NATIONAL ENDOWMENT FOR THE HUMANITIES

DIVISION OF PRESERVATION AND ACCESS

Narrative Section of a Successful Application

The attached document contains the grant narrative and selected portions of a previously funded grant application. It is not intended to serve as a model, but to give you a sense of how a successful application may be crafted. Every successful application is different, and each applicant is urged to prepare a proposal that reflects its unique project and aspirations. Prospective applicants should consult the NEH Division of Preservation and Access application guidelines at https://www.neh.gov/grants/preservation/humanities-collections-and-reference-resources for instructions. Applicants are also strongly encouraged to consult with the NEH Division of Preservation and Access staff well before a grant deadline.

Note: The attachment only contains the grant narrative and selected portions, not the entire funded application. In addition, certain portions may have been redacted to protect the privacy interests of an individual and/or to protect confidential commercial and financial information and/or to protect copyrighted materials.

Project Title: The American Soldier in World War II

Institution: Virginia Polytechnic Institute and State University (Blacksburg, VA)

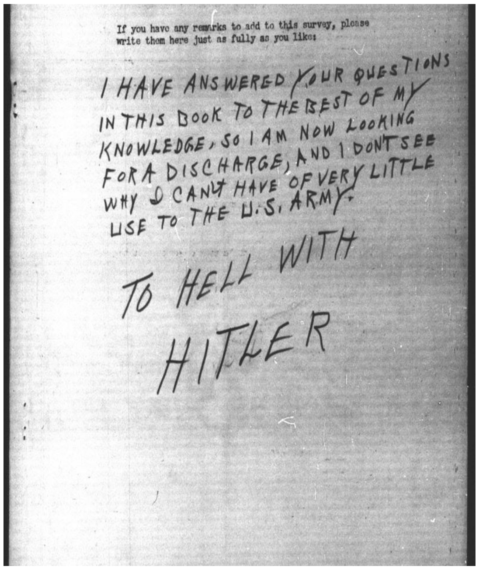
Project Director: Edward J.K. Gitre

Grant Program: Humanities Collections and Reference Resources

III. Narrative

A. Significance

After the US Congress passed the Selective Training and Service Act of 1940, the War Department created an in-house applied social and behavioral sciences research division to understand and assimilate the millions of Americans who were to be added to the Army's roll call. Under the leadership of Frederick Osborn, a family friend of the Roosevelts, the Research Branch played an important role in helping to modernize the US Army, attracting some of the most respected social and behavioral scientists in the country, among them Samuel Stouffer and Carl Hovland. While facilitating the social adjustment of millions of citizen-soldiers to military service, the US Army Research Branch (ARB) and its parent organization, the Information and Education Division (I&ED), helped to create a transportable culture for American servicemembers to enjoy at points across the globe, complete with informative and entertaining films, shorts, and newsreels; books, cartoons, newspapers, and magazines; radio programs that were beamed around the globe; camp shows; correspondence courses; musical instruments and sports equipment; and other leisure-time resources.



If you have any remarks to add to this survey, please
write them here just as fully as you like:

I HAVE ANSWERED YOUR QUESTIONS
IN THIS BOOK TO THE BEST OF MY
KNOWLEDGE, SO I AM NOW LOOKING
FOR A DISCHARGE, AND I DONT SEE
WHY I CANT HAVE OF VERY LITTLE
USE TO THE U.S. ARMY.

TO HELL WITH

HITLER

Sample "free-comment" response

Certain high-ranking officers were originally wary of the "long haired professors" who staffed Osborn's Research Branch. Still, Osborn and his researchers had the indispensable support of the Army's chief of staff, George C. Marshall, and of other progressive commanders who appreciated the monumental challenge of standing up the largest citizen army in US history. With Marshall's consent, the Branch solicited data and opinions on myriads of topics and concerns, from the quality of rations and of specific entertainments to the views of blacks and whites on racial issues. "Tradition" continued to guide the Army's institutional culture, to be sure. Yet the Branch's efforts to adjust civilians to military service helped to channel and intensify the Army's rapid modernization, most controversially toward racial integration.

During the conflict, ARB undertook the most comprehensive and systematic attempt in human history to capture the wartime attitudes, opinions, and experiences of a national army. All told, ARB solicited and amassed information from approximately half a million servicemembers, representing a cross section of the nation. Today, the fruit of the Branch's labor is best known by a four-volume set of books that ARB personnel published after the war, known collectively as *The American Soldier* (1949-50). The books contain a wealth of information on the experiences of soldiers, social science methods, and the US Army as a complex, modern organization. But their comprehensiveness has led to neglect of the exceptional historical records that the Research Branch produced while the world was at war. In the 1970s, the Department of Defense contracted with the Roper Center for Public Opinion Research to preserve ARB's survey data in a more modern format, transferring the data from punch cards to tape. ARB did more, however, than save its survey data for reprocessing and repurposing: it also detached and photographed over sixty-five thousand survey pages containing the responses of soldiers to an open-ended question, a "free-comment" prompt, that asked them to record their opinions, attitudes, concerns, and suggestions on any topic or issue of their choosing. Only a smattering of quotes from these commentaries made their way into the four volumes. Our project aims to take this extensive and unparalleled but hardly known collection of humanities sources and to make it available through a free, open-access website to scholars, students, educators, enthusiasts, and the general public.

The relative inaccessibility of these historical sources accounts for their present obscurity and neglect. The Army photographed the soldiers' open-ended answers separately in 1947 as orphaned records, essentially, and those images were transferred to a single set of forty-four microfilm rolls, which are stored in the National Archives in College Park, Maryland. To our knowledge, these rolls have never been duplicated, whereas anyone who wants to access and download ARB's survey data can do so directly through the National Archives Catalog or the Roper Center website with a membership. (See Appendix A for a list of these surveys and the topics they covered.) ARB's historical collection, especially the anonymous commentaries, absolutely deserve the widest dissemination possible. Guided by the promise of anonymity, the military personnel who responded to the Branch's prompt for their insight wrote without inhibition. "The negro soldier would easily be one of the best and loyalist men in the army if given a half way chance. But the way this army is working you have no chance," one ostensibly black respondent observed. One week, he was a wing assembler on the B-17; the next, he was a hospital porter—a "job that a 70 year old man or woman could do." He concluded, "The co. commander says I will never see a gun. Do you think I feel as if I was doing anything in this war. Hell No." Respondents wrote candidly about their training, equity, and the obstacles of advancement; about leave policies and practices, living conditions on bases, and the Army's caste-like ways; and about the persistence of Jim Crow in an Army fighting for democracy abroad.

There were other opportunities for servicemembers to record their thoughts and experiences. Nevertheless, soldiers could expect that their communications might be censored, or they self-censored. The proliferation of postwar memoirs and oral histories have added immeasurably to our collective understanding of the war. Yet these depend on memory, on the reconstruction of dates and places, on the revivification of feelings, attitudes, and opinions, and on the need to narrate and find meaning in retrospect. Legal restrictions on access to military records have preserved the privacy of individual servicemembers yet at the expense of systematic analysis of the soldier's experience. Finally, tons upon voluminous tons of records were quietly destroyed after the war, as they were deemed to possess no historical significance, to be far too costly to preserve, or both. The soldiers who responded to the Branch's surveys may have done so individually and anonymously; however, the Branch's collective enterprise was by design more systematic, intended to offer the War Department a comprehensive, unvarnished portrait of the state of the US's armed forces, in real-time. The humanities sources that our digital project is making accessible and will disseminate to the public are unparalleled not only because of their singular quality, not only because each soldier's response was composed without fear of consequence, but also because they truly constitute in their totality a cross-section of America's citizen-soldier army.

We believe, though, that students, scholars, educators, and family members of veterans who visit our project's website will value these historical sources for their intimacy and essential humanity. "If this report means anything, and I think it must, I can still believe in Democracy for I, as a soldier, may still express a personal opinion that counts," wrote another soldier. "I was beginning to feel that I was nothing at all, just a uniform. I can maintain some hope for myself as long as I feel that I have some say in the matters which run my own life." Respondents wanted to be heard. The majority of them came to accept that they had a stake in how well the organization functioned, how rapidly it improved and overcame obstacles, how it modernized itself, or didn't, for life, death, or loss of limb might depend on it. These sources tell us much beyond what it took to endure personally and what was required of the Allies to defeat the Axis powers, however. They shed light on what the American citizenry thought about allies, enemies, and non-American cultures; about social adjustment, race relations, mental health, and meritocracy; about education, labor, and recreation; and about the concept of American-ness itself and the meaning and limits of democracy in a time of total war. We anticipate the sorts of people who are likely to visit and use this digital archive, why these humanities sources might be of interest to them, and how these documents might be used in Methodology and Standards as well as Dissemination.

B. History, Scope, and Duration

History

The American Soldier in World War II was initially a curriculum innovation, that of using digital tools, practices, and principles to reimagine the study of World War II. Dr. Gitre had two of the forty-four ARB microfilm rolls digitized in May 2015, intending to design and incorporate a digital project into his World War II courses that focused on the transcription and analysis of these records. Virginia Tech is one of the six US Senior Military Colleges and one of two such programs at a predominantly civilian university. The Department of History and its War & Society course offerings often attract ROTC students. By having students transcribe these handwritten responses, not only did the Project Director's students—many of whom are cadets—assist in making them more accessible, but in turn, his students profited from the no-holds-barred perspective of these uniquely revealing historical sources, which humanized World War II soldiers and helped cadets to put their own military service and careers into historical perspective.

In early 2016 Gitre entered into a collaborative partnership with Dr. Kurt Luther, director of the Crowd Lab and member of the Center for Human–Computer Interaction, both at Virginia Tech. His lab was developing a crowdsourcing plugin, called Incite, for the open-source web-publishing platform Omeka. The plugin extends the platform's capacity for transcribing and annotating digitized historical sources with a focus on classroom application (http://incite.cs.vt.edu). Luther included Gitre in his lab's proof-of-concept testing, using these digitized microfilm rolls, and over the next year and a half, students in six of his World War II courses transcribed, tagged, and analyzed well over 1,500 pages of handwritten commentary, garnering the Project Director a teaching award for the innovative use of technology in the classroom. Their collaboration produced a successful Foundations-level NEH HCRR grant application (PW-253766-17), awarded in the spring of 2017, which provided our team the time and resources needed to advance this project.

During our year of planning, we focused on the following objectives: (1) digitize the microfilmed handwritten documents that were composed anonymously by soldiers; (2) assess a crowdsourcing web-based platform for transcribing these documents and create a site for this purpose; (3) ascertain the required resources and protocols to extract the social survey data accompanying these responses; (4) determine the best methods for reconnecting these humanities and social science sources; and (5) create a full implementation plan. During this past year, the Project Director conducted additional research on ARB's history, determined IP rights, located archival materials relevant to the project, communicated with other institutions holding digitized ARB sources, cultivated external institutional support, and secured agreements that advanced the project. Throughout planning, we strove to maintain an emphasis on the educational usage and value of these humanities sources. We compare proposed activities for the Foundations-level grant to those accomplished in Appendix B.

Digitization of Verbatims

In our original NEH grant application we proposed using internal Virginia Tech funding to digitize all of the microfilm rolls with open-ended responses. In January 2017 Virginia Tech initiated the process with Ms. Denise Henderson, the Internal Digitization Coordinator for the Office of Innovation at the National Archives. Responding to our team's interest in the collection and its inherent public value, the Office of Innovation offered to digitize the entire collection at no cost and in the summer of 2017 scanned all 65,000-plus images as 400-DPI JPEGs. NARA transferred the collection via an external hard drive to Virginia Tech. Although our inspection of the documents found that images from a quarter of the rolls had been cut off and therefore required rescanning, NARA reprocessed defective images and shipped us another hard drive. NARA is also preparing to upload all the images into its public catalog. In the meantime, we have created multiple copies for preservation purposes and have uploaded document sets to the crowdsourcing platform we are using to get the entire collection transcribed.

Transcriptions of Handwritten Open-ended Answers

We originally intended to use the crowdsourcing plug-in Incite to support the transcription and annotation of the open-ended answers. While using Incite in the classroom yielded important insights, during the HCRR grant period we tested the viability of the plugin in a more structured fashion with students in World War II courses taught by another Virginia Tech faculty member. Students produced 520 high-quality, peer-reviewed transcripts and annotations, yet we concluded that Incite's learning-oriented "class-sourcing" approach, which is optimized for classroom usage, was unlikely to scale up to process the complete collection in a reasonable timeframe. After exploring various options, we selected Zooniverse, which describes itself as "the world's largest and most popular platform for people-powered research" with a community of more than one million registered volunteers. Originally built for citizen-science projects, the platform has expanded in recent years to support crowdsource projects in the humanities and social sciences, including several related to the history of warfare. The platform leverages redundant user contributions, which can be aggregated with algorithms, to ensure the quality of crowd contributions meets or exceeds that of trained specialists.

Our Foundations-level grant allowed us to build and test a fully functional Zooniverse crowdsourcing site, working with Zooniverse's design team (see Appendix C for screenshots). We uploaded the scanned images; created workflows and tutorials, field guides, help documentation, and other supporting materials; and tested the site with student transcribers and key project stakeholders before submission to Zooniverse for further testing and review. The feedback we received from Zooniverse community reviewers echoed the enthusiasm that Virginia Tech students enrolled in World War II course often expressed in their evaluations of the project. "I loved this project—I went through a few pages and it was fascinating reading the opinions," commented one reviewer. "I am a genealogist and being able to 'hear' the voices from the past is wonderful. I think sharing this with the public would give people a much deeper understanding of the military in WWII. This is information that is not available in books or anywhere else." The Zooniverse community reviews who successfully completed our transcription workflows unanimously approved the project for public release on the platform.

Having passed Zooniverse's rigorous review, our transcription portal was launched on May 8, 2018, with an in-person and online "transcribe-a-thon," scheduled to coincide with the anniversary of V-E (Victory in Europe) Day. Within the first 24 hours of release, our Zooniverse portal saw 700 new contributors, who transcribed 8,100 documents. Within one week, the number of pages transcribed topped eleven thousand. Commented one citizen-archivist on the platform's message boards, "[G]oing through these you hear the real story from the soldiers [sic] point of view. This has been one of the most incredible projects that I have ever done on zooniverse, this is always going to stay with me." Wrote another volunteer, "I just can not express how much I LOVE this project. I think it is wonderful. I am so glad you got the funds to do what you are doing. Absolutely fantastic how you are planning to use the project. Good job! Bravo!" With eight rolls now transcribed in triplicate, we have initiated the process of transcription reconciliation to produce optimal commentaries. We can confidently estimate, based on our experience so far, that the entire corpus will be transcribed and that the transcriptions will be reconciled well within two years of the Zooniverse launch. Appendix D highlights one poignant commentary the Zooniverse crowd has transcribed, this one on military service from a Black soldier.

Data Transformation

Soldiers who shared their insights and experiences on ARB's surveys did so at the end of a long battery of multiple-choice questions. Historically valuable as the soldiers' commentaries are in their own right, they would be significantly enhanced were they reunited with their parent surveys. Soldiers were encouraged to write freely about any of their concerns, yet their responses were implicitly and explicitly guided by other portions of the survey as well as by the surveyors—by the sorts of multiple-choice questions they posed, by the way questions were worded, by the concerns that motivated survey design, so forth and so on. The multiple-choice questions were designed to capture a considerable amount of information about each respondent. Soldiers were routinely asked not only about their military service, but also about their personal backgrounds, about their education, their families, and religion, about where they lived and what kind of work they did. Too, ARB staff coded surveys during distribution to differentiate units and survey locations, and other demographic characteristics—namely, race. By reuniting the qualitative and quantitative survey data, we could learn far more about these soldiers, beyond their opinions, attitudes, and expectations, even in their anonymity.

Recognizing this, our team decided it was imperative that we extract all that we can from the 138 extant data files. The National Archives and Roper Center hold duplicate copies that capture individual responses from some 50,000 servicemembers. We first engaged Roper to assess their collection and determine the resources and protocols to convert their ASCII-formatted files to a modern statistics software format. Though the Center is eminently capable and was eager to undertake the work as a collaborator with Virginia Tech on the Foundations-level grant, we concluded that its membership model and prevailing terms and conditions would restrict our ability to create a truly open-access website. We then approached Virginia Tech's Social and Decision Analytics Lab (SDAL) to do the same work. SDAL has developed world-class statistical and data science capabilities and has serendipitously established a collaborative partnership with ARB's modern-day successor, the US Army Research Institute for the Behavioral & Social Sciences, to carry out research resembling what ARI's predecessor undertook during the Second World War. Sample survey files were shared with the senior SDAL data research scientist Dr. Aaron Schroeder to determine the best processes for extracting data from the files and to estimate the resources and time necessary to transform the entire collection. SDAL, having successfully extracted sample survey data, estimates it can process the collection during the first year of an implementation grant and has agreed to carry out the work.

Historical Context and External Partnerships

For all this discussion of data and technology, this is, and will continue to be, a humanities project. To appreciate ARB's contribution to the war effort and the humanistic value of ARB's historical records, this little-known Army unit and its applied research must be placed in historical, institutional, social, and political context. Our sense of this necessity was confirmed by reading the NEH HCRR white paper for the "History of Women's Education Open Access Portal Project," which highlighted the "commonly expressed wish" that collection items needed "more robust contextualization" (Pumroy et al., 2015). During the Foundations-level grant, the Project Director continued to collect ARB-related historical sources and with the assistance of a Virginia Tech graduate student started to collate information on and from the individual surveys. He also worked with external institutions to secure additional ARB-related digital assets, including access rights. UNT Libraries, for instance, has granted us permission to use their extensive holdings of I&E Division "Newsmaps." The George C. Marshall Library is contributing high-resolution scans of ARB reports, entitled *What the Soldier Thinks*. We worked with the Social Science Research Council (SSRC) and HathiTrust, which already possesses digital scans of the published volumes of *The American Soldier*, to make all four of the volumes freely available on our website (see Appendix E for a copy of our agreement). In addition to digitizing ARB microfilm rolls, NARA's Office of Innovation also helped to determine usage rights for ARB records as well as assisted in the promotion our V-E Day transcribe-a-thon, as did SSRC.[1]

[1] In accordance with Federal policy, NARA received no funds from the NEH Foundation-level HRCC grant in exchange for its contribution to this project; neither will NARA receive any NEH funds from an HRCC Implementation grant, should this project receive additional funding. Also, as NARA's Letter of Support affirms the activities we are proposing do not overlap with the work the agency is authorized to undertake through its own appropriations.

Scope and Duration

We will create a fully functional, free, public website to house, disseminate, and provide historical context for the work of the wartime Research Branch, anticipating two years to completion. Our Work Plan organizes implementation into two phases, designed more or less sequentially though with some concurrency: 1) Data Transformation and 2) Data Contextualization & Presentation.

Data Transformation

We expect to conclude our transcription drive by the end of year one of an Implementation grant. During this first year, SDAL will concurrently extract survey data from ARB survey files and save the information into file formats amendable to modern application software. The National Archives Catalog provides some metadata for these ARB records, and we have identified survey topics from ARB documentation. But not until the survey data has been extracted from the extant ASCII-formatted files, and not until handwritten documents have been transcribed, can we reunify and optimize the full ARB corpus. Extracted data will help us create controlled vocabularies, identify salient subjects, catalog metadata, and so forth. To reconstruct the Research Branch's portrait of America's citizen-soldier Army, our technical team has planned to implement techniques leveraging the combined strengths of human and artificial intelligence, outlined in Methodology and Standards.

Data Contextualization & Presentation

Once the data have been transformed, we can then focus attention on the website. Our technical lead, Dr. Luther, has already extracted interaction design and system requirements and created "personas" to identify probable site users, and we will continue to refine these. During the first year of an implementation grant, Gitre will conduct additional, though not extensive, research in preparation to oversee the writing of complementary essays to help place ARB and its wartime research into institutional, intellectual, political, and social context. These essays will offer website visitors an overview of the Branch's history and its critical contributions to the war effort, provide short biographies of major ARB staff members, highlight significant issues and themes, as well as furnish information on the nature, scope, and administration of the surveys. Possible themes include race and gender relations, meritocracy and equity, morale, entertaining and leisure, medical care and mental health, leadership and training. Historians on the advisory board will provide editorial oversight and assistance, solicit contributions, and even write some of the site's historical qualitative content.

At the end of the Foundations-level grant, our team secured the Portland-based firm Cast Iron Coding as our web designer and developer. Cast Iron Coding has extensive experience designing open-source web-based applications and websites for universities, non-profits, and other public-sector institutions, as well as for corporate clients. The firm has a record of success with digital humanities projects, such as the collaborative, Andrew W. Mellon-funded, open-source digital-publishing platform Manifold Scholarship, which Cast Iron Coding developed and designed in a three-way collaboration with the University of Minnesota Press and CUNY GC Digital Scholarship Lab. Cast Iron Coding will work with the Project Director and our technical team to develop, design, test, and launch an open-source website using a user-centered design process, one that will allow site visitors to search, browse, and explore transcribed documents, annotations, and other historical ARB sources; interact with the survey data using compelling visualizations; and download historical ARB documents and datasets.

We expect people who visit our website to come with myriad motivations, purposes, technical abilities, and subject-matter knowledge. And while site visitors may not be able to find a particular servicemember by name, we want to create a website that will let users search and explore using proximate information, such as a unit name, demographic characteristic, military installation, or geographical marker. Much like the soldiers' uncensored commentaries, personalized self-exploration can help to humanize a conflict whose global scale easily conceals individual expressions of heroism, sacrifice, and service. Below we outline the preliminary user-centered website planning that was undertaken during the Foundations-level grant.

C. Methodology and Standards

Data Transformation

Data transformation is already underway with the transcription drive to render the soldiers' handwritten responses into a full-text searchable format. Zooniverse's philosophy, backed by empirical evidence, is that crowdsourcing and redundancy can achieve equal or superior results to an expert peer-review process. After consulting with other humanities scholars and social scientists using the Zooniverse platform, we set a triplication redundancy rate for our project. Our CS team is testing reconciliation processes on a batch of completed transcriptions leveraging the National Institute of Standards and Technology's ROVER algorithm. While the algorithm was designed to merge multiple transcriptions from Automatic Speech Recognition (ASR) systems, our technical team believes it can be effectively deployed to reconcile human-produced transcriptions, augmented with human peer review. The results thus far have been promising. Post-reconciled transcriptions will be batch-imported into our backend database/CMS, likely to be Omeka. Zooniverse's Data digging repo scripts can then be customized by our CS and SDAL team members to link transcribed documents to their corresponding JPEG images as annotations.

SDAL, as noted above, has successfully tested data extraction procedures using a sample survey. Dr. Schroeder wrote and executed a first iteration of code to read metadata from the survey's codebook and then use this metadata to parse the corresponding survey file. The file was found to be more complex than initially thought. For instance, while each line of answer codes (per respondent) in the data file corresponds to one "card" of questions administered, each of these coded lines can have different column widths and must be parsed accordingly. After adjusting code to account for this variance, Schroeder extracted the soldiers' responses from the sample file. He then created additional code to pull the question text from the survey codebook, aligned the question text with associated columns in each survey, and verified that the imports had been handled correctly. He and his SDAL colleague Daniel Chen, a data engineer, will apply this semi-automated approach to the remainder of the surveys.

During our Foundations-level grant, and with this information in hand, we developed a hybrid plan to reunify the extant corpus of quantitative survey data and open-ended responses. While prototyping our Zooniverse project, we discovered that many of the free-response documents include intriguing clues as to the individual surveys to which they may relate, such as serial numbers or handwritten codes ARB added during analysis. We have designed our Zooniverse workflows to ask the crowd to capture and transcribe these clues as well as multiple-choice answers that appear on these orphaned pages. We believe these answers and codes may well allow us to reconnect sets of open-ended responses to their parent surveys, or at least to narrow the pool of potential respondents, by correlating SDAL-extracted survey data with crowd-recorded answers. See Appendix F for additional information on re-unification.

Tantalizing as these clues are, we anticipate that many commentaries will remain orphaned, and while Zooniverse transcriptions will render soldiers' responses text-searchable, the unstructured content of the open-ended reflections will be incredibly heterogeneous. Presently we have an incomplete understanding of what these 65,000-plus commentaries have to say at the individual-response level, let alone in terms of broader inter- and intra-survey patterns, themes, and trends. The same applies to the demographics of respondents and to the remainder of the social data contained in the survey files. The challenge is this: how to extract meaning from the transcripts and survey data and how to provide annotations so that this incredible corpus is accessible to our target user groups. Our approach to data transformation requires neither probability nor direct, identifiable linkages between records, nor structured narrative content. Instead, it combines crowdsourced human intelligence with AI intelligence.

Natural Language Processing (NLP) with a topic modelling algorithm, such as the widely employed Latent Dirichlet allocation (LDA), will be used to perform a first pass on the handwritten responses to discover hidden structures for the purposes of extracting salient topics. Expert-generated and crowdsourced tags, solicited through the website, can then be used to improve and enrich these AI-generated topics with human intelligence and domain knowledge. These tags can also be correlated with topics that are delineated in existing survey documentation. Not only will this hybrid approach facilitate the extraction of topics within documents. but will also help us determine relatedness across documents in a scalable way. Dr. Gizem Korkmaz, an SDAL research faculty member, has expertise in developing network-based models and statistical methods using text-based data sources and will provide guidance on implementing this strategy in collaboration with our crowd intelligence expert Dr. Luther and Cast Iron Coding's Zach Davis. Technical details on meaning extraction are presented in Appendix G.

Platform and Hosting

Our Foundations-level technical team, in consultation with Zooniverse, conducted an initial environmental scan of online content management systems (CMSs) for our digitized archives. Our considerations included *cost and maintenance* (the system should be free/open source and have an active base of contributors); *usability and functionality* (the system should provide user-friendly features for searching, viewing, tagging, and categorizing items in a collection); and *portability and interoperability* (the system should use industry-standard metadata formats and provide robust options for exporting data). Based on our current goals, Omeka, an NEH-funded platform developed at George Mason University's Roy Rosenzweig Center for History and New Media (RRCHNM), is our preferred CMS, particularly for the transcriptions and associated image files. Omeka is open source, enjoys a vibrant developer community with dedicated staff, offers a well-designed user experience, and has native support for standard metadata formats with a broad range of externally developed plug-ins. Additionally, our team can draw on deep platform experience, having created and released the Omeka plugin Incite and having taught Omeka workshops as faculty at RRCHNM.

It is important to note, however, that we are continuing to explore other hosting platforms besides Omeka. We need to consider the trade-offs between, on the one hand, using a relatively turnkey solution like Omeka and, on the other, developing a custom site on top of lower-level technologies and languages. We are also mindful of Omeka's limited built-in search functionality, although we are also aware that many projects have successfully integrated Omeka with Solrsearch and Elasticsearch to provide stronger discovery capability. We may find that it is more cost effective to develop more-precise data models in an ORM and expose those to a client application over open APIs than it is to rely on the paradigms built into Omeka. With this in mind, we have added to our work plan an additional discovery phase that allows us to incorporate Cast Iron Coding's essential input into the decision-making process. Cast Iron Coding will consult with SDAL's data scientists and with the rest of our technical team while the firm performs a careful study of CMS options guided by the considerations noted above and by the nature and scope of the ARB corpus and will make a recommendation to the larger group. The backend of our project's standards-compliant, visually appealing, user-focused website will store all 65,000+ 400-DPI scanned JPEG open-ended answers as items with associated annotations. It will also model and house other digital ARB assets, including digital editions of the SSRC's four-volume series *The American Soldier*, along with ARB data sets.

This backend data will reside on high-performance computing servers owned and operated by Virginia Tech's Advanced Research Computing (ARC), while Cast Iron Coding's purpose-built web-interface/frontend will be hosted by VT Web Hosting. This interface will 1) engage our four distinct user groups (defined below); 2) link the handwritten responses to the multiple-choice survey data; 3) leverage crowdsourcing, indexed search with Elasticsearch and NLP to generate meaningful topics and tags; 4) create visualizations of structured data sets as well as user-curated content; and 5) allows users to download all of the Research Branch's open-source records, including complete survey data sets. Presented in Appendix H are preliminary wireframe interfaces that were designed by our technical lead during our Foundations-level grant. Additional technical details on hosting also appear below.

Data Contextualization and Presentation: Searching, Exploring, and Learning

Target Audience and Personas

We aim to reach a broad, diverse audience. And yet, a site designed for "everyone" is not particularly well-designed for anyone. During Foundations-level grant planning, we engaged in the user-centered, Human-Computer Interaction (HCI) design approach called *personas*. Personas are fictional users who represent concrete, real-world needs and attitudes of intended audiences.[2] Project stakeholders were led by our Technical Lead, Luther, himself an HCI specialist, in the development of four such personas, drawn from the insights we gained during Incite testing and Zooniverse review:

- Carol, a military historian researching gender relations and female WWII soldiers (WACs) for a book, representing the *Scholar* user-group of professional historians, gender studies scholars, social scientists, et al.;
- Amy, an undergrad history major representing the *Student* user-group, who use the site to complete assignments but have diverse levels of interest and background knowledge;
- Dave, a high school teacher teaching a US History course, representing the *Teacher* user group, whose assignments must align with state standards of learning; and
- Bill, a retired healthcare worker researching his father's military service, representing the *Enthusiast* user-group, including genealogists and amateur historians.

Moving forward, these personas will help guide our planning, specifically with regard to interaction design and system requirements, complementing the input of our advisory board subject specialists. (See Appendix I for additional details on personas.)

Interaction Design Requirements

Informed by our personas and subsequent stakeholder meetings, we extracted a set of preliminary interaction design requirements and generated corresponding system requirements during our Foundations-level grant. Detailed in Appendix J, they fall under three major categories:

- *General browsing and searching* address the different information needs users have when they first access the site, depending on their goals and background. For example, users who aren't sure what they are looking for should find interesting hooks to start exploring data, such as suggested and featured topics on the home page.
- *Survey navigation and discovery of related content* focus on how to help users interact with the hierarchical structure of the American Soldier survey data. For example, users should be able to navigate easily between surveys, questions, and responses, assisted by the rich metadata we have available.
- *Data inspection and portability* focus on the user experience surrounding *The American Soldier in World War II* as an archive and data repository. For example, users should be able to easily inspect the raw data online or download and export it for later analysis.

Data Management, Metadata, and Storage During Project

The data used within this project provides a historical record and portrait of the soldiers in the US Army during World War II and will be freely available to the public through our website. The execution of this data management plan complies with NEH guidelines and is vital in providing wide access to and

[2] Lene Nielsen, *Personas – User Focused Design* (*Human–Computer Interaction Series)* (New York: Springer Science & Business Media, 2012).

preservation of these records for scholars, historians, students, and the public. For specific data products and standards, see Project Deliverables.

All project data are currently stored in multiple locations and will continue to be for the duration of the project. Upon receipt of the scanned microfilm images, our team made multiple copies on local hardware and on Google Drive with 2-factor security authentication. We uploaded an additional set to the Zooniverse crowdsourcing platform. Completed transcriptions and other crowdsourced classification information are downloaded regularly in a consolidated csv-formatted file and uploaded onto the project's secure Google Drive. After triplicate transcriptions have been successfully reconciled, they will be uploaded into the project's backend and, for preservation purposes, pulled together in groups of 400 images, with corresponding transcriptions, and a README text file with metadata documenting the structure of the data into a ZIP folder, which will be uploaded to Virginia Tech Libraries' data repository VTechData. (Additional details on VTechData follow.) Beyond overseeing this transfer, the Libraries will also port transcribed responses back to the National Archives Catalog to encourage further distribution.

During the time SDAL will be working with the Research Branch's historic survey data, the data will be stored on a new project-dedicated encrypted Logical Volume Management (LVM) partition on their servers. The LVM partition is encrypted using Linux Unified Key Setup (LUKS). Each SDAL data server is housed within one of three highly-secure, high-performance computing (HPC) data centers located in the Virginia Bioinformatics Institute (VBI) in Blacksburg, VA, and maintained by VBI Information and Technology Computing Services personnel. Direct access to data sets for loading and management purposes will be restricted to the Project Director and data managers and will be accessed via SSH using RSA encrypted key pairs. (See Appendix K for additional details.) Virginia Tech Libraries will oversee the uploading of extracted data, completed transcriptions (both the images and associated transcribed text), and supplementary ARB assets to the CMS/backend that is selected by our technical team based on Cast Iron Coding's recommendation.

ARC will host this data on its DragonsTooth cluster, using an on-premises, cloud-system configuration. DragonsTooth cluster runs OpenStack while using a Ceph file system. The combination of OpenStack and Ceph on ARC's hardware gives data availability and integrity assurances in the 99.99+% range. Designed to support general-batch HPC, the cluster is a 48-node, high-throughput, two-socket system equipped with two 12-core Intel Zeon "Haswell" CPUs with 256 GB of memory. To allow I/O-intensive jobs, DragonsTooth nodes are each outfitted with nearly 2 TB of solid state local disk. Jobs running on ARC systems can perform fast I/O to 3.1 PB of parallel storage and have access to permanent storage via a 250 TB Qumulo NFS Home file system. One of several benefits of using ARC is the unit's capacity for high-performance data visualization. Secure SSL web hosting for the project is being provided, as noted, by VT Web Hosting, with an anticipated PHP and MySQL LAMP setup. VT Web Hosting conducts multiple daily backups of sites it hosts and provides support 24 hours a day.

All metadata will follow the Data Documentation Initiative (DDI) standards as this standard aligns closely with data produced from surveys. Metadata for the written responses will be pulled in from the NARA records and mapped to the appropriate DDI fields. Additional metadata fields will be added to the schema as required.

Ethical and Legal Compliance

The original surveys were de-identified with minimal risk to the subjects. The National Archives considers ARB historical documents, including the survey data and open-ended survey responses, to be public domain government records. The project director has confirmed public domain status and out of an abundance of caution requested that Virginia Tech's Institutional Review Board also vet the project owing to inclusion of historic survey data. That office has determined the project is exempt from IRB oversight, as it does not involve human subject research (see Appendix L). To reiterate, survey records were anonymized by the Research Branch at inception. We also in the appendices include a copy of the SSRC-authorized Creative Commons license to reproduce the four volumes of *The American Soldier* on our project's website.

D. Sustainability of Project Outcomes and Digital Content

Preservation of and Access to Data After Project

VT Libraries will use Archive-It and the Wayback Machine to archive the project website. Library personnel are providing advice and assistance on organizing, documenting, and otherwise curating research data to improve its discoverability and reusability. Original and curated datasets are archived according to best practices developed by the Libraries (including conversion of file formats to non-proprietary options where appropriate) and accepted by the disciplinary communities.

As noted above, project data will be stored and archived in VTechData,. Highlighting, preserving, and providing access to data generated at Virginia Tech, VTechData ensures that published datasets receive persistent digital object identifiers (DOIs) while allowing researchers to assign Creative Commons licenses according to their public data-sharing interests. The system relies on item and dataset-level metadata as the primary building block to data discovery, access, and reuse. Use of the repository guarantees long-term preservation of project assets while also providing a secondary portal through which these valuable records can be discovered and accessed. (See Appendix M for more information.) Digital scans of the written responses along with our Zooniverse-produced transcriptions will continue to live at NARA and will be available via their online catalog. Both ICPSR and the Roper Center have expressed interest in obtaining the survey datasets, which would increase distribution as well as ensure preservation by two industry leaders. VT Libraries will arrange and conduct distribution.

E. Dissemination

Our dissemination planning began with the public launch of The American Soldier Zooniverse transcription portal on May 8, 2018, the anniversary of V-E Day. If lessons from the Getty's "Mutual Muses" project are any indication, the most important phase of any project on Zooniverse is its initial launch. Our Project Director worked with the Digital Humanities Coordinator at VT Libraries, Dr. Christopher Miller, and with communications staff from several Virginia Tech units to plan, publicize, and execute a daylong, multi-site, in-person and online "transcribe-a-thon" in commemoration of V-E Day. In addition to the standard, broad approach to dissemination and notification of this participatory DH event—print, social, and news media—we engaged known stakeholders in the digital humanities, military history, and crowdsourcing communities. The Project Director also coordinated promotion efforts with digital media and communications staff at the Social Science Research Council and National Archives. A press release prepared by SSRC went out, for instance, to more than 5,000 members of the SSRC community working in sociology and history and related disciplines and to about 500 press contact covering military affairs.

To emphasize the community-driven nature of this daylong transcribe-a-thon, our Digital Humanities Coordinator employed strategies used in similar contexts to increase participation, such as gamification. High-level contributors were awarded event-specific stickers, based on the number of transcriptions completed with a ranking system analogous to military rank (sergeant, general, etc.). Brief performances, resonating with the content of the transcribe-a-thon, were interspersed across the day, including a dramatic reading of already-transcribed soldier responses and a small ensemble that played period-specific music. The VT Libraries dedicated digital humanities space, the Athenaeum, hosted the transcribe-a-thon along with the performances. Events were simultaneously streamed online, for individual participants to join in virtually, and also to connect remote sites that were also hosting transcribe-a-thon sessions, such as College of the Holy Cross Libraries and Westfield State University.

We have plans to host additional transcribe-a-thon events. The next one is scheduled on or around V-J Day, followed by another soon thereafter honoring Veterans Day. These events will serve as a model for outreach and dissemination for the project after the transcription drive has been completed. (See Appendix N for our distribution list for DH events and project launch and venues for project promotion.)

We intend to maintain the project's initial pedagogical orientation. Dr. Gitre will continue to incorporate the project into his World War II courses and is making arrangements to have students at other institutions contribute. Dr. David Hicks, a Virginia Tech professor of history and social science education, has agreed to help us create a curriculum that use digital archive primary sources and follow state-level curriculum standards. Dr. Hicks, who created the resource site Historical Inquiry, has previously worked with Dr. Luther on another digital history project, Mapping the Fourth of July, funded by the National Archives/NHPRC, supervising the development of 10 assignment guidelines for the high school level and college level that are freely available online and employ a historical source analysis scaffold, called SCIM-C strategy, to support project assignments and assessments. Drs. Hicks and Gitre will work with the project's graduate student to design a similar series of free, web-accessible inquiries using documents from the digital archive. As we outline in Appendix O, our assignment guidelines and activities will be flexible in scope and sequence.

F. Work Plan

Phase	Task	Status	Team Lead
Data Transformation	1. Create, launch, and complete Zooniverse transcription project	To-Do - Y0 - Y1	Luther, Library (Miller)
	2. Extract survey data from ASCII formatted files	To-Do - Y1	SDAL (Schroeder)
	3. Re-unify survey data and orphaned transcribed open-ended answers; deploy hybrid approach to topic modelling and meaning extraction	To-Do - Y1-Y2	SDAL (Korkmaz)
	4. Extract metadata vocabulary and design metadata standards	To-Do - Y1-Y2	Library (Guimont)
	4. Select platform, configure CMS/databases, ingest data	To-Do - Y1	Contract Developer
	5. Identify other historical records and data on ARB activities	To-Do - Y1	Gitre
	6. Port transcriptions to NARA Catalog, deposit extracted survey data into VTech Data for long-term data storage	To-Do - Y0-Y2	Library (Guimont)
Data Contextualization & Presentation	1. Extract interaction design and system requirements	To-Do - Y0 - Y1	Luther, Developer
	2. Produce contextual essays to accompany digital content	To-Do - Y1-Y2	Gitre
	3. Build and test presentation site	To-Do - Y1-Y2	Developer

	3a. Refine topic modelling using human intelligence	To-Do - Y2	SDAL (Korkmaz), Developer, CS
	3b. Implement visualizations for survey responses	To-Do - Y2	Developer, Library (Stamper)
	4. Launch and publicize site	To-Do - Y2	Stakeholders

G. Staff

College of Liberal Arts and Human Sciences, Virginia Tech

Dr. Edward J.K. Gitre, Project Director. An assistant professor of history, Gitre will administer the grant; ensure compliance with NEH, Federal, and University policies; provide overall intellectual direction; and work with the advisory board to produce contextual content for the website. He will hire and direct a History graduate student (GRA) who will assist in project administration, including the scheduling of stakeholder meetings; conduct research in support of writing of contextual content for the site; serve as a community manager for the transcription drive and digital archives; create project-related curriculum; assist in importing images and transcriptions into VTechData; as well as assist the project in other ways as needed. Gitre specializes in the history of the Second World War, war and society, interdisciplinary social sciences, and has held fellowships at the Center for Cultural Analysis at Rutgers University and the Institute for and Advanced Studies in Culture at the University of Virginia. He has been published in the *Journal of the History of the Behavioral Sciences* and *History of the Human Sciences* and elsewhere. His research has been supported by, among other institutions, the Rockefeller Archive, American Philosophical Society, and NEH. 40% FTE.

Michael Hughes, Social Psychology Consultant. Hughes is a professor of sociology at Virginia Tech whose research has been on how social structural factors such as social integration and racial inequality are related to psychological well-being. He will help to ensure the integrity of ARB quantitative data sets post-extraction. Current work focuses on understanding the racial paradox in mental health, with particular attention to racial identity. He is an author of over 70 professional articles, comments, replies, and book chapters, including work appearing in, among other journals, the *American Sociological Review*, the *American Journal of Sociology*, *Social Forces*, the *Journal of Health and Social Behavior*, the *American Journal of Public Health*, and the *Archives of General Psychiatry*. Recent articles have appeared in *Social Science Research, Social Psychology Quarterly*, *Society and Mental Health*, *Psychological Trauma: Theory, Research, Practice, and Policy*, and *The Annals of the American Academy of Political and Social Science*. He is co-author of *Sociology: The Core, 11th ed.*, and has served as Editor of the *Journal of Health and Social Behavior* (2000-2004), as President of the Southern Sociological Society (2004-2005), and from 1992 to 1994 worked on the National Comorbidity Survey at the Institute for Social Research, University of Michigan. 5% FTE.

Center for Human-Computer Interaction, Virginia Tech

Dr. Kurt Luther, Crowd Intelligence Consultant. Luther is an assistant professor computer science and (by courtesy) history at Virginia Tech, where he is also a member of the Center for Human-Computer Interaction and directs the Crowd Intelligence Lab. He will advise the project on user interfaces and software architectures for crowdsourced transcription, annotation, and exploration of survey data. His research group, the Crowd Intelligence Lab, previously developed the NHPRC-funded Incite crowdsourced transcription plugin for Omeka. Luther is active in the digital humanities research community, a contributing editor for *Military Images* magazine, and regularly speaks and writes about how technology can support historical research, education, and preservation. 5% FTE.

University Libraries, Virginia Tech

Marc Brodsky, Archivist Consultant. Brodsky is Assistant Professor and Public Services and Reference Archivist at Special Collections, Virginia Tech. He will work on project metadata and controlled vocabulary to ensure compliance with prescribed standards. Active in the Society of American Archivists, he is engaged in work that incorporates the use of primary-source material in the emerging field of Veterans studies. 5% FTE.

Corinne Guimont, Project Coordinator for VT Libraries. As Digital Publishing Specialist in the VT Libraries, Guimont focuses on disseminating open educational resources and DH projects. She will coordinate and oversee the Libraries' contributions to the project. She will ensure that the data management plan is carried out and that data is stored in the appropriate repositories for long-term preservation. With experience in project management, cataloging, and open-access publications, Guimont is interested in the production of digital scholarly output that utilizes a variety of tools and technology. Her graduate work in information science focused on digital preservation and digital humanities. 10% FTE.

Christopher A. Miller, Digital Humanities Specialist. Miller is Digital Humanities Coordinator for the VT Libraries. He is an archivist, digital collections manager, and computational ethnomusicologist. He will host transcribe-a-thon events at Virginia Tech's digital humanities hub, the Athenaeum, and will liaise with the project's community manager to build and sustain crowd volunteers. Miller has previously worked as SE Asian Studies Bibliographer for Arizona State University Libraries; Curator of e-kiNETx and Cross-Cultural Dance Resources in ASU's School of Dance; Curator of Audiovisual Collections for the Musical Instrument Museum; and Radford University Archivist. Miller's audiovisual and document digitization projects have been funded by the US State Department Ambassador's Fund for Cultural Preservation, the British Library Endangered Archives Programme, the National Science Foundation, and Center for Burma Studies. 5% FTE.

Nathaniel Porter, Data Consultant. Porter is Social Science Data Consultant and Data Education Coordinator in the Informatics Lab at VT Libraries. He will coordinate the porting of data back to the National Archives Catalog, will assist with transcription reconciliation, and work with Guimont on other aspects of data management. He specializes in non-traditional data collection, social network analysis, missing data techniques, and sociology of religion and culture. His data collection work includes surveys, interviews, web scraping, complex online experiments, crowdsourcing, and administrative records. His current research includes crowdsourcing best practices, Indian religious demography, and using online purchasing patterns to study the informal relationships between religious groups. 5% FTE.

Michael J. Stamper, Data Visualization Consultant. Stamper is Data Visualization Designer and Consultant for the Arts at VT Libraries. He will assist in assessing user requirements and in designing the wireframes, interface, user interaction, and display of data and other content. At Virginia Tech, he advises and supports administrators, faculty, and students with their data and information visualization and design needs, helps to define requirements for projects, performs user research, specializes in user interface/experience (UI/UX) design, and integrating the arts, design, and sciences into effective, meaningful, and insightful visualizations. He has taught as an Assistant Professor of Graphic Design at Minnesota State University-Moorhead. 5% FTE.

Social and Decision Analytics Laboratory, Biocomplexity Institute, Virginia Tech

Daniel Chen, Research Associate and Data Engineer at SDAL and Doctoral Student in Genetics, Bioinformatics, and Computational Biology at Virginia Tech. Chen completed his Masters in Epidemiology by developing a computational simulation that looked at the spread of ideas and beliefs in networks. Chen, under Schroeder's direction, will work on the extracting of data. As a data engineer, Chen works on improving the flow of data at SDAL by working on the various technical infrastructure solutions in the Extract, Transform and Load (ETL) process. He is active in the open source R and Python communities and has a book on using the Pandas library for Python data analysis called *Pandas for Everyone*. 40% FTE (Yr. 1).

Gizem Korkmaz, Research Scientist. Korkmaz is Research Assistant Professor at the Biocomplexity Institute and also Adjunct Assistant Professor in the Department of Agricultural and Applied Economics. Her research focuses on social and economic networks, involving mathematical and computational modeling, and empirical analysis. Her PhD dissertation, completed at the European University Institute in 2012, spans game theory and network theory and focuses on the interplay between the network structure and strategic decision-making. She is the principal investigator of the 2016 Minerva research project titled "The Dynamics of Common Knowledge on Social Networks: An Experimental Approach." She was selected as the 2016 Outstanding New Faculty by Virginia Tech Northern Capital Region Faculty Association. The hallmark of her research is to blend her knowledge in traditional economics with big data using tools from social network analysis and machine learning. 5% FTE.

Aaron Schroeder, Senior Research Scientist. An Information Architect and Data Scientist at the Biocomplexity Institute, Schroeder is responsible at SDAL for planning, securing and executing major research projects focused on the techniques, methods, and theories related to the integration, storage, retrieval, sharing, and optimal use of policy-relevant data, information, and knowledge for the purposes of policy analysis and program evaluation. Schroeder will oversee data analysis and extraction of ARB survey data. Schroeder's research focus has been on the integration and analysis of education, health, social service and non-profit administrative data streams for the purpose of conducting policy analyses and program evaluations impacting a wide range of constituents, from pre-K child social and health service recipients to US veteran health and social service recipients. 10% FTE (Yr. 1).

Advisory Board

Dr. Beth Bailey, Foundation Distinguished Professor and Director, Center for Military, War, and Society Studies, University of Kansas

Dr. Kara Dixon Vuic, LCpl. Benjamin W. Schmidt Professor of War, Conflict, and Society, Texas Christian University

Dr. Amanda French, Director of the Mellon Foundation-funded project "Resilient Networks for Inclusive Digital Humanities," George Washington University

Dr. Thomas A. Guglielmo, Associate Professor, Department of American Studies, George Washington University

Dr. G. Kurt Piehler, Director of the Institute of World War II and the Human Experience and Associate Professor of History, Florida State University

Dr. Jeff Pooley, Associate Professor and Chair of Media & Communications, Muhlenberg College

Advisory board members will have reviewed and approved this plan. They will participate during an Implementation grant via teleconference in tri-annual project meetings; review progress under the terms of this proposal; advise on site content, performance, and usability; test prototypes; review the scholarly quality of contextual essays, with some providing essays themselves; and will promote the digital archives. Dr. French will take a more active role in dissemination through community management. Dr. Piehler has included the Project Director in a proposed World War II-related NEH Summer Institute, has invited the Project Director to participate in a NYC teachers in-service training event on Veterans Day 2018 to promote the project, and is also incorporating the transcription drive in an upcoming course at Florida State University.