Ensuring Scholarly Access to Government Archives and Records

Report of project activities from February 2020 through October 2021

Virginia Tech University Libraries in partnership with Virginia Tech Center for Humanities and the U.S. National Archives and Records Administration

Contributors: William A. Ingram and Sylvester A. Johnson

Contents

Preface

Background

The Andrew W. Mellon Foundation provided grant funding to Virginia Tech in support of a collaborative planning project, with the goal of ensuring public access to the massive and ever-growing collection of government records in the digital catalog of the National Archives and Records Administration (NARA). Virginia Tech and NARA planned to convene a diverse group of archivists, librarians, humanists, technologists, information scientists, and computer scientists for a two-day planning workshop at the Virginia Tech campus in Arlington, Virginia, in April of 2020. Because of COVID-19, our plans for the workshop were put on hold. Since we were unable to hold an in-person workshop, we submitted a modification request to the Mellon Foundation and were approved to hold the workshop in the following spring of 2021 as a series of five, two-hour, online meetings.

During the workshop, participants identified requirements, developed conceptual models, and discussed a work plan for a subsequent pilot project that would apply state-of-the-art tools and technologies to increase the effectiveness of archival programs and broaden public access to the important content in the NARA catalog. The workshop focused on humanistic and equitability issues of artificial intelligence and developing ethical, human-centered technology that promotes the public good. As such, the topic of intentional mitigation of AI bias was a thread that ran through the entirety of the workshop.

Review of Workshop Objectives and Deliverables

The ultimate goal of the workshop was to devise and articulate concrete recommendations for developing a pilot project to test and demonstrate the effectiveness of the state-of-the-art in machine learning and other forms of artificial intelligence for automatic metadata creation from NARA’s digital collections.

Workshop Participants

Moving online allowed us to increase our original number of invited experts to participate in the workshop. We deliberately limited participation in the workshop by invitation-only. That way, we committed to ensuring that at least 50% of the invited participants would be from under-represented racial and ethnic populations, and we sought parity in the number of men and women.

We selected participants with expertise in a wide range of relevant subject areas, including cutting-edge technologies such as computer vision, machine learning, text mining, information retrieval, information extraction, natural language processing, deep reinforcement learning, information visualization, and human-computer interaction. Participants also included digital humanists, academic scholars employing archival research, librarians and archivists, data scientists, and scholars of algorithmic decision making, computational social science, and algorithmic bias.

Given the tendency for AI-related projects to reflect an overwhelmingly white and male constituency, which is not representative of society and has contributed to biased and inequitable societal outcomes, we were deliberate and intentional in funding a diverse and inclusive set of attendees to participate. We also included participation from historically Black universities and other predominantly minority-serving institutions. Half of the funded attendees were women. And more than half of funded attendees were from underrepresented groups (BIPOC). Mellon-funded attendees were as follows:

1. Jason Baron, University of Maryland
2. Marina Del Sol, Howard University
3. Jessica Farrell, Educopia.org
4. John Foley, Middlebury College
5. Batya Friedman, University of Washington
6. Kimberly Gay, Prairie View A&M University
7. Michael Hemenway, University of Denver
8. Josh Honn, Northwestern University Libraries
9. Atiya Husain, University of Richmond
10. Gabbrielle Johnson, Claremont McKenna College
11. Ida Jones, Morgan State University
12. Kymberly Keeton, Austin History Center
13. Adriana Kovashka, University of Pittsburgh
14. Alvin Lee, Florida A&M University
15. Liz Lorang, University of Nebraska-Lincoln
16. Richard Marciano, University of Maryland
17. Mark Matienzo, Stanford University
18. Raeshawn McGuffie, Hampton University
19. Tanushree Mitra, University of Washington
20. Sherod Moses, Virginia State University
21. Darby Orcutt, North Carolina State University
22. Chris Prom, University of Illinois at Urbana-Champaign
23. Howard Rambsy, Southern Illinois University
24. Kenton Rambsy, University of Texas at Arlington
25. Roger Schonfeld, Ithaka S+R
26. Kayla Siddell, Xavier University
27. Joshua Sternfield, Independent AI expert
28. Kalinda Ukanwa, University of Southern California
29. Ximena Valdivia, Florida International University
30. Jian Wu, Old Dominion University
31. Seungwon Yang, Louisiana State University
32. Andromeda Yelton, Harvard University

In addition to these funded participants, more than 50 additional attendees participated on an unfunded basis. Most of these attendees were from Virginia Tech and the National Archives and Records Administration. Because of our focus on the broad relevance of social justice and equity, issues that have commanded urgent public and expert attention, we invited a keynote talk by a female scholar of color, Tanushree Mitra[1] from the University of Washington, specializing in AI fairness and the mitigation of algorithmic bias.

[1] https://ischool.uw.edu/people/faculty/profile/tmitra

Schedule and Objectives

The workshop consisted of five online workshop sessions, each of which lasted two hours and was devoted to a specific theme. Most sessions began with a presentation about the day's theme, followed by a plenary discussion. We then divided participants into smaller interdisciplinary breakout groups, using Zoom's breakout room features. The smaller breakout groups allowed participants to discuss the topic in greater depth, brainstorm, and generate ideas and solutions through shared ideation. To leverage the use of the digital meeting platform, we created shared Google documents for all attendees to share ideas and insights in each breakout room across the five meeting days. As a result, we collected a high volume of rich inputs, resources, and strategies related to each day's theme and to the larger subject of leveraging ethical AI for digital records search. Each of the five sessions was aimed at producing a deliverable related to the topic. We briefly describe each session below and include the entire workshop agenda as Appendix A. to this report.

Session One: Welcome and Orientation. We spent the beginning of this first session explaining the goals and objectives of the workshop series, as well as the format for shared ideation and breakout groups. Bill Ingram presented a procedural overview of the workshop followed by a welcome address from David Ferriero, Archivist of the United States.[2] Sylvester Johnson led a plenary discussion of the workshop topics and goals, a review of the document packet, and sample datasets we prepared beforehand. We then went into our breakout rooms and asked the participants to introduce themselves and discuss each other's areas of expertise. We concluded the first session with a presentation from the National Archives' Chief Innovation Officer Pamela Wright about NARA, its mission, and goals. Our desired outcome and deliverables for this session were to introduce our agenda and goals, and to ensure that all participants had the necessary knowledge and information for a successful workshop.

Session Two: Avoiding AI Bias. We began session two with a presentation from Tanushree Mitra, assistant professor in the Information School at University of Washington. Professor Mitra is an expert in social computing and computational social science. She presented her work on measuring online misinformation through an audit study of YouTube. We followed with plenary and breakout group discussions. We charged participants with the task of generating concrete recommendations for avoiding algorithmic bias.

Session Three: User Stories. We devoted session three to coming up with user stories in order to identify high-priority needs and to foreground equitable design processes and mitigation of bias in AI-driven archival processing. We began with a presentation from NARA staff on the content of their online catalog and their users. They presented a set of user personas

[2] A recording of the Archivist's welcome address can be accessed at https://youtu.be/uSPsgGadkOI.

they generated beforehand. We devoted the majority of this session to working in breakout groups, focusing on generating user stories.

Session Four: Exploring the Solution Space. Session four consisted of a plenary discussion of the state of the art in AI and machine learning tools and techniques and their applicability to the problem of describing digital records. We then broke up into small groups, charged with creating a plausible work plan. This was the only session without a plenary speaker.

Session Five: Implementation—Feasibility and Sustainability. Patricia Hswe, program officer for Public Knowledge at the Andrew W. Mellon Foundation, gave the plenary presentation for session five. She spoke about the Foundation's goals and objectives and the importance of sustainability planning. We then went into breakout rooms to discuss the feasibility of the possible solutions that resulted from earlier brainstorming sessions and how such solutions could be maintained and sustained in the longer term. Participants generated a focused set of concrete recommendations for next steps. We concluded the workshop with a survey, intended to assess the workshop activities, solicit ideas for disseminating workshop findings, and gather any post-workshop thoughts and ideas from participants.

Acknowledgments

This project is supported by a generous grant from the Andrew W. Mellon Foundation. We thank the Foundation for their support and their patience and flexibility in allowing us to move the workshop online, and we especially thank program officer Particia Hswe for her excellent plenary presentation on sustainability.

Our utmost thanks to NARA's Chief Innovation officer Pamela Wright for all the time and energy she spent with us planning and running the workshop. We also thank the staff from numerous areas of the National Archives for their time and active participation. We thank all the invited expert participants for taking time out of their busy schedules to spend five Fridays with us, discussing AI in public archives. We give special thanks to Tanushree Mitra for presenting her work on algorithmic bias. The workshop would not have been possible without the Virginia Tech faculty members, staff, and students, who managed logistics, took notes, and led breakout sessions: Bipasha Banerjee, Kira Dietz, Cathy Duerbeck, Joe Forte, Dominique Francesca, Alex Kinnaman, Kara Long, Alan Munshower, Shu Pan, Stephanie Ramirez, Jenna Slezak, and James Tuttle. Dean Laura Belmonte (College of Liberal Arts and Human Sciences) and Dean Tyler Walters (University Libraries) were especially critical in supporting this effort. Angelica Rivera's precision, efficiency and generous talent greatly facilitated this project's completion.

Finally, we reserve special gratitude for Archivist of the United States David S. Ferriero for his support and participation in the workshop, and for his years of service and commitment to expanding access, openness, and inclusion at the National Archives.

Key Findings

In sections below, we summarize key findings from the collaborative workshop based on input from the 85 participants. We address each of the four workshop themes: Avoiding AI Bias, User Stories, Exploring the Solution Space, and Implementation—Feasibility and Sustainability.

Avoiding AI Bias

Bias may imply intent; it can also be unconscious. Therefore, a multi-faceted approach is needed to handle bias. Attempting to address AI bias requires further insight into how algorithmic search and recommendations drive problematic content on online platforms such as misinformation, online extremism, and conspiracy theories. Adopting a descriptive approach to data is best (as opposed to a normative approach) as it focuses on what is in the data set rather than allowing the data set/algorithm to claim how things should be. Search algorithms are a type of decision-making, thus requiring caution. Audit studies of system behaviors may be helpful. However, audits can often fail to detect algorithmic bias that everyday users of the systems are easily able to see once the system has been deployed to the public.[3]

Algorithmic bias reflects the latent biases in the data used to train AI systems. Curators tasked with assembling collections need to understand how algorithmic bias may impact the search results used to produce those collections. Unfortunately, the subjects of such bias are often not involved in conversations surrounding bias and its effects. The labels used to train AI systems reflect the biases of people who created them and the societal norms present when the work was done. The demographics of the library and archives profession historically have not been diverse. Discriminatory cataloging reflects the need for more diverse staff as it relates to the demographics in the profession.

Big conversations are taking place in archival literature, conference panels, and Slack channels about harmful description and reparative language work. Institutions are issuing statements about their collections as they are working to improve archival descriptions or change them. But aren’t new biases added by reparative description? The records of government archives serve as documentary evidence of past transactions of government policy and process, as well as their impact on citizens and society. The role of the archives is not to shape or impact policy beyond the levels of record keeping. Moreover, the most harmful archival biases are reflected in who and what are excluded from archival memory, the marginalized and silenced voices of history. But with mass digitization, optical character recognition, algorithms and computational power, new information may be extracted directly from the full text of archival records, potentially minimizing the effects of biased description, reparative or otherwise.

[3] Hong Shen et al, “Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors," arXiv preprint arXiv:2105.02980 (2021). http://doi.org/10.1145/3479577

There are potential dangers of eliminating bias to rewrite or reframe history. Archives exist to provide access to evidence of past action. They reveal a complex story of government action, its policies, and its impact on the citizenry. The application of AI on this huge body of historical material will potentially uncover connections that could expose evidence of racism embedded in the laws and actions of our government. This might be considered distasteful to some Americans. But to others, it may provide answers to deep underlying questions about our country and its history. The challenge is to ensure that the AI agents are not trained to perpetuate historical bias. Algorithmic biases arise when the data used to train the algorithm are more representative of some groups than others. Careful attention must be paid to ensure fair representation of the marginalized and underprivileged. Therefore, it is crucial to carefully analyze the data beforehand to understand the distribution of bias and identify the gaps that exist in the data—the ability to record and hand over to a researcher or user a clear outline of what information is incomplete; what could have contributed to the introduction of bias; the efforts to correct bias; and what actions have not yet been tried. Transparency is paramount.

User Stories

Viewing user stories from three categories: persona, want/need, and purpose helps identify high priority areas, ensure a user-friendly interface design, and mitigate bias in AI-driven archival processing. For example, the persona assessment aims to understand the person utilizing the system—beyond their job function or title. The want/need describes the person's intent—not the features they use. This category seeks to answer what the user is trying to achieve and should be implementation-free. Finally, the purpose looks at how a person's immediate desire to do something fits into their bigger picture. What overall benefit are they trying to achieve? What is the big problem they need to solve?

Creating user stories starts with persona development. Starting with user personas keeps the focus centered on people and communities, while brainstorming about the possibilities of technology. This human-centered approach to design thinking focuses on the question, What are the community needs and how can we use technology to help meet them? A few novel user personas were introduced by this exercise (e.g., National History Day student, rap musician, mortician), and many of the participants wrote user stories from their own points of view (e.g., archivist, researcher, student, teacher). We include the full list of the user stories generated at the workshop as Appendix B. We highlight a few common themes below.

A few stories arose from archivists, who wanted to be able to use an AI agent to assist them with analyzing, organizing, and locating archival records. Their stories included the following:

- As a processing archivist, I would like to apply an ML/AI toolset to records without embedded metadata so that I can create metadata for review.
- As a records manager, I would like to user Ml/AI tools to assist in decision making so that I can speed up the classification/declassification process.

- As a reference archivist, I would like to apply topic analysis, entity extraction or ML/AI tools to records having only series level descriptions, so that I can respond to reference requests for records lacking object level descriptions.

Other stories involved using facial/image recognition for searching.

- As an archivist I want to be able to take advantage of integrated facial recognition software to perform searches in photographs for people not mentioned in captions or other descriptive metadata.
- As a highschool student/history enthusiast, I would like to find photos matching a face I submit from another historical photo so that I can illustrate my history month presentation with additional content.

Several stories involved using AI to link from an item of interest to similar or related materials (i.e., a recommender system). For example:

- As a curious explorer I want to see related topics and other suggestions that may be related to my records search, because I am not a trained archivist and am not aware of the full universe of related documents (guide me down the "rabbit hole" of information).
- As a researcher in computer science, I want to find semantically similar documents in multiple modalities, such as text, figures, and tables. It would be better if the user portal provides links to external resources so I could explore other resources containing data that are not available from NARA.
- As a researcher of Black history, I want to follow up on a lead from one document and find other documents that might have something in common with it so that I can expand my understanding of the Government's interaction with Black people - even if the race of the people in the catalog is not identified. (I want "more like this.")

Another common theme among the user stories was related to explainability, transparency, and reproducibility of the AI methods used for ranking search results or making recommendations of related content. For example:

- As a first-time NARA user, I need to understand the algorithms and why I'm searching.
- As a history professor teaching an online graduate seminar on historical methods, I want all of my students to be able to conduct a series of search queries using similar terms, so that we can have a discussion about how to evaluate the search results. It would be useful if search results based on similar terms are identical for all of the students
- As an auditor, I would like to see some of the training data that was used to generate the AI/ML model so that I can endorse or provide recommendations to the training data.
- As an instructor interested in my students finding relevant information, I want to be able to establish and share search strategies that will work identically and reliably for my students.

Not surprisingly, many user stories were written from the point of view of AI/CS researchers. A common desire among this group was to download large amounts of NARA data in bulk. For example:

- As an AI researcher, I want to be able to construct a bulk-downloadable data set from a search, so that I can obtain a data set suitable for a specific ML training goal.
- As a computational researcher, I want to discover and download large datasets in order to analyze on my own computer/server.
- As a data researcher, I would bulk download textual records to determine if the records contain terms pertaining to a particular subject.

One participant proposes referring to this last group in aggregate as the “Give me all the data” user—someone (e.g., student, professor, journalist, rights advocate, political analyst) who knows what to do with the data and has local analytical tools much more powerful and specialized than NARA can provide. In fact, NARA does allow its data to be downloaded in bulk. The NARA dataset can be found in the AWS Registry of Open Data.[4]

Exploring the Solution Space

The goal of this portion of the workshop was to generate ideas for how AI technologies could be applied to the problems and user stories described in the previous section. The solution space can be broadly divided into two major categories: improving metadata and improving search and browse. These are reciprocal needs—each informs the other. Search functionality is determined by the feasibility of automatically extracting metadata at a sufficient level of granularity. Likewise, as more is learned about how researchers search for and discover information, that will inform what metadata is needed to support that functionality. We explore both areas in the sections that follow.

Improving Metadata

Workshop participants identified four major difficulties NARA faces with its catalog: 1.) Lack of granular metadata—challenge in finding a dataset with both scope and size; 2.) Absence of customizable search options—many results offered to users do not fulfill their research needs; 3.) Exploratory experiences for users are nonexistent—need for a better user experience, one that allows members of the general public to explore information NARA has from a broad perspective instead of having a specific research project or research topic; and 4.) Diversity of the dataset on top of the scale—NARA could benefit from providing a more diverse perspective on presenting data and working with records in general.

Putting aside the diversity problem temporarily, the remaining three difficulties revolve around a central issue: the lack of sufficient description at the item level. To understand why, we must

[4] NARA publishes its National Archives Catalog dataset biannually to the AWS Registry of Open Data. https://www.archives.gov/developer/national-archives-catalog-dataset

understand the long-held archival focus on contextuality. The uttermost goal of the archives is to get intellectual control over its records, which means understanding where a record sits in the context of its place within a file unit. The latter has a place within a series, which is from a particular organization. In the archives, a record has a contextual importance—the archivist's concern is mapping the provenancial interrelationships between the records and their creators.[5]

For instance, suppose a researcher is interested in a letter from a U.S. president. The context could be that the letter was sent by the president or it could be that the person who received the letter was a soldier in the army—that difference in context is the archival understanding of the record. The "aboutness" of the record may be described in a scope and content note, but mostly, archival work is concerned with its context, with the record's relationship with its creator's functions or activities and the act or process that led to its creation. Records are not arranged by subject, theme, time, or place. They are arranged in series, based on their shared function, activity, or use. As a consequence, there is a mismatch between expectations NARA's user community has for search and browse functionality (as exhibited by our user stories) and the level of description or arrangement necessary to facilitate this functionality.

People want to search and browse records by name or by geographic location, but that metadata has not been brought out for most records. Important data is buried inside the records—entity information such as names of people and geographic locations. This data could be used to connect entities across various series. The information is there, but it is buried in series that are only described at a very high level. Of course, manually identifying, disambiguating, and labeling entities at the item level for millions and millions of records would be impossibly tedious. An archivist would have to read through each record, page by page, to discover these hidden entities. This is where computation and AI could make a huge impact.

Applying optical character recognition (OCR) to records will enable more entities and relationships to be discovered. But it is not a panacea. What can be extracted from the OCR of a textual document is very different from what can be produced from an image or a map or from audio or video. There are also many hand-written or hand-annotated textual documents that are not receptive to OCR. On the bright side, there are billions of born-digital and OCR-amenable documents from which accurate text extraction is possible using off-the-shelf technology. The application of AI for entity extraction on these records alone could be hugely impactful, and it would allow the archivists to focus their manual efforts on the records from which automated text extraction is not yet realizable.

Neural network architectures have been shown to achieve state-of-the-art results for named entity recognition. Given the diversity of content in the NARA catalog, a divide and conquer approach is recommended, wherein a classifier separates records into several subsets of

[5] Terry Cook, "The concept of the archival fonds in the post-custodial era: theory, problems and solutions," Archivaria (1993), https://archivaria.ca/index.php/archivaria/article/view/11882.

documents and routes each to a domain-specific entity extraction model. Current classifiers are not yet sufficiently reliable to be run without a human in the loop. Workflows will need to be carefully designed to efficiently use human time since there is such an abundance of unlabeled data and manual labeling is expensive. One promising solution is active learning, an iterative supervised learning technique in which an algorithm incrementally receives labels from a human oracle. Another challenge will be how to aggregate labels provided by multiple people. This may be best approached by only choosing labels from a controlled vocabulary.

It takes a vast amount of labeled data to train a classifier. It may be prudent for NARA to utilize existing controlled vocabularies to bootstrap labeling efforts. But this brings us back to the diversity problem. Algorithmic biases can almost always be traced back to data used to build the algorithm. As discussed at length in the sections above, when deciding how to label records for use in training AI models, careful attention must be paid to ensure fair representation of the marginalized and underprivileged. Examining researcher behavior (e.g., by mining search logs) can surface ideas for new labels. NARA might explore how advancements in social media data analytics have led to diverse, validated techniques for predicting labels in scenarios where ground truth data are scarce.

Several workshop participants suggested that the best way for NARA to learn about the capabilities of AI would be to begin by picking and defining a specific problem to solve and start experimenting. A development team might start with a classification problem that does not need a highly accurate solution (i.e., where low accuracy could still be useful or informative). Or they might use currently available information for coarse-grained classification and then incrementally refine the models. Once the records have been coarsely divided, they might apply entity extraction to documents belonging to a single category. It is important to narrow down the problems to a specific level, in order to fully leverage the power of AI. Later, specific solutions can be assembled into a broader system. Several workshop participants suggested using NARA data to host a competition at relevant computer science venues that focus on document analysis and evaluation, such as the IEEE International Conference on Big Data, the International Conference on Document Analysis and Recognition (ICDAR), or CLEF: Cross-Language Evaluation Forum.

The potential for automated metadata extraction is not limited to bibliographic description and named entity recognition. Metadata related to content, level of aggregation (record group, series, record), documentary form, provenance, and original order may also be extracted. For records that are digital objects, there is technical metadata that can be extracted such as fixity and checksum. There also may be valuable provenance metadata embedded in digital objects such as Exif metadata encoded in image and sound files created by digital cameras and scanners.

Improving Search and Browse

Various implementations of “Serendipitous Information Retrieval”[6] may be particularly useful for improving the catalog’s search and browse capabilities for novice researchers, students, or any number of curious users wanting to discover new information but unsure what to search for. A pedagogical approach to search could be embedded in the technology. Search query auto-completion techniques, which are used to predict likely completions for user queries, could be used to guide researchers toward more successful searches. These models generally base their predictions on popularity of past searches. But query auto-completion could also be personalized to recommend search queries for related content based on the current browsing session.

Great progress has been made in recent years in image retrieval, especially those based on deep convolutional neural networks. Image retrieval technology can be used in archives to cluster together groups of similar images or assign category labels, and can also be trained to identify objects and faces in images. A popular application of image retrieval is visual search (a.k.a. reverse image search), which retrieves images that are relevant to a query image. There will, of course, be cases in which the AI mislabels or miscategorizes an image. In such cases, corrections could be gathered from users in real time, which could be fed back into model training, thus crowdsourcing training data. One example of image retrieval used in archival work is Carnegie Mellon University Archives’ CAMPI[7] (the Computer-Aided Metadata Generation for Photoarchives Initiative), used internally to aid in the processing and management of the university archives’ vast digital image collection.

Visualization techniques for information retrieval could be developed to present researchers with a visual representation of search results. The goal would be to allow researchers to visualize relevance relationships between search results and the search query, as well as similarities among the retrieved results. The visualization could help researchers to understand which properties of records are used to determine similarity and perhaps allow them to assign higher weight to some properties over others, or to refine or alter their search query graphically.

Several workshop participants brought up the idea of personalized search. One suggested devising a two-layered approach to personalization. The first layer would determine the user’s broad stereotypical persona (e.g., academic researcher, citizen scientist, student, etc). The second layer would learn the users’ particular needs and interests. The system would behave similarly for anyone who fits with the first-level persona, so recommendations can benefit from

[6] Elaine G.Toms, "Serendipitous information retrieval." DELOS. 2000, https://www.ercim.eu/publication/ws-proceedings/DelNoe01/3_Toms.pdf.

[7] Matthew Lincoln et al, “CAMPI: Computer-Aided Metadata Generation for Photo archives Initiative,” (2020), https://doi.org/10.1184/R1/12791807.

what works for others with that persona, but the recommendations are further guided to be individually personalized.

Another participant suggested a “wizard” or chatbot-like search interface in which an AI agent asks the researcher a series of questions, such as: Are you looking for a known item? Do you want information about a particular person, place, or thing? Are you looking for the answer to a discrete question (e.g., What year did *x* happen)? Do you want all information available on a broad issue (e.g., the development of healthcare policy in the U.S. Government in the 1990s)? The answers would lead the researcher down a personalized search path and set of results suited for their particular information need.

Implementation: Feasibility and Sustainability

The workshop culminated with a full session devoted to identifying key strategies and challenges for feasibility and sustainability of leveraging novel AI tools for ensuring digital records search while designing human-in-the-loop protocols to achieve optimal ethical design and public interest.

Having developed an understanding of NARA’s user communities and their needs, we explored the solution space and proposed many AI-based solutions. The next step will be to formulate a plan or blueprint for future work. In doing so, we must consider which solutions are feasible and how they can be adopted, maintained, and sustained. As Patricia Hswe mentioned in her keynote, what often happens in conversations about feasibility and sustainability, is that the hard questions tend to be avoided. Questions like, “What would happen if the project fails?” or “What do we do when the funding runs out?” or even “What happens if the project becomes overwhelmingly successful?” Development of something new is very exciting. But once it is built, who is going to take care of it and who will sustain it? We must also consider what long-term institutional changes will need to come about in order to implement our design. How will these changes be sustained? The workshop aimed to address those hard questions head on.

Several aspects of maintenance and sustainability need to be addressed. Staff will need to be trained to use the new technology. Technical documentation will need to be written and disseminated. Technical staff will need to keep up to date with relevant literature, attend conferences, and perhaps also seek support to pursue their own research agendas. Moreover, the organization will need to develop a program of ethical training and conscious understanding of risks, expected outcomes, and unintended consequences of AI. Large amounts of labeled data will need to be created or adapted from other sources for training algorithmic models. Models will need to be evaluated, augmented, updated, and retrained continually, as new data, improved approaches, and streamlined techniques become available. The data and methods used to train AI models will need to be well-documented and preserved.

AI and machine learning are often seen as labor saving devices. This is not our goal. Our goal is to take advantage of AI's efficiencies for the purpose of improving public access to government records. Rather than cutting costs, the integration of AI into the archives will require a different investment of labor and resources. There is clear public interest and need for algorithmic bias detection and benchmarking of algorithms. One implication is that institutions like NARA will need to develop teams of experts to administer auditing services. Developing models and processes for systematically auditing AI pipelines will require specialized skill as well as active and ongoing collaboration. New organizational structures will need to be put in place to provide the ongoing audit and human-in-the-loop control systems. NARA and other public archives are encouraged to consider developing an internal or external review board to oversee AI issues, including but not limited to bias. The costs of putting these structures into place will be significant and will demand careful consideration for budget planning. Perhaps a not-for-profit company might take up this role. Steep resource requirements may be mitigated by public-private partnerships.

The focus of sustainability should not be on specific tools, but should involve broad planning in terms of workflows and institutional capacity. This is especially important with regard to personnel, specifically the capacity to train and retrain staff as the technological landscape changes. NARA would benefit from fostering a collaborative culture of ongoing incubation of tools and infrastructure and its integration into daily work life. Creating an in-house innovation lab could help staff stay abreast with the current state of the art, and ease friction when adapting workflows to include new technologies. When evaluating new technologies, try to choose those that will improve current practices rather than disrupt them.

Managing an AI infrastructure in production will demand significant IT systems resources and infrastructure costs. Regular CPU-based systems will be sufficient for many high-computation tasks, including traditional machine learning algorithms, but GPUs will be needed for accelerating deep learning workloads. Digital storage capacity will need to grow. There may also be increased demand for network capacity. Cloud services are often recommended for incrementally scaling up or scaling out infrastructure to support AI workflows. Careful attention needs to be paid to cybersecurity. AI systems infrastructure should be safe and tamperproof to ensure the data and results have not been manipulated. Traditional system-monitoring tools are not sufficient for AI models. Software development, operations, legal, and compliance personnel will require additional training on the guidelines, regulations, and best practices for managing risk in an AI infrastructure.

A lesson learned from the agile development methodology is to start with small deliverables. Starting with small, high-impact deliverables allows for more flexibility to respond to unexpected changes or budget fluctuations. Instead of proposing large-scale infrastructure projects, funding should be provisioned for smaller projects or pilot implementations. Institutional capacity for maintenance and sustainability should be built along the way, with governance structures implemented at the start. User testing should be conducted to seek feedback from the community. Progress should be assessed frequently and should not be limited to technical measures. Assessment, moreover, should include the institutional capacity to build a robust program to govern and support the technology. There must also be institutional buy-in and financial support for current endeavors before launching into the next iteration of development. A successful series of small projects can build momentum and attract further funding.

Post-workshop Survey

At the end of the workshop, we asked participants to complete a survey, so that we could assess the workshop activities, solicit ideas for disseminating workshop findings, and gather any post-workshop thoughts and ideas from participants. We began with two quantitative questions, but most of the survey asked for short answers.

Question 1 asked participants to rate various aspects of the workshop. Their responses are summarized in the table below.

	Unsatisfactory	Poor	Average	Good	Excellent
Content/Agenda			5%	35%	60%
Presentations			2%	26%	72%
Breakout groups			14%	49%	37%
Workshop length		6%	7%	44%	44%
Logistics			2%	35%	63%
Venue (Zoom)			12%	47%	42%

Question 2 asked the participants how successful we were in accomplishing our stated goals on a scale of 1 (unsuccessful) to 5 (very successful). Overall, the feedback was enthusiastic and positive.

1	2	3	4	5
2%	5%	35%	40%	19%

The remainder of the survey questions asked for a qualitative analysis of the workshop. To begin with, we asked participants to explain their rating on Question 2. In general, a participants' assessment of the workshop seemed to depend a lot on the experience they had in their breakout group. Several participants commented that their breakout group discussions were engaging and fruitful. However, the comments also suggest that one or two of the breakout groups didn't click. A common notion, both in a positive and negative sense, was the enormity of the problem. Some indicated they thought the topic was too broad or too difficult to sufficiently address in the limited time we had, while others wrote that they were inspired by the great challenges and exciting opportunities to explore and work together on.

A misgiving reported by at least one participant was that, while the workshop's agenda succeeded in working its way up to articulating a plan, they did not think we were ultimately successful in our goal of delivering a concrete set of recommendations for next steps. Another participant wrote that the key to this workshop's success lies in the hands of whoever can pull together the information from all the notes, breakout group discussions, presentations, plenary sessions, and survey responses. Indeed, the workshop generated over 500 pages of documentation, and it has taken us hundreds of hours to organize, transcribe, read, analyze, and distill all of that information into this report. Our set of recommendations outlined in the sections below draw on the rich trove of insights captured during the five meeting days. It is also important to note that this data will continue to pay dividends as a valuable source of strategies and solutions for the full range of challenges that must be addressed to ensure that public knowledge continues to benefit from successful search of digital records.

From the survey, the biggest takeaway of the workshop was that, as is usually the case with technological revolutions, the challenges we face are human problems. The use of AI is quickly expanding into virtually every industry and institution, including government, academia, libraries, and archives. It is not a question of *whether* AI will be used for public archives, but *how* it will be used—the question becomes one of ethics and sustainability. One survey respondent described the workshop as an eye-opening revelation of how algorithmic bias is such a big challenge for public archives. Another respondent wrote that they had no idea of the effects that algorithmic bias could have on search results, and another wrote that their major takeaway was learning that the challenges NARA faces with AI have such powerful implications for our democracy.

We asked participants to leave us some practical suggestions for a subsequent pilot project. Many of the respondents suggested starting with a small proof-of-concept project focused on one type of content. Some suggested building partnerships, criteria for governance, user-feedback mechanisms, and institutional buy-in before building technological solutions. Other suggestions stressed the need to focus intentionally on improving inclusivity, diversity, and interdisciplinarity. These suggestions are reflected in the set of recommendations we propose in the sections that follow.

Finally, we asked how and where we should disseminate our workshop findings. We received many good suggestions for applicable conferences and publications of professional societies, and some suggested we use what we learned to create educational webinars and learning modules. Overall, we received strong encouragement to disseminate our findings widely.

Recommendations

In June of 2020, the Mellon Foundation announced a new strategic evolution toward prioritizing social justice in all of its grantmaking. Their focus is centered on the people and the communities that will benefit from the Foundation's philanthropic support, whose knowledge will benefit, or whose knowledge will be shared as a result. This strategic shift also means a prioritization of content over technology.

At Virginia Tech, we share these values and translate this commitment by working to ensure that technology serves community interests and supports equitable outcomes. In holding these workshops, we wanted to better understand NARA's community needs. We wanted to get a realistic sense of what could be accomplished with AI and estimate its potential benefit to the communities served by the National Archives—specifically, the public interest in the accountability and transparency of public records that preserve the actions and operations of our federal government.

We began the workshop with a deep conversation about the negative societal effects of algorithmic bias. This led to the articulation of a human-centered approach to technology development, which set the tone for the rest of the workshop. By developing user personas and user stories, we developed a shared understanding of the needs of NARA's broad community to discover and access public records. Through shared ideadation, we proposed specific ways in which AI could not only automate processing so that more records could be made available to the public, but also ideas for how AI could link records together in unexpected ways, leading to new knowledge and understanding. Finally, we discussed the feasibility of implementing an AI solution at NARA and suggested some strategies for how an AI program could be sustained.

In the brief sections that follow, we put forward a plan for future work.

Pilot Study

We propose starting with a proof of concept implementation of algorithmic metadata extraction. We recommend limiting the scope of the pilot to named entity extraction, with a primary focus of designing and implementing an ethically-aligned AI system. A successful implementation would achieve high impact and the results would be immediately practicable. It could be put to use and achieve concrete gains for NARA that would not only add value but could potentially cause a paradigm shift in the way records are processed.

A successful pilot project would provide the foundation for a more comprehensive approach to ensuring equitable outcomes with AI. We strongly recommend that the piloting effort be tightly coupled with the development of a governance approach to technology ethics. There should be standards for establishing ethical practices along with new accountability structures. Moreover, we recommend that these accountability structures be set up externally to NARA. We have shown by the constituency of our workshop that convening diverse groups leads to important insights. The most articulated concerns among workshop participants with regard to producing ethical accountability were about racial exclusivity and its implications on minoritized communities. The biases that cause the marginalized and silenced voices of history to be excluded from archival memory cannot be adequately understood and addressed without diverse voices represented in the governance structure.

Therefore, our recommendation is to build into our technology piloting efforts an external governance body made up of a cross section of representatives from NARA and other public memory institutions; academics, especially from minority-serving institutions; members of civic organizations; and practitioners in private industry. We have shown by organizing this workshop that we can bring together a diverse team of people so that we could harness their inputs, their insights, and their experiences. Building on the success of the workshop, we propose convening a similar group for the pilot, one that is inclusive, diverse, skilled, and knowledgeable. In doing so, we will bring together the insights, knowledge, and understanding to constitute a governance model for an effective, ethically aligned AI pilot.

Algorithmic Auditing

In June 2021, the National Institute of Standards and Technology (NIST) circulated "A Proposal for Identifying and Managing Bias within Artificial Intelligence"[8] for public comment, as part of a series of recent efforts toward developing a framework for trustworthy AI. The proposal outlines a strategy for mitigating the risk of AI bias, which it calls a "critical but still insufficiently defined building block of trustworthiness." NARA would greatly benefit from the NIST framework in developing its own approaches to understanding and reducing the harmful effects of algorithmic bias. The challenge for NARA will be how to operationalize an ethical AI framework, and key to this will need to be a systematic approach to auditing the effects of algorithmic bias specific to archival description and digital records search.

We recommend an exploratory commitment to developing an algorithmic auditing service for the benefit of institutions curating digital archives. Algorithmic auditing, as explored and underscored in the second workshop session, has been demonstrated to provide an effective means of mitigating this algorithmic bias. Algorithmic audits test the AI software produced by developers in order to gauge its performance with real-world data. Results of these audits can be used to correct biases in AI software and to create public awareness of problems with the aim of ensuring more equitable outputs. The Algorithmic Justice League, led by Joy Buolamwini, is

[8] Reva Schwartz, Leann Down, Adam Jonas, and Elham Tabassi, "A proposal for identifying and managing bias in artificial intelligence systems," NIST, Gaithersburg, MD, USA, Draft NIST Special Publication 1270, June 2021, https://doi.org/0.6028/NIST.SP.1270-draft

among the most familiar organizations that have used AI auditing to address the harmful effects of biased AI.[9]

Currently, there exists no comprehensive algorithmic auditing service for AI software used in digital records search. The majority of auditing work is performed by researchers focusing on a very small number of specific products. Given the vast potential of algorithmic auditing to identify and address AI bias, it is imperative to structure a systematic auditing of AI software for public archives as public knowledge of history, culture, government, and virtually all aspects of life become increasingly dependent on AI to search the rapidly growing digital records of public archives.

Building Diverse Teams for Design and Implementation

Among the most important outcomes of our collaborative workshop were the critical insights and strategies for solving challenges of digital search. This was a direct result of the deliberate decision to fund a majority of underrepresented stakeholders to attend the workshop (more than 50% of funded participants were BIPOC and women). Because we convened an inclusive and diverse set of participants, the workshop produced robust, substantial engagement with multiple challenges and opportunities for solutions. Effective solutions for implementing ethical AI to support digital records search for public archives will require diverse, radically inclusive teams at multiple levels—software design, human-in-the-loop protocols, dataset curation, algorithmic auditing, technology policy, and so forth.

Because robustly diverse teams are *essential* to developing and implementing solutions, the stakeholders and participants who ensure public knowledge is successfully supported through ethical AI must include a majority of underrepresented team members. It is not enough, for instance, to include merely two or three individuals from underrepresented backgrounds on a team of 20 people, as such ‘inclusion’ might be well-intentioned but will typically fail to move beyond tokenism. When only a handful of underrepresented stakeholders are included, it becomes too easy for the conversations and exchanges to remain within the boundaries of convention and mainstream ideation. By contrast, shifting the balance of constituents closer to parity (i.e., majority-minority) fundamentally transforms the team experience. It is far more likely that the experiences, challenges, and concerns of underrepresented groups will become central to strategies and implementation that result from the team’s work.

This has clear implications for structuring collaboration to implement ethical AI in service of public knowledge. In a nutshell, the teams for future work on ethical AI and public archives must be structured to have a majority of constituents from underrepresented backgrounds.

[9] Inioluwa Deborah Raji and Joy Buolamwini, “Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products,” *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019):429–435. https://doi.org/10.1145/3306618.3314244

A Governance Approach to Technology Ethics

Achieving an ethical, inclusive, just future for AI and public knowledge will require a governance approach to technology and innovation. Our collaborative workshop repeatedly emphasized the central urgency of this principle. A governance framework encompasses a comprehensive set of strategies for managing technology. This includes intervening at multiple points in the technology ecosystem—research and development, human-centered design principles, technology policy, human rights and civil liberties, labor practices (e.g., future of work analysis), market impacts, and education are examples. By anticipating potential impacts on human society and the environment, technology governance employs preemptive measures to ensure innovation outcomes meet societal needs, operate equitably, and address sustainability concerns.[10] The terminology employed to reference technology governance varies—the European Union typically uses "responsible research and innovation," and the American emphasis on "AI fairness" and "trustworthy AI" are specific examples of the more general aim to manage technology's implications for public benefit. All of these terms, however, share a common reference to anticipatory practices, ethical frameworks, and comprehensive strategies that encompass technical and societal dimensions.[11]

Our workshop highlighted the incredible value of recognizing that the government is a consumer and practitioner of technology and not merely a source of law and authority for regulating technology. This is an important insight, one that is often overlooked as public discussions of federal agencies tend to focus on regulating technology and not the government's consumption and practice of technology. Recognizing the role of government as practitioner of technology is especially critical for ensuring the present and future state of public knowledge. Among the key participants in the workshop were the Library of Congress, the Smithsonian Institution, and the National Archives and Records Administration (NARA). All three are public knowledge institutions seeking a viable path to thrive in a future that will increasingly hinge on sound technology practices and consumption. The importance of ensuring public knowledge about America's democracy and threats to that democracy has been dramatically displayed during the tenure of this project—the US Senate's investigative committee on the January 6th (of 2021) insurrection is dependent on NARA for records that will enable US Congress to understand past and possible future threats. The preponderance of digital records held by NARA and the ability to search those records to understand events of the January 6th insurrection constitute a singular demonstration that federal agencies are increasingly dependent on AI and other forms of digital technology to make possible public knowledge in a functioning democracy.

[10] Allan Dafoe Centre for the Governance of AI Future of Humanity Institute, "AI governance: A research agenda," accessed 1/27/2022. https://www.fhi.ox.ac.uk/wp-content/uploads/GovAI-Agenda.pdf

[11] Hilary Sutcliffe, "A Report on responsible research & innovation," https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.226.8407&rep=rep1&type=pdf, accessed 1/18/2022. Sandeep Reddy, Sonia Allan, Simon Coghlan, Paul Cooper, "A Governance model for the application of AI in health care," *Journal of the American Medical Informatics Association* (March 2020): 491–97, https://doi.org/10.1093/jamia/ocz192.

The timing of our workshop produced special resonance with important efforts for technology governance. For several years, the federal government has advanced important initiatives to promote ethical frameworks, of which the NIST framework described above is but one example. In 2020, the Pentagon became the world's first military to adopt a comprehensive set of guidelines for AI ethics.[12] In June 2021, on the heels of our collaborative workshop, the White House Office of Science and Technology Policy (OSTP) and the National Science Foundation (NSF) jointly announced a new formed National Artificial Intelligence (AI) Research Resource Task Force that focuses on AI education, harm mitigation, expanding opportunities to apply AI to solve difficult challenges, and encouraging more collaborative efforts to address the AI as a comprehensive subject that will significantly impact national and global issues.[13]

Also of relevance are efforts by civilian organizations. The Public Interest Technology University Network (PIT-UN), of which Virginia Tech is a member, is a consortium of more than 40 colleges and universities preparing future talent by training civic-minded technologists to ensure innovation serves public interest. The consortium's emphasis on cultivating inclusive, diverse participation to guide the future of technology is the most significant development in higher education related to a governance approach. Private companies are increasingly attentive to "AI fairness" standards and frameworks. Think-tanks such as New America,[14] non-for-profit service providers (such as ITHAKA,[15] which operates JSTOR and many other higher education services), NGOs, industry standards organizations, and professional organizations such as IEEE[16] have all strategically embraced a governance approach to AI and other forms of technology.

[12] https://www.ai.mil/blog_01_05_21-the_ai_ethics_journey_will_hit_new_heights_in_2021.html
[13] https://www.whitehouse.gov/ostp/news-updates/2021/06/10/the-biden-administration-launches-the-national-artificial-intelligence-research-resource-task-force/
[14] https://www.newamerica.org
[15] https://www.ithaka.org
[16] https://www.ieee.org

Appendix A. Workshop Agenda

Session One: April 9, 2021, 1:00 p.m.–3:00 p.m. EDT
Theme: Welcome and Orientation

10 min. Introductions and brief procedural overview of the workshop

- Bill Ingram, PI

5 min. Welcome address

- David Ferriero, Archivist of the United States

20 min. Opening presentation: Why we're here

- Bill Ingram presenting

20 min. Overview of the workshop series:

- Sylvester Johnson presenting
- Discussion of the workshop topics and goals—defining the problem scope
- Review of pre-workshop document packet and data sets
- Development of a common understanding among participants
- Emphasis on feasibility and sustainability

25 min. Ice Breaker

- Using Google Docs for shared note taking
- Breakout group activity: identifying areas of expertise and areas of interest

30 min. NARA presentation and discussion

- Introduction to the National Archives (NARA 101)
- What does this group of experts need to know about the NARA catalog in order to get started?
- What are the operational challenges?

10 min. Wrap up

Expected outcomes and deliverables:

- Onboarding participants and explaining data sets
- Identified participant areas of expertise and breakout team assignments

Session Two: April 16, 2021, 1:00 p.m.–3:00 p.m. EDT
Theme: Avoiding AI Bias

10 min. Reflections on prior sessions and goals for today
30 min. Speaker: Tanushree Mitra
40 min. Plenary Discussion
30 min. Breakout activity
10 min. Wrap up

Expected outcomes and deliverables:

- Concrete recommendations for avoiding bias

Session Three: April 23, 2021, 1:00 p.m.–3:00 p.m. EDT
Theme: User Stories
10 min. Reflections on last session and goals for today
50 min. NARA presentation

- Overview of the NARA catalog, content, tech, users - Jason/Erica/Mike
- User personas
- Description of datasets (Jason)

50 min. Breakout sessions

- Writing user stories
- Ideas:
  - Self-describing records, automated description
  - Searching and information retrieval
  - Recommendation
  - Features to facilitate collaboration
  - New use cases made possible by ML/AI

10 min. Wrap up
Expected outcomes and deliverables:

- Set of user stories to identify high-priority needs and to foreground equitable design process and mitigation of bias in AI-driven archival processing

Session Four: April 30, 2021, 1:00 p.m.–3:00 p.m. EDT
Theme: Exploring the Solution Space
10 min. Reflections on prior sessions and goals for today
50 min. Plenary session: Demystifying the state of the art

- Entity/authority linking
- Bibliographic metadata extraction
- Photo facial recognition

50 min. Breakout activity

- Creating a plausible work plan

10 min. Wrap up
Expected outcomes and deliverables:

- Specific tasks for future work

Session Five: May 7, 2021, 1:00 p.m.–3:00 p.m. EDT
Theme: Implementation—Feasibility and Sustainability
10 min. Reflections on prior sessions and goals for today
20 min. Speaker: Patricia Hswe, Andrew W. Mellon Foundation
20 min. Plenary session
60 min. Breakout activity

- Feasibility of solutions

- Sustainability and maintenance requirements
- Training
- Personnel
- Cost models, revenue models
- Sustainability
- Funding of concrete recommendations for next steps

5 min. Post-Workshop Survey

5 min. Wrap up

Expected outcomes and deliverables:

A focused set of concrete recommendations for next steps

Appendix B. User Stories

1. As a community activist, I want to download data in a format that is widely accessible through a GUI, i.e. CSV
2. As a community activist, I want to find historic data on housing, the environment, and other issues important to me held by NARA without having to know names of specific programs under which the data was collected
3. As a community activist, I want to full text search all available reports, whether digitized or born-digital, from one place
4. As a computational researcher, I want to discover and download large datasets in order to analyze on my own computer/server.
5. As a curious explorer I want to see related topics and other suggestions that may be related to my records search, because I am not a trained archivist and am not aware of the full universe of related documents (guide me down the "rabbit hole" of information).
6. As a data researcher, I would bulk download textual records to determine if the records contain terms pertaining to a particular subject.
7. As a digital humanist I need API access to NARA data to visualize relevant research data within historical research projects. I need clear documentation on API endpoints and parameter options to tune API requests as narrowly as possible.
8. As a doctoral student interested in the last decade of the previous millennium, I want to identify key events and people in that period that had a major impact on the world today, and document clearly what happened of significance, and who were the key leaders in those events.
9. As a documentary filmmaker producing a film on the history of the State Department, I would like to identify all publicly accessible photos of Hilary Clinton before/during/after her time as Secretary of State.
10. As a family historian, I want to find records relating to my grandfather's World War II service so that I can understand more about his experience.
11. As a family historian, I would like to leave documentation for my family of how I gathered information about our ancestry so other family members can continue to do the work when I am not able to.
12. As a first time NARA user, I want the ability to serendipitously discover information because I have no idea where to start and what to look for.
13. As a first-time NARA user, I need resources to learn how to use the archive.
14. As a first-time NARA user, I need to feel connected to and welcome in the institution.
15. As a first-time NARA user, I need to understand the algorithms and why I'm searching.
16. As a FOIA requestor, I would like any type of metadata that would help me narrow my search for responsive records to my request.
17. As a genealogist, I would like to know what databases are relevant to my research, and be able to search within documents, in order to find specific individual's records.

18. As a general user of NARA datasets, I would like to have more flexible and robust software tools and detailed tutorials for using them, provided by NARA, for converting datasets with old formats into more modern/general formats such as ASCII.
19. As a general user, I would like to be able to ask questions (instead of doing keyword search) and get answers so that the archives are more accessible and useful.
20. As a high school student, I want to access primary materials to use in a history course term project so that I do and learn something unique and interesting.
21. As a high school teacher, I would like to present OCR'd handwritten primary sources to my students because I would like them to research [x] topic from the 19th century.
22. As a highschool student/history enthusiast, I would like to find photos matching a face I submit from another historical photo so that I can illustrate my history month presentation with additional content.
23. As a history professor teaching an online graduate seminar on historical methods, I want all of my students to be able to conduct a series of search queries using similar terms, so that we can have a discussion about how to evaluate the search results. It would be useful if search results based on similar terms are identical for all of the students
24. As a lawyer pursuing e-discovery, I would like NARA to use the most efficient machine learning search methods to enable me to retrieve responsive records in the least amount of time (maximizing both recall and precision as much as possible—i.e., so that I find all the records NARA has on the subject and so that I don't have to wade through lots of false positive noise).
25. As a librarian/archivist/museum professional at another institution, I want to be able to create a virtual exhibit at my institution and share that "into" NARA in a way that identifies NARA materials that can be added into the exhibition.
26. As a mortician user, I want to research historic embalming methods used. Specifically to filter out chemicals known to cause body discoloration as it can cause a body's ethnicity to be misidentified.
27. As a National History Day student, I want to find photographs and moving images about my topic because I need to use primary sources in my research project.
28. As a natural language processing researcher, I want to develop tools to help those interested in the NARA content, that will be like a machine translation tool that translates between languages, but in this case translates between the language in use at different time periods.
29. As a new academic researcher, I want to be able to distinguish between archival jargon and terminology used by my major.
30. As a philosopher, I want to be exposed to a diversity of viewpoints, so that I can think critically about the topics that interest me.
31. As a processing archivist, I would like to apply an ML/AI toolset to records without embedded metadata so that I can create metadata for review.
32. As a processing archivist, I would like to harvest embedded metadata from born digital records so that I can save myself processing time.

33. As a public user who uses a screen reader, I want to easily access information, so that I can adapt the resources I find for use within a specific community.
34. As a public/citizen archivist, I would like to use ML/AI tools to develop draft records descriptions so that I can audit and improve them prior to adding them to the public catalog.
35. As a rap musician, I want to incorporate some words from some certain documents in a certain period of time, how do I find these documents and how do I know if I can adapt these words for my purposes as an artist?
36. As a Records Manager, I want the system to automatically identify the Record Group / Collection / Series that a particular description should belong to given the text of the description so that the description can be auto categorized in the appropriate Record Group / Collection / Series.
37. As a records manager, I would like to extract series, file/folder and item metadata from email, so that it can be indexed in NARA's catalog.
38. As a records manager, I would like to user Ml/AI tools to assist in decision making so that I can speed up the classification/declassification process.
39. As a reference archivist, I would like to apply topic analysis, entity extraction or ML/AI tools to records having only series level descriptions, so that I can respond to reference requests for records lacking object level descriptions.
40. As a researcher in computer science, I want to find semantically similar documents in multiple modalities, such as text, figures, and tables. It would be better if the user portal provides links to external resources so I could explore other resources containing data that are not available from NARA.
41. As a researcher in computer science, I want to quickly find the relevant data I am interested in and build a sample used for annotation or testing my software. I can download the samples in a batch with all metadata available.
42. As a researcher in computer science, I want to quickly find the research literature I am interested in by narrowing down the search results by multiple metadata fields, such as year, author, publisher, venue, etc.
43. As a researcher interested in geographical information, I want to find and harvest data using GIS coordinates in order to discover and map relevant information. [AI/ML could provide this metadata?]
44. As a researcher of ancient texts, I want to be able to search multiple different translations of ancient texts, so that I can compare these translations.
45. As a researcher of Black history, I want to follow up on a lead from one document and find other documents that might have something in common with it so that I can expand my understanding of the Government's interaction with Black people - even if the race of the people in the catalog is not identified. (I want "more like this.")
46. As a researcher wanting to understand how government policy on a particular issue developed, I want to search for all records on the topic and then use AI tools to sort through what may be millions of records so that I can figure out what exists and where to start actually reading or analyzing the records in order to use my time effectively.

47. As a Researcher, I want the system to address the needle-in-the-haystack problem so that it is possible to easily search for the records that I'm looking for without having to be deeply familiar with the archival hierarchy.
48. As a researcher, I want to identify how policy changes flow across multiple agencies, and understand about how those agencies relate over time.
49. As a researcher, I would like to distinguish between metadata that originated with the agency from any that may have been revised
50. As a researcher, I would like to narrow my focus and have the same search results appear each time.
51. As a researcher/public historian, I want to access the military records and federal records that pertain to women in the military as well as African Americans in military service—scanned letters, reports, Congressional hearings and reports, photographs.
52. As a researcher/public historian, I want to explore the role of Black women in the military WWII to Gulf War because there are potential oral historians and cultural tourism opportunities for collaborations between nonprofits and universities.
53. As a retiree, I want to connect old pictures from the youth of my parents with pictures from the same time period, so I understand the environment in which they lived then.
54. As a scholar of African American language and literature, I want to have guidance identifying materials that relate to distinct cultural topics but also items that specifically address racial exclusion practices.
55. As a teacher of philosopher, I want my students to be exposed a diversity of viewpoints but with some level of uniformity, so that I can think critically about the topics that interest them while also maintaining a common ground for class discussion.
56. As a teacher, I want API access to NARA data to teach students how to query and use historical data in research projects for viz and analysis.
57. As a teacher, I want to have a sharable platform that hosts live documents, so that students can collaborate on various original texts, reconstruct arguments, and mark-up particularly interesting excerpts.
58. As a teacher, I would like to show my middle school students how to search and find resources on NARA's site in real time and in a way that is understandable and interesting to them.
59. As a veteran, I would like to find my separation document more quickly so that I can get quicker access to the benefits to which I'm entitled.
60. As a young person growing up in an extremely poor household, I want to learn about my community but my technology is really, really old and I'm not very skilled with using IT.
61. As an adult who was adopted as a child, I would like to review military records to find my biological grandfather who I know was a veteran so I can learn more about his family and possibly connect with them.
62. As an AI ethics researcher, I need access to NARA metadata on archival records to help explore the ways in which categories or labels might be leading to biased search results.
63. As an AI researcher, I want to access models that have been pre-trained on NARA data, so that I can apply them to my own data, and/or fine-tune them for my use case. (I can

retrain off-the-shelf models on NARA data, but they're likely to struggle as cultural heritage data can be so different from the data sets those models can be trained on—if other people have already trained models on large-scale NARA data sets it saves me a lot of time to be able to use them!)

64. As an AI researcher, I want to be able to construct a bulk-downloadable data set from a search, so that I can obtain a data set suitable for a specific ML training goal. (This search might be topic-based, format-based, etc.—whatever delimiters the catalog supports, I want to be able to bulk-download my results.)
65. As an AI researcher, I want to be able to navigate rate-limiting, so that I can download very large data sets. (so I don't want my query to stop after 20K objects, regardless of the Elasticsearch limit; I don't want to have my access cut off if I download "too many" objects, without some kind of indication of how to avoid that problem (apply for access token, throttle queries but not TOO slowly, etc.)
66. As an AI researcher, I want to bulk download digital objects (not the records or a web page containing objects + records, but the actual objects), so that I can train an ML system on them.
67. As an AI researcher, I want to get the metadata associated with a given digital object by its filename, so that if I have previously downloaded the object, I can readily associate it with the correct metadata.
68. As an AI researcher, I want to get the metadata associated with a given digital object, so that I can construct labels for my training data.
69. As an AI researcher, I want to have extremely rigorous application of authoritative terms, so that I don't have to spend forever cleaning/reconciling variant forms in my data before I can do training.
70. As an AI researcher, I want to have super-easy ways to construct API queries that correspond to particular searches, so that I can work efficiently & without frustration. (This might come from excellent documentation, or a sandbox that lets me construct test queries, or the ability to use whatever URL has resulted from a catalog search as part of my query.)
71. As an AI researcher, I want to read clear technical documentation, so that I can make progress.
72. As an AI researcher, I want to read documentation about the schema of the returned metadata, so that I will be able to parse it computationally.
73. As an AI researcher, I want linked data (e.g., if I have a list of names associated with a photograph, I want to know whether they are people depicted in the photo vs the photographers, et cetera. This will enable me to do things like face recognition (for which I need to know who is in the photo) vs style recognition (who likely took this photo).
74. As an algorithmic bias scholar (applied economics/social scientist/marketing researcher), I want the archival records to be user friendly in its search capabilities, targeted, yet a little more expansive in interpretation and meeting.

75. As an algorithmic bias scholar (applied economics/social scientist/marketing researcher), I want a search to show counts of how many other records/objects were found in search results with the same associated search.
76. As an algorithmic bias scholar (applied economics/social scientist/marketing researcher), I want to be able to access meta-data and labeling on all objects, not just the collection of objects.
77. As an algorithmic bias scholar (applied economics/social scientist/marketing researcher), I want to be able to find records whose meta-data, labeling, text is not only searchable, but is also translatable based on today's context and meaning.
78. As an algorithmic bias scholar (applied economics/social scientist/marketing researcher), I want to include in meda-data the demographic characteristics of the archivist (while maintaining privacy of the archivist him/herself—no names needed) who entered the record.
79. As an algorithmic bias scholar (applied economics/social scientist/marketing researcher), I want a search that avoids the issues seen in Google's algorithmic bias where searches are driven by everyday users' preferences.
80. As an amateur researcher, I would like to see the results not just from the records but also from current/latest research so I can gauge the bias in the records.
81. As an amphibious intellectual, I want to demystify the concept of research as well as training and/or honing the skills of community and university populations.
82. As an archivist I want to be able to take advantage of integrated facial recognition software to perform searches in photographs for people not mentioned in captions or other descriptive metadata.
83. As an attorney [or a historian], I want to identify a piece of legislation or regulation and with one click identify all the records that speak to its history and development.
84. As an auditor, I would like to see some of training data that was used to generate the AI/ML model so that I can endorse or provide recommendations to the training data
85. As an entrepreneur interested in building on the content of the National Archives, I would like to access data using easy APIs in order to incorporate it into my own app/alongside my proprietary content to enhance its connectedness (and therefore value).
86. As an external software developer, I want to access a linked data representation of objects and metadata in the Catalog, because I want to integrate NARA objects and metadata into a linked data resource with related objects in other collections.
87. As an external software developer, I want to access a simple and well documented API through a web service, because I want to integrate NARA objects and metadata into my service in a way that enhances discovery across collections.
88. As an historian interested in climate change, I want to trace White House policy decisions through email correspondence, because I would like to write the history of U.S. actions from 2000-2020.

89. As an indigenous elder, I have concerns about access and representation of indigenous people, artifacts and knowledge—that these will be consistent with indigenous practices/ideas/norms/worldview.
90. As an Information Retrieval researcher, I would like to use NARA's public datasets in my research in order to better understand user needs around historical search.
91. As an instructor interested in my students finding relevant information, I want to be able to establish and share search strategies that will work identically and reliably for my students.
92. As an open government advocate, I want to be able to know what records have been released in response to Freedom of Information Act (FOIA) requests so that I can inform my own future FOIA requests.