Costello, will work with Digital Archivist, Shiraz Bhathena, to identify collections with significant or potentially significant LGBT content and audiovisual materials.
- Based on results of Phase 2: Pilot, staff in DC&I will prepare digital audiovisual files to meet specifications necessary (if any) for best results using speech-to-text algorithms.
- Project leads will review text outputs from Phase 2: Pilot and document steps for text clean-up for text outputs.

Phase 4: ***Workflow implementation:***

- *File preparation*: Audiovisual files prepared for analysis will be stored in the campus networked storage system and permissions set to allow access for key team members.
    - Files will be named according to existing criteria managed by DC&I with any additional identifying criteria deemed necessary during the pilot and data preparation phases
    - Files will be organized for optimal analysis while maintaining the integrity of collection origin; readme.txt files will be used to describe any deviation from existing file management protocols and why.
- *Speech-to-text analysis*: DC&I staff will share files with L&S Web and Data Services to begin speech-to-text analysis processes. This will utilize existing pre-trained algorithms such as Mozilla's *DeepSpeech,* an open source speech-to-text engine, and other applications and methods identified during the pilot phase. L&S will utilize the campus high performance computing cluster as appropriate.
- *Text output and analysis*: Text analysis processes will utilize the R programming language and environment, as well as investigate use of tools such as NLTK, MALLET and Gensim. We will utilize best practices for text analysis and incorporate approaches for corpus analysis such as cluster analysis and topic modeling.

Phase 5: Interactive discovery interface:

- *RShiny visualization and asset delivery*
    - Leverage a web accessible RShiny interface to provide interactive visualization of topic models, cluster analysis and other results of our text analysis, with the ability to then link to those digital items in the archive that are associated with a topic.
- *Delivery of data sets code, and workflow via web-accessible solutions*
    - Wordpress, RShiny, Github, and Sharepoint are all possible solutions for delivery of digital audio and text data sets at scale. We will determine the best method for asset delivery based on existing infrastructure, sustainability, and ease of specifying multiple parameters for export (item-level, series-level, collection-level; across collections based on specific criteria, etc).
    - The publication of original data sets and workflow solutions, including R text analysis code, would allow for reuse and reproduction at other institutions or against other archival collections.

Overview of the material to be made available as data

We have identified a preliminary list of twelve collections from the UWM Archives LGBTQ+ collections that include AV materials. These include:

1. Shall Not Be Recognized Exhibition Records (digitized)

2. Gay People's Union Records (digitized)
3. Milwaukee Gay/Lesbian Network Records (digitized)
4. Ray Vahey Papers (digitized)
5. Miriam Ben-Shalom Papers (digitized)
6. ACT UP Milwaukee Records (digitized)
7. James Liddy Papers (digitized)
8. Cream City Foundation Records (partially digitized)
9. Oral History Interviews of the Milwaukee Transgender Oral History Project (digitized)
10. AIDS Resource Center of Wisconsin Records (digitized)
11. Oral History Interviews of the Milwaukee LGBT History Project (digitized)
12. Milwaukee PrideFest Records

Of the twelve collections identified, ten have already been digitized. The undigitized material in the *Cream City Foundation Records* and in the *Milwaukee Pridefest Records* comprise six video cassettes in total. Digitization of these materials will be part of the data preparation phase. Additionally, we will identify small portions of the WUWM and WTMJ collections to include in the corpus for the project, based on existing documentation and description that identifies content likely to have LGBTQ+ references. We anticipate additional digitization of reels and audiotape for these collections, though portions of both have already been digitized. We will not include the AV materials from the *Gay Farm Boys Oral History Project* due to the sensitive nature of many of the interviews.

File preparation will consist largely of converting files to appropriate digital formats if that format is not already available. File naming conventions are already in place that serve to organize materials by collection and in some cases, reel numbers or other identification. All files are stored in the campus storage area network (SAN). Preservation files will continue to be stored here, along with copies of access files for AV materials, as well as text data sets and other deliverables. In cases where we determine that the materials meet our criteria for higher level digital preservation, additional copies will be stored in our instance of Preservica, a digital preservation software and storage application.

Plan for the care and continued use of the collection after the funded portion of the project ends

The UWM Libraries place great value on continued use of digital collections and DH projects, with a strong tradition of continuing to build from funded projects and special collections well after their initial launch. Examples include the continued work on digitization and description of negatives and slides from the AGSL photograph collections, continuing work that began under two NEH funded digitization projects in 2010 and 2013. In this case, over 50,000 additional items (color slides) from the *Harrison Forman Collection* have been digitized - completing the digitization of a collection in which nitrate and safety film formats had been digitized as part of the NEH grant. The *March on Milwaukee* digital project is another important example – launched in 2007 to commemorate the 40th anniversary of civil rights marches in Milwaukee, the collection was refreshed with additional materials, a new interface and improved

functionality in 2017 for the 50th anniversary, and continues to find additional applications in the classroom and for research. Likewise, the Libraries will continue to build on work and maintain collections established as part of this initial Collections-as-Data grant.

Importantly, the project itself will develop from existing workflows in Archives focused on digitization of AV materials as a product of patron request and a byproduct of processing workflows. All infrastructure developed during this project will be developed using existing tools and open source applications to ensure continued support and safeguard opportunities for continued improvement. The published data sets, code, and workflows will all be made available via open source and existing infrastructure that also supports other campus activities. The model developed for the *LGBTQ+ Audio Archive Mining* project will inform development of additional speech-to-text workflows for creating useful text data sets based on spoken-word AV materials in the Archives to provide greater discoverability, but also to expose those datasets to researchers.

Finally, the award affords our staff training opportunities that will position the UWM Libraries to continue to develop and use (and promote use of) data sets from our LGBTQ+ collections, and future LGBTQ+ collection acquisitions, as well other digitized and born-digital collections. An important component of this training will include methods not only for mining, analyzing and releasing but also protecting that data. As we move forward with programs that promote data mining, analysis, visualization, and machine learning techniques, we also plan to become better informed and more critical users of those tools as well. Focusing our efforts on the LGBTQ+ collections means we will be exposing hidden histories that deserve to be heard, but also requires us to be sensitive to the voices hidden within that never intended to be broadcast. As such, the *LGBTQ+ Audio Archive Mining Project* is an exemplary testbed for promoting the advantages of a collections-as-data approach and for building in a critical and ethical component.