Data Management Plan

Data: All of the metadata on the ESSSS database is located on the Vanderbilt University Website, stored in a MySQL database which uses a Dublin Core based structure. This requires the production and storage of five versions of each image in order to produce the display. It is stored on network servers with multiple layers of protection and regular backups, and all files are copied onto virtual tapes. For additional security, the system is mirrored at an off-site location. In order to ensure the long-term preservation of these images and their metadata, however, we must encode them according to the Metadata Encoding and Transmission Standards (METS) in XML and then move them to the dark archive in the Digital Preservation Network (DPN), which allows indefinite storage in a failsafe repository. Only three versions of the images are created and stored for processing, allowing them to be optimized to require less storage and fewer resources. If funded, our staff will have sufficient support to encode all of the images and convert them for the DPN.

Interface: The current ESSSS website utilizes two distinct technological frameworks. The content management system is based on PHP (OmniUpdate) and the second uses a locally-developed PERL-based framework for presenting the actual images. Most other humanities databases and archives in use today have moved away from these individual frameworks and toward standardized, open-source systems using TEI or MySQL. This means that without significant manual labor, data from the ESSSS database cannot be effectively integrated with those of others. Despite this, Google Analytics demonstrates that the ESSSS database has a wide readership and that scholars from all over the world use this data in their research. In order to allow optimized data-sharing our staff will transfer the existing files with enriched and

standardized metadata to new and improved software platforms. ESSSS team members will transfer the two existing frameworks into a single system utilizing SobekCM open-source software (http://sobekrepository.org), which utilizes C# to process metadata stored as METS but also provides a crosswalk to TEI. Therefore, all of the information in our current system will be integrated into one software that is designed to work with the types of software the most current Digital Humanities databases utilize, making them interoperable.

Sharing: Every aspect of the database, from the metadata, to the imagines, interface, and the software used will be made freely available to anyone in the world. All source code developed in connection with this grant project will be released under the General Public License Version 3.0 (GPLv3) (http://www.gnu.org/licenses/gpl-3.0.en.html) and publicly distributed through Github. All metadata will be dedicated to the public domain via the Creative Commons Public Domain Declaration (CC0 1.0 Universal) (https://creativecommons.org/publicdomain/zero/1.0/legalcode). All images, both the high-resolution as well as the access images, will likewise be dedicated to the public domain via CC0 1.0 Universal (unless prior licensing restrictions have been attached). Metadata will be made available through Github with direct access to images.

Data Management Plan

Roles and Responsibilities

This data management plan is implemented and managed by the University of Virginia Library's Department of Digital Research and Scholarship. Eric Rochester will assist in transferring the source code to the University of Virginia who will have responsibility for the long-term storage of the source code. All source code will be publically available.

Expected data

We are preserving both the source code for the project and its accompanying documentation. During the course of the project, the source code will be managed through a distributed version control system on the developer's machines, and publically through the Scholars' Lab's open GitHub website: https://github.com/scholarslab/neatline. Documentation is also managed in a similar way, with the original source documents (and their histories) being managed through distributed version control. The end-products of this project will be uploaded to the Omeka add-on list which is backed up on a nightly basis.

Period of data retention

All relevant data will be contributed to the University of Virginia's institutional repository. All source-code will be publically available throughout the development process and afterwards. No data will need to be retained for other purposes.

Data formats and dissemination

The software code and documentation is what will be retained for this project. Both of these sources of data will be maintained as text files publically accessible from the project's GitHub repository as well as a downloadable software package from the Omeka add-on directory.

Data storage and preservation of access

All public data (project documentation, workshop data, and project history) will be deposited in the University of Virginia's Libra institutional repository that has capabilities to manage, archive and share digital content. Libra allows access to the public via persistent URLs, provides tools for long-term data management, and permits permanent storage options. Libra has built-in contingencies for disaster recovery including redundancy and recovery plans.

Data Management Plan

Expected Data

Type of Data	When Shared?	Under What Conditions?
Open source computer code.	Code will be publicly available as soon as the project begins and will undergo development in public.	Code will be available on the project website and GitHub, under the terms of the GNU General Public License.
User documentation and tutorials.	After beta launch.	User documentation and tutorials will be freely available on the project website and with the redistributable software package.
White Paper.	At the conclusion of the initial grant-funded period.	Dissemination of the final report will be the responsibility of the NEH.

Data Formats and Dissemination

Software will be written in the PHP scripting language, with associated files in JavaScript, HTML, and CSS, and possibly other scripting or markup languages. Themes will also contain image files in formats such as .png (Portable Network Graphics). Existing open source resources such as WordPress or the JQuery library may be included with or without modifications, always in accordance with the licenses governing those resources. This code will be released under an open license and hosted on both the CBOX website at http://commonsinabox.org/ and on GitHub.com for perusal, retrieval, and modification by the public. Formal reports will be made publicly available in PDF format on the project website at the end of project completion.

Period of Data Retention

The Commons In A Box does not collect data from users of the software. The software is distributed by the WordPress Plugin Repository and is downloaded through the WordPress Dashboard onto servers run by individual users (be those users single instructors or entire institutions).

The software itself is archived under version control on Github; all software produced through this grant will be freely available on the project website or, in the unlikely case that Github is unavailable, on a similar code-sharing website, for a period of no less than five years.

Data Storage and Preservation of Access

All computer code, including versioned records of all changes, will be licensed under the GNU General Public License and available via Github alongside the existing code for Commons In A Box. User documentation materials and tutorials will be hosted on the Commons In A Box server at http://commonsinabox.org, which is administered and housed on site by the Information Technology (IT) Department at the Graduate Center, CUNY (GC). The IT Department creates a data backup every day at 6am by exporting all databases and all files on the Commons server to a co-located backup server at the GC. Each daily backup is kept for 28 days. Additionally, the Graduate Center has an identical fail-over physical server located at Baruch College, to which it synchronizes data daily at 5am. The Domain Name Service (DNS) record for the Commons, commons.gc.cuny.edu, is hosted through the cloud-based Amazon Route 53 DNS service, so the Commons can be switched over to the Baruch College servers should GC facilities become unavailable.

Development of CBOX will continue after the grant period ends, but at the end of the grant period, CUNY Libraries will store complete and current (at that time) versions of CBOX as compressed tar archives (.tar.gz files) in the Graduate Center's institutional repository to ensure long-term access to the code.

Revitalizing *Mission US*
Data Management Plan

The proposed project will generate several different types of data, which we plan to manage and disseminate as follows:

***For Crown or Colony?* game source code and assets**

- Management: All source code and assets for the Unity game engine and new game features, including the digital storyboard tool, will be stored in a password-protected cloud-based SVN repository. (SVN is an industry-standard version control system.)
- Period of retention: Source code and assets will be retained indefinitely.
- Dissemination: During the final six months of the project, the updated version of the game will be made widely available for free via the mission-us.org website and as a native app on iPad and Android tablet devices. Source code for the new digital storyboard tool will also be made freely available through GitHub (https://github.com/) upon release of the updated game. We do not plan to make source code for the full Unity game engine publicly available, as the amount of extra documentation and support required to create a version that would be usable by outside developers would exceed the scope and resources available within our project budget.

***For Crown or Colony?* user registration, game state, and web usage data**

- Management: All user registration and game state data are stored/hosted by Amazon Web Services. Google Analytics tracks and stores web usage data. Electric Funstuff also gathers data on user gameplay decisions that are stored in a private Google doc. In accordance with COPPA regulations, no personal identifying information is gathered for student users. Educators/teachers have the option to submit an email address if they choose to opt in to the *Mission US* email list; a copy of the email list is stored on WNET's internal Dell server, backed up internally on a daily basis, and secondarily backed up to an off-site tape archive on a weekly basis.
- Period of retention: Data are retained indefinitely. Daily backups on Amazon Web Services are retained for five days.
- Dissemination: User registration, game state, and web usage data are used for internal analysis to inform game design and outreach. These data are not shared externally, with the exception of broad metrics (e.g., total number of registered users or pageviews over time) that may be shared in fundraising or promotional materials. For those users who opt in to the email list, email addresses may be used to send updates specifically about *Mission US* games.

Formative testing data

- Management: Data gathered during formative testing of new game features among small groups of users will be aggregated by researchers at Education Development Center and stored on their internal servers. These will include observation notes, records of game play moves, and transcribed interviews. These data will be collected under protocols approved by EDC's institutional review board, and will be anonymized in accordance with human subjects privacy requirements.
- Period of retention: Formative evaluation data will be retained for three years on a secure server and then deleted.
- Dissemination: Aggregated findings of formative research will be synthesized and shared as part of the white paper that will be disseminated via the *Mission US* website, presentations at conferences and webinars, and the final report to NEH. No information will be shared that could identify individuals participating in the assessment process.

Project evaluation data

- Management: The *Mission US* team will assemble documentation of upgrades implemented, why these upgrades were deemed important, challenges the team encountered, results of formative testing, final reviews by project advisors, and best practices/lessons learned. Data will be aggregated and stored on WNET's internal Dell server and backed up to the internal server on a daily basis with Symantec Backup Exec vRay Edition software. A secondary backup is made to tape archive at an offsite storage facility on weekly basis.
- Period of retention: Project data will be retained indefinitely.
- Dissemination: These findings will be compiled in a white paper to be made freely available via the *Mission US* website at mission-us.org. Information will also be disseminated via presentations at conferences, webinars, and the final report to NEH.

6 Data management plan

6.1 Roles and responsibilities

The project director and the postdoctoral fellow will have primary responsibility for data management during the grant period. The project director will be responsible for data management after the expiration of the grant. The director and postdoctoral fellow will be assisted by members of the cyberinfrastructure team headed by Chris Sweet within the Notre Dame Center for Research Computing for issues related to data management during the grant. For long-term preservation issues, the CurateND team based in the Hesburgh Libraries will assist with repository issues.

6.2 Expected data

As noted in the project narrative, we will produce data of two types: geographic data associated with digitized texts and code used both to produce that data and to implement the project interface site. The anticipated total size of the data generated by the project is estimated to be several terabytes, almost all of which is accounted for by the dataset itself. Processing and interface code occupies no more than a few gigabytes.

For purposes of compliance with OMB Circular A-16 and Executive Order 12906, we certify that the geospatial data products we propose will be produced in compliance with applicable guidance from the Federal Geographic Data Committee.

6.3 Period of data retention

Data will be retained on all platforms through at least 2023 and in Notre Dame's institutional repository indefinitely.

6.4 Data formats, storage, and preservation of access

Data will be made available in multiple formats as appropriate for its intended uses. Geographic data will be accessible through the project interface and downloadable in CSV and JSON formats. Project source code will be managed on GitHub or a comparable service in the event that GitHub becomes unavailable or unsuitable at a future date.

We already own the hardware on which the project is hosted. The server and disk array have substantial overhead for increased use; as noted in the project narrative, we have 64 cores, 64 GB RAM, and 34 TB of RAID storage collocated in Notre Dame's high-availability data center. Network connectivity and data backup are provided on an ongoing basis as indirect costs.

We see two aspects of the project that will require resources beyond the end of the grant period, and have made provisions for both.

First, the geographic dataset will need to be stored and made available to users in minimally processed form. This need will be met via three independent channels: the project site, Notre Dame's institutional repository (CurateND; https://curate.nd.edu), and the HathiTrust Research Center. The project site allows us the greatest flexibility in formatting, subsetting, and interactivity. Its costs after the grant period are minimal and will be covered

for at least five years by Notre Dame funds already allocated for that purpose. CurateND is specifically designed for ongoing, long-term preservation of research products and is provided to Notre Dame-originated projects on a no-cost basis. Finally, our collaboration with the HTRC is designed to make both our data and our computational methods available to the largest possible community of users. The HTRC is committed not only to hosting our full dataset in a form compatible with their existing products, but also to implementing our ingest pipeline so that it can be applied to volumes that are added to the HathiTrust Digital Library in the future. This ensures the preservation of grant-produced data, extension of that data to newly acquired materials, and additional documentation of the processing and ingest methods.

The second element of the project that will require support beyond the end of the grant period is the user interface site. This site is hosted on the production server described above and consists of code written primarily in Python and JavaScript. We have three channels of long-term sustenance for this aspect of the project. As in the case of the underlying dataset, the server that hosts the project will be supported for a period of at least five years through funds granted by Notre Dame. We will deposit the full source code for the project in CurateND for long-term storage, with plans to provide updated, versioned releases at major milestones. Finally, we will make all source code available on GitHub (which currently hosts the development code) or a comparable platform under an open-source license for public use. Using GitHub helps to ensure easy practical access to the project's code by interested users and facilitates user contributions back to us, thereby extending intellectual collaboration around the project.

6. Data Management Plan

Expected Data

A variety of data types are expected to be generated as a result of this grant. The primary data to be generated by this project will be the programming code of the interface plugin itself—a BuddyPress-compatible WordPress plugin (PHP) with accompanying JavaScript and CSS, comprising approximately 10,000 lines of code. All project staff will check the code in and out of a Git repository, hosted centrally through the MLA's private GitHub account. Documentation of this code will be produced alongside the code itself in PHPDoc format, with accompanying Markdown-formatted configuration and installation instructions. Some further data may also be collected to assess the usability of the user interface plugin, using bibliometric, qualitative, and click analytic methods. Minimal data generation is expected using these methods. Both the collected and the analyzed data will be anonymized or aggregated to the extent necessary to obscure any individual participant's identity before they are shared. Prior to sharing, collected and analyzed data will be kept in private spaces on the MLA and CUL networks, backed up securely, and accessible only to project staff.

Period of Data Retention

In accordance with the Columbia University Retention and Access to Research Data policy, the implementation project staff agree to retain all project data for a minimum of three years after the period of the grant. All project data to be made publicly accessible will be deposited in public repositories (such as *CORE*) within 30 days of the submission of the final report to the NEH. At the project outset, no embargo periods are expected.

Data Formats and Distribution

Final project data connected to the *Humanities CORE* software along with its associated documentation will be fully released upon project completion through open repository and open-source software distribution platforms. Where permissible, anonymized and analyzed assessment data will also be published openly in connection with the final project deliverables. Data reuse conditions will be made explicit upon publication, with a preference wherever possible for maximum potential reuse. This includes data in all expected formats: source code; documentation in text, word-processing, and printable formats as necessary; and documentation in all other formats aggregated over the course of the project period. Preference will be given at all times to platforms and formats that favor open distribution and technology free from proprietary limitations, in pursuit of the goal of data freedom upon project completion (within 30 days of the submission of the final report to NEH). No embargo periods are expected at the project outset, and it is not anticipated that any of the project data will be subject to confidentiality concerns. Access to project data will be facilitated by the policies of the open repositories that will be used to publish them. No access arbitration will be necessary for project data.

Data Storage and Preservation of Access

The investigators will use *CORE* as the primary preservation and access platform for the final project data. Domain-based repositories may also be employed where identified as trustworthy and beneficial to project data during the project period. Deposit in *CORE* provides a persistent URL, secure replicated storage (multiple copies of the data, including onsite and offsite storage with verified checksum procedures), accurate metadata, a globally accessible repository and the option for contextual linking between data and published research results. Any file type may be deposited in *CORE*. Files are available for public consumption and reuse, and for machine extraction.

Roles and Responsibilities

Throughout the Implementation grant period, access to the project data will be limited to the named project staff and the CUL system administrators. Responsibility for data management will remain with CDRS. Adherence to this Data Management Plan (DMP) will be the responsibility of the designated Project Manager, who will ensure the following with respect to software developed: (1) that project staff are following the agreed-upon practices for code versioning and development, (2) that the data archives are comprehensive throughout the period of the start-up grant, and (3) that the source code for the final production code is published to an openly accessible data repository. The project manager will likewise ensure appropriate practices with respect to data produced during the assessment portions of the project. If the Project Manager should need to be replaced midway through the period of the start-up grant, it will be the joint responsibility of the Principal Investigators (PIs) to periodically audit the data practices of the project staff until a new Project Manager is named.

DATA MANAGEMENT PLAN

Six Degrees operates under the LOCKSS principle--Lots of Copies Keeps Stuff Safe. Chris Warren is the owner/registrant of *Six Degrees'* shared accounts and emails, with ultimate control over the Google Drive folder and Github repositories. The postdoctoral fellow is responsible for active data management.

Operational Data

All of our operational data--including text documents, spreadsheets, images, and videos associated with the project--is stored online in a shared Google Drive, which all current and former collaborators on the project may access. In addition to Google Drive's versioning system, this data is manually backed up on a monthly basis and these versioned backups are stored on four hard drives maintained in private offices and residences in three separate locations--two within the city of Pittsburgh, PA, and one near Washington, D.C. All hard drives stored in residences are password protected for security purposes.

Datasets

Since all of the subjects of the project are long dead, no privacy issues inhibit data sharing. While copyright prevents us from sharing the textual corpora that are the basis for the project's statistically-inferred data, the datasets themselves are stored online in our main project website. Registered users may already freely download the datasets in .CSV format, under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 Unported License.

All our users agree to Terms of Service that allow their contributions to our datasets to be downloaded along with our statistically inferred data. These Terms of Service also make clear that usernames will be publicly visible in order to give users credit for their contributions. Special user classes can further download user-contribution-related metadata--such as timestamps--in the same .CSV files as the rest of the datasets.

The datasets are manually backed up on a weekly basis and these versioned backups are stored as .CSV files on the same four hard drives, in three separate locations, as our operational data. These datasets are also backed up to an account in Box Cloud Storage--a commercial service licensed by Carnegie Mellon University for campus file storage--which the project co-PIs and postdoctoral fellow can all access. The Folger Shakespeare Library, in Washington, D.C., also wishes to host versioned copies of our dataset in their current digital media repository, which will ensure the datasets' long-term preservation. Further programming is necessary to automate the database backups and establish a remote backup with the Folger Shakespeare Library. We consider this a medium-level project priority.

Software Code

All R code for statistical aspects of the project are available online through a public GitHub repository (github.com/sdfb/sdfb_network). JavaScript code for a preliminary version of our website interface--containing most of our network visualization features and using simple, humanist-friendly Google Sheets as the database framework--is available online through a public GitHub repository (github.com/sdfb/sdfb_spring2014IS). The JavaScript and Ruby on Rails code for our current website interface is stored online in a private GitHub repository and will be released to the public at the end of the grant period, a delay which is necessary to ensure the stability and security of the user contribution component of our interface. Prior to the end of the grant, the code may be released to known researchers on request. In addition to GitHub's versioning system, this code is also manually backed up on a weekly basis. Versioned backups of all code - R, JavaScript, and Ruby on Rails - are stored on the same four hard drives, maintained in three separate locations, as our operational data.

Articles and Scholarly Publications

To ensure public access to publicly-funded research, the full text of all articles and similar scholarly publications will be made available via the project website in .PDF format as well as deposited in arXiv.org, which has been in existence since 1991, and is more permanent than many commercially-published journals. These articles will also be archived through the Modern Language Association's Common Open Repository Exchange (CORE) (commons.mla.org/core/) and Carnegie Mellon University's institutional repository (repository.cmu.edu). These steps will be taken as soon as possible after publication in order to ensure that, even if the project's website folds, there will be permanent public access to the papers.

The full text of all our scholarly publications are also are stored online in our shared Google Drive and the same four hard drives, maintained in three separate locations, as our operational data.

DATA MANAGEMENT PLAN

Expected Data

The data generated by this project will fall into six major types: documentation (procedure manual, ImageJ plugin documentation), captured and processed images, RTI images, annotation data, web pages, and software code. Email correspondence among the team will not be considered archival data. Additionally, instructional videos may be created.

Data Format

The procedure manual will be archived in PDF format. The documentation for the ImageJ Spectral RTI Toolkit plugin will be archived in HTML5. The captured and processed images will be archived as uncompressed 16-bit TIFF files with EXIF metadata describing contents, capture, and processing. PNG and JPEG compression may be used for web dissemination of images that will also be available without compression. RTI images will be archived in PTM (Polynomial Texture Map) and/or HSH (Hemispherical Harmonics) format. WebRTI, as a form of web dissemination, will use PNG and JPEG compression. Annotations will be stored in XML and HTML, with TEI standards for transcribed text. Shared Canvas manifests will use JSON. Image metadata will adhere to Dublin Core standards. Authority data will comply with MADS/RDF. The proposed Digital Public Library of America record will use OAI-ORE. Webpages will use HTML5 with CSS and Javascript. The ImageJ plugin will be written in Java. Developments to WebRTI will be in CSS and Javascript. Instructional videos will use the H.264, MPEG-4 AVC format at 1280x720 or 1920x1080 resolution (used by BluRay and web browsers).

Access to Data and Data Sharing Practices and Policies

The primary means of access to all data will be palimpsest.stmarytx.edu. This site will include the IIIF compliant image repository, navigable web pages for each manuscript page with image and annotations, the complete data archive (mostly TIFF images), documentation, and software code. Static URIs will follow the principles of Linked Ancient World Data. Code will also be available through GitHub. The white paper will also be available through neh.gov. Instructional videos will also be accessible on YouTube. The complete data archive will also be available on magnetic disk in the St. Mary's University Library Archives. The description of Spectral RTI will be submitted to Wikipedia. Other major discoverability and aggregation engines will be Google, WorldCat, Digital Public Library of America (conditional), and the Online Critical Pseudepigrapha. Access to the data will be completely unencumbered; no account registration or other impediments will be required.

Policies for Re-Use and Re-Distribution

The manuscript remains the property of the Biblioteca Ambrosiana and permission will be required for reuse in print media, but permission has been granted for all non-profit use (see Letter of Commitment). All other products can be used for any purpose with attribution (CC BY-SA).

Archiving of Data

The project server hosted by Amazon Web Services includes RAID redundancy and off-site backup. Additional copies will be kept on magnetic disk minimally by St. Mary's University Library Archives, the project director (Hanneken), and the Early Manuscripts Electronic Library (Phelps). Data integrity will be verified with MD5 checksums throughout the project. Syntax of filenames will be documented, as will correlation with IIIF identifiers.

Data Management Plan

Since 2008 the website, www.slavevoyages.org that this project is designed to sustain and enhance, has provided free and unrestricted access to three databases that have become standard sources for anyone interested in the transatlantic slave trade. The first is a set of nearly 35,000 transatlantic slave voyages (termed "TSTD") containing 99 variables, the second, a database of 11,709 annual/regional estimates of the slave trade (termed "Estimates") and the third provides the personal details of 92,000 Africans rescued from slave vessels in the last half century of traffic (termed "Names"). Not only are these data viewable, but they may be searched, selected and analyzed, and any part of the data or the output of subsequent analysis may be downloaded into csv or Excel files. Given that the transatlantic slave trade ended in 1867 there are no legal or ethical restrictions on the use to which these data may be put and there are no retention requirements. TSTD and Estimates contain imputed variables as well as data variables. In each case the website offers full explanations as to how the imputed variables were generated and these explanations, too, may be downloaded without restriction. The downloads page at http://www.slavevoyages.org/tast/database/download.faces provides versions of the three databases in a variety of formats. The refurbished website will continue these data management policies.

This project will result in the creation of an additional database of slave voyages. These sailed between one port in the Americas to another and date from the same era as TSTD. It will be termed the Intra-American slave trade database. Its structure will be identical to TSTD in that it will have a mix of data and imputed variables and the algorithms for calculating the latter are also the same. By 2007 we estimate that it will comprise at least 10,000 records and contain 99 variables. The project team will spend seven months preparing the dataset and will make it available for use on its own interface that will be built as the database itself is created. Once the new site is launched then it will take its place with the other three data sets and be subject to the same unrestricted public access. A copy will be made available on the website's download page in Excel, SPSS and csv. Also available on this page is the SPSS codebook. Of the 99 variables in the new database, 32 are coded.

The website will be stored and maintained on servers managed by the Emory University Library & Information Technology Services (LITS). The re-engineering of the site will include provisions to enable effective web crawling for archiving purposes, which will significantly improve on the current site. As of now, the site cannot be effectively archived via web crawlers due decisions made about Java implementation. The re-coded site will address these web archiving hindrances, allowing the website to be included in regularly scheduled crawling and archiving of university domains, and archived in the Emory University Archives Web Archives. The datasets and all accompanying documentation (codebooks, methodology and guides from the website) will be archived in the Emory Dataverse, which participates in the Dataverse Network, an open source framework for archiving and disseminating data developed at Harvard University. The Dataverse Network utilizes a secure and geographically distributed digital preservation system using the LOCKSS software created by Stanford University and under the guidance of the Data-PASS project. Copies of the site and the data it draws on are also located at the Wilberforce Institute for the Study of Slavery and Emancipation, Hull University, UK, (under the guidance of Richardson), at the Fundação Casa de Rio Barbosa of Rio de Janeiro, and at Victoria University of Wellington (under the guidance of Behrendt). Emory LITS is negotiating participation in an established national digital preservation network, which would provide a long-term multi-site dark copy of the data and website content, in the event of total failure of the university's systems.

Data Management Plan

The *Pleiades* data management plan conforms to, and in many respects goes beyond the requirements of, New York University's *Policy on Retention of and Access to Research Data,* dated March 1, 2010.[1]

Pleiades manages a variety of data classes that are necessary to support the description of ancient geography and to manage the activities of our scholarly community.[2] Upgrades planned for this project will refine the data model by adding support for relationships between places and associated documentation, but all existing content will be migrated with no loss of information. Data objects are versioned now, will continue to be versioned, and the versioning history of all data objects will also be migrated with no loss of information. Other datasets subject to this plan include blog posts, help documents, and other content published on the *Pleiades* web site, as well as conventional server access and error logs. *Pleiades* code is published on-line, and is also subject to this plan.[3]

In accordance with NYU policy, copies of all *Pleiades* research data will be retained by the PI for a minimum of 5 years following the final reporting of the project. Copies of some datasets will be retained for longer periods under various custodies (see below).

Server access and error logs will be retained in their native formats. Although they do not include personally identifiable information, they contain IP addresses and other sensitive information, and so will not be released publicly. Under terms of the NYU data retention policy, they will be retained and made available by the PI to appropriate individuals in the context of an audit, dispute, sponsor request, FOIA request, or other matter as directed by the Senior Vice Provost for Research (or a designee).

Pleiades geographic data is published over the web from the system in a number of standard formats, and we will maintain the full suite of these going forward. Individual data serializations for each *Pleiades* place, as described in the "Innovations" section, are disseminated using HTTP via stable URIs. The current nightly export to CSV format will continue to be disseminated via the *Pleiades Downloads* page at http://pleiades.stoa.org/downloads/. It will be enhanced to include all published data items and their attributes, thereby constituting a complete, application-neutral copy of the gazetteer in a widely used, non-proprietary open format that is easy for digital repositories and collaborating third parties to maintain and use. Textual information in all files is encoded using the standard UTF-8 (Universal Character Set Transformation Format 8-bit). This export package is currently accompanied by a ReadMe.txt file explaining the content and its expected use. The PI, in collaboration with the Managing Editors, will enhance this metadata with the addition of appropriate documentation in accordance with ISO 19115:2003 "Geographic information – Metadata (corrigendum 1)" as recommended by the Federal Geographic Data Committee (FGDC).

Pleiades is currently hosted on a leased server in a redundant data center located near Denver, CO. Our present contract provides for a dual, high-performance disk array configured as RAID 1, which provides

[1] NYU policy: https://www.nyu.edu/about/policies-guidelines-compliance/policies-and-guidelines/retention-of-and-access-to-research-data.html.
[2] See further: *Pleiades Data Model:* http://pleiades.stoa.org/help/pleiades-data-model.
[3] http://pleiades.stoa.org/software/.

for operational redundancy. The vendor provides comprehensive daily backups that are tested for restoration readiness on a quarterly basis (we see the raw test results on these). The contract also provides for 15-minute response time on-call emergency service and repair/restoration in the event of hard drive or other equipment failures. We believe that similar arrangements, adjusted as necessary over time to support performance of the web application and manage costs, are adequate to ensure data integrity and access to live *Pleiades* content.

In addition, we will implement a preservation strategy for *Pleiades'* geographic content by putting processes and agreements in place to effect the regular alienation and backup of the comprehensive *Pleiades* export in CSV format in multiple external hands. These provisions will ensure long-term access even in the event of a catastrophic event involving our datacenter, the demise of the project, or changed priorities on the part of New York University. The Ancient World Mapping Center at the University of North Carolina at Chapel Hill already copies the *Pleiades* CSV export to its own servers daily, and will continue to do so during and after the period of performance. The Perseus Digital Library at Tufts University has pledged to download and maintain copies in a similar manner, alongside its backups of its own data, beginning in September 2015.[4] ISAW's digital programs team will continue to make quarterly deposits of the latest version of this file to NYU's Faculty Digital Archive (http://archive.nyu.edu), a formal, institutional archive operated by the NYU Library system that provides automated bit fixity testing and repair, geographically separated replication, and the assignment and maintenance of persistent identifiers for deposits. FDA deposits will be made with the addition of Dublin Core metadata conforming to NYU specifications.[5] We will make parallel deposits of the CSV export and of all our software releases to the Zenodo.org research output repository, which is funded by the European Union and managed by CERN. Zenodo.org will preserve and assign a Digital Object Identifier (DOI) to each. The Managing Editors are actively cultivating additional, complementary partnerships, likely to include both informal holding of copies and formal archival deposit.[6]

As described in the "Significance" section, *Pleiades* assigns stable identifiers to all objects of interest in its dataset and serializes these identifiers into stable HTTP URIs. As a matter of policy, the *Pleiades* Managing Editors are committed to the persistence of these URIs with the support of ISAW, which budgets for the on-going registration of the pleiades.stoa.org domain. As a hedge against longer-term technological changes that might render the resolution of HTTP URIs and Internet domain names obsolete, *Pleiades* embeds the full URIs in every dissemination and archival format, thereby enabling future computational or manual reconstruction of connections to other datasets.

[4] Crane letter.
[5] FDA URIs conform to the *Handle Specification:* http://www.handle.net/. Dublin Core: http://dublincore.org/.
[6] Discussions underway with the German Archaeological Institute and The Digital Archaeological Record (tDAR: http://www.tdar.org/).

VII. Data Management Plan

Scalar is hosted by a virtual machine running on VMWare ESX VMM software load-balanced over two dedicated HP ProLiant DL380p Gen8 servers, with all new equipment installed 2014. Administered by the University of Southern California's Dornsife College, Division of Information Technology, additional virtual machines can be set-up to host other projects, including Critical Commons, which is in the process of migrating to this environment. Scalar is part of a group of USC servers that share a 1-Gb/sec connection, although Scalar's stake of the shared port is variable depending on current and future needs.

Server software management is shared between ANVC staff with expertise in server software and members of USC Dornsife College IT staff who handle core operations (such as critical updates to the operating system and Apache web server). The Scalar software itself is reviewed weekly in a development environment by ANVC developers before updates are placed online, while its databases are likewise supervised for speed and reliability.
The two servers include an automated failover solution to ensure uptime. Also, protocols are in place to backup filesystem and database content both locally and to the Disaster Recovery Center at Arizona State University in Phoenix. While ANVC presently has a tacit agreement with USC Dornsife College IT for long-term support, in 2015 both parties will enter into a renewable Service Level Agreement (SLA), with one-year renewable terms, under a larger five- and ten-year Memorandum of Understanding (MOU).

Planning the future of digital landscapes beyond ten years is very difficult, of course, but in any case, ANVC and Scalar team PIs, plus our Partners in this ODH proposal, Hypothes.is and the University of California Press, are committed to pursuing all available means to ensure the permanent archiving and accessibility of all Scalar publications, by working closely with the presses, and digital librarians in the ANVC Alliance, at USC, and around the world.

6. Data Management Plan

Roles and Responsibilities

The implementation of the data management plan during the course of the grant will be the responsibility of Hannah Alpert-Abrams. Three parties are responsible for maintaining the data. Taylor Berg-Kirkpatrick at UC Berkeley will maintain an open-access version of the modified Ocular code. The Early Modern OCR Project (eMOP) at Texas A&M University, under the direction of Laura Mandell, Anton duPlessis, and Matt Christy, will be responsible for maintaining a second open-access copy of the Ocular code. The University of Texas Libraries will provide preservation services, as well as web access, for the transcriptions. This will be mediated by Ladd Hanson, Associate Director for Information Technology Architecture and Strategy, and Aaron Choate, Head of Technology Integration Services at UT Libraries.

Expected Data

Our project will produce data in three forms: software code, transcriptions, and language models. All data and transcriptions will be stored and backed up daily on GitHub, where they will be made publicly available via the web. They may be viewed as individual files, or downloaded as a zipped archive.

In addition, the software code will be preserved and made available on the personal website of Taylor Berg-Kirkpatrick, the developer of Ocular. It will also be made available to anyone using eMOP, the Early Modern OCR Project hosted by Texas A&M University. eMOP will maintain a copy of the software on its servers and provide access via the eMOP website and Github code respository. The transcriptions will be produced in the form of XML files using the ALTO standard. These files will be preserved by UT Libraries.

This project draws on data in the form of PDF scans of books from *Primeros Libros*, currently preserved and made available at the UT Libraries-hosted project website. Our OCR system uses language data collected from the open-access text repository Project Gutenberg, and from private collections. All language data is preserved as plain-text files and backed up daily on GitHub. Though some of the data in these models is proprietary, language models based on the data will be made available for out-of-the-box use with the OCR system.

Data Formats and Dissemination

Format: All transcriptions will be made available in their final format as text files, with markup and metadata encoded using ALTO, the Library of Congress XML format for OCR transcriptions. This will include language tags for all transcribed words, along with document-level metadata drawn from the *Primeros Libros* website. Transcriptions will also be published on the *Primeros Libros* website.

All data produced during the course of this project will be made freely and publicly available. The OCR code will be held under the GNU General Public License version 3, in accordance with the original Ocular code. The transcriptions, like all content associated with the *Primeros Libros* project, will be public domain.

Dissemination: All data stored on GitHub is made publicly available to anyone with web access, and does not require any form of registration or user account. The same is true for the *Primeros Libros* website.

Likewise, all code developed for and by eMOP and its workflow is available via Github for download and use.

The only proprietary data in this project are transcriptions of early modern Nahuatl documents made by scholars in the field. Though we cannot make this data available, our project will produce new digital corpora of Nahuatl that will be made freely available. We will also be able to provide a Nahuatl language model based on the proprietary data, which can be used to run our OCR system on Nahuatl documents.

Data Storage and Preservation of Access

All data stored on GitHub will be made available indefinitely, and can be maintained over time; it will also be available for user collaboration in the form of bug fixes and other features. The University of Texas Libraries has a long-term commitment to the preservation of all *Primeros Libros* content,, including the transcriptions produced as part of this project. The Initiative for Digital Humanities, Media, and Culture at Texas A&M University is committed to the long-term storage, maintenance, and upkeep of data related to the Early Modern OCR Project.

Following consultation with the University of Texas Libraries staff members, we plan on depositing the research data in the University of Texas Digital Repository (UTDR). We will submit the necessary metadata and other resources to make the data accessible for future users. The UTDR will preserve the data indefinitely and is committed to responsible and sustainable management of submitted works as well as associated descriptive and administrative metadata, by employing a strategy combining the following: nightly secure backups, storage media refreshment, file format migration (including possible migration to standard formats during submission), and assignment of a unique and persistent URL.

Data Management Plan

Roles and Responsibilities

During the duration of the grant, the co-PIs (John Wall, David Hill, and Yun Jing) will take primary responsibility for managing any data generated as part of the project and for hosting the project website. In consultation with NCSU Libraries staff, they will ensure that data is regularly backed-up to university servers. Once the grant is completed, the PIs will transfer responsibility for permanently archiving and managing the project data to the NCSU Libraries through the Libraries' standard records transfer process. The PIs will continue to be responsible for maintaining the project website.

Expected Data, Data Formats and Dissemination

The project will primarily generate two types of data: proprietary project files generated by the 3D modelling (Google Sketch-Up) and acoustic modelling (Open Source software, to be developed in the course of the Project) applications and the exported audio, video, and image files. Any proprietary files will be both published and archived in their original format. Exported files will be made available in the following formats:

- Images will be published in JPG format as well as the PSD (Adobe Photoshop) files from which they were generated. A TIF version of each image will be generated to serve as an archival copy.
- Audio files (both original recordings and processed versions) will be published as MP3 and as high-quality WAV files. The WAV versions will also be used as archival copies.
- Video files, such as fly-throughs of 3D models and promotional videos, will be linked to from the project website and made available through YouTube in any format offered by YouTube. The FFV1 format will be used for archival copies.

Any metadata, such as settings for the acoustic modelling software, will be stored in cloud-based Google Spreadsheets. At the project's conclusion, this information will be exported and archived as CSV (Comma Separated Value) text files. In addition, a Dublin Core XML document will be generated for each media file that will include this and other information about the file.

Any software developed as part of this project will be made publicly available on Github (http://github.com).

Added to this will be media files that were generated during the first NEH-funded stage of the Paul's Cross Project and that are available through the currently existing website (http://vpcp.chass.ncsu.edu/). Any new files will be published without access or usage restrictions through the project website as soon as they have been prepared for publication. It is expected that the final project will consist of five to six GB of data.

Period of Data Retention

During the duration of the grant, data will be stored both on local desktop computers and on university servers. Upon completion of the project, all generated data, including a snapshot of the project website

and a README file describing the relationship between the various components, will be transferred to the NCSU Libraries' Special Collections Research Center, where the NCSU Libraries will take responsibility for the permanent preservation of any project data.

Data Storage and Preservation of Access

Data will be stored on one of the Libraries' data servers, mirrored to a second server, and backed-up to tape at a separate facility on the NC State campus. Full data backups are done on a monthly basis, with incremental back-ups occurring in-between. In addition, the project will be considered for submission to the Libraries' instance of DuraCloud, a cloud-based digital preservation system (http://www.duracloud.org). Public access will be provided through the project's website.

Current Activities

Continuing to use the Virtual Paul’s Cross Project as a proof-of-concept project and to bring that project into line with current best practices, the PIs are now developing an archive of data for that project according to the plans and guidelines outlined here. Any learnings from doing so will be incorporated into the Data Management Plan for the Virtual St Paul’s Cathedral Project before its implementation.

Data Management Plan

Immigrant Stories will generate data as both video and text files, and the project already has measures in place to manage and store its data files in perpetuity. All digital stories are video files that participants must submit as either QuickTime or MPEG-4 files. Video files must be original and uncompressed, in accordance with best practices for long-term preservation. Text files include participants' scripts (which will be displayed as video transcripts in our online repositories), metadata (including country of origin, ethnicity, date of migration, date of story creation, and themes), and donation forms which authorize the IHRC Archives to preserve and share the data. All users participate in *Immigrant Stories* with the knowledge that their videos, transcripts, and metadata will be publicly available. (The fact that *Immigrant Stories* is a public platform is often a key part of the project's appeal.) As part of the donation form, users will explain the copyright permissions for media that they have included in their videos, verify that they have permission to use these materials, and affirm that they have included all necessary attributions in their videos. Users will donate their materials to the IHRCA, and the institution, not the Project Director, will preserve and manage all project data.

The online digital story creation and submission platform will preserve users' scripts, donation form, and metadata as individual PDF and XML files. These file formats are most suitable for long-term preservation and data management, including dissemination through the Minnesota Digital Library (MDL) and Digital Public Library of America (DPLA). All files will be reviewed by IHRCA staff to ensure that the files are functional and meet project guidelines. IHRCA staff with authorize University of Minnesota Libraries staff to transfer these files from the *Immigrant Stories* web application to both the University's preservation environment and the MDL discovery interface. We estimate that it will take 1-2 months for submissions to be reviewed, archived, and made discoverable via *Immigrant Stories*' online repository.

All *Immigrant Stories* data will be shared under a Creative Commons license which allows non-commercial reuse, with attributions, and forbids the distribution of derivatives of the data (CC BY-NC-ND 4.0).

All *Immigrant Stories* video, text, and metadata will be preserved and maintained by the University of Minnesota Libraries. The University of Minnesota Libraries' storage environment consists of data center class hardware provided by Sun Oracle. The hardware includes Storage Area Network (SAN), Network Attached Storage (NAS) and a modular tape library system. Hardware specifics include Sun ZFS Storage 7410, Sun ZFS Storage 7420, Sun Storage 2540 M2 Array, and a Sun SL500 modular tape system. The Libraries are currently adding the Sun 7420 and migrating to the Sun 2540. This movement is from an installation of Sun 35XX fiber channel equipment. As further backup measures, the Libraries keep a minimum of three copies of each file which are geographically separated and include off-site tape storage.

All University of Minnesota archives and special collections are part of the Libraries, including the IHRC Archives. The Libraries' Digital Preservation and Repository Technologies (DPRT) department, part of the Libraries' Data & Technology Division, provides contemporary and enduring access to digital objects under the collection stewardship of the University Libraries. The DPRT works with the University's Office of Information Technology (OIT) and other partners to apply professional digital preservation methodologies, standards, and technologies to Libraries collection through the delivery and support of robust high quality discovery, access, management, and preservation systems. The DPRT department already works with a range of data types, including images, audio, and video, and they assist in ongoing policy development concerning format specifications for preservation purposes.

The University Libraries have previous experience managing large collections of digital video files. They have existing commitments to manage and sustain upwards of 40TB of digital video related to the Libraries' performing arts archives. They have also committed to support the preservation and discovery of video files in numerous archival collections, including the

University Archives, Northwest Architectural Archives as well as the IHRCA. All of the digital media are managed and maintained in accordance with the Libraries' existing data management procedures.

4.5 Data Management Plan (DMP)

Data collection

The data that are used in this project comes in the form of standardized text in “Canonical ASCII Transliteration Format” (C-ATF).[1] The text itself is encoded in UTF8 and the original language transliterations are restricted to simple ASCII characters. This notation system has been in use for 15 years and because of its simplicity and high level of standardization, all research projects that use large quantities of cuneiform texts will base their work on the Cuneiform Digital Library Initiative (CDLI) database which hosts these texts, or will use a derivative of C-ATF. The ATF notation created by the CDLI is the widest-used standard in the field. In the case at hand, the project will use approximately 24,000 lines of text and their translation, which will be augmented and pre-processed for use as a training data set. Because the CDLI is a long-lasting initiative, there are already quality checks and versioning systems in place. Each time a change is saved in one of the texts, a backup copy of the previous version is saved in the database. There are also a number of tools in place which are used to verify the quality before commit, such as compliance to the C-ATF standard and to a list of preferred sign readings (each cuneiform sign can have more than one reading), depending on genre and time period.

Documentation and Metadata

As a companion endeavor to the Cuneiform Digital Library Initiative there exists a wiki[2] which documents all aspects of the CDLI in a range of articles on history, specific inscribed artifacts, and genres, as well as discussions of processes and data acquisition. On the CDLI website, there are also articles discussing the museum collections holding the physical documents[3] and also the terms of use of the data[4]. These tools will be used to help document the project and its outcomes. Individual texts in the CDLI have an Open Context ark number assigned. Moreover, we will collaborate with a French research group[5] for the alignment of the texts’ metadata with the CIDOC-CRM ontology. The linguistic and semantic information generated in the translation process and by information extraction will be linked with linguistic open linked vocabularies. The software created will be thoroughly commented and a github Jekyll website will serve as a documentation hub for each new software module.

Licensing

New software and derivative data generated by the project will be both released to the public domain by using the Creative Commons CC0 license “Public Domain Dedication” (CC01.0).[6]

Storage and Backup

During the research, GitHub will be used as a versioning system for the code base of the project. The sample text will also be joined to the code, exceptionally as all text of the CDLI is usually backed-up daily in SQL, text and disk image formats. This is in order to keep a controlled sample since the CDLI data change every day. The Center for Digital Humanities at the University of California in Los Angeles gives us technical support and external backups that increase the security and recoverability of the data in case of a problem. We also have a mirrors of the servers both through the Max Planck Institute for the History of Science, Berlin (MPIWG; and through them to the Max Planck Society’s persistent storage hub in Göttingen) and through the University of Oxford. Teams in Toronto and Frankfurt will each use a development server of which the relevant data will be periodically sent to the CDLI and the code on GitHub.

---

[1] <http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html>

[2] <http://cdli.ox.ac.uk/wiki/>

[3] Take for example the page of the British Museum <http://cdli.ucla.edu/collections/bm/bm.html>

[4]< http://cdli.ucla.edu/?q=terms-of-use>

[5] <http://triplestore.modyco.fr:8080/ModRef/>

[6] <https://creativecommons.org/publicdomain/zero/1.0/>

Preservation
By renewing periodically our agreements with the Center for Digital Humanities, the MPIWG-Berlin and the University of Oxford, we are convinced that the CDLI offers optimal storage security and web server longevity; CDLI is in fact a model of data persistence—the longest lived digital humanities project in the field of Assyriology, with its predecessor the Uruk Project at the Free University of Berlin now 26 years in existence. Since the new data produced answers to a need in the study and teaching of cuneiform cultures, its usage will only increase. For the eventuality of any risk to the preservation of the software or the data, we will put copies of our work in official repositories to maximize their preservation.

Data Sharing
The code and data produced by this project will be released in the public domain and we will encourage anyone to use, modify and reuse any of their components. Our code will be available on Github at all times, the data will be viewable and searchable on the CDLI website, it will also be accessible to download in full as an archive from the same website. There will be no intermediary between the user and the data, no account verification or login. With our strong communications plan, we expect that a large proportion of people who might be interested in our results will hear of us one way or another. Because the process of translation involves the internal tagging of the text, it is possible to leverage these text notations and export them into a variety of formats, and due to the high level of standardization of the data, it will be possible make them compatible with other projects like ORACC, but it will also be provided in an RDF edition compliant with Linked Open Data (LOD) principles. An important contribution of this LOD interface is facilitating interoperability from other philology portals for which LOD-compliant components are currently being devised,[7] as well as with museum collections in the LOD cloud[8], it also will help discoverability from search engines.

Responsibilities and Resources
Because the CDLI has been running for many years, we are fortunate most of the lab material is already in place. We are planning on updating the actual software and on building upon it to be able to host and serve the new data produced by the project, but the costs involved in these upgrading and maintenance tasks are comprised in the workload of those participating in the project. We are, for example, using a public Github code repository as opposed to paying for private repositories. The IT team of the MPIWG-Berlin, the Frankfurt Team, the Toronto Team and the Center for Digital Humanities (CDH) at UCLA will be responsible for maintenance and backup of their respective servers. Once the project is completed, the CDH and the Berlin mirror will maintain the CDLI server copies. Any translation generated by the project that attains our quality standards will be merged with the current CDLI data into their respective text entry in the database and thus available to view on the CDLI website at all times, and also downloadable in part or in total at all times. The translations and the various information extracted from the analyzed texts will also be available to consult on the web interface that will have developed to this effect.

[7] E.g., Homer Multitext <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/blackwell-smith/>, Perseus <http://dlib.nyu.edu/awdl/isaw/isaw-papers/7/almas-babeu-krohn/> and SAWS <http://www.ancientwisdoms.ac.uk/method/ontology/>

[8] E.g., the British Museum (<http://collection.britishmuseum.org/>

Data Management Plan

Assessment of existing data

Explanation of existing data sources used by the research project: Dig that Lick uses metadata from the J-DISC Online Jazz Discography provided by the Center for Jazz Studies at Columbia University (CU) in the City of New York (`http://jdisc.columbia.edu/`) as well as the MusicBrainz Database (`http://musicbrainz.org/`), DBpedia (`http://wiki.dbpedia.org`) and other Linked Open Data resources. We analyse audio recordings in the J-DISC collection at CU (see Letters of Commitment).
Analysis of the gaps identified between the currently available and required data for the research: The available data consists of audio recordings and discographic metadata; our interests lie with the relation between the two. In particular, we use methods from Music Information Retrieval to analyse the audio content in order to discover the information that can be garnered from automatic analysis of the recordings, and relate this to the available discographic, historical and geographic metadata and external background knowledge, in order to perform a data-driven study of the creation and spread of new musical forms.

Information on new data

Data produced or accessed by the research project: Dig that Lick will produce content-based metadata, in particular automatic transcriptions of melodic material from the audio recordings, as well as links between the various data sources used in the project. The internally used formats will consist of Sonic Visualiser project files and SQL database files and the data will be published in these formats and as Linked Open Data in a dialect of RDF.

Quality assurance of data

Procedures for quality assurance carried out on the data collected at the time of data collection, data entry, digitisation and data checking: One of the challenges of large-scale (Big Data) analysis is that it is not possible to check the correctness of automatic analysis outputs for the complete data set. Our methodology is to use existing manually created and checked transcriptions from the Jazzomat project (`http://jazzomat.hfm-weimar.de/`) in order to estimate the reliability of automatically generated data. Other automatic tests of data consistency will be performed, as well as manual checks of small random samples of the data, plus checking particular samples which give rise to the most interesting results. In addition, all provenance data will be saved and published with the data to remove any ambiguity about claims to data quality.

Backup and security of data

Data back-up procedures adopted to ensure the data and metadata are securely stored during the lifetime of the project: Data will be backed up using standard procedures in each of the partner institutions, as well as being mirrored across multiple institutions to ensure that no data will be lost during or after the project. Software will be developed using the Sound Software repository (`https://code.soundsoftware.ac.uk/`), which provides version control and redundant storage for source code.

Management and curation of data

Plans for management and curation of primary or third party data: Data will be deposited in a long-term institutional repository. QMUL provides up to 1TB of storage free of charge for its projects, using a DSPACE repository, with a guarantee of hosting the data for at least 10 years from its time of last access. We plan to publish the data as Linked Open Data using established ontologies such as the Music Ontology (`http://musicontology.com/`), in order to ensure that the data can be understood and re-used by others. The published data will contain provenance information, such as unique identifiers for source data, software details including

version numbers and parameter settings, plus details of methodology, assumptions made, and the formats and file types of the data.

Difficulties in data sharing and measures to overcome these

Obstacles to sharing data and measures to overcome these: In any project involving commercial audio recordings, it is not possible to share the recordings which we analyse. Metadata, on the other hand, is not subject to such restrictions. In order to make the outputs of *Dig that Lick* reusable and the results reproducible, we will publish links containing identifiers pointing to the analysed recordings, such as MusicBrainz IDs (MBIDs), or proprietary identifiers (e.g. YouTube URLs or Spotify identifiers). These identifiers will allow users and other researchers to access the audio recordings legally.

Consent, anonymisation and strategies to enable further re-use of data

Procedures to handle consent for data sharing for data obtained from human participants, and/or how to anonymise data, to make sure that data can be made available and accessible for future scientific research: We do not intend to gather any data directly from human participants.

Copyright and intellectual property ownership

Who will own the copyright and IPR of newly generated data: The data generated from *Dig that Lick* will be owned by the project partner who generated it. In the case of joint work, partners will own the data in equal shares, unless an agreement to the contrary is made in advance of data production. Data will be published using a Creative Commons Attribution (CC-BY) licence, in order to ensure that it can reused and extended freely.

Responsibilities

Responsibilities for data management within research teams at all partner institutions: Each partner will be responsible to comply with the data management requirements of their respective funding bodies. The Principle Investigator at each site will delegate this responsibility as appropriate to a member of their team. Data management will be a standing item on the agenda of PI meetings, to ensure that best practice is shared and followed throughout the project.

8. DATA MANAGEMENT PLAN
T-AP DiD "Responsible Terrorism Coverage (ResTeCo): A Global Comparative Analysis of News Coverage about Terrorism from 1945 to the Present"

This data management plan will be implemented and managed by PI/Althaus for data stored or analyzed at the Cline Center, by PI/Wessler for data stored or analyzed at Mannheim University, and by PI/van Atteveldt for data stored or analyzed at Vrije Universiteit Amsterdam. Overall responsibility for compliance with the data management plan will be overseen by PI/Althaus.

Types of Data and Software to Be Used, Produced and Distributed

Two types of data will be used in the proposed project: (1) copyrighted full-text news data, and (2) non-copyrightable metadata and extracted features derived from full-text records (examples include word frequency tables, lists of named entities appearing in news texts, sentiment scores, etc.). All copyrighted full-text news data that will be used in the proposed research are already in the physical control of collaboration team members, with all required permissions already secured. The proposed project will also develop software tools and algorithms that can be deployed by other researchers on other textual corpora to replicate the analyses generated by the proposed research activity. These tools will be publicly released and disseminated under a permissive open-source licenses such as the MIT license. As far as possible, individual modules will be published separately to enhance community participation. All code will be published on github or a comparable platform.

Raw Textual Data: Format and Content

The Cline Center text data includes over 85 million news articles, most of which are protected under US copyright law. All these articles have metadata associated with them, indicating news source, title, date, etc., as well as extracted features stored and available as part of this project, such as named entities and geocoded place references (see Appendix Table 3). Cline Center full-text news data is stored in MongoDB and is documented with metadata files stored in SOLR indices. Exportable metadata files and extracted features will be distributed in ASCII text, CSV, or JSON files. Metadata attributes drawn from the Dublin Core Metadata Initiative will be used to define the schemas associated with the data stores.

The Cline Center corpora are complemented by over 10 million Dutch news articles stored in AmCAT (only a subset of these articles will be used in the ResTeCo project), an open source text analysis infrastructure. Internally, text and metadata are stored in a PostgreSQL relational database and are indexed using an Elasticsearch cluster. All data is accessible through an API that can export to CSV, JSON, and other formats. Metadata is based on Dublin Core and includes a URL that points to the original source, which for the longitudinal Telegraaf links to the freely accessible Dutch Royal Library (KB) Delpher application that contains both the OCR'd text and a scan of the original newspaper page. The Mannheim data are saved in a relational MySQL database. They are saved in fulltext together with metadata such as the origin URL, the date the article was published and crawled, the article title and, if they exist, the article authors and categories. These data can be exported into a CSV or TXT/JSON format.

Metadata and Extracted Features Derived from Full-Text News Data

Exportable data for public release will include metadata elements that identify the specific source article as well as extracted features that were derived from the full-text article at the levels of documents, actors within documents, and statements within actors. All metadata will be published under a permissive license (CC-BY or comparable) and will be accessible through an API linked from the project web site. Moreover, all data will be published as Linked Open Data, allowing easy access and combination with other data sources. Finally, all metadata will be linked to the original textual data, both from the original source (e.g. URL for online news or open archives) and the textual data entry stored in the respective

institution as detailed above. As far as permissible, headlines and snippets will be accessible through the project web site API to enhance validation and qualitative understanding of the metadata.

Copyright limitations prevent the project PIs from redistributing or making publicly available the copyrighted news text from which metadata and extracted features are drawn. However, much of the original source texts are either directly available on the Internet (in the case of web-crawled news stories) or available to the research community through standard news aggregation vendors such as Nexis-Lexis or ProQuest Historical Newspapers. In order to accommodate the needs of researchers who want to validate the quality of extracted features against the original copyrighted text, the ResTeCo project is committed to publishing metadata adequate to tracking down the original source material in a large number of cases (e.g., URL, title, source publications, date of publication, etc.). Such metadata cannot itself be copyrighted, so distribution of this metadata at a level of detail that allows researchers to track down original source records on their own should satisfy the validation needs of most users.

Three categories of metadata and extracted features will be publicly distributed:

1. *Document-level metadata/extracted features*: (a) date of publication, source of publication, title, URL, etc.; (b) PETRARCH events derived from the document; (c) classifier output on relevance for containing information about terrorism; (d) topics and subtopics derived from LDA analysis; (e) presence of episodic / thematic framing elements; (f) named entities / referenced actors; (g) ingroup / outgroup cues.
2. *Additional actor-level metadata/extracted features within documents:* (a) originating document; (b) associated organizations / groups; (c) actor labeling (terrorist, freedom fighter, militant, insurgent, etc.).
3. *Additional statement-level metadata/extracted features within actors:* A semantic network analysis consisting of (a) originating document; (b) source of statement (actor, or journalist if none); (c) target of statement (actor); (d) evaluation (sentiment); (e) ingroup / outgroup cues; (f) statement topic derived from LDA analysis.

Provisions for Archiving, Preservation, and Distribution of Data and Software

Since legal restrictions prevent the copyrighted full-text news data from being publicly released by the research team, only the non-copyrightable metadata and extracted features will be publicly released and disseminated after the conclusion of the proposed research activity. All derived metadata and extracted features will be publicly distributed through the Illinois Data Bank (https://databank.illinois.edu/), which assigns DOIs to all data files and maintains a stable and policy-compliant environment for preservation and public distribution of a wide range of data forms for a minimum period of five years past the date of original publication. All software tools and algorithms developed for this proposed research will be publicly distributed via GitHub.

The Cline Center has been running and developing cyberinfrastructure for preservation of news data since 2006. The Cline Center data store is actively updated, managed and maintained by dedicated Cline Center staff on an ongoing basis. These activities are independent of the proposed budgeted activities and will continue long after the proposed research has been completed.

Gold Standard Data

The final data deliverable of this project is the gold standard data that will be used to validate the text analysis methods. These data will be published under a permissive license (CC-BY or comparable) As far as possible, these manual annotations will be conducted on publicly available source material so the raw text can be published or linked together with the annotations.

DATA MANAGEMENT PLAN

Raw Audio Data and Meta/Derivative Data

Given that the raw audio data for this project is intrinsically identifiable data (i.e. recordings of day-to-day activities from young children who wore audio recorders), extra steps are taken to ensure that these raw audio files are only available to authorized researchers, as specified by the data guardians, based on families' consent as regulated through each relevant institution's Research Ethics Board.

All members of the research teams working with raw audio data will undergo mandatory ethics training as dictated by their host institution (e.g. CITI Certificates in the US, CORE for Canada). Indeed, most PIs in our group already stipulate this lab-internally, and are well-versed with the infrastructure for compliance. All researchers on this project will be or are authorized HomeBank researchers; HomeBank has an approval process in place to facilitate ethical raw-audio and metadata access. The Argentina and LuCiD Corpora (not presently on HomeBank) will be shared over secure-server connections (sftp with unique logins) before being packaged into the virtual machines (see below).

Raw audio data will be stored on encrypted, password-protected machines kept in locked offices; each primary data guardian retains a copy of raw data in the home lab, ensuring mirrored archiving. All manual annotations will be backed up nightly over secured network connections maintained by Duke University IT (***Bergelson*** Lab).

Github, Virtual Machine, and OSF storage of Code, and Documentation

All code for each tool, and the manual and tool-derived annotations will be packaged into a 'child language module' via virtual machine platform, using the already-existing infrastructure developed by the Virtual Speech Kitchen (***Metze***). This will allow standardized computing environments across labs, and facilitate bug-fixes and version-control. All code will be shared both group-internally and with the research community, and will be written in either free or standalone formats (e.g. python, R).

As the tools team (***Dupoux, Metze, Räsänen, Rudzicz, Schuller***) improves their code, as the datasets team provides further annotations (***Bergelson, Soderstrom, Rosemberg, Cristia***), and as the group analyzes human- and machine-annotation for publication (**all PIs)**, each will "push" their work to a shared private github repository, to be "pulled" by the other groups. This is for intermediary control of training and testing data, and analyses, before code is ready for public use and feedback. At each "code release" landmark, and at the end of the project, code and annotations will be released publically through the project github page, linked to PI, lab, and HomeBank github code repositories for long-term storage (Spring, Fall 2018, Fall 2019, Spring 2020, see Appendix 3 & PMDC). Documentation of the developed annotation process will be maintained, stored and released longterm via the Open Science Framework.

Long-term Storage of Data

Once they reach acceptable levels for distribution, both hand-coded and automated annotations will be stored in the long term together with the original raw audio (i.e. in HomeBank and at LuCiD, in addition to the individual laboratories) so that these codes can be used for subsequent analyses by other researchers. A short-term moratorium on use of these derivative data (specified in HomeBank's dataset-specific fair-use policies) may be imposed until the main research findings have been published.

Results Dissemination

As described in the Data Dissemination section of the PMDC, results, workflows, and data will be made available to the research community and the general public via project blog, wiki, conference presentations, and open-access publications. We will take extra efforts beyond the requirements of publication to make our analysis code available via R Markdown scripts coupled with sharing of summarized de-identified data on the project website, so that other researchers can replicate our analyses directly, and apply them to their own data, which may not be sharable outright.

Our group is committed to open-source pipelines, for increased dissemination, replicability, reuse, and extension.

5 ISEBEL Data Management Plan (DMP) DMP.PDF

Given the close connection between the analytics and data storage and preservation aspects of ISEBEL, data management is integrated into the overall architecture of the project. The respective institutions (Meertens, UCLA, Rostock) have long-standing, institutional investment in the preservation, management and dissemination of the underlying data, particularly given the central role these collections play in the two main European institutions, while the UCLA data is managed through the UCLA digital library and deposited with the Danish Folklore Archives, at the Royal Library in Copenhagen.

The Meertens Instituut guarantees that the Dutch, Danish and German databases will be accessible online permanently. UCLA library also provides a consistent repository for the underlying data, and will continue to mirror the ISEBEL system for a period of at least ten years past the expiration of the challenge. Existing protections in place at the Meertens Instituut for the Dutch materials will be equally applied to the Danish and German materials for access to copyrighted material (newspaper clippings, press photos, the international catalogue of folktale types by Uther 2004), materials that are subject to the privacy protections of the EU (personal data about living narrators and collectors).

At University of Rostock, the Wossidlo archive (WossiDiA), is managed and curated by the university library. Sustainability is part of the so called “Rostocker Modell”, an agreement among the university library, the computer centre and the computer science department. The library runs the digital library and repository systems and provides persistent identifiers for published resources. The computer centre provides hardware and platform as a service and the computer science department incorporates new technologies, develops, evolves and migrates software solutions with external contractors/software companies on a project basis. The WossiDiA systems is also part of long-term preservation funded by the German Federal Office of Civil Protection and Disaster Assistance (BBK). This holds for the German database part of ISEBEL, too.

Access, Sharing and Re-use of Data
The researchers associated with this study are not aware of any reasons which might prohibit the sharing and re-use of the data beyond the standard provisions of the governing copyright regimes. There will be no additional restrictions or permissions required for accessing the data beyond those stipulated above. Findings will be published by the researchers based on this data. The data sources are not only made accessible as they are but will be enriched by a set of meta-data and according to the linked open data (LOD) guidelines.

Access, Sharing and Re-use of Code
All of the code for the project will be published on git.hub with a GNU-GPL license. R packages will be stored on CRAN. An overview of the software requirements and hardware configurations for ISEBEL will be available through the “about” page on the main project site. The hypergraph database systems used as a software platform for developing the integrator and hypergraph analysis component will be made available separately under GNU-GPL as an generic graph database system deployable apart from ISEBEL.

6. Data Management Plan

Responsibilities

Project Directors Caroline T. Schroeder and Heike Behlmer will oversee the data management plan, ensuring that all processes are implemented.

Expected Data, Collection Methods, Data Formats, and Data Dissemination

Digitized Text from Shared Corpora: The digitized text will be raw data in the form of text files, XML files, Microsoft Excel files, PAULA XML files in English, German, Greek Unicode and Coptic Unicode (using the Antinoou font created by the International Association of Coptic Studies.) We anticipate no changes to the Unicode standards for the Coptic character set that would affect our work.

The digitized text files, Excel files, html files, and draft XML files will be stored on version-controlled servers on Göttingen, Münster, and Georgetown Universities using Git or Subversion, as well as the Coptic SCRIPTORIUM GitHub site at http://www.github.com/CopticScriptorium. The INTF text files are already stored and versioned using a Git server. Some of the project text data will be drawn from ancient and medieval manuscripts. Under intellectual property law in Germany and the United States, the text from the manuscripts is in the public domain; editorial work can be under copyright. The project will not be publishing online editorial work under existing copyright. Other text data, such as the *TLA* lexicon, will be drawn from digital-native editorial work, for which the *TLA* will give project partners permission to distribute under an open source, Creative Commons attribution license (CC-BY). (See Dr. Frank Feder's letter in Appendix 10.) Prepublication files will be stored on these servers and GitHub. Digital publication of shared data occur on the Göttingen Coptic Old Testament project site, Coptic SCRIPTORIUM's site, and the INTF Virtual Manuscript Room.

Image files from the NT.VMR sample corpus of manuscripts of John's Apocalypse (the book of Revelation) manuscripts are free to view by written permission of the holding institutions and the transcriptions are made available under a Creative Commons license (BY-SA). These image files are available as imaged pages alongside with their electronic transcriptions.

Tools and Technologies: The digital tools to annotate and format the text files will be written in Java, Python, Perl, JavaScript or other scripting languages. The NT.VMR suite of tools and Coptic SCRIPTORIUM tools are already open-source, and the development of future tools during the project period will be open-source, as well. We will distribute the tools as free public downloads under open-source licenses, such as the Apache 2.0 license, or a GNU license. The software will be distributed on GitHub, with links on project partners' sites.

Documentation: The project will provide documentation of the tools and technologies develped on partner project websites. Documentation will be labeled with date and version information and disseminated under open-source licenses, such the Apache license, GNU Free Documentation License, and the Creative Commons Attribution (CC-BY) License.

Detailed documentation on data curation, methods, and standards for digital Coptic Studies will also be distributed in this manner. But in addition, project participants will present papers at the appropriate conferences (such as the International Association of Coptic Studies in 2016 and 2020) and will publish standards in relevant journals.

Period of Data Retention

The project participants embrace the principles of timely, rapid, and open-source data distribution. Tools, methodologies, and documentation will be released as they are created. The tools and standard developed during the project period will be stored on the Göttingen Coptic Old Testament servers and site (due to their long-term funding) and in the Georgetown University Institutional Repository (http://www.library.georgetown.edu/ir/policies).

Data Formats and Dissemination

The project participants embrace the principles of timely, rapid, and open-source data distribution. Tools, methodologies, and documentation will be released as they are created. The tools and standard developed during the project period will be stored on the Göttingen Coptic Old Testament servers and site (due to their long-term funding) and in the Georgetown University Institutional Repository (http://www.library.georgetown.edu/ir/policies). Georgetown University's Institutional Repository can hold files in .pdf format and datasets. Repository Staff will assist Faculty if the format of the data or publications require conversion.

Project participants will also disseminate their results in journal articles and at professional conferences and symposia. The most important of the latter events are the *Society of Biblical Literature* annual meetings, the quadrennial international congress for the International Association of Coptic Studies, and the annual international Digital Humanities conference.

"Modeling semantically Enriched Digital Edition of Accounts" (MEDEA)
NEH/DFG Bilateral Digital Humanities Program
September 2014

Data Management Plan

The bilateral MEDEA project will create data as test sets for the data models developed in the meetings. This data will be created by human experts in the fields of scholarly editing and economic and social history, who will transcribe from already existing images of original sources. Thus the leading questions, individual problems, and suggestions for solutions must be documented extensively during the data creation process. The project will store these kinds of comments on the transcription process, annotation and structuring of the data together with the created data, preferably in headers and as comments directly in the data.

Data and code created during the project will be managed locally under the responsibility of the individual contributor. These files will contain enough metadata to identify the sources from which the data is drawn (archival references) and the responsibilities of persons creating and modifying the data. The metadata for the test sets will preferably be organised following the guidelines for the TEI header section. During the publication process the project will create a representation of the metadata in the Dublin Core elements set.
Data created in other projects and reused in the project will be managed at the originating institutions. Suggestions for modifications will either be submitted to the originating institutions or added to the documentation of the project with the consent of the originating institution in accordance with their intellectual and other property rights.

The major format to store the data in the project will be XML. CSV may be used for tabular data considered on a case by case basis. RDF files can be stored as N3/Turtle triples.
In cases of complex objects integrating multiple files, the project will rely on METS.
The complete documentation of the project will be published at the http://encodinghfrs.org/ site maintained by Kathryn Tomasek and in the Humanities Asset Management System at Graz University (GAMS, http://gams.uni-graz.at, see Sustainability Plan). The identifiers created in the GAMS will be used as permanent identifiers.

The publication process will be managed by the principal investigators of the project Kathryn Tomasek, Mark Spoerer, and Georg Vogeler who will evaluate the scholarly quality of the data created in the project and submitted by external parties. For collaborative production of code and data the project will make use of git technologies and use github (http://github.com) as a repository where the information will be made public after quality assessment by the Project Directors. The Project Directors will be supported in their assessment of data and reports by the Advisory Board.

Data management plan

This data management plan was created on September 9, 2014, for submission to the Office of Digital Humanities (ODH), National Endowment for the Humanities as required by ODH Guidelines in the interest of securing funding for this project under the NEH/DFG Bilteral Digital Humanities Grant program. This is the first version of the data management plan associated with this data.

Roles and Responsibilities
US Project Director Hayim Lapin will be responsible for implementing this data management plan in consultation with Tal Ilan, the Germany Project Director. They will oversee regular collection and curation of data over the life of the grant period. At the end of the grant period, the Directors will deposit the data in the digital repositories of the University of Maryland and the Freie Universität Berlin. At the final data management meeting of the grant period, the Directors will finalize plans for the permanent deployment of a live version of the application developed.

Types of Data
This project will produce a data set comprising transcriptions of manuscript witnesses to classical Jewish texts along with contextual information describing these texts and their scholarly interpretation. These data will be created by trained transcribers employed by the project or by external scholars cooperating with the project. In addition, the project will generate three data tables in collecting (a) alignment data for Mishnah witnesses and (b) for Tosefta witnesses, and (c) synopsis data for the Mishnah and Tosefta. Along with the text and alignments, information about the provenance and condition of manuscript witnesses, previous editorial interventions will be captured during the course of this project.

The project will also generate or refine algorithms for matching strings in corpora, programs for aligning texts in two distinct works, and the software for managing the web application

Data and Metadata Formats
Project data will be stored, managed, and archived in a custom XML format compliant with the Text Encoding Initiative (TEI) P5 Guidelines, version 2.6, for encoding of electronic texts. TEI is the most-widely adopted standard for scholarly representation of texts in digital form. This format is platform-independent and open source and freely-available. TEI XML is suitable for long-term archiving and preservation in its current form; no transformations are required for preservation.

Standard bibliographic metadata about the digital files will be stored in the header of the TEI files along with information about the provenance of the manuscripts on which the transcriptions are based. Additional metadata that captures in detail important facts about the transmission of the text, such as damage or editorial intervention, will be represented in specialized tags intended for these purposes throughout the body of the text.

The project's customization of the standard is documented in a TEI ODD file. This enables generation of prose documentation for the project's tag set and schema files for technical validation. Derivative files in HTML format, which capture various presentational aspects of the data will be managed and archived along with the XML data. The data set will also include programs written in the eXtensible Stylesheet Language (XSLT) to allow all derivative files to be regenerated as needed.

Other types of data may be in the form of C#, javascript, XQuery, SQL, Java, or HTML.

Access and Dissemination
All data and software will be available through the project's Github repository, which allows other researchers to download and build upon the work of this project immediately. Presentational versions of the data will also be available through a website designed by the Maryland staff. All data from this project will be immediately shared under the terms of open licenses approved in consultation with the University of Maryland's Office of Technology Commercialization.

Data Storage and Backups During the Active Life of the Project
At least two types of copies of the data will be actively managed and stored during the life of the project. One copy will be stored on servers managed by GitHub. The subscription supporting the storage of this copy will be paid by the Project Directors. A second, local copy of the data will be stored on shared server space managed by the College of Arts and Humanities and service the Division of Information Technology at Maryland, which has a proven record of and commitment to secure data archiving for the University.

Long-Term Preservation
Within three years from the end of the grant period, a copy of the data will be permanently archived with the University Libraries at Maryland and at the Freie Universität Berlin. Copies of the data may also be deposited with other suitable repositories as identified by the Project Directors. Data will remain publicly available through the two principle repositories.

6. Data Management Plan

The programs resulting from this project will be registered as open-source projects at sourceforge.net, as well as the LREC language resources repository. The textual data will be stored in an extended version of the TalkBank format[1] that will also be compatible with the ANNIS3 framework. This will allow us to include the data in TalkBank, ANNIS, and other open repositories such as the Virtual Linguistic Observatory (VLO) or the Linguistic Data Consortium (LDC).

Consistency of data format will be achieved by use of the TalkBank *Chatter* program that checks XML data for validity against the schema and outputs text formats in the required display formats. To further guarantee validity, it creates a round trip in which text in the display format is reprocessed into XML and compared against the original to make sure that nothing has been altered during formatting and that all codes are accurately stored in the XML archive. TalkBank also periodically runs custom Unix-based tools for automatic checking of the overall database for annotation linkage, integrity of directory structure, proper harvesting of metadata, mirroring of data to other websites, and availability of resources. All methods are part of the process of obtaining the Data Seal of Approval for TalkBank described at http://talkbank.org/share/preservation.html and http://talkbank.org/share/workflow.html.

[1] http://talkbank.org/software/talkbank.xsd

6. Data Management Plan

Expected data:
The projects intermediate and final result will contain predominantly program source code and text-based documents (e.g. DOC, LaTex). Furthermore, any descriptive or technical meta-data produced based on the provided use-cases and objects are used for further analysis.

During the project all intermediate results and findings on a dedicated project management system (redmine) including a WIKI system for text-based content and a GIT versioning system for source code.

Period of data retention:
There will be no specific retention period for the project's output.

Data formats and dissemination:
Any final result (i.e. public deliverables) will be published on the projects web site on completion.
Source code will be maintained will be publish for public access using the public github.org service.

Data storage and preservation of access:
All textual output will be publish at *e-LIS* an international digital repository for library and Information science (LIS) and the University of Freiburgs Freidok repository.

Data Management Plan

Products of the Research:
All the data used for this project comes from newspapers and periodicals published in the era 1889 to 1893. All of the newspapers are either in the public domain because of their date of publication (prior to 1922) or they are available to researchers affiliated with Virginia Tech through subscriptions through the University Libraries. The articles used for this research are available in one or more of the following formats and databases:

Digitized newspapers, available through a public database:
- Chronicling America, hosted by the US Library of Congress
- State Library of Berlin
- University of Bonn
- Bavarian State Library
- Austrian National Library (ANNO)
- University Library of Freiburg

Digitized newspapers in a subscription database, from VT University Libraries:
- Proquest Historical Newspapers (select titles)
- America's Historical Newspapers

Digitized newspapers in subscription databases, but **not** from VT University Libraries:
- Proquest Historical Newspapers (Library of Congress or other libraries)

Digitized medical periodicals:
- HathiTrust, consortium of university libraries, including Virginia Tech
- Medical Heritage Library, Wellcome Trust and National Library of Medicine
- Internet Archive, publicly accessible library materials

Periodicals from the Medical Heritage Library allow access to public domain digitized collections through the Internet Archive. Periodicals from the Hathi Trust are available to Virginia Tech researchers through the partnership agreement. Separate agreements to allow access to the German scholars from Hannover University will be negotiated as needed.

Data Formats:
All of the research data for this project began as printed editions of newspapers or periodicals. In all cases, newspapers were filmed for preservation as microfilm. Newspapers that have been digitized are already available in pdf formats. Periodicals were scanned from the bound copies held by university libraries. Articles will be saved in pdf format. Databases that allow readers to access the OCR text make it possible to save articles as text files. The saved articles will be stored on a project site, using Scholar, an open source software adopted at Virginia Tech for instructional and research uses, or the university's licensed google folders. Only registered users of the site will have access to these materials during the research process. Data that is already publicly available, such as articles from the Chronicling America collection, can be made available to the public by linking to these online versions. Articles from newspapers in subscription databases can be made publicly available as needed for research purposes. The estimated amount of data secured through these methods will be less than 200 gigabytes.

Access to Data and Data Sharing Practices and Policies:
All data for this project originated in the public domain, and therefore confidentiality, privacy, security, and intellectual property issues are not relevant. In the case of materials available through subscription, data sharing arrangements will be developed in consultation with University Libraries and the vendors, on

terms consistent with the subscription and licensing agreements already in place. Source code for algorithms and format converters, plus preprints of papers, presentations, technical reports, and educational material will be posted on a dedicated project website, which will be updated regularly during the project. Preprints will be posted soon after acceptance; software will be posted after successful testing in trial version. The software will be provided as-is, requiring only proper acknowledgment; limited technical support will also be provided by the co-PIs and their students to qualified research groups, at no cost. Resources needed for web page creation and maintenance are minimal and readily available. Students and PIs have the necessary experience, since they design and maintain their own webpages.

Policies for Re-Use, Re-Distribution, and Production of Derivatives.
This research will generate derivatives in forms appropriate to the fields of humanities and computer sciences, including publications in scholarly journals, online research updates, and online postings. Access to these research materials will take forms appropriate to their form, including subscriptions to journals, the book will be available online, in libraries, and for purchase, and online postings will be freely available. That said, we believe in no-cost access to publicly-funded research data, software, and preprints, and we will go the extra mile to ensure the widest possible dissemination of our research results and the means to (re)produce them. Following completion of the project, we will consider offering a mature version of the code under a GNU or Apache license.

Archiving of Data:
Data will be archived in the Scholar site, hosted by Virginia Tech Information Technology services. Additional copies will be preserved by University Libraries. Virginia Tech Libraries is a member of MetaArchive Cooperative, a digital preservation consortium that uses a LOCKSS (Lots of Copies Keeps Stuff Safe) preservation strategy. Archival copies of files are contributed to a secure, closed-access network of dark archive servers set up between the MetaArchive Cooperative's institutional members. All servers are stored in different geographic locations and maintained by different systems administrators. When content is added to the institutional repository, VTechWorks, it is visited by seven of the network's servers, each of which replicates and preserves a copy. Servers are selected and assigned to content on the basis of their widespread geographical location. All seven servers revisit the content source on a regular basis to find content that has been added. The seven servers also check in with each other regularly to make sure that all copies are identical. If a mismatch is detected, the servers come to quorum regarding which copies are correct and which do not match, and then the network repairs the files. Repaired files are stored alongside the originals so that no file version is ever lost/replaced within the system. Content in MetaArchive is regularly migrated to new storage media in order to maintain its integrity. All of the material produced by this project (including source code, documentation, preprints, technical reports and other non-copyrighted publications) will be preserved for at least three years beyond the end date.

Data management plan

Expected data

Type of Data	Format	When shared?	Location
Open source computer code associated with interactive, embeddable site extensions or widgets	HTML, CSS, PHP	At the conclusion of the project.	GitHub and Graduate Center institutional repository
Platform content	HTML, CSS, PHP	At the conclusion of the project.	Internet and Graduate Center institutional repository
White paper	PDF	At the conclusion of the project.	NEH website and Graduate Center institutional repository
Final Report to NEH	PDF	At the conclusion of the project. It may be necessary to redact some financial information from the report for privacy reasons.	Graduate Center institutional repository

Period of Data Retention
Data and reports will be publicly available within one year of project completion, if not sooner. Data will be retained for a minimum of five years beyond the completion of the start-up phase.

Roles and Responsibilities
At the end of the grant period, the Co-Project Director will assist the Project Director with depositing copies of the data in the Graduate Center's institutional repository, *Academic Works*, and any additional repositories designated by the project. She will also create contextual information to be deposited including information about the provenance and condition of the data.

Data Formats and Dissemination
All data, including code and website content (unless otherwise specified on the Beyond Citation) will be made available under a Creative Commons B-NC-SA license. Code for the interactive, embeddable site extensions or widgets will be available on GitHub, a publicly accessible code repository, as well as *Academic Works*.

Data Management and Maintenance
Code will be maintained on GitHub by members of the Commons-in-a-Box programming team. Long-term management and maintenance of data deposited with *Academic Works* is the responsibility of staff employed by the Graduate Center Library.

CUNY Graduate Center institutional repository
The Graduate Center's new repository is *Academic Works* (http://works.gc.cuny.edu) which is administered by the Graduate Center Library. As a bepress repository, *Academic Works* can accept any type of file, including data files, without a size limit. Bepress repositories are hosted on Digital Commons network servers with service and support from bepress and are easily discoverable on Google and Google Scholar. The Graduate Center Library is committed to making *Academic Works* a central and sustainable service. In addition, CUNY is making a major institutional commitment to a CUNY-wide repository, which will be overseen by a newly hired scholarly communications librarian. The GC repository and future CUNY-wide repository will be included in search results on CUNY OneSearch, CUNY libraries' soon-to-launch discovery tool (from Primo).

THE SEARCH FOR HARMONY

Data Management Plan

Expected Data

Type of Data	**When Shared**	**Under What Conditions?**
Open-source computer code including Javascript, PHP, HTML and CSS, including all algorithms and design templates.	At the conclusion of the start-up grant period, following quality assurance testing. Additional versions as project progresses.	Code will be freely available.
A playable, web-accessible version of *The Search for Harmony* proof-of-concept, which demos the tool.	At the conclusion of the start-up grant period, following quality assurance. Additional versions as project progresses.	Website will be freely accessible.
Sample database and WordPress installation for *The Search for Harmony* to be used as test data for end users.	At the conclusion of the start-up grant period.	Freely available as a GitHub code repository linked to the primary tool.
Exported event-driven Google Analytics data for impact metric analysis.	At the conclusion of the start-up grant period.	Included as a report with code and linked to from any ancillary media, e.g. Stone Soup website.
White Paper	After the project has been completed.	Freely available on the Stone Soup website.
Progress reports including multimedia.	From the time of writing, throughout the duration of the project.	Freely available on the Stone Soup website.
Documentation of reusable tool.	At the conclusion of the start-up grant period. Additional versions as project progresses.	Freely available, included with the GitHub code repository.
Final Report to NEH	After the project has been completed.	Dissemination will be the responsibility of the NEH.

Data Formats and Dissemination

Two components comprise the bulk of data generated during the course of this Start-Up Grant Period: Web-based computer code for the WordPress plugin that is being developed to facilitate game creation in the humanities, and sample code in the form of *The Search for Harmony* proof-of-concept level. This plugin and the demo will be shared on GitHub. The plugin will also be shared in the WordPress Plugin Repository, and will adhere to WordPress coding best practices.

Also included will be plain text and markdown documentation on the plugin and how it can be used, and information related to its implementation in *The Search for Harmony.* The project itself will be freely available for download so it can be duplicated and studied as a use case.

Progress reports and the White Paper will also be made available in PDF format, downloadable through the Stone Soup website.

Analytics data will be collected using Google Analytics' event-based Javascript code, which will be used to measure engagement and user progress anonymously for research that will foster development of the game. Once the project has reached its conclusion, this anonymous analytics data will be made freely available as a sample data set for future humanities media makers interested in measurement methods related to web-based games.

The playable version of *The Search for Harmony* will require a server capable of running WordPress, which is a common requirement for most websites.

Data Storage and Preservation of Access

All materials will be hosted using existing code repositories such as GitHub, extending the duration of their access without incurring additional cost. GitHub allows for free repositories that are open-source.

Period of Data Retention

Data will be retained for a minimum of five years, although no effort will be made to remove the code from GitHub or the WordPress Plugin repository following this period. The Lead Developer owns cloud servers and will support the playable version for a minimum of five years, removing the normal cost of hosting from long-term storage.

Ancient Graffiti Project (AGP): Data Management Plan

Expected data

We expect to produce several types of data:

- A graffiti-specific metadata schema
- A custom vocabulary for describing figural graffiti, published according to SKOS specifications
- An openly available database of graffiti in Herculaneum that adheres to the metadata schema and applies the controlled vocabulary
- An open RESTful API for searching for graffiti by all metadata fields
- The source code for a web application interface to search for graffiti using the search API and visualize the graffiti locations at multiple geo-spatial levels

Period of data retention

The graffiti metadata will be published in, i.e., accessible from, the AGP and EDR and EAGLE web interfaces and APIs. As we complete our fieldwork, we will improve the entries (e.g., adding a graffito's letter height). We favor partial entries over not publishing. By the end of the grant period, the metadata of the graffiti found in Herculaneum will be complete.

We will release the source code for each component of our platform after we have completed a version of the component. For example, we will release the code that drives the RESTful API when we have implemented and thoroughly tested searching by location. Again, we favor frequent releases with only partial functionality implemented rather than waiting to release. At the end of the grant period, we will release our fully functional code, with documentation on how to use the code.

Data formats and dissemination

Our data will be disseminated in common formats and will adhere to best practices. The graffiti metadata schema will be published on AGP's web site. The graffiti metadata will be available from our open API in JSON as well as from the web interface. Since Cultural Heritage belongs to Italian State, EDR will host photographs of all inscriptions in accordance with EDR's memorandum of understanding with the Ministero per i Beni e le Attività Culturali (Ministry of Culture), accordo Mibac-EAGLE, dated 21 November 2005. The vocabulary to describe figural graffiti will be published according to SKOS specifications. The source code will be released under the Creative Commons license Attribution-NonCommercial-ShareAlike 2.5.

Data storage and preservation of access

AGP's Graffiti Metadata
Our graffiti metadata database is running on an Ubuntu virtual server. By using a virtual server, we can more easily upgrade the server's capabilities, thus providing better service to those accessing the data.

W&L's Information Technology Services (ITS) will provide resilient, high-speed storage, secure access, and regular backups to the graffiti data, which will be stored on mirrored systems in two data centers. The Richard A. Peterson Center, the primary data repository, is a state-of-the-art facility utilizing current best practice technologies to monitor and manage the equipment. Redundant HVAC, multiple power distribution units, uninterruptable power supply, automatic transfer switch, and generator provide a high

level of resiliency for the supported equipment and data. Regular 24-hour monitoring of equipment and services provided in this data center are also in place. The backup data center in Wilson Hall provides similar capabilities.

ITS uses Symantec Backup Exec 2010 to manage data backups using a direct to tape, monthly full backup/weekly incremental tape rotation. Tapes will be stored off site for archive purposes. Monthly tapes will be available for 12 months. After that period, an annual archive will be the primary restore point. Off-site storage of back-up data includes direct-to-disk, direct-to-tape and disk-to-tape duplications. Backup rotations vary depending on the data type and determined need for retention period, recoverability and change volatility. We will continually work with ITS to determine which backup rotation is most appropriate for the generated data. In addition to the backups at W&L, some of the graffiti data will be either housed (e.g., images) or duplicated (e.g., id number, find spot) on EDR's servers.

W&L's ITS uses Vmware ESXi on HP blade servers to support the virtual server environment. Each virtual machine is able to run on any of eight different blade servers in the separate data centers. Storage systems in use are Tegile hybrid storage arrays. These systems use both solid state drives and spinning disks to provide high input/output operations per second. Each storage system uses dual/redundant controllers to provide access to the data. Daily backups of the virtual machines are run and stored to separate storage systems. Those backups are then moved to tape for longer term, off-site storage.

AGP's Source Code

The source code that drives the web interface and open API will be released periodically on GitHub: https://github.com/AncientGraffitiProject/AGP . In addition, we are using local Subversion repositories of the source code.

Data Management Plan

Project Objective:

The Black Book Interactive Project (BBIP) proposes to create, develop, and integrate a metadata schema that allows for access and discovery of African American literary texts, making the archive more useful for research and scholarship. BBIP will use approximately 75 novels from the Project on the History of Black Writing digital archive as our model texts to produce: (1) a metadata schema that fully accounts for race in a digital context; (2) enhanced access to little known African American texts; and (3) interdisciplinary discussions among professionals to foster a greater understanding of text analysis and related computational research in African American literature. Our goal is to develop and foster a "best practices" schema that is sensitive to race and to establish a relational coordination between the producers of content (humanities researchers and teachers), academic librarians, and African American archivists.

Roles and responsibilities:

This data management plan will be implemented and managed by the Project manager (Will Cunningham), assisted by two Graduate Student GRAs, under the project supervision of Maryemma Graham, PhD. The Project manager will assist with all phases of the work, transferring final project artifacts and data to the University of Kansas (KU) institutional repository KU ScholarWorks. KU Libraries will have long-term responsibility for the permanent storage needs of the data.

This project will continue to work in partnership with two entities: (1) The **Chicago Text Lab (CTL)**, led by Hoyt Long, PhD and Richard Jean So, PhD. The CTL group explores the quantitative and computational methods for conducting macro-scale comparative inquires that have a direct correlation to the work we are doing. The partnership has already supported the full digitization of HBW's novel collection and will house the collection. (2) **The Institute for Digital Research (IDHR – KU Libraries)**, Brian Rosenblum and Erik Radio. Erik is the Metadata Librarian at the University of Kansas. His work in the libraries and with BBIP focuses on curating data and enhancing its discoverability. Brian Rosenblum, co-director of IDHR and associate librarian for Digital Scholarship, will assist and support BBIP in areas of access, production, and technical support.

Finally, Amy Earhart, PhD, who co-founded the Institute of Digital Humanities, Culture, and Media at Texas A&M, and teaches African American literature and the Digital Humanities, will serve as a special digital consultant. She has written extensively on matters of race in DH and most recently published *Traces of the Old, Uses of the New: The Emergence of the Digital Humanities* (2015).

Expected Data:

The metadata schema is intended to promote access to and analysis of African American literature. There are 5 levels of expected data:

- Academic papers or reports
- Project Reports
- Documentation of schema creation
- Digitized texts
- Metadata schema and records

Currently, data is stored on a networked hard drive. Project files and artifacts will be moved from the KU campus networked folder. Project artifacts will be stored on KU campus servers, as will any reports documenting the text analysis and project processes. At the project's conclusions, any finalized project files will be converted to PDFs and saved on the KU ScholarWorks server.

Period of Data Retention:

All final versions of the data resulting from the project (artifacts) will be deposited in KU ScholarWorks for perpetual storage and preservation upon completion of the project study. Once data is transferred to ScholarWorks, all data will immediately be made publically available.

Data formats and dissemination

Data Type:	Data Format:
Academic Papers or Reports	PDF, DOC, RTF
Project Reports	PDF, DOC, RTF
Documentation of Schema Creation	PDF, RTF
Digitized Texts	PDF, TXT
Metadata Schema and Records	XSD, XML

Through its commitment to long-term preservation of scholarly output, KU Libraries managed ScholarWorks Repository will allow easy sharing and accessibility. This repository contains scholarly work created by KU faculty, staff and students, as well as material from the University Archives. As an open access repository, ScholarWorks gives access to a wide audience and ensures its preservation.

Data storage and preservation of access

All public data will be deposited in KU ScholarWorks, which manages, archives, and shares digital content. ScholarWorks allows access to the public via persistent URLs, provides tools for long-term data management, and permits permanent storage options. KU Libraries has built-in contingencies for disaster recovery including redundancy and recovery plans.

Additionally, data not designed for open access storage will be stored in the University's Research File Storage (RFS). This data is accessible both within the KU network and remotely via VPN. Data is RAID protected and backed up on a rolling 30 day period.

HBW's Digital Novel Archive will be store both at the University of Chicago and KU's RFS.

Data Management Plan

Types of Data. The project manages and produces the following types of data:

a. Unstructured text data for "One Book" literary works,
b. Circulation and holdings records from the CPL ILS system,
c. Chicago-area demographic data extracted from census records,
d. Software for text analysis, predictive modeling, and interactive tools,
e. Features, models, visualizations, and other outputs from analysis and modeling activities,
f. Social media data (from Twitter, Facebook, and other sites) as accessed through the public APIs of these services,
g. Textual project updates, meeting notes, and final report(s),
h. Physical materials from Chicago Public Library (loose papers, CDs, still picture files) and reproductions of the same.

Data Sharing/Retention/Dissemination. All applicable electronic materials produced will be made available to the public through DePaul University's open access repository system, Via Sapientiae. (Via Sapientiae is a secure, permanent, read-only repository freely available to all at http://via.library.depaul.edu.) Appropriate metadata for project data will be created according to international cataloging standards and loaded into shared catalogs such as OCLC's WorldCat (a shared catalog containing the holdings of over 10,000 libraries), in order to facilitate discovery of the data sets by researchers worldwide. Data will be made available only after appropriate steps have been taken to protect intellectual property. Confidential material will be handled according to policies and protocols for human subjects, FERPA, and any other applicable regulations and restrictions.

The source code for our analyses, models and tools will be made available as open-source software and distributed using GitHub, thus making it available to the broader research community.

Our use of in-copyright works is confined to non-consumptive processing of the full text of each work within the secure HathiTrust capsule. The project does not have to manage these copyright-protected works. Descriptive features extracted from each text within the computing capsule can be released to our project team after they are vetted by HathiTrust to ensure that they meet fair-use conditions.

When necessary, the project team intends to secure permission (free or paid) to reproduce OBOC-associated texts or images for use in published articles, lectures, presentations at professional conferences, and classroom instruction. However, depending on the content, its owner, and the agreement governing each work, the rights to use content may not extend to other researchers or to follow-on or derivative research by others. Materials will be disseminated only to the extent that content owners will allow, and reproduction or reuse may be restricted.

Data Formats. It is anticipated that data will be in text and numerical form (.txt or other common machine-readable formats). Specific file formats and encoding standards are continually determined in relation to the selection of tools and ongoing needs of DePaul and Chicago Public Library.

Data Curation Plan

The data that will be produced as a result of this project will include:

1. **The open source code for the prototype lighting design emulator**

The code for the prototype lighting design emulator (likely written for the free-to-use Unity development platform) and the data conversion tools will be published on a public GitHub and made freely available to anyone who wants to use it for as long as GitHub, Inc. continues to host it under their current terms of service. The code will be public throughout the development process as we will be working entirely in public.

2. **The initial specification from the first meeting**

The list of properties of a design specified by the lighting designers and theater historians as significant and important to reproduce in an emulation will be recorded and published on NYPL's blog within one month of the first meeting.

3. **Work produced at the final hackathon**

At the public hackathon, interested members of the public will work with the code and the findings of the team to extend the work of the project over two days of coding and brainstorming. We will strongly encourage all of the groups to commit their code to a public repository (GitHub, Dropbox, etc.), and we will index all of these repositories in the final report.

4. **The final report**

As required by the NEH, we will write a report describing all of our findings which will be published on the NEH servers. We will also promote the report using NYPL blogs and social media.

Preservation

As long-term digital preservation is very expensive, and resources are scarce, libraries and archives must be judicious in their choices of what is accessioned for long-term preservation. We may find that the results of this feasibility study merit such long-term preservation, but we will dedicate our immediate efforts towards ensuring that the content is known and available to as many people as possible over the years immediately following the conclusion of the project.

Data management plan

The *TourSites for WordPress* project will generate four primary types of data: Multimedia tour content in the form of still images, video and audio; code; support documentation; and evaluation data. These will be produced, maintained, stored and shared in various ways, however in all cases the Project Team will work to ensure high standards of development and production value are upheld and that the produced data created for public use is maintained and unencumbered by legal, ethical or technological considerations.

Digital assets

Given the goal of producing media-rich tour experiences for evaluating the project, a large volume of raw data will be created by this project. All video content will be recorded in 1080p resolution to SDHC card media during production. Audio-only content will be recorded as WAV files to SDHC card media. Still images will be recorded to SDHC card media in a Camera RAW format or, in cases where existing two-dimensional materials are to be included, they will be scanned into an uncompressed TIFF format at a minimum of 600dpi. Once created, this raw multimedia material will be copied to hard drives in two locations. One will be used for post-production and the other will serve as an off-site backup for redundancy.

Video will be output in HD resolution (1920x1080 at 30 progressive frames per second) and uploaded to YouTube. Audio files will be output as MP3 files at 44.1kHz and uploaded to a cloud hosting solution, such as SoundCloud. Still images will either be uploaded into the WordPress content library or, in the case of accessioned digital content, uploaded as individual records in the Ohio Memory digital library and incorporated into the project via the CONTENTdm shortcode plugin.

During the grant period, all of these digital assets and their associated metadata will be maintained on redundant storage arrays in two physical locations. Once the project is published and the grant period has ended, all digital assets will be archived on a secure, redundant storage array at OHC for posterity. The digital assets will be available to the public through the free tour experiences and will also be used to promote the project through print, electronic and social media channels.

Code

The project will use GitHub to create, edit and share code between Project Team members during the grant period. Once the grant period has ended and the TourSite and CONTENTdm plugins have been finalized, they will be shared with the larger community in a GitHub repository as open-source projects. Copies of the code will also be considered a digital asset and stored on a redundant storage array at OHC for posterity.

Support documentation

Given that this project is intended to create a new opportunity for cultural heritage of all sizes, it is imperative that this information is shared with as wide a contingent of humanities professionals as possible. Data on workflows, methods and best practices will be compiled throughout the grant period as the Project Team decides what methods allow for the most effective use of the platform to deliver engaging, media-rich tour experiences. After the user experience has been evaluated and any changes made to the codebase, the Project Team will

generate supporting documentation for institutions to adopt the plugins for their own use. Upon completion of the grant period, this data will be shared through several methods:

- Final grant report
- Conference and workshop presentations
- Articles and blog posts on the digital tour experiences
- Personal conversations

In all cases, the Project Team will refer interested parties to the plugins and support documentation to enable the ready adoption of *TourSites for WordPress* by museums and other institutions.

Evaluation data

The evaluations performed as part of this project will be aligned with current best practices in the field of visitor studies and human research. Informed consent will be required for participation in the evaluations, no names or sensitive personal information will be collected, and data will be aggregated into a quantitative analysis of the usability and success of the project's development goals.

The analyzed results of this data will only be shared outside of OHC and the Project Team in their final, aggregated form to conform to the ethics of the field. These results will be included in presentations on the project through:

- Final grant report
- Conference and workshop presentations
- Articles and blog posts on digital tour experiences
- Personal conversations

Given the innovative nature of this project and the need it will meet, we expect these results to be of great use to the digital humanities field. Mr. Pierce will complete a final report on the project immediately following the grant period and begin sharing the results of the project at that time.

7. DATA MANAGEMENT PLAN

Data to be Generated

Type and Format of Data	*When Shared?*	*Under what conditions?*
Open Source computer code associated with the prototype module under construction	At the conclusion of the start-up phase	Code will be freely available
Textual and chronological data manually extracted from Priestley's chart in Rich Text Format (.rtf) and .csv	At the conclusion of the start-up phase	Data will be freely available
Images of historical objects in archival format provided by originating library or TIFF v.6 format if originated locally	At the conclusion of the start-up phase	Viewable on the Web, some are proprietary
Original historical research and analysis in Rich Text Format (.rtf)	At the conclusion of the start-up phase	Readable on the Web, Creative Commons license
Meeting notes and other relevant project records in Rich Text Format (.rtf)	At the conclusion of the start-up phase	Scholars' Bank, Creative Commons license
Final "lessons learned" white paper in Rich Text Format (.rtf)	At the conclusion of the start-up phase	Freely available through NEH
Alpha web module for Priestley in HTML5 and JavaScript	At the conclusion of the start-up phase	Available online through uoregon.edu domain

Data Management and Maintenance

Computer code, including HTML5 and JavaScript, and data extracted from historical sources including charts and books for **Time Online** will be open source and public domain. All code and data will be available online on the **Time Online** website in the uoregon.edu domain upon completion of the start-up phase. Additionally, code and data will be deposited in Scholars' Bank, the University of Oregon's Institutional Repository, and appropriate GitHub locations, so that they will be freely accessible to anyone and preserved long-term by the UO Libraries. Both the **Time Online** website and the Scholars' Bank will be open to crawling by the Internet Archive. Additionally, the uoregon.edu domain, where the project will reside, is regularly preserved by University Archives using Archive-It.

During development, **Time Online** will rely principally on computer systems administered by the InfoGraphics Lab, and hosted by the College of Arts and Sciences IT division (CAS IT). The Lab maintains an "always on" enterprise-level ESRI ArcSDE Geodatabase, ArcGIS Server, web servers, and file server to manage the University of Oregon Facilities GIS Infrastructure for campus mapping, emergency management, and facilities maintenance, etc. This system also houses and serves all of the geographic and historical data for previous

digital projects treating the cartography of Rome and other research in Stanford's Spatial History Lab. Data may be input simultaneously by multiple users from any location over a secure connection, and several levels of user rights management may be implemented thus facilitating remote collaboration between our venues at the University of Oregon and Stanford University. The systems of the InfoGraphics Lab and CAS IT Data is snapshotted twice daily for instant recovery. Data is sent offsite once per month for backup.

Meeting notes and other relevant records will be preserved by the Principal Investigator and redundantly backed up to a local hard drive, the .uoregon server, and dropbox.com. At the end of the start-up phase, these records will be saved to the open-access University of Scholars' Bank, as will the final report to NEH, which will also be available through NEH.

Images of historical artifacts presented in Time Online will remain property of the institutions (libraries, museums, private collections) granting rights for reproduction and publication, and access to these images will be determined by the policies of the collaborating libraries. In cases where full download is permissible, our system will permit it. Original physical artifacts (books, charts, and so forth) will be maintained by their proprietors (mostly university libraries). Digital reproductions will be preserved according to the protocols listed above.

Time Online will produce a final report for NEH at the end of its start-up phase. This will be published and preserved via Scholars' Bank and the NEH itself. Peer-reviewed research emanating from the project will be published through standard scholarly journals. Any work produced as a result of this NEH grant will remain in the public domain regardless of future publication venues.

Period of Data Retention

An alpha version of the start-up module described in this application will be made available online in the formats and locations discussed above at the end of the start-up phase. Data not available through the Time Online website itself will be stored and retained for at least 5 years beyond the completion of the start-up phase in the uoregon.edu domain and in the University of Oregon Scholars Bank.

7. Data Management Plan

A. Roles and responsibilities

This data management plan will be implemented and managed by Benjamin Brochstein, with consultation of Chad Shaw. The Rice Digital Scholarship Archive (RDSA, http://scholarship.rice.edu/) will have long-term responsibility for the permanent storage needs of all data. All transferred data will be made publicly accessible, any computer code or packages will also be posted on GitHub.

B. Expected data

We are developing a set or parameters and requirements for developing software objects. Therefore, our data is at two levels: the collation of the data into a set of parameters and objects, and the documents from which the data is collated including questionnaires, white board preservation, and workshop notes. The data from preservation of objects will include:

- a comprehensive list of potential uses for DTA
- a comprehensive list of DTA techniques and the advantages and disadvantages of each
- a set of requirements for developing an intuitive user interface to access
- answers to the pre-workshop questionnaire
- photos and/or video of the whiteboard

C. Period of data retention

All relevant data will be deposited in the Rice Digital Scholarship Archive (RDSA, http://scholarship.rice.edu/) for long-term storage upon completion of the project study. Once data is transferred to the RDSA, all data will be made publicly available immediately for a period no less than five years. No data will need to be retained for other purposes.

D. Data formats and dissemination

Specifications, requirements, and answers to the questionnaire will be in Word format and posed on the RDSA. The distributable package will be an R package for small corpus DTA tools posted on GitHub.

E. Data storage and preservation of access

All word documents will be deposited in the Rice Digital Scholarship Archive (RDSA, http://scholarship.rice.edu/), that has capabilities to manage, archive and share digital content. The RDSA allows the public access to the stored material via persistent URLs, provides tools for long-term data management, and permits permanent storage options. The RDSA has built-in contingencies for disaster recovery including redundancy and recovery plans. The distributable R package will be stored at no cost on the Comprehensive R Archive Network (CRAN) that also has built-in contingencies for disaster recovery including redundancy and recovery plans.

7. Data Management Plan

Data to be generated

Type of data	When shared	Under what conditions?
Open Source computer code associated with tool, interface, and server-side component development	When the prototype has been made	Code will be freely available
Data for the website, including pictures of clothing artifacts from the Beeman Historic Costume Collection, Digital generated patterns and imagery, historical information about the garments	When the prototype has been made	With consultation from Ball State University's Copyright and Intellectual Property Office and Section 107, Fair Use, of the Copyright Law, Title 17 of the United States Code, Librarian Jim Bradley, and Sophie Gervais-Trammell Senior CAD Specialist, Lectra Brand Software
Assessment data generated throughout the project	Aggregated data will be shared via the white paper and final report to the NEH. There may be the potential to publish the data in referred publications.	No information will be shared that could identify individuals participating in the assessment process. Prior to completing the assessments, the methods will be reviewed by the Institutional Review Board at Ball State University (Office of Research Integrity).
White paper	After the project has been completed	The white paper explaining the process and research results will be freely available to the public via the project website.
Final report to NEH	At the conclusion of the project	Dissemination of the final report will be the responsibility of the NEH.

Period of data retention
Data will be retained for an unlimited time period beyond the completion of making the prototype. Aggregated data will be publically available for an unlimited time period beyond the project completion on the project website. Copies will be stored long-term on WordPress (or Omeka) using a complementary Ball State University PHP.

Data formats and dissemination
Computer code will be made available as open source in a publicly accessible code repository (e.g. GitHub). Reports will be made available in PDF format and disseminated on the project website. All other data associated with collection visuals, patterns, and text, will be freely available on the Fashion Fusion website. Only copyrighted cleared material will be accessible to the public.

Data management and maintenance

The data will be temporarily managed and stored on various secure hard drives on the Ball State University campus during the developmental stage of the project. During the website development, the data will be transferred to a developed server. The easiest solution to the data while creating and housing the website would be to use a WordPress (or Omeka)-based workflow. Using WordPress would require a PHP server. Ball State University's server will host and manage the data of the final version of Fashion Fusion. Computer codes will be stored in a publicly accessible code repository (e.g., GitHub). The website will be available long-term. The Principal Investigators plan to maintain the website at no additional costs. To build on the database it may be necessary to apply for grants from other organizations, such as Ball State University, The National Leadership Grants offered by The Institute of Museum and Library Services, and Costume Society of America.

7. Data Management Plan

i. Roles and Responsibilities

Data management and maintenance will be administered by co-PIs Martin and Pitti of the University of Virginia with support from Institute of Advanced Technologies in the Humanities staff including database and web developers attached to the project. Martin and Pitti will oversee data management and will monitor preservation and dissemination efforts for the grant period. Wicker of the University of Mississippi and Kopár and Koivisto of The Catholic University of America will maintain the responsibility for managing additional project data in the short term, but will ultimately deliver all draft and final data to IATH for long-term preservation. As IATH's data preservation policy accounts for data migration, project data will be updated as is necessary to ensure ongoing use and access based on the judgment of IATH staff.

Project Andvari data hosted at IATH will include ingested metadata records from third-party collections, images, maps, controlled vocabularies, registered user information, and the web-accessible interface.

ii. Expected Data

a. Data Types

Project Andvari will generate XML records consisting of ingested metadata records from third-party collections and SKOS-formatted controlled vocabularies. The project will also generate HTML files that comprise the Drupal CMS-based web interface, project website, WordPress project blog, and the project Twitter feed. Image files including TIFF, JPEG, and PNG formats will also be included. Additional data including

- Email discussions between project staff, advisory board, and data contributors
- Workshop preparatory materials, notes, and resultant products or reports
- NEH interim reports and white papers

will be generated in native formats and ultimately migrated to preservation formats (i.e. XML).

b. Data Management and Maintenance

All *Project Andvari* systems and data will be hosted, supported, and managed by IATH to ensure continuity of access and ongoing preservation. The web interface, all associated records and data sets, and the project web site will be stored on IATH web servers which will be maintained by IATH staff, including a full-time systems administrator. Email communications will be archived by the email servers at UVA, UM, and CUA. Additional project materials such as workshop notes, blog posts, and any supplementary records will be transferred to IATH for archival processing and storage.

c. Data Access Restrictions

For the duration of the grant period, the project team and advisory board will maintain full access to all project data whether generated by the project team or ingested from third-party data contributors. During development activities, all web accessible data will be password protected and restricted to read-only access by public users. After the grant period and the initial pilot launch, the *Project Andvari* interface and data records will be made open source and publically accessible. Certain data editing/updating

features will be included but will be limited to registered users who have been reviewed and approved by the project team or advisory board.

d. Data Sharing

All *Project Andvari* data for web interface coding, data records, images, and controlled vocabularies will be publically available through either the *Project Andvari* interface, project website, or WordPress blog. Project staff and any workshop participants may freely share email communications, but are not required to beyond the transfer to IATH for archival storage.

iii. Period of Data Retention

In order to assure ongoing preservation, all *Project Andvari* data will retained by IATH indefinitely. Email archives will be retained in accordance with the UVA, UM, and CUA email server archival schedules.

iv. Data Formats and Dissemination

All XML, HTML, image, and map records will be disseminated via the *Project Andvari* platform through graphic user interface transactions or open source bulk data acquisition.

v. Data Storage and Preservation of Access

IATH, in conjunction with the UVA Information Technology and Communications department, maintain several server arrays that will provide a sustainable preservation platform for all *Project Andvari* interfaces and data. All servers at IATH will provide public access to all project data and ensure consistent and effective processing. A full-time system administrator will provide additional support for data dissemination, storage, and preservation activities.

6. Data Management Plan:

Responsibilities:

Project Director Dr. P. Gabrielle Foreman and Gregg A. Silvis, Associate University Librarian for Information Technology & Digital Initiatives, will oversee the data management plan which will be implemented by CCP Coordinator James Casey, DSpace Administrators Molly Olney-Zide and Jordan Howell, and a CCP team.

Expected Data, Collection Methods, Data Formats, and Data Dissemination:

Type of data	Data format and long-term preservation	Conditions of dissemination
Digital facsimiles of convention minutes, which consist of previously printed and documented speeches and debates at the various conventions.	Digital facsimiles are in PDF format. Facsimiles will be converted to PDF/UA (or PDF/A) and stored in DSpace.	Facsimiles will be made freely available via *ColoredConventions.org*.
Digitized photographs and prints.	Digitized photographs and prints are in JPEG format. These same items will be stored in DSpace as TIFF files.	Photographs and prints will be made freely available via *ColoredConventions.org*.
Transcriptions of historical documents.	Plain text transcriptions of historical documents are in .txt format, and stored in DSpace.	Transcriptions will be made freely available via *ColoredConventions.org*.
Audio and video files.	Audio/video files will be delivered in the m4a/mp4 formats using the AAC (CoreAudio) audio format and the H.264 video codec.	Audio and video files will be made freely available via *ColoredConventions.org*.
A database of convention attendees will be generated from the transcripts. The database will document biographical information and the roles and responsibilities of convention attendees.	The database will be in MySQL format, and preserved in an XML-DBML format.	The database will be available to the general public upon request.
White paper.	After the project has been completed.	Available on the project website and UD's institutional repository.
Multimedia progress report.	Duration of grant period.	Available on the project website and UD's institutional repository.
Final report to NEH.	Conclusion of the project.	Dissemination of the final report will be the responsibility of the

		NEH. Will also be available on the project website and UD's institutional repository.

All materials will be publicly available either on the Colored Conventions website, the University of Delaware Institutional Repository, or both. All historical documents and images are in the public domain and will be licensed under Creative Commons. Additional material created by the Colored Conventions Project will be considered under copyright, but licensed under Creative Commons. The Creative Commons license used will be the CC BY-NC-SA (Attribution-NonCommercial-ShareAlike).

"This license lets others remix, tweak, and build upon your work non-commercially, as long as they credit you and license their new creations under the identical terms."

Data Storage and Preservation of Access:
All material will be deposited and permanently stored in the University of Delaware Institutional Repository. The repository consists of the DuraSpace DSpace (*http://www.dspace.org*) software running on fault tolerant server grade hardware, backed up offsite nightly. The material will be permanently stored in the repository. Twenty-five fields of the Dublin Core Metadata Element Set (NISO Standard Z39.85) will be used to store associated metadata for each repository item. Items stored on DSpace will be made available to the general public through CCP's content management system, Omeka.

Period of Data Retention:
Under an agreement between the University of Delaware Library and the Colored Conventions Project, data will be retained for a period of 40 years.

Data Management Plan

1. Roles and responsibilities
The Co-Directors will share overall responsibility for verifying the implementation of this data management plan, including ensuring all data management practices follow Rensselaer's Institutional Review Board requirements. For the purposes of this plan, "project personnel" refers to: the Co-Directors, the graduate student researcher(s), undergraduate research assistants, and the Advisory Board members.

2. Expected data types
The project will collect data in the form of photo, video, observational notes of 3D printing practices and results well as participant discussion of and reflection upon those practices. Project data may include interviews with individuals and groups (audio recordings and transcripts). It is unlikely that any data gathered will be considered sensitive by participants in the project or will put them at risk in any way.

3. Data de-identification
The project includes no research subjects beyond project personnel so data will not require de-identification at any point. Project personnel may choose to include their material as attributed or unattributed at any stage of the research project.

4. Securing data
All in-process data will be maintained on Rensselaer's secure servers, which require authentication to access, and password-protected personal computers of project personnel. In-process data access will be limited to project personnel.

5. Data sharing
This project will strive toward open sharing of project data and research findings. Research findings will be disseminated via standard professional venues, including presentations, conference proceedings, and journal articles. Additionally, much of the data collected in 2 above will be made available on-line via the project website. Project participants may choose to remove attribution of their data or to remove entirely any data that identifies them at any time in advance of dissemination or after it has been included on the project website.

6. Period of data retention
Given the public dissemination of most project data, data retention timeframes will be left to the discretion of individual project participants. Data that participants wish to withhold from dissemination will be deleted not more than one year after the project's funding period ends.

Data Management Plan

As part of the TAPAS project, TAPAS Classroom's data management planning is framed within the expectations of TAPAS itself. TAPAS is a long-term, repository-based publication framework for TEI data, and in addition to serving the data curation needs of its member projects, it also serves as a repository of TEI data for the TEI community as a whole. Because the TEI itself is designed with the goal of supporting long-term data curation (and has already been in wide use for over 20 years), TAPAS project staff is deeply committed to supporting that goal by managing and curating community created TEI and data.

All data submitted to TAPAS for archiving is stored in a Fedora/Hydra repository hosted by Northeastern University Library. This hosting relationship is governed by a memorandum of understanding which stipulates that the Northeastern University Library will provide long-term storage of and access to unlimited quantities of TEI-encoded data for 20 years at no cost, with the option to extend this period by mutual agreement; Northeastern also provides storage of non-TEI data (including image files) on a cost-recovery basis. The TAPAS repository is maintained by Northeastern University's Library Technology Services staff and Information Technology Services staff on the same basis as Northeastern University Library's own Digital Repository Service (DRS), with respect to backups, data integrity checks, and other storage and preservation practices.

TAPAS Classroom data falls into three categories. The first category is TEI files uploaded to the TAPAS "sandbox" by users with free accounts (including students), which will generally be simply test data: for instance, TEI files uploaded to see how they look on the web, or as part of a short-term assignment. To avoid encumbering the TAPAS repository with ephemera, we will provide an option for users to identify such files as test data, in which case such files will be deleted after a short period (to be determined in concert with our test community), and the standard "test" workflow will alert users to the short-term retention. Users will also receive an email alert prior to the deletion, with the option to move the data into the TAPAS Commons for long-term storage.

Within TAPAS Classroom proper, a second category of data includes course assignments, schemas, documentation, and other supporting files. These will be treated as repository items, and will be stored and preserved according to existing DRS preservation policies. The third category is TEI files uploaded by students within the formal framework of TAPAS Classroom. By default, the work completed by students as assignments will be retained for a period of three years to permit the completion of a degree in which these materials might remain relevant. At the end of three years, individual students may choose to submit their TEI Classroom materials to the TAPAS Commons or to create a TEI membership of their own and migrate their class work into a TAPAS project. They may also download their work and retain it personally.

During the year-end course transition process, instructors may select which parts of the course to carry over to the next course by:

- downloading the entire course unit (including course materials and student content);
- migrating the entire course into a TAPAS project (in which case all student materials would be treated as repository items); this option would be appropriate if the class collaborated on a project such as a digital edition; or
- retaining the course framework for reuse while clearing out the student assignments.

Additional options for managing TAPAS Classroom data may be added during development of the system specification.

All TAPAS program code, including any future code developed for TAPAS Classroom, is maintained in a GitHub repository under an open-source license.

Data Management Plan

Data formats, maintenance, and dissemination
The following table summarizes the principal types of data to be generated during the grant period, along with an approximate schedule for dissemination:

Type of Data	When Shared?	Under What Conditions?
Source code and documentation for intertextuality search tools.	As soon as testing is completed, and no later than end of grant period.	Freely available on Github.
Web implementations of intertextuality search tools.	As soon as developed, and no later than end of grant period.	Freely available on project website.
Searchable Greek-Latin thesaurus.	With release of cross-linguistic tool.	Freely available on Github.
List of intertexts in the Argonautic tradition.	With publication of associated paper.	In the supplemental information, and freely available on project website.
White paper.	At conclusion of project.	Freely available on project website.

The intertextuality search tools will be written in Python (as are the current trial versions). All code and documentation will be made freely available on a Github repository dedicated to the project. Git will be used for version control of software in development. The tools will be run on texts digitized by the Perseus Project and in the public domain (see Final Product and Dissemination for details). The searchable Greek-Latin thesaurus compiled for use in the cross-linguistic source tool will be made freely available as a standalone resource for potential use in other projects. It will consist entirely of material from editions of texts in the public domain. Supplemental files containing relevant raw data will be included with all papers published in science journals.

The tools will be made available to users via a web interface created by the software developer hired to work on the project. The implementation will be planned and designed in consultation with the Neukom Digital Arts Leadership and Innovation (DALI) lab based at Dartmouth. Beyond the grant period the co-PIs will explore possibilities for incorporating the tools into new databases of classical texts currently in development, such as the Digital Latin Library at Oklahoma University.

Responsibilities
The co-PIs will be responsible for implementation of the data management plan. Collaborators, research staff, and students will provide documentation for any code that they write, which will be reviewed regularly by the co-PIs. The software developer will be responsible for creation and initial maintenance of the project website, with input from the DALI lab.

Period of data retention
All data generated will be retained for at least five years from the conclusion of the proposed grant period.

White paper
By the conclusion of the grant period, the co-PIs and primary collaborators will prepare a document recording the state of progress in all areas of the project, comparing the outcomes to those projected in this proposal, and explaining any significant successes or failures in planning or execution. The document will also include an updated environmental scan to take account of ongoing work in the field, and a step-by-step plan for further development, which will be based on the application for implementation funds submitted during the course of the grant.

7. Data management plan

Introduction

This plan for Level II Start-Up funding describes the management, dissemination, retention, and archiving of **unique** research data produced during the proposed project. Collaborating with Indiana University (IU) Libraries, this project will provide for sustainable discovery, access to, and preservation of these data for use by other researchers, instructors, and interested members of the public for the length of this project and beyond. This will be facilitated through data and publication deposits in existing open-access disciplinary and/or institutional repositories. A second repository for the data, when fully aggregated, will be the National Archive of the Republic of Tatarstan (Russia).

Responsibilities

The PD will oversee the DMP with responsibility for ensuring that all requirements are fulfilled and will supervise the creation and management of the CEMPP public web platform in both its English and Tatar-language versions. For the latter, Iuri V. Pivovar, Chief Technical Specialist at the National Archive of the Republic of Tatarstan (NART), will assist him. Technical Director, Vincent Malic, will create and manage the project database and English-language public web platform; he will also work with Indiana University's Data Management Service for the successful migration of data to its repository and public server.

Expected Data

The data generated by this research, for which there are no precedents, are derived from a massive set of **MBs** compiled in the Russian Empire between 1828 and 1918 containing demographic data about approximately 25,000 Muslim inhabitants in the city of Kazan. The information recorded in these registers is organized in tabular format, with rows documenting particular demographic events (e.g., birth, death, marriage, or divorce) and with columns containing feature information per row. Experts in Kazan will transcribe these tables into digital format while retaining their original Tatar language. Transcription is regulated by a set of explicit guidelines that standardize the migration of physical data from source to database and the notation of exceptions and document defects. A complete transcription of an annual metrical book produces a single Excel file containing all of that book's tabular data.

Data Formats

Using a lightweight Python script, we will transfer the data contained within the Excel files to a MySQL *preprocessed* relational database that preserves the data in the form they take on the physical pages of the **MBs**. Doing so will ensure a high level of fidelity. Though the data are divided into table cells, for many column headings the data in a cell contain multiple, distinct pieces of information. For this reason, we will run the data in the preprocessed database through a processing pipeline that sifts discrete pieces of data from complex table cells. This processing pipeline will depend on the development of a Tatar Natural Language Processing framework that uses machine learning techniques supported by domain expertise to extract entities and relations from table text. We will develop this Tatar-language framework by expanding on existing NLP tools, such as Stanford's CoreNLP suite and the Natural Language Toolkit. Though this framework will be calibrated to meet the needs of the CEMPP, it will also constitute a valuable contribution to the advancement of NLP technology for Tatar, a threatened language. We will make all Tatar-language NLP tools so developed freely available under the Creative Commons Attribution 3.0 license.

The processing pipeline output, or *processed* database, will be organized around entities identified in tabular form. Most of this information will concern identified individuals and the circumstances of their births, deaths, marriages, and divorces, or their status as third-party witnesses to or participants in these events. The processed database will be a MySQL relational database implementing the Intermediate Data Structure (IDS) for Longitudinal Historical Microdata v. 4. Using this robust and widely adopted format for demographic data ensures that the processed database will be fully accessible to researchers world over and that the CEMPP database can be seamlessly integrated with other demographic databases employing the same model. A suite of Python scripts will periodically synchronize the processed MySQL database with an online-facing RDF/XML Semantic Web database that preserves the IDS schema but makes the data available for SPARQL queries and integration with other Semantic Web databases.

In the later stages of the project, team members will develop a set of rules for inferring whether two items in different records refer to the same entity. The software for making these inferences will be written in R and Python and will also be made publicly available under the CCA 3.0 license. Potential entity matches and their associated level of confidence will be stored in a separate table in the processed MySQL database and will also be made available as supplementary triples in the online RDF/XML dataset. Metadata regarding the transcription, digitization, and processing of all data, such as the date the data were created and the framework used for processing, will be recorded and stored alongside the data in dedicated MySQL tables.

Data Storage and Preservation of Access

Indiana University provides cloud storage systems through the Research Technologies division of IU Information Technologies Services as well as open access repository services through the University Libraries. The initial Excel files, the preprocessed database, and the processed database will all be stored in IU's Research File System (RFS). RFS data are regularly backed up and stored in physically secure environments in Bloomington and Indianapolis, ensuring robust data longevity. While the project is being built and expanding, access to the files may be granted to researchers beyond the project team. The data will be managed and synced in consultation with IU Data Management Service**.** For long term preservation, the database will be deposited and made accessible through the open access IUScholarWorks Repository. The IU Webserve publishing service, providing server space and scripting environments, will host the public website and online SPARQL endpoint.

Data Dissemination

The preprocessed and processed MySQL databases will be made available in the form of downloadable MySQL dumps on the project website. Both of these databases grow with the input of transcribed **MB** data; as a result, all versions of the database will be listed on the website alongside timestamps and basic content statistics. The RDF/XML mirror of the processed database will be available as downloadable triples and accessible via a SPARQL endpoint to the latest version. All code developed for processing and analysis of the data will be available with extensive documentation on the project's GitHub repository.

Property Rights, Ethics, and Privacy

There are no copyright issues for this project nor is protection of human subjects of concern, because all of the persons identified as being born in the last year of the metrical books examined (1918) or before are more than likely to be dead.

Data Management Plan

We will make all the data and methodological procedures generated in the proposed investigation easily accessible to the research community via the web, using standard formats. This will be done by posting the data and methodology on the websites of the USC Shoah Foundation, the UCREL website at Lancaster University, the Spatial History Lab at Stanford University, the Holocaust Geographies website at Texas State University, and the Maine Dataverse Network. These universities will provide basic support and data storage. We will also post our publications on those websites, as well as on the Digital Commons at the University of Maine.

A. Expected data:
This project will propose and test analytical and interactive graphical exploratory tools for understanding and gaining insights into the oral histories contained in video interviews of Holocaust survivors from the USC Shoah Foundation's Visual History Archive (VHA). These capabilities will be harnessed through an extensible framework and toolset geared especially at the humanities research community. The system and its ancillary data will result in datasets and analytical results that will be published in research publications and made available through the project websites. The investigation will also produce digital and non-digital visualizations that we will make available and will use as figures in publications.

The following significant artifacts will be produced in the course of this research:

- A working dictionary of spatial and relational terms that we will use for further analysis of the spatiality of Holocaust survivor testimony. The dictionary will be developed for release to the larger research community and interested parties (such as the USC Shoah Foundation) through research publications and the above websites. Data and query processing capabilities will be made available to the research and education communities while ensuring that users' privacy is preserved. We will release a detailed description of our methodology, including the names of particular corpus linguistics (CL) and natural language processing (NLP) software and analytical modules proved useful in our exploratory analysis, through the above websites. New or custom modules that will be developed as part of this project may also be released with an open source license.
- Any new software or methodological procedures that are used to generate the data dictionary, and the pilot project's hybrid methodology combining those methods and procedures with close listening with manual and computer visualization, will be shared with the broader research community through the above websites, research publications, and research presentations. We will also share the names of CL and NLP methods that were developed by other researchers, along with how scholars and the public can access those methods, to preserve the rights and authorship of those who created and who maintain and update the websites and other means of access to those software programs.
- Results produced by the proposed research will be made available to participating government agencies (e.g., the United States Holocaust Memorial Museum) and to the broader research community through the above websites, research publications, and research presentations.

The following types of data will be retained, utilized, and archived, including:

1) Analyzed data (e.g., digital information that is published, including digital images, published tables, and tables of the numbers used for making published graphs).
2) Metadata that define how these data were generated (e.g., data that will be published in theses, dissertations, refereed journal articles, supplemental data attachments for manuscripts, books and book chapters, and other print or electronic publication formats), as well as full documentation of our source materials, including unique identification information for each VHA interview used in the investigation, and full bibliographic information about all other sources consulted in this investigation.
3) Raw data (e.g., data derived from VHA video interviews in order to create the data dictionary) and

transcripts of the interviews. Due to the nature of the project, the collected data will not need to be anonymized before publication. The VHA interviews used in the study are available to other researchers and the public through the USC Shoah Foundation.

For the **evaluation** of different phases of research, we will prepare and use the following datasets:

1) Datasets to test the validity of the data dictionary will be generated from the VHA video interviews: Such datasets will be constructed by considering as many available properties as possible, such as the location of the speaker and other individuals, time, locations of points of interest, location of concentration camps, railway stations, etc.
2) Preprocessed offline datasets from the Shoah Foundation testimonies and the Holocaust Historical GIS applications generated during the NSF-sponsored research (award nos. 0820487 and 0820501) (for Budapest, Italy, Auschwitz, and the SS camps system): Both GIS and oral histories data, as well as other contextual historical data, will be used to develop, validate and demonstrate the validity of our approach.

These data will be made available to interested researchers and educators in the field upon reasonable request.

B. Standards to be used for data and metadata format and content:

Data formats that will be used to make data available to others, including any metadata, will include ASCII, XML, JPEG, etc., selected according to the standards in publishing each specific type of data. Collected or used video and image data will be available in standard formats such as MPEG and JPEG, respectively. Geographical data sets will be described according to standards endorsed by the Federal Geographic Data Committee (FGDC), including the Content Standard for Digital Geospatial Metadata (CSDGM), and others as applicable.

C. Access to data and data sharing practices and policies:

All participants in this proposal will conduct research and publish the results of their work. Papers will be published in peer-reviewed academic journals or as peer-reviewed conference proceedings. Beyond the data posted on the above websites, primary data, samples, and other supporting materials created or gathered in the course of work will be shared with other researchers upon reasonable request, at no more than incremental cost and within a reasonable time of the request. The project websites named above will be maintained by the participants for a minimum of three years after the conclusion of the grant.

1) Policies and provisions for re-use, re-distribution, and the production of derivatives:
   It is the policy of all institutions participating in and supporting this proposal to encourage, wherever appropriate, research data to be shared with the general public through Internet access. Public access will be regulated by the partners to protect privacy and confidentiality concerns, and respect proprietary or intellectual property rights. Administrators will consult with the University's legal office to address any concerns on a case-by-case basis, if necessary. Terms of use will include requirements of attribution along with disclaimers of liability in connection with any use or distribution of the research data, which may be conditioned under some circumstances.

2) Plans for archiving and for preservation of access:
   Data and other research products will be made available immediately after publication. Final peer-reviewed journal manuscripts, and supplemental information such as data tables for graphical information in manuscript figures, which arise from NEH funds, will be posted at the websites named above no later than twelve months after publication (if publishing agreements allow it; if not, the above websites will provide a reference to the journal articles coupled with the supplemental information.) These records will be durable, accessible through web protocols, and made safe from tampering or falsification. The storage media will be updated as necessary to keep it current.

Data Management Plan

Expected Software: Our work will result in the creation of novel machine learning algorithms for visual stylometry and analysis of artwork. We will also create workflow fragments for the WINGS semantic workflow framework. All software resulting from the project will be released with an open source license on Fitchburg State University (FSU) servers and simultaneously released on GitHub.

Expected Data: Our work will also result in the data collection of artworks from various partners. We will publicly make available a selection of the collection which transfer rights to FSU. The rights and permissions of each of these collections will be clearly advertised for potential researchers, along with contact information to gain rights for those that do not have a sharing agreement with FSU. We view these as extremely valuable datasets for the digital humanities research community which could contribute to further research work in this area. Sample datasets, along with the corresponding processing packaged as WINGS exportable linked data workflow fragments in Open Provenance Model (OPM) format, will also be released on FigShare along with their own unique Digital Object Identifiers (DOI). Where necessary, we will ensure all data is appropriately anonymized and we have CITI training from FSU.

Data Formats and Dissemination: Our project will result in digital image (PNG or GIF), video (MP4), and linked-data web objects (OPM). All participants in this proposal will conduct research and publish the results of their work. Papers will be published in peer-reviewed scientific journals, conferences, or Annual Meetings. Beyond the data posted on the FSU website, primary data, samples, physical collections and other supporting materials created or gathered in the course of work that are releasable to the public domain will be shared with other researchers upon reasonable request within a reasonable time.

Period of Data Retention: Datasets, samples, and other research products will be made available immediately after publication. Final peer-reviewed journal manuscripts, and supplemental information such as data tables for graphical information in manuscript figures and for statistically processed averages, which arise from NEH funds, will be posted in the FSU Digital Repository and will be available on this publically available website no later than 12 months after publication. Authors will ensure their publishing agreement allows the paper to be posted to the archive; alternatively the FSU website will provide a reference to the journal articles coupled with the supplemental information. These records will be durable, accessible through web protocols, and made safe from tampering or falsification. The storage media will be updated as necessary to keep it current.

Data Storage and Preservation of Access: We plan to share all our data on FSU websites. This public access will be regulated by the university in order to protect privacy and confidentiality concerns, as well to respect any proprietary or intellectual property rights. Administrators will consult with the university's legal office to address any concerns on a case-by-case basis, if necessary. Terms of use will include requirements of attribution along with disclaimers of liability in connection with any use or distribution of the research data, which may be conditioned under some circumstances.

Data Management Plan

Types of Data

The data produced in the first prototype will be in the form of Markdown and derivatives (Word, PDF, HTML). In addition we will produce spreadsheets documenting the cost of production of our workflow. The second component will produce an unpredictable set of data depending on the exigencies of the Scalar prototype we will use to disassemble into the repository. This may include images, video, text, sound, maps, etc. In addition to the individual media-specific digital assets, we will count the database itself, a snapshot of the site, and a tarball of all assets as discrete data units.

Data Standards and Capture

For our first prototype, Markdown files encoded in UTF-8 will be mostly produced by authors after receiving adequate training, using text editors recommended by us. We will then process these .md files using Pandoc in order to generate all other derivatives. Eventually these will go into github for production of this part of the site. In the second prototype we will set standards for media types that conform to the Library of Congress "Recommended Format Specifications," in combination with the data standards of CDRS and Academic Commons.

Metadata

In both prototypes, we will charge authors with the task providing certain metadata and will adhere to the metadata standards of Academic Commons. In the case of the second prototype, which may generate additional metadata, we will encourage the use of Dublin Core. This metadata in turn will become an added discrete data unit for us.

Legal Policy

Our authors will retain copyright, inasmuch as they agree to make their research available to the public, once it is in a state that pleases all parties, using an appropriate Creative Commons license. Specifically, we will encourage them to use a CC BY 3.0 US license. Our own content will be available to the public with that license. In the case of the material depositing in Academic Commons as a result of our collaborations, that content will be available under the existing provisions of the repository.

Period of Data Retention

One of the main goals of our project is to prolong the life of digital scholarship affordably and realistically, whether textual or multimodal, to the greatest extent possible. To that aim, we are building systems that ultimately align with the university repository as the best guarantors of longevity available to academics at the moment.

Access Sharing and Reuse

All data generated in this project, aside from the spreadsheets detailing private salary information, will be released to the public in the form of a white paper, a series of research essays published using our minimal computing workflow, and the Scalar prototype with its disassembled components and other discrete data units necessary for its reconstruction.

Data Management Plan

Roles and responsibilities: This plan will be implemented and managed by Della Pollock, Hudson Vaughan and Heidi Dodson.

Types of Data: This project will largely generate data at two levels: data associated with the archival files we are collecting, curating, sharing, and preserving and documentation of the Omeka site development.

The data associated with project archival content include digital files of oral histories, audio documentaries, images, textual narratives, and historical textual materials to be shared in the digital commons. The project will also generate digitized digital exhibits within the Omeka platform, which includes photographs, audio clips, video and textual responses. Other data include text files: documentation of project planning and processes, a Northside Digital Commons user manual, project and historical context for digital collections, and documents associated with oral history processing, including life history forms, consent forms, abstracts, field notes, proper word lists, tape logs, and transcripts. Archival item metadata will be also be generated.

Documentation of software development includes source code documenting the development of the Omeka content management system.

Data Management Prior to Dissemination: Oral history audio files and associated files such as photographs are recorded and collected at the interview site (varied locations) and are then renamed according to the Jackson Center's protocol and uploaded to a Dropbox account that is synced to the hard drive of the password-protected Jackson Center computer. Within Dropbox, the original files are uploaded to a master folder. Copies are made and transferred to a working folder. These files are also backed up on a weekly basis to both an external hard drive and a shared server. The files are thus saved in four locations. Volunteers work from copies placed in a separate designated shared folder and do not have access to the working or master folder. As they process the interviews, Project Staff transfer materials to the working and master folder. Textual planning and process documents are stored the institutional Google Drive account and are backed up to the external hard drive and shared server. Source code for the Digital Commons is maintained in a version-controlled repository hosted by Bitbucket.org, and also backed-up locally on developers' computers. The Digital Commons site itself is stored on servers operated by Reclaim Hosting, and the site and associated database are backed-up regularly to an off-site location.

Factors that might impinge on ability to manage data: If an interview or associated item does not have a consent form giving permission to share the data, it will remain in storage until a consent form can be obtained. If an interviewee requests part of the interview be restricted, the original audio file will remain in the designated digital storage locations, and an edited audio file copy and/or transcript will be created to provide access to the non-restricted material.

Data Formats and Dissemination: Audio interviews are recorded as uncompressed wav files, which are converted to mp3 files for sharing in Omeka. Photographs are typically jpeg format, but archival scans made at the Jackson Center include uncompressed tiff files. Textual documents will be stored and disseminated as pdfs. Interviews and photographs will be uploaded to the Omeka site as soon as an abstract and keywords have been created. These minimum elements will facilitate public access to the interviews. Additional materials, such as transcripts, photographs, and tape logs will be uploaded as soon as they are created by Jackson Center Fellows. Paper consent forms and other hard copies of photographs or textual data will be stored in an office file cabinet, filed under the interviewee name. All interviewees and their families will be given CD copies of the interview for their own use. In addition to access and

download interviews through the digital platform, neighbors will be able to access and download files or make cd copies at the Jackson Center. The lowest level of aggregated data that project directors might share would be an mp3 or wav audio interview file, digital photographs, and processed materials from community documentary initiatives.

Metadata
Descriptive metadata for audio and image files materials are stored in one main Excel workbook on Dropbox, in Omeka, and on the local computer hard drive, backup external hard drive, and shared server. Metadata elements for audio and image files are created using the original Dublin Core Metadata Element Scheme (15 elements). For audio oral history interviews, two additional elements, length of recording and interview processor, are also captured in Omeka. The Excel workbook contains administrative metadata, such as interview processing status, that is not included in Omeka. Fellows and consultants will create and apply controlled vocabularies for subject terms and keywords added in Omeka and will use consistent content conventions for other fields in order to improve discoverability.

Procedural documentation
The Jackson Center's manual will be updated to include the Data Management Plan so it is accessible to staff, Fellows, and consultants. As we work on the re-design and expansion of the Omeka site, we will document our changes in a separate, in-house Omeka manual and on Github. The in-house manual will include technical instructions as well as explanations for our decisions regarding design, public accessibility, and metadata.

Long-term Data Storage and Preservation:
The Jackson Center will maintain the archival components of the Northside Digital Commons using best practices established by the Southern Historical Collection at University of North Carolina-Chapel Hill. All unrestricted oral history files and associated metadata will be stored in the Carolina Digital Repository through submission to the North Carolina Digital Heritage Center on a biannual basis. The Community Review Board will be tasked with identifying the appropriate partner for long-term oversight and storage if, at any point in the future, the Jackson Center were to cease to exist. The Digital Commons site is stored on Reclaim Hosting's servers, and backed up regularly.

Legal Policy: Interviewees will retain their copyright to interviews and photographs submitted to the Jackson Center, but will agree to license public use of the materials through the Creative Commons Attribution-ShareAlike 4.0 International License. This permission is granted through the Jackson Center's Interview Agreement form. The interviewees have the option of including restrictions on access.

Holocaust Ghettos Project Data Management Plan

Roles and responsibilities: The Advanced Computing Group at the University of Maine (ACG@UMaine) will provide computer and data services, including backup and long term archiving. The ACG@UMaine's primary data repository is the Maine Dataverse Network (MDVN). Data produced by the individual researchers on this project will be stored on this resource. ACG@UMaine services include training and support for all technical aspects of the data produced during the course of this project necessary to maintain security, dissemination, and preservation. The PI will have decision-making authority over all data management. Those who generate data as a result of this project are responsible for adding it to the MDVN repository immediately after publication. It is the responsibility of the individual PI, Co-PIs, and Consultants to NOT add data to the repository that violates privacy, confidentiality, security or intellectual property concerns. The archival life cycle and retention policy for archived data will be managed by ACG@UMaine. Data will be retained indefinitely after the end of the project until or unless doing otherwise is necessary to meet legal obligations or adhere to an established data management policy. The PI will check adherence to this plan at least 90 days prior to the expiration of the award. Adherence checks will include review of the MDVN content, number of studies released, availability for each study of subsettable/preservation friendly data formats (possibly embargoed, but listed); availability of documentation (public); and correctness of data citation, including an integrity check.

Expected data and period of data retention: The data generated by this project will be in a variety of formats, including program code, databases, images, shapefiles, text, and binary data. Once cleaned and finalized, all data will be shared with the scholarly community via the MDVN. All associated metadata will also be shared, including sources, analytical methods, and rules for data entry. Several database management systems are available on the data server. We plan to use MySQL for this project, although alternative DBS solutions will provide flexibility if needed. All data are stored native in journal-based file systems ext4 with a configuration of RAID-5 underneath. For program codes, version control systems, such as git and cvs, are available to record the history of changes. There are no known legal or ethical restrictions on access to non-aggregated data generated in this project. During the project, all data files will be stored, maintained, and regularly backed up daily on Seafile, UMaine's cloud storage system, which is maintained by ACG@UMaine and will be accessible to all team members.

The primary tool for providing data retention on this project is the Maine Dataverse Network (MDVN). The MDVN is a public repository, which runs on secure servers hosted and maintained by the ACG@UMaine. Data will be deposited into the MDVN as soon as it has been cleaned and checked by the PI and the Co-PI or Consultant responsible for its creation, in every case within six months of the completion of data analysis. Data will be retained indefinitely, never destroyed (unless required by law).

Data formats and dissemination: The primary tool for providing access to shared data produced by the individual researchers on this project is the MDVN. The MDVN facilitates access to data through descriptive and variable/question-level search; topical browsing; data extraction and re-formatting; and on-line analysis. All data will be deposited within six months of the completion of data analysis. Such data may be embargoed until the publication of research based on the data or until three years after the expiration of the award, whichever is sooner. Users will be required to agree to click-through terms at the discretion of the PI that prohibit unlawful uses and intentional violations of privacy, and require attribution. Use of the data will be otherwise unrestricted and free of charge.

Data generated by this project can be released without privacy restrictions. The data extracted from the U.S. Holocaust Memorial Museum's *Encyclopedia of Camps and Ghettos* does not constitute private information about identified human subjects. The data extracted from transcripts of Holocaust survivor video testimony comes from sources for which the collecting organizations (USHMM and the USC Shoah Foundation's Visual History Archive) have permission to distribute the oral histories to the

public. The data will not be encumbered with intellectual property rights (including copyright, database rights, license restrictions, trade secret, patent or trademark) by any party (including the investigators, investigators' institutions, and data providers); nor are the data subject to any additional legal requirements. Depositing with the MDVN does not require a transfer of copyright, but instead grants permission for the Maine Dataverse Network to re-disseminate the data and to transform the data as necessary for preservation and access. Access to restricted data at any point in the data life cycle may be granted through MDVN request for access function. At the discretion of the PI, data will be shared on the MDVN and an email sent to the requestor with the access information. Requestors are required to agree to click-through terms and conditions of reuse and re-distribution (e.g,. authorship, acknowledgment, citation use). Terms and conditions are included in associated data files on MDVN and determined by the PI on a case-by-case basis. The PI, Co-PI, or Consultant responsible for each part of the project will create documentation detailing the sources, coding, and editing of all data, in sufficient detail to enable another researcher to replicate them from original sources; and descriptive metadata for each study including a title, author, abstract, descriptive keywords, and file descriptions. The PI will coordinate metadata creation to ensure consistency. The project will include bibliographic information for any publication by the project based on that data. Sound, videos, and image files will be maintained in WAV, JPEG2000, and TIFF formats. Other documentation will be deposited in PDF/a, or plain-text formats, to ensure long-term accessibility, including a readme file in each directory to describes detailed metadata information, such as data formats, access methods, usage, authors, notes, and revision history. Quantitative data will be converted to CSV and ESRI shapefile (SHP) format. These formats are fully supported by the Maine Dataverse Network (MDVN), which performs archival format migration; metadata extraction; and validity checks. Deposit in these formats will also enable on-line analysis; variable-level search; data extraction and re-formatting; and other enhanced access capabilities. The MDVN repository system's “templating” feature will be used for consistency of information across studies. The MDVN system automatically generates persistent identifiers, and Universal Numeric Fingerprints (UNF) for studies; extracts and indexes variable descriptions, missing-value codes and labels; creates variable-level summary statistics; and facilitates open distribution of metadata with a variety of standard formats (DDI v 2.0, Dublin Core, and MARC) and protocols (OAI-PMH and Z39.50).

Data storage and preservation of access: The archiving and preservation of all data will be managed by ACG@UMaine, whose primary data repository is the Maine Dataverse Network (MDVN). The MDVN is hosted by ACG@UMaine on a secured shared storage array acquired under the funding support of a previous NSF MRI grant. The storage array is fast and reliable, with a total of 120 TB of raw storage interconnected with multiple 4 Gb/s Fiber Channel paths and offsite backups. ACG@UMaine commits to good archival practice including 24x7 machine operation support versioning and deaccession compliant, regular off-site backup, and regular content migration to ensure that data are available for access consistently and kept secure. All data will be stored, during and post project, within the UMaine network infrastructure, which employs firewalls and secure authentication and authorization methods for login and access. Data deposited into the MDVN will be retained indefinitely, never destroyed (unless required by law), backed up on a daily basis and replicated across multiple locations for long-term access. Long-term access to data will be provided in accordance with the permissions, terms of use, policies and procedures established by the PI in consultation with the research team at the time of archiving. ACG@UMaine services include personnel training and support for all technical aspects of the data produced during the course of this project necessary to maintain security, dissemination, and preservation. Should the archiving entity be unable to perform, transfer agreements with the University of Maine System are in place to easily migrate to another entity within the university system.

9. Data Management Plan

Responsibilities

PI Zeldes will oversee the data management plan. He will oversee the development of digital tools, manage the servers described below, and manage the GitHub repositories. Co-PI Schroeder will oversee the project website and digital document and collection management. At the end of year 1, the DH Specialist will take over the maintenance of the project website (including documentation) and GitHub repositories. Staff from Georgetown's Institutional Repository (IR) and University of the Pacific IR will assist with long-term data retention (see *Sustainability*).

Expected Data

Our expected data fall into the following four general classifications: digitized texts, tools, corpus database, documentation. The data types within each classification are described in more detail below in *Data Formats and Dissemination* (see also *Final Product and Dissemination* for more details on data content).

Data Formats and Dissemination

The project participants embrace the principles of timely, rapid, and open-source data distribution. Open source and open access practices assist in data management and dissemination by creating resources that can be used and disseminated beyond the life of the project. For example, Coptic SCRIPTORIUM published a pilot Coptic language treebank for linguistic research as part of the Universal Dependencies project (http://universaldependencies.org/), which coordinates universal, cross-language standards to facilitate multi-lingual research using the same annotations. Our work now appears alongside those of over 50 other languages, although the resource is still in development and only at a pilot stage.

Tools and Technologies: The digital tools described in this proposal will be written in Python, Java, JavaScript, and other languages. We will also pursue the adaptation of existing open-source tools. The digital tools will be developed and distributed as open-source software and free public downloads under open-source licenses, normally the Apache 2.0 license or similar (depending on usage of imported library licenses, when those are incompatible with Apache 2.0). The software will be developed and distributed on the project's GitHub repositories, with links to these resources provided from the Coptic SCRIPTORIUM website and from Georgetown University Computational Linguistics webpages.

Digitized Text: The digitized text will be raw data in the form of text files, TEI XML files, PAULA XML, SGML files in English, Greek Unicode and Coptic Unicode (displayed in the Antinoou font created by the International Association of Coptic Studies.) We anticipate no changes to Unicode standards for Coptic characters that would affect our work. The digitized text files will be stored on version-controlled servers at Georgetown University (e.g., the Corpus Linguistics server at http://corpling.uis.georgetown.edu/) using Git or Subversion, as well as the Coptic SCRIPTORIUM GitHub site at http://www.github.com/CopticScriptorium.

Some of the project text data will be drawn from ancient and medieval manuscripts or scans of books out of copyright. Under intellectual property law in the United States, the text from the manuscripts is in the public domain; editorial work can be under copyright. The project will not be publishing online editorial work that has been published in print or in digital formats under existing copyright. Prepublication working files will be stored on the above-named servers. Images of manuscript pages will not be circulated outside of the direct project participants unless authorization has been provided by the image providers.

Digital publication of textual data will occur on Coptic SCRIPTORIUM's site and the Georgetown Corpus Linguistics server and site. The subdomain of Coptic SCRIPTORIUM's site, data.copticscriptorium.org, publishes annotated digitized text and is hosted on an Amazon Web Services server. Published text, corpora, and textual annotations will be distributed under the

Creative Commons attribution (CC-BY 4.0) license whenever possible (https://creativecommons.org/licenses/by/4.0/legalcode).

Documentation: The project will provide documentation of the tools, technologies, methods, standards, and data models developed and used during the grant period as text files and pdfs. Video tutorials with screenshots may be made available on YouTube under public licenses. Documentation will be labeled with date and version information and disseminated under open-source licenses, such the Apache license, GNU Free Documentation License, and the Creative Commons Attribution (CC-BY 4.0) License. News and information about project progress (with links to the tools or more detailed formal documentation) will appear on the project blog; text of blog posts will be licensed Creative Commons Attribution (CC-BY 4.0).

Period of Data Retention

Tools, documentation, and database files will be released as they are created on our public GitHub repositories, with links provided on the project website and the Georgetown University Corpus Linguistics website. Digitized text files which contain edited, annotated text will also be released as soon as the editorial process is completed and made available indefinitely (see below).

Data Storage and Preservation of Access

Long-term storage and access for tools and published annotated digital text will be provided by the Georgetown University and University of the Pacific Institutional Repositories (IRs). Both of these IRs are open access. Georgetown's repository at Lauinger Library, called DigitalGeorgetown (http://www.library.georgetown.edu/digitalgeorgetown), is powered by DSpace, an open source software solution, and hosts the scholarly output from faculty members, institutes, centers, publishers, and round tables across all of Georgetown's campuses, with the goal of providing long term open access to this material. The repository includes working papers; journal articles; theses and dissertations; data sets; citation and image databases; and much more.

In addition to helping disseminate the scholarly output of the University, DigitalGeorgetown also serves as a robust digital preservation platform. Each item within the repository is assigned a handle, which acts as a persistent identifier and permanent URL to the resource. This ensures that links to items within DigitalGeorgetown will always resolve, even if the software and underlying architecture of the repository change over time. In addition to the use of the Handle System (https://www.handle.net/), DigitalGeorgetown also utilizes the Academic Preservation Trust repository as a back-end dark archive for digital preservation. APTrust (http://academicpreservationtrust.org/) is a consortium of higher education institutions, including Georgetown, which are committed to providing a preservation repository for digital content. Georgetown University Library's Digital Preservation Policy ensures that there is a commitment to preserving all content from DigitalGeorgetown into APTrust, irrespective of the bitstream and data format and storage requirements. Metadata is also harvested by the Digital Public Library of America (DPLA), thereby increasing the findability of resources produced in the project. The University of the Pacific IR can hold additional copies of our files in all formats as well. Repository Staff will assist Faculty in ensuring proper documentation and metadata for the IR.

The final White Paper will be disseminated on the NEH website, the two university IRs, and the project website (under Reports at http://copticscriptorium.org/reports, along with prior NEH reports and White Papers). Project participants will also disseminate their results in journal articles and at professional conferences and symposia. The most important of the latter events are the Society of Biblical Literature annual meetings, the quadrennial international congress for the International Association of Coptic Studies, and the annual international Digital Humanities conference. In Computational Linguistics, technical advances will be presented at regional and international conferences of the Association for Computational Linguistics (ACL) and Corpus Linguistics conferences (for example the Language Resources and Evaluation Conference, Treebanking and Linguistic Theories, or the international Corpus Linguistics conference).

Picturing Urban Renewal
Data Management Plan

Roles and Responsibilities. PI David Hochfelder will have overall responsibility for implementing and managing the data management plan, in consultation with digital librarians Mark Wolfe and Greg Wiedeman of the University at Albany Libraries' M.E. Grenander Department of Special Collections and Archives.

Expected Data. We anticipate generating images, GIS, descriptive metadata, and a website.

- Digitized photographs.
- Data pertaining to Historical GIS. This includes geographical coordinates for maps, property parcels, and photographs; shapefiles; and map layers created for data visualization purposes.
- Metadata. This includes information about photographs, oral history interviews, and other assets collected for this project.
- Interactive website. This consists of the map layers and metadata presented to the public through an online web application

Data Capture. The interactive website will be crawled and preserved as Web Archives using both the Internet Archive's Archive-It service and Webrecorder under the supervision of the M.E. Grenander Department of Special Collections and Archives. The captured site will be disseminated through the UAlbany's Archive-It page as well at the Internet Archive's public Wayback Machine, ensuring long-term public access. Master files for the Web Archives captures will be preserved as ISO-compliant WARC files. We will store all data at multiple redundant sites, including a University at Albany server that is backed up daily, Dropbox cloud storage, and a public Github repository. Finally, for long term preservation all metadata and GIS data created during the project as well as the master WARC files will be deposited with the M.E. Grenander Department of Special Collections and Archives. Digitized photographs will be given back to the originating repositories with identifier links in accordance with professional archival standards.

Metadata. In all cases, we will conform to Dublin Core or New York State Archives standards. At a minimum, metadata for each photograph will include a unique identifier, title, subject, description, source, format, scanner capture settings, creator, date photographed, date digitized, street and address location, geographic coordinates, and rights-holder. Metadata will be linked to digital photographs via the unique identifier, and to GIS data though geographic coordinates. In addition, New York State Archives photographs scanned by project photographer Mike Wren will have additional technical metadata as requested by Archives staff, including data pertinent to the scanning process.

Legal Policy. We have obtained non-exclusive, irrevocable, worldwide, and royalty-free permission for all photographs and other materials used in this project, including the right to publish them in print and online. Photographs and other material held at the New York State Archives are already in the public domain. Organizations and individuals who have contributed photographs and other assets to this project will retain their copyright and other intellectual property rights. The Picturing Urban Renewal website will employ a Creative Commons "Attribution—ShareAlike" license, similar to that used by Wikipedia.

Period of Data Retention. The M.E. Grenander Department of Special Collections and Archives is committed to ensuring long term access to the GIS Data, metadata, and Web Archives captures generated for this project.

Access, Sharing, and Reuse. All data generated from this project will be made available for download from a public Github repository or by contacting the project director. Long term access will be ensured by the M.E. Grenander Department of Special Collections and Archives, using web servers managed by the University's Tier III data center. No additional permissions will be required to download or reuse data. However, organizations and individuals who have provided photographs or other assets still retain their full intellectual property rights over these materials.

Data Management Plan

Expected Data
We are developing a web application and platform to collect and archive crowd-sourced digital media content and associated metadata. Therefore, the data generated by this project will include source code and both software and participatory design process documentation for the DD platform and apps as well as the digital media content and metadata submitted via the DD app.

The data will include:

- PD process documents: interview transcripts, focus group notes, UI/UX design storyboards, etc.
- Software code
- Text and image files of correspondence, notes, academic papers, design sketches, storyboards and planning documentation from project team members
- Digital media content and metadata collected via the DD app: images, scanned documents, audio recordings, videos, etc.

Data Management and Maintenance
All computer code and software documentation will be stored and made available to the public via Georgia Tech's enterprise GitHub code repository where it will be maintained by project team members. The *Digital Drawer* platform components, including web applications, media and metadata database and API, as well as all project development and planning documentation will be hosted on server hardware at IPaT at Georgia Tech. IPaT's servers are continuously maintained, backed up and kept current with security updates by dedicated IT staff. IPaT will be responsible for long-term data management and maintenance.

Period of Data Retention
Data and formal reports will be publicly available within 1 year of project completion. Data will be retained by Georgia Tech for a minimum of 5 years beyond the completion of the project and will be migrated to servers maintained by HRCGA during this period of retention.

Data Formats and Dissemination
Computer code (i.e. NodeJS, Javascript, PostgresQL) will be freely available via Georgia Tech's GitHub repository, a publically accessible code repository. Our crowd-sourced media dataset and corresponding metadata will be deposited and maintained on servers at the Institute for People and Technology (IPaT) at Georgia Tech. Reports will be made available in PDF format and deposited in Georgia Tech's institutional repository, http://www.library.gatech.edu/. A project website will serve as a portal to all project data.

Data Management Plan

Both the Kerameikos.org ontology and concept data are stored in openly accessible Github (https://github.com/) repositories. Data revisions are committed to Github in an automated process each night, but bulk downloads of the data are also available, linked from the project's home page. Since the data are open, they may be reused and redeposited in any other system without reservation. Since Github is a commercial entity and its persistence cannot be guaranteed in the future, we intend to deposit the ontology and data in several other information systems for long-term preservation:

- The ontology will be deposited into Linked Open Vocabularies (https://lov.okfn.org/dataset/lov/), hosted by the Open Knowledge Foundation.

- The concept data will be deposited quarterly into Datahub (https://datahub.io/), also hosted by the Open Knowledge Foundation.

- Both the ontology and data will be deposited quarterly into infrastructure hosted by the University of Virginia Library, mostly likely Dataverse (https://dataverse.lib.virginia.edu/), its data repository.

Data Management Plan

Data Generated

This project will generate several forms of data, beginning with the acoustic impulse responses measured in the present-day Thomaskirche. These will be in the form of wave files (.wav), which record the entire acoustic response of the church to any musical performance, and may be of use to other scholars studying the church as it exists today. From these impulse responses we will calculate many different acoustic parameters used to classify different aspects of the space's sound field, such as the late reverberance (T30), the early decay time (EDT), or the clarity index (C80). These will be stored as text files (.txt) with embedded tables that can be exported into a Word document or PowerPoint presentation.

During Braxton Boren's research at the Bach-Archive, he will also scan and save image or pdf files of any prints or drawings of the Thomaskirche, as well as documentary evidence describing earlier orientations of the church.

After returning from Leipzig, the computer modeling process in Sketchup will first generate geometry files (.dxf) describing the layout of the present Thomaskirche, the Bach-era Thomaskirche, and the pre-Reformation Thomaskirche. These geometry files will be used to create acoustic simulations in the modeling software CATT-Acoustic, which generates its own auxiliary files including geometry files (.geo), source/listener location files (.loc), and several other internal data formats for the simulation data. The final product of the simulation will include simulated impulse responses for both the Bach-era and pre-Reformation Thomaskirche, which will again be wave files (.wav).

During the recording of Bach's cantata BWV 102, the musicians will listen to themselves with the reverberation of the past versions of the Thomaskirche via realtime convolution with the simulated impulse responses. This recording session will yield both dry recordings of the musicians and auralized versions which combine the dry audio with the reverberation of each time period's Thomaskirche. All this data will be stored as wave files (.wav).

Data Management

The acoustic measurement files will be stored on Davide Bonsi's and Braxton Boren's computers, as well as backed up on a separate external hard drive and a remote cloud backup server. The impulse responses from these measurements in the present-day church will be available through the project's website, both as convolution filters (allowing listeners to add the church's reverberation to their own recordings) and as direct downloads of the impulse responses themselves. This will aid in dissemination of this data to scholars interested in the current acoustic properties of the church.

Any image files or pdfs from Braxton Boren's research at the Bach-Archive will be stored on his computer as well as a remote cloud backup server. Images or text judged to be appropriate for the public narrative of the project will also be used on the project website.

All Sketchup and CATT-Acoustic files will be stored on Braxton Boren's computer as well as an external hard drive and a remote cloud backup server. These files will be available for download from the project website if other researchers wish to validate our simulations or build on these models for other projects related to Bach's music. The simulated impulse responses for the Bach-era and pre-Reformation Thomaskirche will be available on the project website, both as convolution filters and as direct downloads of the impulse responses themselves.

The dry recordings of Bach's cantata BWV 102 will be stored on external hard drives and a remote cloud backup server. These will be available to other scholars should they request them, but because of the large file size they will not be posted online, as dry audio is often thought to be unpleasant to listen to, and is not meant for casual listening but only as a research tool. The auralized versions of the cantata in the Bach-era Thomaskirche and the pre-Reformation Thomaskirche will both be available on the project website via streaming stereo audio. The cantata will be auralized at different locations in the church, and listeners will be able to select their position on a visual map of the church in each time period.

IX. Data Management Plan

Indiana University's Cyberinfrastructure

Indiana University is a national leader in the deployment and use of advanced information technology and cyberinfrastructure in support of research and education. IU's Pervasive Technology Institute is the umbrella for six major centers that provide systems, tools, and services furthering IU's research mission. That includes the University Integrated Technologies Services (UITS)'s Research Technologies, which run IU's computing and storage resources and help build scholarly communities. All online development work we do runs on Indiana University Webserve and benefits from significant cyberinfrastructure: computing, data storage, archive and restoration, database, and network access and control resources. Major components of IU's cyberinfrastructure include the following:

Supercomputers: IU's supercomputers provide IU researchers with access to some of the most powerful supercomputers in the US.

Scholarly Data Archive: IU's massive data storage system can hold up to 42 petabytes of data. With mirrored tape silos in Indianapolis and Bloomington, this very secure storage system ensures data are stored securely and reliably. Backup and replication is done within SDA's High Performance Storage System (HPSS). Data are written to a fast, front-end disk cache and migrated over time to IBM TS3500 tape libraries on the two main campuses, providing highly reliable disaster protection. The SDA is backed up nightly. All data are retained permanently.

IUScholarWorks: This is an open access repository service provided by the IU Libraries for disseminating and preserving the intellectual output of IU scholars. The repository is designed to hold and deliver scholarly materials in digital form (text, data, image, etc.) that will not change over time and that are described with standard keywords and descriptors. IUScholarWorks makes scholarly research materials freely available at a stable URL, and maintains them over the long term.

IU Web Framework: IU provides at no cost a set of standard tools to build feature-rich web sites where contributors can create, maintain, and publish web content.

Box @ IU: The IU Box is a no-cost cloud storage and collaboration environment that provides a secure way to share and store non-critical files and folders online. Account quota is unlimited.

Data generated by our research

Envisioned as a scholarly community of editors and inquirers, the Peirce Project's production platform (STEP) will produce several categories of data.

(1) The software and other artifacts that drive STEP and its companion suite of applications STEP TOOLS.

(2) Real scholarly data, including digitized manuscript images, raw transcriptions, encoded transcriptions, rounds of proofreadings and corrections, rounds of critical editing and corrections, rounds of apparatus documents and corrections, rounds of layout documents and corrections, thousands of annotation documents, bibliographical data, associated databases, technical reports, associated non-confidential professional correspondence, all connected to the development, testing, incremental implementation, and refinement of the platform for later public release.

(3) Online project website, online video tutorials; digital publishing tools (including online display format); custom software code along with archiving of open source software tools integrated into the project to maintain compatibility; abstract custom digitized workflow algorithms; digital tools for generating scholarly content, documentation and end-user education material; relational databases.

(4) Scholarly contributions to the critical edition, peer reviews of contributed scholarly content, and high-quality large resolution digital image files of original material.

Data management plan

Open source, open access, and limitations

STEP development uses open source software released under various free licenses, along with the creation of new software. This allows for most data listed above under (1) to (3) to be released under an open

license approved by the Open Source Initiative and made available for download, modification, and use by any party, except where governed by other open source copyright licenses. Data listed under (4) will be open access except when subject to university press contracts, copyrights, or when release could result in an invasion of personal privacy.

Data retention

STEP and STEP TOOLS will be released under an open license compatible with CDDL license once stable release versions have been reached that meet our benchmarks. Such stable versions should (a) successfully process the creation of potential project data, (b) enable customized workflow, variable editing guidelines, and modifiable access-control lists, (c) facilitate the creation of structured data, (d) incorporate data backup and restoration solutions, and (e) provide a method for unique website theming. Produced content used as both pilot data and scholarly work will be made available after an initial STEP release once the content will have been approved for dissemination.

Using components of IU's cyberinfrastructure stated above and below, all data produced during STEP and STEP TOOLS development and pilot testing will be both accessible *as soon as produced* and protected from disaster, assuring present and future access. Redundant access configurations will afford constant access with emergency outages being covered by enterprise-level restoration processes, systems, and personnel. All data are automatically saved to UITS's cloud storage server as well as to a local server. Data include database files (SQL), XML files, image files (jpeg, gif, tiff), document files (MS Word, PDFs, html), TEI files, STEP-generated files, and zip files. Files generated by STEP users are stored wherever wanted, on IU servers or elsewhere.

Data formats and dissemination

The software data category will contain a mixture of file types, all with third-party open-source products available for their viewing and editing. Scholarly work data, aside from being both generated and rendered from within STEP and STEP TOOLS, will consist of digital image files and structured data marked up according to platform and/or edition requirements, including TEI-XML mark-up compliance when that applies (one virtue of that mark-up is to ensure long-term interoperability). All STEP-related code will be posted on GitLab.

Pilot scholarly work and other subsequent STEP users will be afforded the customization options necessary for control access to produced data, assuring any legal or private data is available only to those with permission while guaranteeing public access to available humanities collections.

Toward a cloud-based operation for data storage, access, and preservation

The Drupal framework and STEP components are currently installed on IU's Linux virtual machine (VM) server in IU's Intelligent Infrastructure, where a standalone server hosts all services (web server, database, and git); we also use one of IU's cloud storages (box.iu.edu) for extra storage and backup.

We plan to keep the current standalone web service in the IU VM server coupled with a separate cloud server to manage increased storage. We will thus be able to support multiple projects in the current system, projects that can share and integrate the data in the cloud anywhere around the world. No local data are stored in VM itself: all the data are stored in the cloud storage system *(see diagram in Appendix 4)*.

By integrating the current system with an affordable cloud service (e.g. box.iu.edu, or Google Drive), we will ready ourselves to move later on to a globalized cloud solution (cloud + extra support for the standalone web server). Primary service will then come from a cloud interface similar to Google docs, Google sheets, and Google presentation, while global service will support multiple localized cloud services based on distinct languages or the various interfaces of preferred services. STEP will consist by then of a standalone TEI editor, and of a virtual STEP Platform management (workflows) and TEI online editor system. This will allow any number of editorial projects to use STEP. Saving all data to the cloud provides easy access globally, along with unlimited storage space, robust security, considerable scalability, and especially ease of maintenance and thus appreciable savings: only one person will be needed to maintain the whole structure instead of multiple maintenance staff in different places.

8. Data Management Plan

8.1. Expected Data

In the course of the project, we expect to collect or generate three types of data. The first type is social media text data, which will be downloaded either from public social media platforms or from private sources through a paid subscription or with the owner's permission. We will keep a copy of all of the collected social media text data on our project server so that we can reuse them for further analysis. The second type is the software of our tools for generating evolution models, for analyzing dissemination patterns, and for parsing and identifying texts that have evolved from the same text but have linguistic or structural variations. This type of data includes source code, binary executable, design documents and tutorials. The third type are the evolution models for identifying social media texts belonging to the same discourse/dialogue and constructing their temporal evolutionary paths. Variations of the evolution models with annotations of relevant events, such as whether they are deemed "real news" or "fake news," which could inspire particular transformations of their grammatical structure, lexical forms, or semantic content, will also be saved for verification purpose.

8.2. Period of Data Retention

All relevant data will be retained on our data storage server permanently at the University of South Carolina. As new texts are added to the dataset over time, previous texts will still be needed for the computation and refinement of the evolution models.

8.3. Data Formats and Dissemination

All data will be stored electronically. The archived social media texts as well as state files will be converted into an XML file format in both text and binary forms and stored on our server. The source code of the crawling, parsing, and evolution model construction tools will be stored in the version control system (CVS) installed on the server. The binary format will be stored on the server as well. Moreover, the source code and binary format will be posted on open source repositories such as http://sourceforge.net. All tutorials and documents will be in PDF and HTML format and will be available on our project website. The sources of our collected social media text data will be clearly stated in our publications as well as in our software.

Users will be able to download social media text data obtained from public sources and to redistribute them to other parties without restriction. Since we use public sources to build our tools and social media discourse evolution model, we believe there should not be any restrictions on distributing our model. However, if a user wants to access restricted data (such as social media text data obtained from private sources), then the redistribution of such data and generated results will be subject to restrictions imposed by the owner of restricted source. We will put this requirement as disclaimers on our website. In the disclaimer, we will also require the users to cite our Evolutionary Analysis of Social media Texts (EAST) project if they publish results generated

from the data hosted by our project. We will also acknowledge all of the sources, including financial support that made the development of system and tools possible.

Importantly, in order to provide timely access to both the research community and the general public, we will set up an Evolutionary Analysis of Social media Texts (EAST) website to interface users and the data storage server. Although it may take a substantial amount of time to establish a meaningful evolution model of any given event's social media discourse, we will update our website on a weekly basis such that public users can access most recently archived related texts and the latest version of evolution model constructed by our tool. Moreover, we will make the website interactive by designating specific well-known events as its focus and soliciting contributions of related social media texts from general public users. Other data, including social media discourse evolution models and source code of developed tools, documents and tutorials will also be available for free on our website. In subsequent phases, we will expand the scale and construct a public cloud to house all the data.

8.4. Data Storage and Preservation of Access

We have budgeted for one workstation and two 10TB external drives, which will be dedicated for the storage and maintenance of collected and generated data. Moreover, the IT team at University of South Carolina will provide enterprise backup solutions to perform daily backup for all of the data on the servers we use, and the backup will be stored in separate locations that are fire and waterproof. We will require all of the project participants to check in their documents and source code to CVS at the end of the day so new developments can be saved and backed up. The backed-up data are typically stored for up to 30 days, which should be sufficient to recover the data in the event of a data loss.

9. Data Management Plan

Expected Data

The project expects to produce six types of data: (a) software source code, (b) website source code, (c) reports and articles, (d) metadata from case studies, and (e) closed captioning data and (f) still images extracted from television episodes.

Data elements (a) through (d) will be saved as git repositories and published on GitHub. All of these data will be publicly available as they are built by the team. The commit history in the git repository provide a complete summary of the process used to construct all of the data elements. Also, the code for creating the extracted metadata (d) is included in the examples directory of the source code in (a).

Data elements (e) and (f) are being stored using cloud storage provided by Box under an institutional license from the University of Richmond. These elements are under copyright protection and we will therefore not distribute them outside of our core team. If other researchers require access to this material, there is an easy approach. We provide reproducible code in (a) for extracting the copyrighted material from commercially available media formats (typically DVDs). External groups can then recreate (e) and (f) after acquiring a copy of the raw source material.

Period of Data Retention

All of the data that will be made public, elements (a) through (d), will be accessible on GitHub as it is created. There will be no embargoing of the data and no need for a separate process of publishing the data.

At the end of the grant phase, a machine readable version of elements (a) through (d) will also be placed on the University of Richmond Scholarship Repository. The UR Scholarship Repository is a service of the University Libraries at the University of Richmond. Published materials and data included in the digital repository reflect the research and scholarly work of the university community and are openly available to the general public for download and use. The University Libraries provide this service to University of Richmond faculty, staff, and students free of charge and are committed to providing perpetual access to deposited content.

Data Formats and Dissemination

The source code (a) and (b) will be written in code readable by any plain text editor. The DVT Toolkit will utilize Docker for ease of use, but could also be run without Docker by manually installing all software dependencies. The research reports and articles (c) will be written in markdown; this is also readable by any plain text editor and can be converted into a large number of other formats using the open source software *pandoc*. Markdown is also seamlessly converted to HTML by the GitHub platform.

Extracted metadata (d) is saved as plain text files using a comma separated value scheme. This format can be read and parsed by any programming language and most statistical and visualization software. The closed captions (e) are also stored in plain text files. Still images (f) are stored in the JPEG format, which is readily readable and convertible by all major web browsers, image programs, and email clients. We anticipate no need to convert these to another format inside of the next 10 years. All text files will be saved using the UTF-8 encoding.

Data Storage and Preservation of Access

Elements (a) through (d) will be stored on GitHub, with a duplicate copy of the entire repository stored on our internal Box account and on GitLab. The copyrighted material will continue to be hosted on our institutional account with Box. Migration to a new cloud platform in the long-term, if necessary, will be managed with the help of the University of Richmond's IT department. A machine readable version of elements (a) through (d) will be placed on the University of Richmond Scholarship Repository for long-term storage.

Data Management Plan

Roles and Responsibilities

This data management plan will be implemented and managed by Matt Shoemaker, under the project supervision of Peter Logan. Mr. Shoemaker will oversee maintenance, backup, and archiving of data generated by the project. And he will manage the transfer of data for project research to Drexel's Metadata Research Center. He will also be responsible for final project data transfers to the OTA and Humanities CORE repositories. All repository data will be publically accessible. If Mr. Shoemaker leaves Temple University during the course of the grant, his role will be taken over by his successor as Coordinator of the Digital Scholarship Center.

Data Storage Hardware

During the production phase, all data is stored locally in the DSC on its internal server, consisting of a networked pair of 6TB hard drives in a RAID1 configuration. Access to the DSC server is limited to project participants authorized by Dr. Logan. Project files are automatically archived daily to a separate 4TB external hard drive. Files are also synchronized daily with an online Box service provided by Temple University, with remote access limited to team members.

Data Formats

1. Image files. These are duplicates of external image files downloaded from the Hathi Trust or the Internet Archive (see Appendix D).
2. AFR Project Files. These are file folders generated by ABBYY FineReader to organize large amounts of page images for scanning.
3. AFR User Files and Dictionary Files. These are proprietary files that store custom information used to fine-tune the OCR process.
4. HTML files. There are two types:
    a. Those output by AFR, with one file for each print page of the Encyclopedia.
    b. Final files generated by XSLT from the project's TEI-XML data files with full metadata, for uploading to the OTA at the completion of the project.
5. XSLT scripts.
6. TEI-XML files. These contain the core textual data of the project and the generated metadata.
7. Oxygen Project Files. These are small data files for use by Oxygen XML Editor to organize large numbers of XML files.
8. TXT files, generated by XSLT from the project's TEI-XML data files for internal use in testing the viability of using online topic-modeling to identity concept drift.
9. Project spreadsheets documenting the creation of all files and any modifications made to them.

Data Organization Plan

Core Data and Metadata Organization

In order to output consistent metadata with the textual data, our 110,000 individual entry files are organized within a hierarchy of file dependencies, with entry files at the bottom level (see Appendix F). Above them are container, or "wrapper" files for each volume, followed by wrappers for each edition, and finally a corpus wrapper containing basic encoding to all files. Files at lower levels of the hierarchy inherit the attributes of those above them. They also have unique metadata describing the content of each entry. This structure allows us to keep the encoding of entry files as simple as possible, while the series of dependencies means that each entry will have a comprehensive set of metadata associated when output into its final format for repository storage.

Production File Organization

A comprehensive data organization plan for use by project participants is spelled out in the "Data Organization" section of the online Project Manual, specifying the storage location for all forms of project data (https://tu-plogan.github.io/#source/data_organization.html).

Expected Data for Preservation

The project will generate approximately 110,000 different files of textual data. Each file represents one entry in the four Encyclopedia editions. These files serve as "master files" that are used to generate output in other file formats for end-users, including researchers. From the files, we will generate final "digital edition" files TEI-XML format the include all of their critical metadata within each file, and so serve the needs of researchers working from file metadata, rather than the textual data alone.

Permanent Preservation and Access

One copy of these "digital editions" files will by uploaded to Humanities CORE in bulk form, with all files for each print volume contained within a single ZIP or RAR archive. The four editions contain a total of 89 volumes, so the archive will consist of 89 ZIP or RAR files. The textual data will also be output in two other formats: HTML and TXT, the most useful formats for researchers who want to work with the complete data set. The data set in these alternate formats will be uploaded to Humanities CORE in the same ZIP or RAR archive form. Humanities CORE promises permanent storage and open access for data deposited with them.

A second copy of the data will be made publically accessible in perpetuity by the Oxford Text Archive, when they are uploaded at the end of the project. OTA will make them publically available for free as HTML files readable online. Additional details will be negotiated with them, such as whether or not they wish to supply alternate formats, like EPUB or TXT, and the DSC can provide them with those formats.

Five-Year Preservation

All AFR Project, User, and Dictionary files, plus the XSLT and Python scripts used in the production process, will be preserved internally in the DSC for a minimum of five years, to allow us to regenerate the raw textual data at any time. The XSLT and Python scripts used to modify that data and convert it to TEI-XML will also be uploaded to the project GitHub site, for free download by anyone interested (https://github.com/TU-plogan/encyclopedia-project) during the same time period.

Test Data

In the final stage of the process, Dr. Logan and Dr. Greenberg will trial two different methods for identifying concept drift in the data set and write a journal article explaining their results. The data for those tests will be preserved and posted on Humanities CORE, for use by readers of the journal article or others interested in the outcomes. Some of it will also be shared at the DH2019 conference during our presentation.

Continuing Research

The full textual data set and master files will also be retained by Dr. Logan for continuing research on nineteenth-century knowledge. Dr. Greenberg also will retain a copy of the TEI-XML digital edition files for use in her future teaching.

Data Management Plan for *National Breath of Life 2.0: Creating a 'Second Breath' for Indigenous Language Revitalization (BoL 2.0)*

This project results in two major data deliverables: (1) the *Indigenous Languages Digital Archive* (ILDA) software and (2) the language research artifacts produced by the user communities. The following is the data management plan for each deliverable.

ILDA Software

Intellectual Property Agreements

The relationship between Miami University, the Miami Center, and the Miami Tribe of Oklahoma is governed by a perpetual Memorandum of Agreement, which outlines proprietary interests over products produced through the Myaamia Center, such as Miami-Illinois Digital Archive (MIDA) and ILDA. Article Four - Intellectual Property, states:

> *WHEREAS, It is understood by the terms of this agreement that the Miami Tribe, by right of self-determination, has control over its cultural and intellectual property on behalf of the citizens of its nation. Neither the Myaamia Center nor Miami University may copyright materials produced through and/or by the Myaamia Center. No reprinting or distribution of materials produced by the Myaamia Center may occur without the express written consent of the Miami Tribe of Oklahoma. (Memorandum of Agreement signed March 31, 2016)*

The ability of ILDA to meet the needs of other communities who are revitalizing their languages from documentation was tested during the 2017 National Breath of Life Archival Institute for Indigenous Languages (BoL). Given the success of the pilot as outlined in Appendix 2, the proposed BoL 2.0 will offer ILDA through future workshops. Permission is granted by the Miami Tribe of Oklahoma to continue developing ILDA for the purpose of sharing it with BoL alumni (see Miami Tribe letter of commitment). Based on the outcomes of BoL 2.0, additional long-term permission will be required and will be handled through appropriate legal advisement.

ILDA Software Access

ILDA will be made available through a publicly accessible website which will provide read-only access to language research artifacts as the default. Accounts with secure credentials will be provided to BoL workshop trained researchers that enables them to maintain their language artifacts. Each account allows users to maintain the data for their respective research projects only.

ILDA Software Architecture

ILDA (and MIDA) are implemented using the industry-standard web service solution commonly referred to by the acronym LAMP. The 'L' in LAMP refers to the operating system, Linux. The web server is Apache, the 'A' in LAMP. The underlying database is a relational database management system using mySQL (the 'M' in LAMP). The web pages, search logic, and database management code are programmed using a combination of the HTML, PHP (the 'P' in LAMP) and JavaScript programming languages. The use of these industry-standard languages,

operating systems, and applications assures ongoing software support, maintainability, and technical documentation.

ILDA Hosting Environment

ILDA is hosted by Miami University's Information Technology Services, the provider of the University's Internet and Web infrastructure. Miami's IT Services infrastructure includes a fully redundant data center located in Oxford, Ohio. Internet and Internet2 access is provided to the data center by OARnet on redundant connections. Computer servers and associated storage are provided by a VMware vSAN cluster. The ILDA software and language research data is backed up to another physical location on campus using IBM Spectrum Protect. Miami's IT Services provides continuous operating system and select application security updates to the entire infrastructure. Included are Linux operating systems, Apache web servers, and supporting middleware to assure future compatibility and security.

ILDA Software Development Procedures

The ILDA software development environment utilizes industry-standard best practices and tools. GITlab is used for the source code repository and version control. The development process follows the DevOps industry standards. DevOps requires a development site for the programmers, test site for pre-deployment, and a production site for the release version. The test site is used for user testing before updates are released to production. This process is consistent with Miami's IT Services development and operational policies. It assures that ILDA is developed following robust programming, testing, and release procedures.

ILDA Project Management

To mitigate risk and be responsive to user needs, the ILDA developers use an Agile project management approach. Agile practices include close communication with users, iterative development, frequent releases, and continuous improvement of work practices. Dr. Troy is an accredited Agile instructor by the International Consortium for Agile (ICAgile).

Language Research Artifacts

Three main types of data constitute the language artifacts for each community: (1) digital image surrogates of archival language documents, (2) alphanumeric data such as transcriptions, translations, and linguistic analyses, and (3) digital audio recordings. The management of each is described below.

Intellectual Property Agreements

The management of intellectual property rights for language research artifacts, including obtaining permissions from archives, rests with and is the responsibility of each of the participating user communities and will be defined in an end user license to be developed to support this project.

Digital Image Surrogates

Digital images are high resolution JPEG images of archival source documents. Each user community will be trained in the process of preparing and uploading the images corresponding

to each document to be analyzed. Images are indexed by source document, page, line, and phrase. Once uploaded the images are stored in the ILDA Linux server directory structure, which is organized by document within each participating community section of the public website.

Alphanumeric Data

This data consists of transcriptions of source documents, translations, and linguistic analyses. This data is stored in the ILDA mySQL database. ILDA provides for bulk upload and download capability, allowing users to export this data from the database at anytime. This capability can be used to migrate information out of ILDA, if needed. It can also be used to perform bulk edits, with many changes being subsequently uploaded in bulk.

Audio Recordings

Audio recordings are stored as MP3 (MPEG) files. Audio recordings are stored in the ILDA Linux server directory structure similar to the way that digital images are stored.

Language Artifact Storage and Backup Procedures

As described above, the The Miami University hosting environment provides secure data storage in its fully redundant data center. The Linux file system and the mySQL database are backed up daily.

Conclusion

The ILDA software will be available to the research community through a public website. The ILDA software and database is implemented, hosted, and backed up using industry-standard practices with support from Miami University's Information Technology Services. Data produced by the community researchers, stored in the ILDA database and Linux file system, is secured through password-protected accounts and backed up daily. Bulk download capability allows community researchers to migrate data out of ILDA if necessary.

DATA MANAGEMENT PLAN

INDIGENOUSMAP: MAPPING INDIGENOUS AMERICAN CULTURES AND LIVING HISTORIES

Data Description

The data produced by this proposal will include an open-access, digital map of indigenous nations in the region of the United States, including names, boundaries, languages, and other data of the Osage, Modoc, and Pomo/Miwok nations. Data will be stored and made accessible in the form of geographic coordinates with additional data attached. These coordinates may be points or polygons depending on the needs of the data. The data attached to these coordinates may vary in content but will relate to different cultural and historical data gathered during research. A number of different files will store collections of geographic data relevant to different areas of research.

Access and Sharing

The data is intended to be available for open source public usage through an interactive map application on a website. Any restricted data will not be included in the public-facing application. Files in form of geoJSONs will be available from a Github repository that also contains the interactive map, similar to http://native-land.ca (https://github.com/tempranova/Native-Land). Once the data is standardized, ESRI shape files will also be made available on the website itself. Data will be made public after processing into categories of files and after the interactive map is fully developed and released. As data changes during research and work is saved, past archives of files will be automatically stored in the Github repository.

Metadata

Documentation explaining the categorization of files and the types of data available in each file will be available in a readme in the Github repository, as well as on the website. The data will be stored as fully standardized geoJSONs and ESRI shape files.

Intellectual Property Rights

The data gathered will be available to all users for public use under open source licensing such as GNU and Creative Commons.

Format

The public-facing mapping application will use the open source mapping library, Leaflet, to build a full-window application that allows the user to toggle between different views of the gathered data. This will be built using ReactJS and hosted with a university server. The data powering the map will be saved and accessed as geoJSONs.

Storage and Backup

Throughout the research process, data will be backed up daily on local machines and shared using Google Drive and Github. As additions are made to geographic data, the Github repository will be updated, allowing access to older versions if they are needed later.

Responsibility

Janet Hess will be responsible for research and obtaining data to attach to geographic coordinates through interviews and other methods. Victor Temprano will organize and standardize the data in line with geographic norms, and will develop the front-end website and interactive map.

Existing Data

There has been some work done previously on Indigenous land use and cultural history, such as nativemaps.org, hundreds of individual tribal websites, and large works like native-land.ca. All of these are potentially useful, but the scope of this project is more tightly confined to gaining deep research into specific tribes with tribal collaboration. This project will be released into a growing ecosystem of geographic data related to indigenous history, and has the future potential to consolidate existing data systems in a central site.

Data Management Plan

The Roy Rosenzweig Center for History and New Media (RRCHNM) at George Mason University and partners at the World History Association and Monash University (Australia) are committed to open-source software development and open access for all data and content. The Project Director and Principle Investigator, Dr. Kelly Schrum, will ensure that all of the software and digital assets created for this project are published for public use as soon as they are available, and will be responsible for assuring that the project team adheres to this data management plan throughout and beyond the grant period.

RRCHNM has a long history of institutional commitment to open-source software development, open access, and long-term sustainability. NEH funding for the last of the *World History Matters* projects <worldhistorymatters.org> ended almost a decade ago, but RRCHNM has continued to host and maintain all of these world history websites as Open Educational Resources for the scholars, teachers, and students who visit more than 7 million pages annually. RRCHNM will make the same commitment to *World History Commons*; users will have access to all project data and resources after the grant funding ends. *World History Commons* will be licensed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0), ensuring a free, open resource for all users.

The *World History Commons* database-driven website will generate three types of data: software code, documentation, and digital content, including text, images, and multimedia.

Software Code and Documentation: RRCHNM is a major contributor to open-source initiatives as well as a proponent of code re-use internally and externally. *World History Commons* will be developed using open-source languages and platforms that are modular, robust, secure, and well supported, including Drupal, JavaScript, HTML 5, CSS, PHP, and Apache. The source code will be documented and made freely available for download, reuse, and modification in a public repository at GitHub, a cloud repository service widely used in software development. All documentation will also be available for download, reuse, and modification in a public repository at GitHub.

Digital Assets: All existing digital assets located in the RRCHNM *World History Matters* projects as well as in the *Global History Reader* will be reviewed for revision and inclusion in *World History Commons*. This includes primary sources, case studies, teaching modules and guides, multimedia historical thinking and scholar interviews, annotated bibliographies, and in-depth topical essays. New digital assets will be created throughout the life of the project as well. All digital assets will be freely available and the project as a whole, including all digital assets, will be licensed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)

White Paper: The White Paper as well as any reports will be available on the *World History Commons* website. They will also be shared with a broad audience via the RRCHNM blog.

Data Management Plan

1. Roles and Responsibilities

The data management plan will be implemented and managed, at different phases, by all members of the project team. Scott Graham and Dave Clark will be responsible for executing the data management plan during the development of the T2V toolkit. Ann Hanlon will assist with transfer of toolkit documentation and data to the UWM Digital Commons which will ensure permanent storage.

2. Expected Data

Data for this project falls into three broad categories that include: 1) the prototype dataset on conflicts of interest in biomedical publishing, 2) user-experience data developed during T2V toolkit prototyping, and 3) the T2V toolkit itself and associated documentation.

1. The prototype dataset includes conflicts of interest disclosure statements that have been made publicly available on biomedical journal websites and corresponding bibliographic information. This textual data has already been curated and is currently hosted on the PI's lab server.
2. User-experience data will include notes on and correspondence from test users at various phases of the T2V toolkit development. All user-experience data will be text files.
3. Expected data also includes the T2V toolkit code and corresponding documentation including the white paper and associated academic publications.

3. Period of Data Retention

All data for this project with the exception of user-experience data will be deposited with the UWM Digital Commons in perpetuity. User experience data will be retained until the end of the funding period and then destroyed.

4. Data Formats and Dissemination

A working T2V prototype will be made available on the project website which will be hosted on the UWM web servers. This website will also serve as a portal to all project data. T2V documentation, data, and metadata will be deposited on the UWM Digital Commons. T2V source code will also be deposited and freely available at GitHub

5. Data Storage and Perseveration of Access

All project data with the exception of user-experience data will be deposited with the UWM Digital Commons. The UWM Digital Commons is a virtual showcase for research and creative profiles, administered by the UWM Libraries. The UWM Digital Commons provides a dedicated infrastructure for long-term preservation (including redundant storage and data recovery protocols) and worldwide electronic accessibility.

9. Data Management Plan

Data Sharing

For each research project, typically an excavation at a particular site, DRC collaborators enter context and artifact data into the DRC Application and its PostgreSQL backend. The primary means by which data are shared with scholars and the general public is the DAACS website (www.daacs.org), which is built in WordPress. The Principle Investigator (PI) for each project or site works with DAACS staff to create a "home page" for each project and link to it discursive background historical information, archaeological chronologies, stratigraphic summaries, and overall digital site plans, sections, and photographs (*e.g.*, https://www.daacs.org/sites/sugarloaf/#home) After final checks are complete and the site's PI has signed off, data for the project in the PostgreSQL backend are made available to the query module on the DAACS website (https://www.daacs.org/query-the-database/).

The query module gives users access to a point-and click interface that generates custom SQL queries against the PostgreSQL backend. Users can query on chosen sites and artifact classes and chosen variables that describe them and have the data returned at different levels of aggregation, ranging from individual excavation contexts to entire sites.

Based on user choices, the query module can deliver fine-grained data on artifacts and their excavation contexts from related tables, in which each record represents a different attribute value on an artifact, and single artifacts, identified by a unique DAACS artifact ID number, span multiple records. This is useful, for example, in the study of correlations among decorative motifs on ceramics. The query module can also deliver coarse-grained data as well. For example, site-level assemblage summaries, in which each record represents a broad artifact class (*e.g.*, cut nails and its total count). The new "DRC Silver" and "DRC Bronze" cataloging modes will require catalogers to record values for smaller subsets of the complete set of "DRC Gold" fields. The DRC Advisory Committee will help us identify the fields that will be required for each mode. Finer grained data will not be available for sites catalogued in Silver and Bronze modes. But the benefit of the tradeoff is the ability to catalog larger assemblages on tighter budgets.

Data Types

Below is a summary of the kinds of data types cataloged into, and managed, by the DRC Application and the DAACS website.

Archaeological Field Records: DRC-certified archaeologists enter all information found on the original archaeological field records, including sediment descriptions for layers, and the stratigraphic relationships among those layers, into the DAACS database's context tables and related tables. This process creates easy to search records that are linked to each artifact from that context. DRC partners link photos and scanned paper field records to their corresponding context records. All context data are available for download in .html format through the *Query the Database* section of the DAACS website.

Artifact Records: The structure of the DAACS database allows for recording of detailed information about individual artifacts. DAACS classification and measurement protocols are described in online manuals. Detailed guidelines in written form ensure consistency among catalogers and provide researchers with an opportunity to understand how the data they seek

to use were generated. DRC catalogers and analysts undergo specialized training and testing in these protocols to ensure that they are implemented accurately across all collections. All artifact data are available for download in .html format through the *Query the Database* section of the DAACS website (https://www.daacs.org/query-the-database/).

Artifact Images: All unique, illustrative, or diagnostic artifacts and all artifacts exhibiting any sort of post-manufacture modification are digitally imaged. A set of DAACS cataloging protocols specifies how these artifacts should be imaged, named, and stored. Depending on size, select artifacts are either scanned or photographed using a professional camera setup. Images are uploaded and image records are linked to the appropriate artifact and/or object record. They are served to the public through the Query-the-Database section of the DAACS website.

Context and Other Excavation Images: Existing photographs and slides are scanned at 300 dpi, saved as archival .tiffs and .jpgs, and are linked to the appropriate project and context records in the database. They are served to the public through the Query-the-Database section of the DAACS website.

Site Maps, Excavation Plans and Profiles: Hand-drawn site maps and excavation plans and profiles are scanned following established DAACS protocols at 300 dpi and saved as archival .tiffs and .jpegs. Plan and sections of individual contexts are linked in the DAACS database to the appropriate context records. Composite site plans are then digitized in vector format using a CAD program from the mosaics of the scans, following standard DAACS protocols to ensure consistency in the depiction of particular subsurface features, above-ground, extant architectural elements, and excavation areas across all sites launched online. Vector plans are saved for delivery on the DAACS website in .dgn (Microstation) and .dxf (AutoCad) formats. Site plans in .pdf (Adobe) also are made available online for download.

Metadata Resources: DAACS maintains detailed metadata on the DAACS data structures and the process for creating the archaeological data in DAACS, such as cataloging manuals, and tutorials on how to use the data. These are shared on the DAACS website (https://www.daacs.org/about-the-database/daacs-cataloging-manual/)

Source Code

As detailed in the Final Product and Dissemination section of the Narrative, all source code for the DRC interfaces and Open API will be documented and made freely available for download, reuse, and modification in a public repository at GitHub.

Storage, Maintenance, and Protection of Data

Currently, the DAACS database backend and all linked artifact and context images are housed on a dedicated PostgreSQL server located at the University of Virginia's Institute for Advanced Technology in the Humanities. The entire server is backed up daily on two Quantum VS1 DLT tapes. The DAACS PostgreSQL database is not directly connected to the Web. The DRC Application website (www.daacsrc.org) and PostgreSQL database are tarred and gzipped to another server not accessible to the public after every update. This process adds a third layer of backup and allows previous versions of the website to be retained for historical purposes. Finally, the DAACS PostgreSQL database is also manually backed up prior to any database maintenance or updates. This

process of data storage and maintenance has worked seamlessly since 2014. We will continue using this process for all software and data produced as a result of the *Expanding DRC* project. The DAACS website is built in WordPress, an open-source content management system. It presents, in HTML, discursive sites summaries and the items listed above.

Data Management Plan

Bodies on the Move: Panorama of Caribbean Carnival will produce data generated from presentations and moderated discussions in workshops and webinars, as well as identified archived and curated materials available in several collections held by the University of Puerto Rico, and the Puerto Rican Humanities Foundation/Fundación Puertorriqueña de las Humanidades. *Bodies on the Move* is being developed specifically to support new research needs in the humanities as made possible through access to these materials. As such, data management is a critical concern for the file level preservation and access, and for preservation and access to the materials in context.

For success, all materials for the project need to be shared openly and as widely as possible, and the infrastructure must be in place for ongoing growth. These supports are in place from both the Digital Library of the Caribbean (dLOC, www.dLOC.com, an international collaborative digital library), as well as the University of Puerto Rico-Rio Piedras Main Library. Both will provide expertise on metadata, and provide a stable and secure digital archive for the files. Accordingly, both institutions are committed to providing a stable and secure website for the materials in context, as well as the guidance and support for the contextualized web resources to be defined in part through the collaboration among the project participants and the consultants during the proposed level one, exploratory stage.

An open-access website generated by the project to showcase the concept of movement in Caribbean carnival will provide the proponents of this project the means to divulgate text, video and curated materials as well academic papers and research findings that result from it. Data from this interdisciplinary project will provide access to archived and curated materials related to the movement and Caribbean carnival to scholars and community members interested in the Humanities, particularly Language, Art, History, Human Communication, Cultural Studies, and Performance, among others. The University of Puerto Rico is currently developing a system-wide repository aligned with open-source licenses (Creative Commons) and is committed to allow the reuse of data produced by researchers.

The investigators commit to openly sharing all data in a timely manner. The Digital Library of the Caribbean (dLOC) will provide an immediate open access repository and long-term digital repository for all project materials including materials from the archives as well as project documentation, tutorials, and other materials. dLOC commits to archiving and making materials accessible on an ongoing basis and at project end. This is in keeping with normal practices with dLOC's commitment to open and speedy dissemination of research and project materials to build the community of practice and resources for Caribbean Studies. dLOC is supported by the over 40 partner institutions whose representatives comprise the Executive Board and two host institutions: Florida International University is the Administrative Host and the University of Florida is the Technical Host. UF dedicates staff time to digital preservation and access from the Digital Library Center staff, Preservation staff, Programming & Web Services Team, Digital Librarian, and others. dLOC provides resources online, a robust training program, and continuous consultation and facilitation for collaboration across the region for conducting digitization and digital curation in adherence to digital preservation standards. The project will generate a variety of data materials, and all will follow existing national and international standards for access and preservation, including for multiple file derivatives as applicable. Training and support materials will be stored in standard formats (e.g, HTML, PDF, AVI, PPTX, etc.) with all files made available in their optimal format for access and preservation.

As the dLOC Technical Host, UF is committed to long-term digital preservation of all materials in dLOC. Redundant digital archives, adherence to proven standards, and rigorous quality control methods protect digital objects. UF provides a comprehensive approach to digital preservation, including technical support, reference services for both online and offline archived files, and support services by providing training and consultation for digitization standards for long-term digital preservation. The UF Libraries support locally created digital resources as powered by and hosted with the SobekCM Open Source Repository Software, including the UFDC which contains over 528,000 digital objects with over 36 million files (as of June 2017). The UF Libraries create METS/MODS metadata for all materials. Citation information for each digital object is also automatically transformed by the SobekCM

software into MARCXML and Dublin Core. These records are widely distributed through library networks and through search engine optimization to ensure broad public access to all online materials.

In practice consistent for all digital projects and materials supported by the UF Libraries, redundant copies are maintained for all online and offline files. The digital archive is maintained as the Florida Digital Archive (FDA) which was completed in 2005 and is available at no cost to Florida's public university libraries. The software programmed to support the FDA is modeled on the widely accepted Open Archival Information System. It is a dark archive and supports the preservation functions of format normalization, mass format migration and migration on request. As items are processed into the UFDC for public access, a command in the METS header directs a copy of the files to the FDA. The process of forwarding original files to the FDA is the key component in UF's plan to store, maintain and protect electronic data for the long term. If items are not directed to load for public access, they do not load online and are instead loaded directly to the FDA (*more information*). Further, this technical processing is supported through the dLOC operations and by-laws, which also stipulate that ownership of all dLOC materials remains with the partners, and that UF commits to serving as the technical liaison and facilitator for partners even in the event that dLOC were to cease to exist.

Because dLOC partners understand that data management is an ongoing part of the data lifecycle including access, especially a project such as this where the materials are to be shared and built upon to create new possibility for research, UF's role as the Technical Host for dLOC also includes facilitating digital scholarship. UF is committed to long-term digital preservation of all materials as well as to digital scholarship projects, including facilitating collaborations and creating new engagement opportunities with Digital Humanities, Public Humanities, and other groups with new and emerging collaborative networks for Caribbean Studies Digital Scholarship.

DATA MANAGEMENT PLAN

Efficient sharing of data, scholarship and research tools is essential to the successful promotion of research in the academic community. All research supported by this grant will be made available in a timely and responsible manner (see Final Product and Dissemination above) through presentations at academic conferences, publishing in peer-reviewed scholarly journals, a White Paper for the NEH website and Digital Commons, and the provision of wide access to and publicity about the open-source tools. During the grant term, at server at Northwestern in Information Technology will host and run Gentle and Drift, linked to the website of Northwestern University's Sound Arts and Industries program. This server will clear all audio files uploaded for analysis every 24 hours, as well as data about those audio files. Individual user-testers on the project team and GLASS members will be responsible for backing up data about their datasets of audio files, and disseminating their research on them through presentations at academic conferences and publishing in peer-reviewed scholarly journals.

At the end of the grant term, the source code and documentation for the integrated package of Gentle and Drift will be available on Github. The package will be available for download and installation with a MIT License (a very permissive open-source license type for free software, commonly used on GitHub) on the website of Northwestern University's Sound Arts and Industries program, which is committed to developing and facilitating access to open-source tools for the digital humanities. Its website will also include links to the project's White Paper, as well as to any resulting research articles and reports, a web-based demonstration version of Gentle and Drift, and the source code and documentation for Gentle and Drift on Github. Links to the source code will also be posted with permission on the websites of PennSound, I-CHASS (Institute for Computing in Humanities, Arts and Social Sciences), HASTAC (Humanities, Arts, Science, and Technology Alliance and Collaboratory), and Digital Commons. We will also ask Alan Liu to include links to the web-based version of Drift and the source code and documentation for both tools on his DH Toychest/Digital Humanities Resources for Project Building website.

Data management plan

Roles and responsibilities

Patrick Murray-John will be responsible for maintaining code in GitHub, and for providing current regular releases of the software generated. The most current stable code will be available on the 'master' branch with the current listed release, while development or experimental code will be available on other git branches. Permissions to modify code and releases will be managed by Murray-John, or another representative of the Omeka team should he no longer be able to do so.

Megan Brett will create and maintain documentation, which will be written and maintained on Omeka S's end-user documentation site on GitHub <https://github.com/omeka/omeka-s-enduser> and published on Omeka's official site <http://omeka.org>.

Expected data

Code:

Omeka-ORCID Integration will produce three Omeka plugins. That code will be posted and maintained, together with all Omeka code, on GitHub, <http://github.com/omeka-s-modules> from the earliest stages of the project development. Developers can follow the progress of the plugin development, fork the code, comment on the development, and submit bug reports. Final versions of the plugins will be posted as downloadable .zip files on Omeka's site <http://omeka.org>.

Documentation:

Documentation will be written and maintained in the Omeka S end-user repository on GitHub
<https://github.com/omeka/omeka-s-enduser>, where it will be available for pull-requests for the community to suggest improvements or corrections. The current state of the documentation will be published on Omeka's site <http://omeka.org>.

Period of data retention

Omeka S, all of its modules, and documentation will be available and maintained on GitHub from the
beginning of the project forward.

Data formats and dissemination

Code:

All code will be written in PHP, javascript, and CSS. It will be publicly available on GitHub <https://github.com/omeka/omeka-s-modules> under a GPL V3 license at all stages of development. Final products will be packaged for easy distribution on Omeka's site <http://omeka.org>.

Documentation:

Documentation will be written in markdown, and maintained on GitHub. User-friendly publication of the documentation will be available on Omeka's site <http://omeka.org>.

Data storage and preservation of access

All code and documentation will persist on GitHub in public repositories under a GPL V3 license.

9. Data Management Plan

Roles and Responsibilities

This data management plan will be implemented and managed by Christine Ruotolo, in consultation with the Content Stewardship team at the University of Virginia Library. The University of Virginia Library will have long-term responsibility for the permanent retention of encoded texts and image files that comprise the bulk of the project data, as well as for preserving a static archival version of the project website.

Expected Data

The following data will be produced in the course of the project:

- TEI-encoded text files in XML format
- Representative page images associated with each text file, in JPG format with accompanying metadata
- A customized TEI schema and project-specific templates for text encoding
- A customized web interface, based on WordPress and incorporating the Anthologize plugin
- Technical guidelines and workflow documents

Period of Data Retention

The TEI-encoded text files created by the project and their associated metadata will be deposited in the University of Virginia Library's digital repository and will thus be added to the Library's permanent digital collections. The project website will be actively maintained for a minimum of five years beyond completion of the project. Workflow documentation and reports will be retained on GiHub for a minimum of five years beyond completion of the project.

Data Formats and Dissemination

The texts created for the project will be encoded in TEI and stored in XML format. They will be delivered online through a customized instance of WordPress, and will be made available to the public as they are completed. Descriptive metadata will be contained in the TEI header, which can easily be converted to other standard formats as needed. Representative page images created for the project will be stored as JPGs and described using Dublin Core or another widely recognized metadata standard.

Web hosting for the project website will be provided by the University of Virginia Library, which is currently investigating long-term hosting solutions for developing and preserving digital projects created by faculty.

Data Storage and Preservation of Access

Upon project completion, the WordPress site will be crawled using the Archive-It web archiving service, creating an archival snapshot of the finished project that will be added to the University of Virginia Library's Fedora-based digital repository for long-term preservation. The TEI-encoded text files will also be

ingested into the digital repository as stand-alone objects, and will be publicly accessible through the Library's discovery interface. The ultimate preservation destination for these materials is the Academic Preservation Trust (APTrust), of which the University of Virginia Library is a founding member.

Other data created by the project will be stored and disseminated via GitHub, a publicly accessible code repository. These data will include page images, schemas and markup templates, customized software, technical guidelines, and workflow reports.

9. Data Management Plan

A major drive of the proposed work is to establish a new platform for intertextual search over large collections of digitized documents. While our proposed work includes our own initial forays into this new space, we anticipate the greatest possible impact will be achieved by fully engaging and leveraging a much broader community of investigators in the spirit of “open science”. As such, a significant component of the proposed work is the creation and development of publicly-sharable material — both datasets and software.

This plan for the NEH is largely inspired by the National Science Foundation’s policy on the dissemination and sharing of research results within a reasonable time. In accordance with this policy, this plan does not include preliminary analyses (including raw data), drafts of scientific or humanistic papers, plans for future research, peer reviews, or communication with colleagues.

Furthermore, data to enable peer review and publication/dissemination and/or to protect intellectual property may be temporarily withheld from distribution and other proposed data management. This plan will make certain that the data produced during the period of this project is appropriately managed to ensure its usability, access and preservation.

Deliverables

Upon its completion, the project will make available the following six deliverables:

1. The core RESTful software service code package based primarily in Python, hosted on the Tesserae Github repository. The service will allow users to input two or more texts for comparison and a set of parameters, and return a list of parallel passages with the similar words or other language features marked. The languages serviced will be ancient Greek and Latin.

2. An operational version of the service hosted at UB, Notre Dame, and UCCS, to be used by the Tesserae project, as well as partner digital collections and interested maintainers of other digital collections.

3. A plugin to the Plokamos annotation framework for making the core Tesserae service compatible with its operation. This will be published on the Tesserae Github repository and presented to the Plokamos project for inclusion in its code base.

4. Interface code for making the core Tesserae service compatible with the Perseids collaborative editing platform via Plokamos. This will be published on the Tesserae Github repository and presented to the Perseids project for inclusion in its code base.

5. Interface code for making the core Tesserae service available to Open Philology, the Perseus Digital Library, and the Digital Latin Library via Perseids. This will be published

on the Tesserae Github repository and presented to each collection for inclusion in collection-specific code bases.

6. Documentation of the new software service, interface modules, and Tesserae front-end, both accompanying the code on Github and hosted on the Tesserae website.

Data and Code Sharing Timeline

We anticipate regular releases of data and code as they are completed. In keeping with the overall schedule presented in the timeline found in the proposal narrative, we anticipate the release of the TIS, TIS interface to Perseids, and the TIS interface to Plokamos to take place in the second year, ideally coinciding with the initial publication of results. We anticipate the initial releases of the interfaces to the Open Philology Project, Perseus Digital Library, and the Digital Latin Library at the end of the second year. Algorithm development will track the progress of each task, with stable code releases at the end of each project year, along with scientific papers describing our advances in intertextual search.

Data Management Plan

Roles and responsibilities
Prof. King-Ip Lin and Prof. Elise King will both oversee the data management plan for the project. Prof. Lin will focus on the software that is developed throughout the project, while Prof. King will focus on the management of the floor plan data, together with the interaction with University of Texas Alexander Architectural Archives for their floor plan data.

Expected data
The following type of data/information will be generated from the project:

1. Source code for BuDAS. A final, working version of the source code of BuDAS will be made available on GitHub at the end of the project. We anticipate BuDAS will be using other public-domain source code. That information, together with the location of the necessary code, will be included in an installation guide that comes with the software on GitHub.
2. Sample data for testing. We will include a set of floor plans as sample data. The data will come in various formats, all of them will be available via GitHub as well as the Texas Data Repository (see section 5):
    1. Original floor plan images (PDFs)
    2. Data that is transformed by BuDAS into the database (stored as a text file containing SQL statements to construct the database)
    3. Ground truth: all the information that is supposed to be captured (e.g. room relationships for each house) will be stored in a text file
3. White papers, reports, and publications will be made available through a web page dedicated to the project. The page will be managed by the Project Directors as well as the system support staff at the Department of Computer Science at Baylor University. We will also make these items available via the Texas Data Repository.

Period of data retention
All the data and any publications will be made available as soon as they are created. The prototype BuDAS system will be made available when tests of stability have been completed to ensure it runs smoothly.

Data formats and dissemination
As mentioned previously, all the data will be made available on-line, via GitHub and the Texas Data Repository.

Data storage and preservation of access
The software will be stored on GitHub, which is a third-party open repository that allows software and data to be stored. We will provide links on our web pages to ensure search engines are able to discover and locate the repository for the project.

In addition, to ensure that the data will be discoverable by the larger research community, our data will be curated in the Texas Data Repository (http://data.tdl.org/) through the Texas Digital Library. The Texas Digital Library (TDL) is a consortium of academic libraries in Texas, providing shared technology services to support secure, reliable access to digital research collections. The Texas Data Repository is a project of the TDL to develop a statewide research data archive for researchers at Texas universities. Data will be curated in the repository following accepted standards. Data in the TDR are assigned DOIs, which

enables the citing of data by anyone who uses it. Metadata for the project data will reflect best practices developed by the Library of Congress (Metadata for Digital Content [MDC]) and will provide information on subject, provenance, authorship, methods and post-processing, and copyright that will support discoverability, curation, and preservation of the collection, as well as accessibility via harvesting and APIs.

9. DATA MANAGEMENT PLAN

9.1 Gallaudet University

Gallaudet Technology Service, as the home institution, will continue to provide their team's expertise in hosting needs, harvesting digital content in an institutional repository, and facilitating the transition from CraigInteractive (current dsdj.gallaudet.edu) to Michigan Publishing (MP). The data from past issues will be stored in three different secure locations: the GTS server, the DSDJ server, and two dedicated external hard drives.

9.2 Michigan Publishing

Michigan Publishing is responsible for hosting, disseminating and preserving the Deaf Studies Digital Journal. All products and deliverables created during the course of this project will be made freely and publicly available.

Expected Data

The data produced by this proposed project will include the hosted website, source code for accessibility features, video files, audio files, text files, and images.

Data Formats and Dissemination

Michigan Publishing and its parent, the University of Michigan Library, will from the outset create all files related to the journal in open and migratable formats, including:

1. The web version of the journal will be rendered in HTML 5. This format is based on an open W3C standard and will be trivial to migrate to new formats in the future.
2. The code base for Fulcrum as well as additional source code developed through the proposed project will be documented and freely available for download, reuse, and modification in a public repository at GitHub, a cloud repository service widely used in software development. https://github.com/mlibrary/fulcrum
3. Video: All videos created for this project will be encoded in MPEG-4/H.264 format, the most commonly used video compression standard on the web. The ubiquity of MPEG-4 video means that commercially-available tools exist, and will continue to exist, for transcoding it to new formats as technology changes.
4. Audio: If audio files are created for this project, they will be encoded using a commonly used format such as MPEG-3, so that they may easily be transcoded to new formats in the future.
5. Text: ePub 3.0, a related W3C standard to HTML 5, will be the file format into which each transcript is encoded, and it will be migrated to future formats as needed.
6. All images will be encoded in JPEG, TIFF, or PNG formats for easy recompression into new formats for future needs.

Data Storage and Preservation of Access

The journal will be preserved as part of the University of Michigan Library's digital collections. Michigan Publishing is committed to preserving content indefinitely, and will continue to migrate content even after a journal has ceased to publish.

Server Infrastructure

The Fulcrum platform, which will host the Journal, is based on the open source Hydra/Fedora technology stack, which is actively developed by more than 30 libraries across the world. This codebase, in this instance, will be run from University of Michigan's enterprise-level UNIX server infrastructure, and all data will be backed up nightly and replicated offsite. Access to this server environment is

controlled by the University of Michigan's Kerberos authentication software with two-factor authentication mandatory.

Fulcrum is hosted on two complete self-contained environments in two different data centers. Our primary data center is a tier 2+ facility off campus with 2+1 redundant generator power. Our secondary data center is a tier 1+ dedicated server room in the library and uses the more reliable campus power generation rather than commercial utility. Our load-balancing is active-active, meaning both sites are in operation, and if one goes down, the other keeps serving content. The sites are kept in sync with a mixture of formal software release processes and automated synchronization, aiming for 99.9% uptime.

Data Management

In addition to the materials digitally repatriated during phase one, the project will develop a significant number of new digital objects. These new materials will be integrated into the digital infrastructure established in each community. Please see below for a list of the digital resources that will be created and the data management plans within each institution (please see appendix 2, "References to Earlier Works," for more information on the materials that were digitally repatriated to each community during phase one of the project). The "Digital Humanities from an Indigenous Perspective" white paper will be designed in Scalar using Murkutu as the content management system and maintained as part of Penn's ScholarlyCommons, which is part of the Digital Commons Network.

Deyohahá:ge: Indigenous Knowledge Centre Digital Archive at Six Nations Polytechnic

- Six Nations Polytechnic (SNP), an accredited post-secondary school in Ontario, maintains redundant servers with its own IT staff. The materials from the National Anthropological Archives and the American Philosophical Society digitally repatriated during phase one, have been processed by Tanis Hill, head archivist for the Dayohaha:ge: Indigenous Knowledge Centre and data management specialist for the current grant proposal.
- All of the digital videos created for this project and the materials digitized at the Canadian Museum of History will be stored on servers maintained by SNP. The Dayohaha:ge archive is open to the public
- The "Haudenosaunee History" DH project will be overseen by project manager Rick Hill and will be created in partnership with the Price Lab at Penn, which will use Scalar for the web design and Murkutu as the content management system. The DH project will be made available through Penn's ScholarlyCommons, which is part of the Digital Commons Network and other social media resources at Penn.

Junaluska Museum at the Eastern Band of Cherokee Indians (EBCI)

- During phase one of the project, the archival materials digitally repatriated from the American Philosophical Society (APS) and the National Anthropological Archive were stored on redundant hard drives located in the Kituwah Education and Preservation Program (KPEP), which is maintained by the Eastern Band of the Cherokee Indians (EBCI). Because the EBCI have a casino, the tribe has the financial stability to support the long-term maintenance of the servers.
- The tribal government is currently building a physical archive where all of the digital assets will eventually be housed.
- All of the newly generated digital materials--the 360-degree images from the Penn Junaluska Museums and the archival materials digitized from the NY State Archives will be housed on KPEP's redundant servers.
- The "Cherokee Stickball" DH project will be overseen by project manager T.J. Holland and will be created in partnership with the Price Lab at Penn, which will use Scalar for the web design and Murkutu as the content management system. It will be made

available through Penn's ScholarlyCommons, which is part of the Digital Commons Network and other social media resources at Penn.

- "*Tsalagi Uweti*: Traditional Landscapes Mapping Project" DH project projected is being funded by a grant from the Cherokee Preservation Fund and will be housed and maintained on KPEP's redundant servers. When completed, it will be made available to the public through Western Carolina University and EBCI, both of whom are partners on the project.

Tuscarora Historical Society

- During phase one of the project, the archival materials digitally repatriated from the American Philosophical Society (APS) were stored on redundant hard drives located in the Tuscarora School and at the Tuscarora Historical Society.
- Both portable hard drives will be replaced with Drobo 5C 5-Drive DAS with Seagate 3TB internal hard drives. EPIC will assume the cost of maintaining and replacing the Drobo system every five years.
- The newly digitized language materials will be integrated into a database that will housed on the Drobo system.
- Discussions are currently under way with Tuscarora School about moving the digital assets onto servers maintained by the school's IT personnel. Storing the materials on portable hard drives is not an ideal situation, obviously, but it is an important part of the study of how digitally repatriated archival resources are used by Indigenous communities, many of which lack funding to maintain sophisticated digital infrastructure. It also reflects the challenges that arise when considering issues of cultural sensitivity and the sovereignty of Indigenous Nations that will form an important part of the white paper.
- Before the database is made publically available, the digital resources will be vetted by the Tuscarora History Committee to protect the culturally sensitive materials related to Tuscarora religion.

Ojibwemowining Center at Fond du Lac Tribal and Community College (FdLTCC)

- FdLTCC supports the Ojibwemowining Center and has redundant servers maintained by the college's IT staff. All of the materials repatriated during phase one and the Frances Densmore database created by a previous NEH grant are currently stored on redundant servers maintained by the College.
- Many of the songs Densmore recorded were part of Midewiwin ceremonies, so they are considered to be extremely sensitive by the community. The vetting process is ongoing, but when complete the approved songs will be available through the Ojibwemowining site and Ojibwemowining's vimeo site
- "History of Ojibwe Women's Music" DH project will be overseen by project manager Lyz Jaakola and will be created in partnership with the Price Lab at Penn, which will use Scalar for the web design and Murkutu as the content management system. It will be made available through Penn's ScholarlyCommons, which is part of the Digital Commons Network. The project will also be accessible through the Ojibwemowining site and its social media resources.

DATA MANAGEMENT PLAN

Data management to date

During the startup phase of the project, data and programs recovered from Lawrence University were documented and deposited in MINDS@UW, the institutional repository of the University of Wisconsin-Madison (where Burkert completed her doctoral work in 2016). The scans Daland produced of the printed code base were deposited in the same location.[5] UW-Madison hosts the items under a non-exclusive license and has agreed to allow them to be saved to the Utah State University repository, as well, in order to maintain a single central access point to project data. In addition, Burkert has maintained detailed project logs since 2013 and keeps the project folders in Dropbox and Box, both of which provide cloud backup and sync to three computers (her personal desktop, personal laptop, and office machine).

Roles and Responsibilities

Hugie (Developer) will work with Betty Rozum (Data Services Coordinator, Merrill-Cazier Library, Utah State University) to ensure storage and preservation in accordance with current best practices. All project participants who edit the data and source code will be responsible for using shared file naming and versioning conventions, updating readme files, and saving incremental progress to shared folders in Box. Hugie will conduct monthly backups to a designated Box folder and to a local hard drive from the beginning of the grant period; he will make long-term storage deposits in Digital Commons and Digital Preservation Network at the end of the project. Utah State University and Merrill-Cazier Library will provide access to all preservation solutions (detailed below) at no cost and will be responsible for decisions about the data over long term.

Expected Data

The project is expected to generate up to 1GB of data, including:

- Flat .txt files representing multiple versions of the flat-file database, from the dirty recovered data through several stages of cleaning and parsing
- XML-tagged data representing all the entries in the flat-file database
- JSON-tagged data, again representing all the entries in the flat-file database
- Relational database (likely in MySQL or an open-source equivalent like MariaDB) populated with the XML- or JSON-tagged data
- Software code (in Python) for cleaning, parsing, and tagging data
- Website code (PHP, XML, JavaScript, HTML, XSLT)
- Documentation, including the database schema and README files, stored as plain-text .txt files
- The 2018 meeting report, peer-reviewed publications, and 2019 white paper, stored in PDF and .txt file formats

Period of Data Retention

USU will provide access to server space for the website and will maintain access to the Digital Commons repository for as long as reasonably possible.

Data Formats and Dissemination

The flat-file data will be maintained as .txt files. It will then be tagged in the preservation-friendly XML and JSON formats. This structured data will be imported into MySQL or its open-source counterpart, MariaDB, and end-user access to the database will be provided through a web interface. The server space will be provided by USU Central IT. Users who wish to download the raw data and documentation in its entirety can do so through Digital Commons, the USU institutional repository. The documentation will include the version(s) of software used to create the database on which the code runs. All code will be made openly available on GitHub. To the extent allowed, publications resulting from analysis of the data

[5] https://minds.wisconsin.edu/handle/1793/71768

will be deposited in Digital Commons as well, either as post-prints or in the version of record. Our goal is to make all project outputs, as well as all procedural and contextual information, open access and open source.

Data Storage and Preservation of Access

All of the data and software code described above (.txt flat-file data, XML-tagged data, JSON-tagged data, database schema, programs, readme files, and additional documentation) will be bundled using BagIT and uploaded to USU Digital Commons repository for preservation. Digital Commons is backed up to Amazon S3. We also have arranged to deposit the files in Open SIUC, the institutional repository of Southern Illinois University, in accordance with our agreement with SIU Press, the publisher of the original reference books *The London Stage*. Long-term preservation will be provided through the Digital Preservation Network, which is one of the most robust digital preservation solutions available. The Merrill-Cazier Library at USU will provide cost-free access to all of these preservation systems.

9. Data Management Plan

Roles and responsibilities
Stanford Digital Repository manager
Global Medieval Sourcebook PI: Kathryn Starkey
Global Medieval Sourcebook project manager: Mae Lyons-Penner
Global Medieval Sourcebook technical expert: Michael Widner

Expected data
Types of data:

- Transcriptions, translations, and critical notes encoded in TEI-compliant XML
- Metadata describing each work
- Audio recordings of works being read aloud
- Computer code (PHP, Javascript, CSS, HTML) to generate the project website
-

Period of data retention
Data will be retained for as long as the Stanford Digital Repository is available.

Data formats and dissemination
Computer code will be available to the public via Github and via the Stanford Digital Repository. All code will be open source using the GNU General Public License.
Reports will be available as PDFs on the project website and in the Stanford Digital Repository.
Metadata and XML-encoded texts will be available on the project website and in the Stanford Digital Repository.
Audio files will be encoded in MP3 format and available on the project website and in the Stanford Digital Repository.

Data storage and preservation of access
XML-encoded transcriptions, translations, and critical commentary will be stored in the Stanford Digital Repository. Audio recordings will also be stored in the Stanford Digital Repository.
The Stanford Digital Repository (SDR) is a service offered by the Stanford University Libraries that provides digital preservation, hosting, and access services that enable Stanford researchers to preserve, manage, and share research data in a secure environment for long-term citation, access, and reuse.

Digital content ingested to the Stanford Digital Repository's preservation core is replicated multiple times and stored in geo-diverse locations on different media types. All content is audited systematically to ensure that the bits are maintained exactly as deposited, and a log of preservation actions is kept to help ensure the content's integrity. The repository is built using open-source software widely adopted across the research community, with dedicated staffing by digital preservation experts. Access is controlled using strict authentication policies and enterprise-level security mechanisms. Metadata describing the content is indexed for searching, and copies of ingested content are provided via persistent URLs to authorized users via Stanford's digital library environment.

Roles and Responsibilities

The data management plan will be implemented and managed by Stacy Reardon, under the project supervision of Bryan Wagner. Reardon will manage data and backup during the development phase and transfer data to institutional repository UC DASH for long-term permanent storage. Patty Frontiera and Susan Powell will be responsible for geospatial data. In the event that one of the responsible parties listed above is no longer involved in the project, responsibility for data management and the project website will fall to Principal Investigator, Bryan Wagner. In the event that Wagner is no longer involved in the project, responsibility for the website and project data will fall to the UC Berkeley D-Lab.

Expected Data

Data will consist of digitized primary sources, transcripts, translations, metadata, digital geographic data, the project website and any attendant custom code and design files, documentation for internal use, and tutorials for public consumption. In addition, the project will produce curriculum materials, conference materials, and scholarly publications.

Period of Data Retention

Processed primary source data will be made available immediately upon launch of the project website. Additional and newly acquired primary source data will be made available in a timely manner upon processing. Monographs or articles resulting from the project will be made available as soon as possible according to the embargo period of the publisher, and open access publishing will be preferred.

Data Formats and Dissemination

Materials for public consumption will be made available on the project website and archived in UC DASH. Image files will be made in TIFF format and stored in California Digital Library's Merritt repository for archiving, and copies will be made in JPEG for web display. Transcripts, translations will be stored in PDF form. Dublin Core metadata will be maintained in CSV and XML. Digital geographic data will be maintained in JSON files, with an eye towards making it extensible to other projects. Custom code will be made available on GitHub and archived in Merritt. Tutorials and curriculum materials will be disseminated in the most appropriate media and stored as PDF when possible; video content will be in MP4 format. All internal documentation will be stored in UC DASH as plain text and CSV files. Scholarly output will be archived in California Digital Library's open access repository whenever possible.

Data Storage and Preservation of Access

During the active project phase, all data will be backed up in UC Berkeley's Google Drive storage. In addition, GitHub will be used for any custom code development.

After the grant period, the UC Berkeley D-Lab will assume web hosting costs and essential maintenance for the project's online learning environment. The D-Lab will pay the campus fee for a Pantheon Pro Account ($900/year; web.berkeley.edu/web-hosting-pantheon). The D-Lab will also pay to have the project website regularly updated, maintained, and repaired. This commitment is calculated at five hours a month ($30/hour or $1800 annually).

A graduate student assistant will ensure that the website is crawled and archived through the Internet Archive at least monthly from the time it is launched through the duration of the project phase, and thereafter the PI will be responsible for periodical, regular web archiving.

In addition, all images, texts, metadata, documentation, and code components will be made available for public download through a free and open-source GitHub site as well as through the UC DASH Repository (dash.ucop.edu). UC DASH allows access via persistent URLs, offers tools for long-term data management, and permits permanent storage options. UC DASH has built-in contingencies for disaster recovery including redundancy and recovery plans.

Data Management Plan

"Documenting the Ethnobiology of Mexico and Central America (DEMCA): A Digital Portal for Collaborative Research in the Humanities, Social Sciences, and Natural Science" will develop a web portal of ethnobotanical and botanical materials pertinent to the aforementioned disciplines. The substantive content of the portal—botanical collections and in-situ photographs; ethnobotanical data on plant nomenclature, classification, and use as well as digital recordings, transcriptions, and translations—will be developed by Amith and Beck from projects they currently direct in Nahuat, Totonac, and Mixtec communities, projects supported by the NSF (Documenting Endangered Languages), the Endangered Language Documentation Programme (SOAS, London) and the NEH (Preservation and Access). Seven additional linguists (see support letters, pp. 18–26, section 9) will provide smaller amounts of similar data to fully test DEMCA functionalities and the viability of the standards for ethnobiological metadata developed for the Symbiota/DEMCA portal.

This data management plan includes management of all materials that will be used in the DEMCA portal even though these materials were produced with funds from other grants. This Level II project will, however, produce new material: an integrated database of botanical and ethnobotanical materials. Brandt will ensure that a system is implemented to produce backups files of the entire DEMCA database and contents on a nightly basis. Each nightly backup file will be preserved for the duration of a one week on a machine separate from the server to ensure several options for recovery if data restoration is ever required. At project end the last backup will be deposited at the Archive of Indigenous Languages of Latin America, University of Texas (AILLA) and at the American Philosophical Society (APS) for long-term archiving (see Letters, pp. 36, 37)

1. Type of materials + indicates material produced by other grants but which will be used as content for DEMCA * indicates material produced by this NEH project

+a. A database of Indigenous nomenclature, classification, and use of plants based on information derived from various sources: (1) Amith's and Beck's work on Sierra Norte/Nororiental de Puebla Nahuat and Totonac; (2) Amith's work on Pacific Coast of Guerrero Mixtec; and (3) contributions by the seven collaborating linguists and anthropologists (see Letters, pp. 18–26).

+b. An extensive set of archival quality (48KHz, 16-bit) digital Indigenous language recordings, the majority from sources (1) and (2) in (pt. a) above. Towards project completion, seven linguists; see (3) above, will provide DEMCA test data from other languages and hopefully include recordings. Amith and Beck will develop some video (mp4) to test DEMCA's video presentation.

*c. Metadata of all ethnobotanical and botanical information built into DEMCA as documented in the Appendix. This metadatabase will provide the informational foundation for the DEMCA web portal.

+d. Voucher specimens of fertile (flowering or fruiting) plants from areas mentioned in 1.a.. These materials will be collected according to herbarium specimen standards with full collection data (e.g., GPS coordinates, habitat, plant description) and ethnographic information (nomenclature, classification, and use in Indigenous languages and cultures). All collected plants will be extensively photographed in situ (flower, leaves, stem, fruit, entire plant) and the mounted voucher specimens will be digitally photographed. Approximately 10,000 collections and 2,500 species will be made.

2. Standards for data and metadata

There are three basic types of data, each of which will be managed with distinct metadata and standards.

a. Ethnographic material includes (1) database of plant nomenclature, classification, and use, linked through metadata to a physical botanical specimen and its occurrence data; (2) digital recordings, transcriptions, and translations of Indigenous experts talking about a particular plant or group of plants. The recordings will be archived in uncompressed .wav format (at 48KHz, 16-bit). The metadata for the recordings will be archived in IMDI format (https://tla.mpi.nl/imdi-metadata/), a shared standard used by many archives for endangered language material. The transcriptions and translations will be in ELAN format (https://tla.mpi.nl/tools/tla-tools/elan/), again a standard XML scheme used by most linguists and endangered language archives. Data on plant nomenclature, classification, and use will be archived in an XML schema developed in this present project.

b. Plant specimens and photographs: The collections will be have unique numerical identifiers; the database will record standard collection data: coordinates, biological form, habitat. Specimen labels

for herbarium sheets will include this information as well as ethnographic data: indigenous name, classification, and use. The Darwin Core standard will be used for all metadata pertinent to botanical collections. The physical specimens will be accessioned at major herbaria (e.g., Mexican and US national herbaria, Missouri Botanical Garden). The photographs will be incorporated into the Tropicos system at the Missouri Botanical Garden (see Letters, pp. 34, 35, 17**)**.

c. DEMCA database: DEMCA will develop an integrated botanical and ethnobotanical database according to the work plan and appendix. This database will contain the metadata and content from the botanical and ethnobotanical material. It will be backed up daily and periodically deposited at project conclusion at AILLA and at the APS.

3. Policies for access, privacy, confidentiality and IPR

Ethnographic material (e.g., recordings) will be gathered with strict adherence to recognized IPR protocols of informed consent and privacy when the collaborator so desires. Amith has secured IRB approval for this project with Gettysburg College and follows a protocol that he has been using for 13 years. Communities in whose lands collections are undertaken will be approached and collection will only proceed if a signed agreement is reached (see example, Letters, p. 27). When pertinent, additional agreements will be reached with community authorities and schools (Letters, pp. 28, 29). All those who record will be given a mutually signed "memorandum of understanding" about privacy issues, potential dissemination, and rights as authors (Letters, p. 30). Both orally and in this agreement it is explained that the material will be used only for education and research and any "commercial" publication would require new written consent with a clear statement that any economic benefit would be for the author. All botanical collections will be made with the proper permits and export from Mexico will be through the national herbarium (MEXU).

4. Policies and provisions for re-use, re-distribution and the production of derivatives

All ethnographic material (digital recordings, transcriptions, translations) may be used for academic, non-commercial purposes as per provisions that will be posted on the open access website (e.g., AILLA, APS, DEMCA). Distribution will follow the Creative Commons License most appropriate to this material: Attribution-NonCommerical-ShareAlike (CC BY-NC-SA) although it is recognized there are some grey areas.

The voucher specimens and collection data will also be freely available through the herbaria in which they are deposited and collection data may be cited. Photographs (both of plants in situ and voucher specimens) may be used under the same CC BY-NC-SA license.

The Symbiota software developed for this project will be open source and available on GitHub.

5. Plans for archiving data, samples and other research products

Ethnographic material: Arrangements have been made to archive all Indigenous language digital recordings, time-coded transcriptions, and translations at AILLA and the APS digital archive. Proper metadata records in the IMDI format (https://tla.mpi.nl/imdi-metadata/) will facilitate discoverability and provide researchers with the necessary information about the recordings (e.g., name, origin, age at recording, of narrators) and about the plants and voucher specimens to which they refer. AILLA and APS will also archive the database of all collections (which will include data on the collection and ethnographic information about plant nomenclature, classification, and use). **Support letters: AILLA and APS (Letters, pp. 36, 37)**

Plant specimens: Of three plant vouchers, one voucher of all collections will be deposited at MEXU (Mexican National herbarium), one at US (US National Herbarium), and one at the Missouri Botanical Garden. The Darwin Core compliant metadata will be provided to each herbarium for porting to their database system. In-situ photos will be accessioned at Missouri and the metadata for the photos, incorporated into the Tropicos database. **Support letters: MEXU, US, MO (Letters, pp. 34, 35, 27)**

DEMCA database and content: At project end, all metadata and content will be exported from DEMCA and deposited at AILLA and APS. **Support letters: AILLA and APS (Letters, pp. 36, 37)**

Symbiota software: This software is open source and will be posted on GitHub. The posted version will include the data structure for ethnobiological information developed by this project.

Visualizing Webpage Changes Over Time

Michele C. Weigle, Michael L. Nelson, Deborah Kempe, Pamela Graham, Alex Thurman

Data Management Plan

Expected Data and Outputs
The main product of this project will be software, specifically for generating thumbnail representations of TimeMaps. The software will be written in either Java (for the Wayback Machine), or in Python (for "pywb" -- the open source Python implementation of the Wayback Machine that is gaining traction in the web archiving community). The software that will write to process third-party TimeMaps (i.e., creating thumbnails of TimeMaps in arbitrary web archives that we do not have root access to) will likely be written in Python. Software written for embedding thumbnail representations in web pages will be written in Javascript and HTML.

ODU has a dark archive of most of Archive-It's collections (totaling 230+TB and > 5.3M mementos, representing Archive-It through April 2013, but updates are in the process of being obtained), which are stored in WARC (Web ARChive) files, which are the official (ISO 28500:2009) and most popular format for storing the results of web crawling activities. For detailed information about the WARC format, see http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

We will also work directly with the WARC files of specific interest to NYARC and CUL that may lie outside of our collection.

We expect the output of generating thumbnails to be stored in a new, yet-to-be-determined format, but it will be serialized in JavaScript Object Notation Format (JSON), the de facto standard format for web processing. The images themselves will likely be Portable Network Graphics (PNG) format, a popular and standard (IETF RFC 2083) format for storing images on the web.

Period of Data Retention
ODU has 120 TB of mass storage available for faculty research data. As a backup we will store all generated source code and example datasets for demonstrating the code on this mass storage system for at least 5 years, though there is no standard time limit on data storage for this mass storage system.

The Archive-It WARC files themselves are part of an ongoing institutional commitment by ODU to being a dark archive for Archive-It, so those files will be maintained and internally available for at least the next 5 years, most likely much longer.

Our research group at ODU now uses GitHub for our code, documentation, and code management. We commit to maintaining public access to our developed code for at least 5 years through GitHub, or other public source code system, if for some reason GitHub is no longer available.

NYARC and CUL are clients of Archive-It, who will be responsible for the long-term archiving of their WARC files.

Data Formats and Dissemination
As stated above, we now use Github for code dissemination; our existing projects can be explored at https://github.com/oduwsdl/

As we have done in the past, we will continue to use this method to make our code publicly available to the web archiving community and the public at large. We use the MIT Open Source license, see for example: https://github.com/oduwsdl/ipwb/blob/master/LICENSE

The WARC files themselves are not publicly available (as per our agreement with Archive-It, and consistent with us being a dark archive). However, there are several commonly used, open-source tools available that create WARC files (e.g., Heritrix, wget, webrecoder) so there will not be a problem in applying our code in new scenarios. Similarly, the Wayback Machine software and pywb are both open source as well.

Publications resulting from this project will be stored as PDF documents in public repositories such as the ACM Digital Library and the arXiv eprint server. These will also be made available from the ODU WS-DL research group's web page, http://ws-dl.cs.odu.edu, promoted on our blog http://ws-dl.blogspot.com, and on our Twitter page http://twitter.com/WebSciDL.

Data Storage and Preservation of Access
ODU has 120 TB of mass storage available for faculty research data. As a backup we will store all generated source code and example datasets for demonstrating the code on this mass storage system.

Our source code and documentation will be stored and maintained at Github, with local copies of the code repositories on research servers at ODU (with several TBs of storage space currently available).

The Archive-It WARC files themselves are part of an ongoing institutional commitment by ODU to being a dark archive for Archive-It.

Data Management Plan

This project will primarily result in three forms of data: computer source code, machine learning models, and documentation.

In keeping with the STUDIO's commitment to open source development, our prototype software tools, models, associated project source code, workflows, documentation, and use cases will be posted on the STUDIO's GitHub site. This will be released under the MIT open source media license or CC BY 4.0 Creative Commons license as applicable to allow for the largest possible application. These will remain available on GitHub for a period of not less than five years, and we anticipate them remaining there for the foreseeable future. We will also host a mirror of the source code, documentation, and models on the STUDIO's website to protect against the event that GitHub is no longer available for hosting.

Our primary data set, the Teenie Harris Archive, consists of images and metadata that are publically available through the Carnegie Museum of Art. CMOA is the custodian of these data sets and will continue to maintain the canonical copy of this data. Throughout the project, our use of the Teenie Harris Archive will generate annotations in the Open Annotation (http://www.openannotation.org/spec/core/) standard that augment that collection. We have agreed to provide these annotations at the completion of the project to the Carnegie Museum of Art for potential integration into their permanent collection. Additionally, through the use of our tools, these annotations can be regenerated as needed.

(It is possible that IIIF will replace Open Annotation with W3C Web Annotations (https://www.w3.org/TR/annotation-model/) before this project can be completed. In that case, our annotations would use the Web Annotation Model.

9. Data management plan

Documentation and Metadata

As a companion endeavor to the Cuneiform Digital Library Initiative, cdliwiki[1] documents all aspects of the CDLI in a range of articles on history, specific inscribed artifacts, and genres, as well as discussions of processes and data acquisition. On the CDLI website itself, there are also articles discussing the museum collections holding the physical artifacts[2] as well as the terms of use of the data[3]. These tools will be used to help document the project and its outcomes. Awaiting the results of a French research group[4] working on the alignment of our texts' metadata with the CIDOC-CRM ontology, we will prepare an XML output for each search result view. It will be machine readable and easily modifiable to include CIDOC-CRM attributes. The software created through the project will be thoroughly commented, and a GitHub Jekyll website will serve as a code documentation hub directly in the same GitHub location as the code.

Textual Data

The transcriptions and some text structural information of cuneiform inscriptions of artifacts are preserved using the "Canonical ASCII Transliteration Format" (C-ATF).[5] The text itself is encoded in UTF8 and the original language transliterations are restricted to simple ASCII characters. This notation system has been in use for 15 years and because of its simplicity and high level of standardization, many research projects base their work on the CDLI or will use a derivative of C-ATF. The ATF notation created by the CDLI is the widest-used standard in the field. With this project, we will continue to offer the input and output of transcriptions in this format but we will also store and offer to view and download data in an updated format where the transliterations lines of the text will also be UTF-8, and appropriate IPA characters will be used instead of derived ASCII transcription. This use of two standards makes the data usable by more people and thus will enhance its preservation. These textual data are currently viewable online and downloadable in text format. Because the CDLI is a long-lasting initiative, there are already quality checks and versioning systems in place. These checkers will be enhanced as part of the work plan. Each time a change is saved in one of the texts, a backup copy of the previous version is saved in the database.

Licensing

New software and documentation generated by the project will be released to the public domain by using the Creative Commons license "Public Domain Dedication" (CC01.0).[6]

Storage and Backup

During the research, GitHub will be used as a versioning system for the code base of the project. The Center for Digital Humanities (CDH) at the University of California, Los Angeles gives us technical support and external backups that increase the security and recoverability of the data. We also have a mirrors of the servers at the Max Planck Institute for the History of Science, Berlin (MPIWG); and through them at the Max Planck Society's persistent storage hub in Göttingen) and at the University of Oxford. Additionally, we expect news shortly from

[1] <http://cdli.ox.ac.uk/wiki/>
[2] See, for example, the page of the British Museum <http://cdli.ucla.edu/collections/bm/bm.html>
[3] <http://cdli.ucla.edu/?q=terms-of-use>
[4] <http://triplestore.modyco.fr:8080/ModRef/>
[5] <http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html>
[6] <https://creativecommons.org/publicdomain/zero/1.0/>

Compute Canada concerning an application for resources allocation in order to set up a Canadian web site mirror and backups. These services come at no cost when allocated.

Preservation

By renewing periodically our agreements with the CDH, the MPIWG-Berlin and the University of Oxford, we are convinced that the CDLI offers optimal storage security and web server longevity; CDLI is in fact a model of data persistence—the longest lived digital humanities project in the field of Assyriology, with its predecessor the Uruk Project at the Free University of Berlin now 26 years in existence. Since the framework update project will increase the access of the data and interface of the CDLI, its usage will increase. For any eventual risk to the preservation of the software or the data, we will put copies of our work in official repositories to maximize their preservation.

Data Sharing

The code produced by this project will be released in the public domain and we will encourage anyone to use, modify and reuse any of its components. It will be available on GitHub with its accompanying full documentation.

Responsibilities and Resources

Because the CDLI has been running for many years, our lab is equipped with the needed material to undertake the framework update. Since some of our operations will be transferred to virtual servers at the Center for Digital Humanities at UCLA, costs will be impacted negatively where we will be able to repurpose two of our physical servers and retire another one, and we will no longer be required to maintain the retired physical server, or the material update of the servers now virtual. Our mirrors and backups are hosted for free at the Oriental Institute in Oxford and at the MPIWG. Each of these services are responsible for the maintenance and backup of their own servers.

Data Management Plan

The intent of this project is to create a working prototype of a functional system using *Microsoft's HoloLens* to be able to document and triage built and movable cultural heritage including those found during archeological surveys and excavations. This will involve the creation of a software system (source code and binaries) and eventually white papers, publications, presentations, and documentation. Project PI, Co-PI and staff are fully committed to uphold the letter and spirit of the NEH's wish to make these products available to the broader community. Software and other products will be licensed under *GNU General Public License* (GPL) (https://www.gnu.org/licenses/licenses.html), *Creative Commons License* (https://creativecommons.org/licenses/), or similar license as applicable. These licenses allow the broadest distribution, use, reuse, and modification of the research products. These research products of this project will be handled in the following ways:

1. All white papers, reports, presentations, publications, and other project generated writing will be posted to a project website to be maintained on University of Central Florida (UCF) computers. Additionally on the conclusion of the grant a copy of all such materials will be deposited in *UCF's STARS* repository (http://stars.library.ucf.edu/) for long term storage and access by researchers. *STARS* is committed to providing access to individuals with disabilities (http://stars.library.ucf.edu/accessibility.html).
2. Source code, documentation, and binaries will be made available on Github (https://github.com/) or another open and publically accessible software repository. At the conclusion of the grant a snapshot copy of these materials will be placed in *UCF's STARS* repository for long term storage and access.
3. Data generated from field testing will be made available as practicable via the project website, other associated project websites, and as a snapshot on *UCF's STARS* repository for long term storage and access. Such data will be made compliant. where applicable, to standards used by *Arches* (http://archesproject.org/), an open heritage data management system which is licensed under AGPL and uses open data standards.
4. Source code for integration with *Arches* will be publicized and made available through the Arches Project's website.

DATA MANAGEMENT PLAN

Expected data:
The "Integrating digital humanities into the web of scholarship with SHARE: an exploration of requirements" project will produce a number of outputs and digital assets. The first phase of the project involves gathering DH requirements to enhance the SHARE data set; this will be conducted via an online survey and in-person focus groups. The survey will produce a survey instrument, raw results, analysis scripts, and the analyzed data. The focus groups will produce notes, code book, coded answers, and the focus group script and question set. The next phase of the project involves a workshop to explore the alignment of the SHARE Curation Associates program and the CLIR Curation Fellowship. The expected outputs of this phase include notes from the workshop, presentations, and a final report. The final phase of the project will results in the development of technical prototypes, including software code, scripts, and wireframes. Finally, the results of the technical evaluation of these prototypes will be a result of this project.

Period of data retention: The project team will make the resulting assets (redacting any personally identifiable information) available to the public as each phase of the project comes to a completion. The project team will use the Open Science Framework to manage the project workflow, collect documentation, and make the assets public upon completion. These assets will also be archived and curated in the Washington University in St. Louis institutional repository, Open Scholarship, where they will be retained and curated for a minimum of 10 years.

Data formats & dissemination: Whenever possible, open data formats or formats that do not use closed proprietary specifications will be adopted as asset accessibility and archiving standards for the project. For example, all text/data will be encoded using Unicode to prevent data loss. Uncompressed TIF (or comparable) will be used for all images. Archival copies and originals of the data will be maintained according to the WU Libraries archiving policies outlined below.

Metadata will be created and saved throughout the lifecycle of the research project and will be in line with the commonly accepted scholarly standards. All collections will have a DublinCore metadata record created. This record includes elements such as author, license, abstract, publication date, funder, and others. As appropriate, the Data Documentation Initiative (DDI) metadata standard will be used to describe data granules (individual files rather than collections) and survey instruments. For continued preservation of the data and materials, metadata elements consistent with the PREMIS data dictionary and data model will be implemented.

A "read-me" file will also be created that includes an asset inventory, general rights for reuse, contact information, a recommended citation, and a synopsis of the project.

To improve dissemination of the research outputs a digital object identifier (DOI) will be assigned to the collection or at necessary granularities of the project. Additionally, metadata records are propagated to metadata harvesters, such as SHARE through the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

The dissemination information package (DIP) will include access copies of the assets, any required analysis code or software, the data dictionary/code book, the DublinCore metadata record(s), the discipline metadata record(s) and the "read-me" file.

Data storage & preservation of access: As previously stated, day-to-day project management and storage will take place on the Open Science Framework (OSF). The OSF will allow the project team to integrate with existing systems, such as box for large file management. Given its design, it also facilitates the creation and storage of various asset types, including text, tabular, images, and more. Once a phase of the project is complete, the OSF project space or component will be made publically available.

Digital assets archiving and preservation is supported through an approach that starts with adequate documentation of data using metadata and other formats appropriate for long-term preservation. Unique digital assets from this project will be redundantly deposited in the Open Scholarship repository, and archived according to the Open Archival Information System (OAIS) framework.

The archival information package (AIP) will contain all the materials previously mentioned in the DIP, a copy of the untreated, original submission information package, the PREMIS compliant metadata, a checksum manifest, and a curation treatment actions file.

Data Management Plan

1. Roles and responsibilities

This data management plan will be implemented and managed by Ron Joslin, under the supervision of project Co-Directors Brigetta Abel and Amy Young. All project data used in the curriculum that is not restricted by privacy and/or intellectual property will be made publicly available under a Creative Commons license. While the project data will be linked and accessible from the project's web site, it will also be transferred to and permanently archived on Macalester's institutional repository, Digital Commons, to ensure it's long-term availability for re-use by others. The repository staff will have responsibility for the preservation of the data and will work to ensure that the value of the data is maintained over time.

2. Expected data

Because this project involves the development of an interactive open-access web-based German-language and -culture curriculum for use with introductory students, our data will be at two levels:

- **Completed learning objects**

These will include web pages, final version audio and video files, interactive exercises, transcripts of audio/video content, image files, among other learning objects.

- **Metadata and documentation of these objects**

These will include descriptive information about the multimedia files created (people, location, topic, etc.), tags/descriptors for video/audio content, field notes, software code, and permission/release forms from video and audio participants. An adaptation of Dublin Core standards will be used for producing the information requirements needed to document content for preservation and possible re-use.

During the project's lifetime, both working and archival versions of software code for web pages will be stored on our web hosting provider's servers which are backed up nightly. Other documentation, including text files, transcripts, as well as notes documenting the preservation process will be stored using cloud-based tools with nightly backups. A copy of all files will also be downloaded and backed up on a Macalester server weekly. Appropriate naming conventions and version control will be implemented.

3. Data formats and dissemination

Standard open source non-proprietary software formats will be used for all archival and shared files. Software source code and multimedia content will have adequate metadata wrapping to ensure that it can be easily reused by others and migrated to another format, if needed, for future use.

4. Data storage, retention and preservation of access

All public data will be deposited in the Macalester College Digital Commons institutional repository which has capabilities to manage, archive and share digital content. Digital Commons allows access to the public via persistent URLs, provides tools for long-term data management, and permits permanent storage options. Digital Commons also has built-in contingencies for disaster recovery including redundancy and recovery plans.

The Development of Digital Documentary Editing Platforms

NEH Digital Humanities Advancement Grant, Level 1

Data Management Plan

The data produced by this project will include the presentations by projects currently using the two systems, presentations by those outlining specific editorial/publication needs, and data generated through the workshop's planning, organization, and discussions. The presentations will be a combination of images and text detailing the project's development process, editorial methodological requirements, publication goals, and specific editorial/publication needs when applicable. Dialogue and discussion during and after the workshop will be an additional source of data. This includes the project white paper, and final report. All data will be made available on the CDE's website (centerfordigitalediting.org) where individuals will be able to freely comment on the materials as well as Libra, the University of Virginia's scholarly repository (http://www.library.virginia.edu/libra/landing/). Furthermore, efforts will be made normalize all presentations into PDF files to ensure long-term accessibility. The white paper and final report will also be made available as PDF files.

The workshop presenters will retain their copyright and other intellectual property rights of material submitted for the workshop. They will, however, agree to license presentations and any other materials prepared for the conference under the Creative Commons license Attribution 4.0 or agree to a similar arrangement to allow for the greatest amount of dissemination and use of the materials. Contributions made to the white paper and final report will fall under the same licensing agreement.

Data Management Plan

The Roy Rosenzweig Center for History and New Media (RRCHNM) is committed to open-source software development and open access for all data and content. The Project Director will ensure that all of the data and content work of the project is published for public use as soon as it is available, and she will be responsible for assuring the project team adheres to this data management plan.

Transcribing and Reviving Early American Federal Records with Scripto and Omeka S will generate three different types of data: software code, software documentation, and digital assets of the *Papers of the War Department*. RRCHNM will maintain all data generated through this project in perpetuity.

Software: RRCHNM's experience with GitHub for software development has convinced us that it is the optimum platform to maintain and manage revisions and additions to *Omeka S* and its modules. All code for this project will be developed in public in RRCHNM's *Omeka S* GitHub repositories <github.com/omeka/omeka-s>. Following current workflow practice, we will create individual repositories for all code developed for the *Scripto* module and the Transcribe theme for *Omeka S* and will begin developing in public through GitHub. Any enhancements to the *Omeka S*, required during the migration of PWD's assets to *Omeka S*, will be merged into the core and available for all users through GitHub. By publishing on GitHub, we make it possible for users to follow the development of the project, to offer feedback early and often by creating issues that raise questions about our approach, to submit pull requests that contribute revisions to *Scripto*, and to fork the code to adapt it for their own use and adaptation. RRCHNM's *Omeka* team monitors issues and pull requests both through GitHub's email notification system, and through the integration of repository activity feeds into our Slack channel. As a result, community contributions get prompt attention and consideration from the developers. The software produced will be release under a GNU GPL 3.0. license.

Documentation: All technical workflows and guides written for advanced developers and for end users will be published with all other *Omeka S* documentation and released under a Creative Commons By Attribution 4.0 license. Developer documentation is currently updated and available in GitHub, https://github.com/omeka/omeka-s-developer, together with the end user documentation: https://github.com/omeka/omeka-s-enduser By writing the documentation in Markdown in GitHub, the *Omeka* team, and all users, can easily locate, update, and pull these documents from GitHub and publish on other webpages while offering version control through one main repository.

All **code** and **documentation** will adhere to existing work flows well established by the *Omeka* team, and as a result will be maintained and incorporated as part of the existing body of code data. The *Omeka* development team is committed to sustaining and maintaining all code and documentation generated by this project through the funds committed by the Corporation for Digital Scholarship.

Digital Assets of the Papers of the War Department Website: All existing and new digital assets located in the *Papers of the War Department* website <wardepartmentpapers.org> developed during the course of the project will be migrated to the *Omeka S* platform. Once PWDs digital assets are moved

to *Omeka S*, the project team can manage the data more easily than in its current state. With the migration complete, PWD's digital assets--descriptive metadata, files, and transcriptions—will be deposited in George Mason University's institutional repository system, http://mars.gmu.edu/. The public website itself will be maintained on the RRCHNM server system, which provides a fall-back array of servers to support its integrity and to remain available to researchers at all times.

9. DATA MANAGEMENT PLAN

Expected data.

1. Scholarly output
    a. With input from the project team, the scholarly chapters produced by the team will be preserved in UGA's Athenaeum open access repository. Athenaeum is built with the widely used DSpace institutional repository application and has been managed for over a decade by the staff of the Digital Library of Georgia which operates as part of the UGA Libraries. This scholarship must be open access compliant in order to be included in the institutional repository.
2. Images and media files
    a. Project team will be responsible for attaining rights for any images or media content on the site.
    b. UGA Libraries will store all image and media files while the project site is available. As part of their normal procedures for support, these files along with the entirety of the site will be backed up on a regular basis. Beyond the life of the project site, the Libraries will minimally provide offline preservation of the associated images and media files for a defined duration.
3. Datasets
    a. The UGA Libraries will preserve any datasets less than 25GB that were created in the course of this project in UGA's Athenaeum open access repository and provide persistent URLs for continued access to content. For larger datasets, the Libraries will work with the project team and UGA's Enterprise Information Technology Services to store these datasets and provide links to them from Athenaeum.
4. Frontend
    a. The project team will determine rights to code developed for this project with designer and developer.
    b. If appropriate, the UGA Libraries will preserve the code produced to create the additional components of the site. Custom code and/or applications will be retained and preserved for a period to be defined beyond the end of the grant and life of the site. Due to unpredictable future changes in technology, the operation of this code or applications is not guaranteed and reuse may require modifications to restore their original functionality.
    c. The UGA Libraries will preserve screen shots of the original site. Screen shots of the site will be preserved from the launch of the project and include any substantive iterations that alter the site during its life.
5. Maintenance
    a. In case of future degradation of the WordPress theme or underlying structure during the course of the project site's life, the UGA Libraries will work with the project team to migrate the site to a new platform that best maintains the originally functionality of the site.

Period of data retention. The UGA Libraries will guarantee this maintenance for two years after the completion of the project. At the end of that time, they will meet with the project team to determine the future of the project and either continue maintenance on a year by year basis or decommission the site.

Data formats and dissemination. The platform will remain available to the public as a scholarly resource and model, hosted and sustained by the University of Georgia Libraries and DigiLab. Students and scholars will be permitted to use and adapt the scholarly content based on terms set in a Creative

Commons license. We will advertise it through MLA Commons, where we have already received a very positive review of our beta site, as well as ModNets and D-LAX (Digital Liberal Arts Exchange). We will use our contacts with modernist scholars and author societies (H.D. Society, Marianne Moore Society) to publicize the project, and will prepare a press release and solicit reviews and articles in local and national news outlets and literary journals such as *TLS*. After the flash mob, we will continue to use Twitter, Facebook, and Instagram to publicize the site. Our institutions will promote the platform through forums such as UGA's annual "Spotlight on the Arts" and Davidson's Digital Showcase. We will exhibit the platform at a travelling Mina Loy exhibition curated by Roger Conover and at the MSA Digital Exhibition in Fall 2018, and will present papers at MLA and other relevant conferences and venues. All source code created for the project will be made publicly available through a GitHub repository, so that other scholars can take our code and build from there. The White Paper and other papers and publications resulting from this work will be downloadable from our website.

Data storage and preservation of access. The UGA Libraries will maintain this project on their servers ensuring security to the site and upgrades to the WordPress installation as necessary. The site itself will reside on a virtual machine and be backed up regularly and redundantly to both LTO tape and hard disk. The project team will be responsible for updating content, ensuring links are live and accurate, and that connections to third party tools are functioning and active. UGA Libraries will work with the project team to resolve issues with third party tools and help find alternate solutions or alternate tools if necessary.

Data Management Plan

The data created by DHQ in the course of this project will include DHQ articles, DHQ bibliographic records, local authority files, and the final white paper. DHQ articles are represented as TEI/XML; approximately 40-60 articles are published each year with a cumulative total of 285 articles as of January 2017. Associated resources (image, video, and audio files) are stored in a limited set of curatable formats (JPG, PNG, PDF, MP3). The bibliographic records are expressed in XML using an internal schema that is designed to map cleanly onto the semantics of the major bibliographic formats, while expressing more fine-grained information about genre, creative responsibility, and publication details that is required for DHQ's intended analysis. The first phase of development funded under our previous grant produced approximately 6000 records; we estimate that this further phase will generate an additional 2000-3000 records. The local authority files will also be expressed as XML. The final white paper will be encoded in TEI/XML and published at DHQ, and will also be circulated in various derived formats including PDF.

DHQ was commissioned and designed as an open-ended and very long-term project, with the intention that it would serve as a record of digital humanities scholarship starting with the foundation of the Alliance of Digital Humanities Organizations (ADHO) in 2005. Data curation is built into the heart of the project and motivated the journal's decision to use only open-source publication tools and to express the journal's core content as TEI/XML. The period of retention of all DHQ data is effectively perpetual. The journal is currently supported by Northeastern University, ADHO, and the Association for Computers and the Humanities. In the event that the journal is decommissioned by its publishers, the source data will be deposited in the Northeastern University Digital Repository Service and will remain freely accessible there and visible via the TEI/XML viewing options now being built into the publication systems through which DRS data is disseminated.

DHQ's publication system is a standard open-source XML publication pipeline that uses Cocoon, XSLT, and Tomcat to dynamically generate HTML from the journal's TEI/XML source. The XML source data is also available for direct viewing and download.

Viral Networks: An Advanced Workshop in Medical History and Digital Humanities
A Proposal for an NEH Digital Humanities Advancement Grant

Data Management Plan

I. Products of the Research: This workshop will result in an online publication freely accessible to all users. The written texts will be available for public access in multiple formats. Scholarly rigor will be ensured by the sequence of review steps built into this workshop, including review of multiple versions of the research drafts by Contributing Scholars, the full review of each draft by the Advisory Board, and review of the workshop drafts by the Consulting Scholars. The data will also be made available in a variety of formats for reference by scholars and the public. The published research and data sources will adhere to professional standards regarding the anonymity of personal data. The flexibility of the digital platform provides significant advantages over conventional publication options, such as journals or academic presses, which are especially necessary for network analysis which usually involves large, complicated, and layered diagrams, maps, and charts. The capacity of an online publication platform will ensure that these visualizations can be integrated into the analytical writing while also linking to full scale versions linked to open data sources, thus allowing for more detailed review and evaluation.

II. Data Formats: The research will be published in multiple formats that allow for reading online, using multiple devices, as well as downloads. The texts and data will be linked with each other to facilitate scholarly review. The digital format is especially appropriate for this topic as it will allow for scholars to include multiple graphs and diagrams as well as data tables in ways that are not feasible in traditional scholarly publication formats such as journals or books.

III. Access to Data and Data Sharing Practices and Policies: Workshop participants will have access to shared work through a collaborative platform (Canvas), shared documents (Google), and a communication tool ( ). Contributing Scholars will post drafts in shared documents, with guidelines for commenting, that are accessible to workshop participants. Data will be shared in draft forms as well as spreadsheets, graphs, and charts, as appropriate. Contributing Scholars will also be invited to share their work in progress with colleagues, particularly those with expertise in the field, whose comments can be integrated into the revised papers. The final papers will be openly accessible in a scholarly platform easily located and accessed by scholars and the public. The open access and wide dissemination will be maintained through an online site hosted by Virginia Tech University Libraries that allows readers from any location to locate, access, read, and download the research in multiple formats without restrictions or subscriptions. This publication will adhere to the emerging standards for open science publication while also exploring innovative approaches consistent with digital humanities and medical history.

IV. Policies for Re-Use, Re-Distribution, and Production of Derivatives.
The data produced by scholars and posted for public dissemination may be used, distributed, and analyzed by other scholars. The chapters published in an open access format may be used, distributed, and assigned for research and teaching purposes. The Contributing Scholars reserve the right to further revise their research and publish it in traditional venues such as monographs and journals, with the understanding that the final products of the workshop remain publicly accessible.

V. Archiving of Data: Research products will be hosted and maintained by Virginia Tech University Libraries, using formats, backups, and security strategies consistent with a library repository.

NEH Digital Humanities Advancement Grant
Project Title: Building a Decision Tree for Watermark Identification in Rembrandt's Etchings—The WIRE Project

Data Management Plan

Data to be Generated

- **Computer code for interrogative watermarks decision tree.** The key data resulting from this project will be in the form of a complete, interrogatory computer decision tree, with coded branches for each of the 54 known types of Rembrandt watermarks and the approximately 500 individual subvariants within these types (with the addition of newly discovered watermarks as they arise). In addition to the decision tree's usefulness to researchers, the value added in data terms will be the individually observed verbal differentiations among these subvariants that identify them as unique, a set of data currently absent from the literature on this topic.
- **Image datasets.** Project participants are currently working with an image dataset supplied by Dr. Erik Hinterding, which includes one digitized radiograph image for every subvariant in the full decision tree, in a combination of TIFF and JPG formats. As new watermarks are discovered, they will be imaged by project staff or institutional partners, and added to the tree. The future addition of further images from museums and other archival sources is anticipated, but will not be enacted until this project is complete and image permissions are negotiated.
- **Algorithms.** Project collaborators Dr. Vikram Krishnamurthy, professor of electrical and computer engineering, Cornell Tech, and his PhD student Sujay Bhatt are developing algorithms in collaboration with project co-director Dr. C. Richard Johnson that will use decision tables extracted from the constructed decision trees to determine potentially alternative branch configurations of decision sequences to minimize the number of questions needed to reach the variant end-points. Algorithms will also be sought to 1) expand the decision trees to accommodate newly discovered watermark types, and 2) allow the tightest group classifications for images of watermark fragments. This research and the specific algorithms devised (typically programmed in Matlab) will be openly presented in the form of web postings (on the websites of the Johnson Museum and of the algorithm creators) and conference and journal papers, and will thus be accessible to all. Separate funding will be sought to accelerate algorithm development activities.
- **White paper.** Upon completion of the decision tree construction phase of the project. This white paper will be made available via the website of the Johnson Museum.
- **Final report to NEH.** When the project is complete, in compliance with NEH procedure. This report will also be made available via the website of the Johnson Museum.

Period of Data Retention

All computer code for the Rembrandt watermark decision tree for all 54 watermark types will be made freely and publicly available within six months of completion of the project end date, and will be maintained through Cornell University/the Johnson Museum through at least the year 2021. Dr. Johnson is now in exploratory discussions with the RKD (The Netherlands Institute for Art History) in The Hague about the possibility of an eventual transfer of archival management of the data, where it may fit most logically alongside the RKD's large store of related data on seventeenth-century Dutch art [see the example of the RKD's Rembrandt Research database: https://rkd.nl/en/collections/other-web-services/rembrandt-db].

Data Formats and Dissemination

All computer code (written in HTML, CSS, and Java) will be made freely and publicly available within six months of completion of the project via GitHub or equivalent third-party repository for open-source code. Also within six months following completion, a fully illustrated PDF instruction manual for the creation and coding of decision trees will be made available as a free download via the Johnson Museum website.

Data Management and Maintenance

All computer code, including embedded image files used for differentiation of watermark types, will be stored during the project period on the file server used by the Johnson Museum, which is maintained by the College of Arts and Sciences, Cornell University, and backed up daily by Cornell University Information Technologies. The code will be maintained and updated by project team staff and ultimately transferred to GitHub or an equivalent third-party repository.

DATA MANAGEMENT PLAN

1. Roles and Responsibilities

This data management plan will be implemented and managed by the Principal Investigator and the Photographic Archivist. The Project Assistant will also assist with data management. Historypin will provide training on how to manage digital tools and how to back up digital mapping data as part of our software maintenance package. The James B. Duke Memorial Library at Johnson C. Smith University will be responsible for data storage, access, and dissemination if the Principal Investigator leaves the institution.

2. Expected Data

We will be collecting several items that have already been digitized and are wrapped in metadata, and will be preserving and digitizing newly obtained items including photographs, letters, and oral histories. We will also be documenting the preservation process, which will include transcripts of oral histories obtained, gift agreements between institutions declaring and clarifying copyright and ownership of material, and text files of correspondence, planning documentation, and meeting minutes generated by the advisory planning team. This documentation will ultimately be preserved as an academic white paper that will serve as the project's final report and outcome. transferred data will be made publicly accessible.

All digital content collected from partner institutions and newly digitized material will be managed with Content DM and displayed on Johnson C. Smith University's Digital Smith online repository, and will be stored on University servers with weekly backup. Notes documenting advisory team meetings and the acquisition and preservation process will be made using Google Drive, and downloaded and backed up on a library computer weekly.

Digitized material will be made available to the public through the non-profit public history website, historypin.com. The Historypin.com website software is proprietary, and use of the site and membership is free. Terms and conditions of use for the site can be found here: http://www.historypin.com/terms-and-conditions/. The access sized images will have CC BY-NC licenses unless in the Public Domain, and will be identified as such on the item level. Images will be uploaded to Historypin.com via bulk upload, and then georeferenced within the Historypin system, which utilizes Google Maps and Google Street View APIs. The Historypin.com website is hosted on Google App Engine and Big Table. It uses a NoSQL database and programming is done in Django and Python.

3. Period of Data Retention

All relevant data will be deposited in Digital Smith for indefinite long-term storage upon completion of the project, and any additional data that is added to the digital mapping platform after the project is over will be added with the same data protocol. Once data is transferred to Digital Smith, all data will be made public immediately and indefinitely.

4. Data formats and dissemination

All digital content will be managed with Content DM and displayed on Johnson C. Smith University's Digital Smith online repository. Many of our partner institutions also use Content DM, and data can easily be analyzed and transferred between our systems.
A Dublin Core schema and Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) will be employed to sufficiently provide detailed and easily discoverable metadata for the project. In order to maximize contributions Historypin requires minimal descriptive metadata, but it will be mapped to MARC when added to collections. The final product will be a mobile app showcasing archival and community content, which will utilize an open source code base using the Historypin API and database to explore and share content directly in the community. While content can be licensed in any number of ways on Historypin, from All Rights Reserved to Public Domain, the backend is interoperable with the Digital Public Library of America, the Internet Archive, as well as other platforms through an Application Programming Interface (API).

Interviews will be for historical purposes only and conducted to Oral History Association standards. IRB approval will be sought and obtained to interview new human subjects identified through the course of the project. Oral history subjects will sign a release form which will be explained to them in detail before the interview is conducted.

Historypin will preserve following data formats where possible: Preservation masters: Photos: 3000 pixels on long edge, 2100 dpi, TIFF. Video: AVI Audio: AIFF files.
Access copies: Photos: 1000 pixels on long edge, 72 dpi, JPEG. Video: MP4 Audio: MP3
Thumbnail copies: Photos: 60 pixels on long edge, 72 dpi, JPEG

Copyright of the included items will be determined during research and items with questionable copyright will not be included. If there are any issues related to rights or ownership of the intellectual products generated from this project, the JCSU University Intellectual Property policy will be consulted. But also mention that the funding agency reserves an exclusive right to get access to, use, share and distribute your project products.

5. Data storage and preservation of access

All public data will be deposited in Digital Smith, which has the capabilities to manage, archive, and share digital content. Digital Smith allows access to the public via persistent URLs, provides tools for long-term data management, and offers permanent storage options. Backup data will be stored on JCSU servers, which have built-in contingencies for disaster recovery.
Historypin data development documentation is recorded in Github. The primary final product will be an offline collection process that will run from software on a computer to record submissions and then bulk upload to Historypin once an internet connection is available. There will be user interface on a locally run database and filenaming process, together with a bulk upload mechanism. Contributions to the Historypin website are stored with redundant file storage, and select contributions will be added to the JCSU Digital Smith online collections as a bulk download of all project contributions.

Attachment 9. Freedom on the Move: Data Management Plan

Expected Data. The project will produce the Freedom on the Move (FOTM) database, its metadata, source codes for data entry applications and scripts used to automate data processing and management. The FOTM relational database will hold all surviving runaway slave advertisements placed in North American newspapers before the end of slavery, gathered and stored in PDF form by researchers participating in this crowdsourcing project and by subscription-based private repositories. The FOTM database contains transcript (text) of the ads, the links to the PDF version of the ad, demographic and physical characteristics of caught and runaway slaves, their children, and their owners; as well as additional information such as newspaper edition and geographic location of the runaway and owners. By including the actual texts along with the researcher-generated codes derived from the texts in the database, researchers have the option to apply computational analysis ("text mining" tools), quantitative research techniques, or both to analyze the data.

Static and dynamic versions of the FOTM database will be produced and made publicly accessible for browsing, local analysis, and downloading for research and analysis with no restrictions imposed on its use.

Source codes of the web-based user interface, so it could be used as a model for other crowd-sourced archival digitization analysis projects, will also be available for sharing.

Period of data retention. It is expected that the FOTM database produced by this project will be dynamic and continue to grow as newfound ads are added, transcribed, and processed. As soon as new records are added to the database, the dynamic version, which is the most up-to-date, will always be available to the public as long as the FOTM website remains online. A dynamic and ever-growing FOTM database is the optimum solution and our preferred strategy. However, to protect prior investments in FOTM and data generated to-date, a static version of the database will be made available at Cornell University as part of CISER's Data Archive. CISER is a world-renowned social science data archive and is directed by FOTM Co-PI William Block.

Data formats and dissemination. To ensure FOTM access to members of the public as well as professional scholars, we will provide multiple mechanisms of sharing and accessing the FOTM database First, a publicly-available static version will be periodically updated and hosted at the CISER Data Archive. Second, as long as FOTM is active, the dynamic version will be accessible on the FOTM website and available to anyone with an Internet connection. Third, and intended for scholars who possess or have access to technical skills, an SQL script will also be provided so that users can execute the script in their own instances of PostgreSQL (an open source data entry program) and produce a replica of the FOTM relational database including all tables, keys, constraints, and data. Last, for scholars and others wishing to analyze FOTM data in various analytical software packages, the database will also be made available in CSV format and the SQL script will be made available in .sql format (which is viewable in any text editor).

Finally, to enable project and data searching, discovery, versioning, sharing, and access, the Data Documentation Initiative (DDI) metadata standard will be used. DDI allows for the discoverability and access of all metadata pertaining to this project across the data life cycle (i.e., from data conceptualization to collection, processing, distribution, discovery, analysis, repurposing, and archiving). The metadata will reside at CISER's Data Archive indefinitely. While no data archive can be guaranteed to exist forever, CISER is a university-supported data archive now in its 35th year of existence with good

support at the highest university levels. The source codes of the data entry applications or web-based user interface and SQL scripts will be made publicly available at Github (github.com), an open repository for collaboration, review and management of codes. The release of this source code will be concurrent with the release of the FOTM database.

Data storage and preservation of access. The metadata and static version of the database will reside at the CISER Data Archive and the dynamic version's FOTM database will be hosted at the CISER Database Production Server. CISER is committed to providing researchers access to the database past the end of the project. Cornell University Library will continue to host the FOTM website and its interface for crowd- sourcing and downloading of dynamic version of the database. Github will continue to be used for sharing of the source codes, the SQL scripts for as long as their use policy does not change. In the event of a change, these codes will be hosted on the CISER Data Archive.

DATA MANAGEMENT PLAN

Digital datasets collected and created by the project, including raw images, processed datasets maps, 3D photo models, illustrations and the regional archaeological GIS, will be deposited in the Digital Archaeological Record (tDAR) at Arizona State University. Comprehensive metadata consistent with tDAR requirements will be developed through the field acquisition and data processing and made part of the archive, and all data will be linked to a persistent URI with associated COinS citation data. For geophysical datasets, the ArcheoFusion processing system which will be used by this project, will preserve the actual raw data and the processing stack, allowing future researches to access the raw data and repeat (or modify) the analytical workflow.

In addition to long-term data archival at tDAR, individual copies of the dataset will be stored on secure servers at Dartmouth, and made available to share with project collaborators at McGill University, University of British Columbia, Southern Methodist University, and University of Northern Florida.

Results of the project will appear in a wide range of publications, including planned papers in *Antiquity, Near Eastern Archaeology, Kiva, Historical Archaeology, American Antiquity, Latin American Antiquity, Advances in Archaeological Practice,* and *Journal of Archaeological Science*. These papers and their findings will be made open access to the extent that is possible in accordance with various publisher's policies. Dartmouth College offers publication subvention funds to support open access publishing in cases where this is necessary.

9. DATA MANAGEMENT PLAN

Inasmuch as the Herodotos Project does not involve any human subject data, there are no issues of confidentiality or privacy or protection of data that need to be addressed here under the rubric of data management in the short or even long term. And, while the long-range goal of the Project is the development of a content-rich database (the information pages on the various peoples of the ancient world) that will occasion a need for management and dissemination, data management for the phase of the Project covered in the immediate proposal requires a concern only for the software for Latin and Greek Named Entity Recognition (and any related matters) that we are focusing on at this point.

Any software developed for the Herodotos Project — and ultimately the content we will be developing later under the scope of the project — will adhere to the following conditions:

(a) the software developed will be available pursuant to an open source license located at www.opensource.org and in a public repository (such as Github or Sourceforge);
(b) the digital content will be made broadly available;
(c) the project will not infringe on third party rights with respect to the development, dissemination, and use of the software and/or digital content; and
(d) The Ohio State University, as the primary grantee, will be the copyright owner of any software developed with grant funds and of any digital files that result from grant-funded digitization.

Nonetheless, we can state here that our plan for managing the informational material ultimately to be put together on each of the ancient groups of peoples that make up the database is to offer it to an existing Classics web-repository such as the Pelagios cooperative.

Data Management Plan

Expected Data

This project will result in a specific series of products, identified below. Outcomes, data, and products will include:

Network of Regional Comprehensive Digital Humanities Practitioners	Infrastructure for the network will be created during the workshop. Membership in the network will be open to faculty, administrators, graduate students, and independent scholars interested in developing digital humanities programs or teaching with digital humanities at regional comprehensive universities. Membership will be publicized through social media, as well as the network's website and email list.
Website	A website will be created to publicize the network and to disseminate digital files developed in conjunction with the workshop and contributed by network members.
Digital Files	Digital files include slide presentations, audio and video recordings, and text files developed by participants in the workshop and contributed by network members.
Surveys and Reports	Survey questions and results from approximately 400 colleges will be collected and analyzed. This material will be available on the network website. The final white paper for the project will also be available on the network website.

Data Format and Metadata Standards

In line with the open, collaborative ethos of the project, all data will be documented, shared, and distributed via the Internet on the website for the Network of Regional Comprehensive Digital Humanities Practitioners. Project co-directors Roopika Risam and Susan Edwards will assume responsibility for updating and sustaining the website. All data will be maintained in that are accessible, adaptable, editable, and portable for all interested parties, including faculty, administrators, graduate students, and independent scholars.

Standard document formats will be used: Microsoft Office, PDF, JPEG, MP3 audio, MP4 video, HTML, XML, and .CSV. Audio and video files will be captured in uncompressed formats and made available through Salem State University's institutional repository, Digital Commons, and also will be converted to MP3 (audio) and MP4 (video), which are suitable for sharing online. The web publishing platform WordPress (self-hosted by Salem State University Library) and

video publishing platform YouTube will also be used for the network website and event videos, respectively. The project will follow standard XHTML and HTML guidelines as set by the W3C's Interaction Domain, along with best practices for meta tags and searchability via Google and other web crawlers. The project co-directors will use the WAVE Accessibility Tool to ensure accessibility of materials created and made available publicly.

Descriptive metadata will be developed for preservation of materials in Salem State University's Digital Commons institutional repository. The Archives uses Dublin Core as its metadata scheme and the metadata will be developed by project co-director Susan Edwards, to ensure consistency and application of controlled vocabulary. Technical metadata will be maintained with video and audio recordings.

Policies for Access and Sharing and Provisions for Appropriate Protection/Privacy

The website will be open access and content will be licensed via Creative Commons (CC-BY).

Policies and Provisions for Re-use, Re-distribution

The project-generated material will be available on the network's website. Contributors will retain their intellectual property rights for materials developed in the workshop but agree to license materials they contribute to the website under a CC-BY license, as stated above. Contributions to the project's white paper will be handled the same way. Participants will sign consent forms for audio and visual capture of presentations.

Plans for Archiving and Preservation of Access

During the life of the project, all distributed materials will be stored digitally in two separate locations and in hard copy (if possible). Daily, off-site backups of all server based data will be implemented and continued through the end of the life of the project, using the Salem State University Library's Microsoft cloud server. Both digital and hard copy materials will be preserved through the Salem State University Archives. Digital materials will be preserved in the university's Digital Commons institutional repository, and paper materials will be housed with the Archives' print holdings beyond the end of the life of the project. Materials will be available on the network's website and in Salem State University's institutional repository within two months of the end of the workshop.

Any issues related to data management that are not directly addressed in the foregoing sections will be handled in accordance with NEH policies and procedures, along with state and federal statutes governing intellectual property.

5 Data Management Plan

5.1 Roles and responsibilities

The execution of this data management plan is the responsibility of the project PIs, David Bamman and Taylor Berg-Kirkpatrick. After transference of the data to the Merritt repository at the close of this grant, the University of California Curation Center will hold responsibility for management of the data.

5.2 Expected data

There are three types of data produced under the scope of this grant:

- Manually created training data for page-level classifications
- Manually created training data for intra-page classifications (e.g., paratext regions, including footnotes)
- Manually annotated data for regions of lacunae in documents.
- Code for performing page-level and intra-page level classificadtion, lacuna reconstruction, and compositor attribution

5.3 Period of data retention

All data (annotations and code) will be immediately posted to Github on the publication of research papers that make use of them.

5.4 Data formats and dissemination

All annotations produced in this project will published as text files in open formats and freely disseminated with a Creative Commons Attribution-4.0 license.

5.5 Data storage and preservation of access

Manually created annotations and code produced under this grant will be sustained in the long term in two locations: on Github (an online version control system), where it will be freely downloadable by others and can be continually updated and improved by other members of the community.

At the end of the grant, we will take a snapshot of all annotations and code and commit them to the Merritt repository of the University of California Curation Center for long-term preservation, including issuing digital object identifiers (DOI) for them.

9) Data Management Plan

Data to be Generated:

Preliminary data will be produced directly from the images of the playbills:

- **JSON data** produced from transcriptions produced by both public crowdsourcing and dedicated transcription by the project assistant using the Ensemble software
- **Text files** produced by OCR, which will be converted, using text mining approaches and manual work, into structured **JSON** files

Note that OCR work will be performed on all the same data as the transcription data as a comparison of the data generated by the two approaches. The OCR work will be much more experimental in terms of how accurately the OCR technology will be able to read the varying fonts of the playbills and how well the resulting text files are able to be converted into structured data. The transcription methods will present the challenge of cleaning up potential human errors and inconsistencies produced when entering the transcriptions.

- This preliminary data as will then be cleaned using Open Refine and converted to **RDF triples**. Portions of the existing metadata contained in **Marc** records for images in the Furness Theaterical Image Collection will also be converted to RDF.

Data Formats and Dissemination:

The final format for the project data-set will be RDF. The final RDF data, as well as the clean versions of the preliminary data (both JSON and the best text files produced by the OCR work) will be made openly available on GitHub. The project will also produce a website where the final data set will be made publicly available. All images used in the project are or will be publicly available on Penn Libraries' website, and digitized images of the playbills will be available on Penn Libraries OPenn repository project.

Data Management and Maintenance:

Upon completion, all data for the project, along with a white paper documenting the projects process and outcomes, will be stored in the Penn Libraries' institutional repository. Long term maintenance of the project website will be transferred to Penn Libraries' IT staff. Penn Libraries provides a $20 million dollar facility, and both Special Collections and Digital Humanities are at the heart of its strategic plan. The Kislak Center staff work between and across the worlds of traditional special collections work and digital innovation, with a team that includes staff with blended expertise in data curation and book and manuscript history, including a curatorial position exclusively dedicated to Digital and Research Services. Both the Penn Libraries Digital Scholarship Center, and the recently founded Price Lab for the Digital

Humanities at Penn also offer technical expertise and resources for ensuring a secure and sustainable environment for Digital Humanities projects at the University of Pennsylvania.

9. Data management plan

During the period of the Phase I Digital Humanities Advancement Grant, V-ESPACE data will consist primarily of 1) historical and literary documentation (including visual resources); 2) conference papers from the public presentations at the second meeting; and 3) the final white paper, outlining research outcomes and future work plans. While some code for the virtual rendering of architectural space and the behavioral modeling of avatars may be generated during this grant period, this is not a primary objective at this time. The data generated during this grant period will serve as the historical basis for game-design decisions around the interaction between space, performance, and audience, as well as serving as a primary content source for research-intensive modes of game-play.

Type of Data	**Available When**	**Accessibility**
Bibliographies (text files)	Prior to first meeting and, as progressively modified, throughout the duration of the project	Available to the public on LSU's Digital Commons
Literary, journalistic, and archival materials relative to fair theater (text and image files)	At the conclusion of the first meeting	Available to the public on LSU's Digital Commons
Eighteenth-century iconography and architectural plans (image files)	At the conclusion of the first meeting	Available to the public on LSU's Digital Commons, subject to copyright restrictions of image owners
Progress reports (text files)	At the conclusion of the first meeting	Available to the public on LSU's Digital Commons and emailed directly to all participants
Conference papers (text and image files)	At the conclusion of the second meeting	Available to the public on LSU's Digital Commons
White paper (text and image file)	After the conclusion of the second meeting	Available to the public on LSU's Digital Commons, and emailed directly to all participants
Computer code relative to modeling space and avatar behaviors	After the conclusion of the second meeting	Available to the public on GitHub

Data Management will be coordinated by Dr. Leichman for all data other than computer code; Dr. Kooima will assume primary responsibility for the management of computer code and associated data.

Working in concert with the LSU Libraries system Technology Initiative, V-ESPACE will benefit from a dedicated space in the LSU Digital Commons, which provides permanent and publicly searchable data storage. LSU is currently developing institutional best practices in Data Management, and the V-ESPACE project will actively further these goals. To that end, we are collaborating with Associate Dean for Technology Initiatives Gina Costello, as well as with the LSU Digital Scholarship Lab, directed by Associate Professor of English Lauren Coats.

We also intend to use these meetings to initiate discussion about the far more significant data management needs that will accompany the implementation of this project, with particular attention to standards of accessibility, responsibilities for storage, and intellectual property concerns. Bearing in mind the potential for differing cultural and legal expectations to impede international collaborative efforts, we

are committed to a robust exploration of these issues in order to position ourselves for a successful continuation of the V-ESPACE project beyond the term of this grant.

*v*HMML Data Management Plan

Data Types and Backup Procedures

The core of HMML's data preservation plan is to have backup copies of critical data in multiple formats and in multiple locations. Since the *v*HMML system is hosted by CSB/SJU IT Services (CSB/SJU=College of Saint Benedict/Saint John's University), HMML relies upon their workflows and processes for data loss prevention and restoration, as well as for system software backup and restoration. The *v*HMML system is tiered on multiple virtual machines (VMs) with the MySQL database and the image data as separate VMs. In addition, there are separate tiers for a *v*HMML Test platform as well as for the Production platform that is exposed to the public. Since the Test platform is regularly deleted and re-created, sometimes with bogus data for testing purposes, the data management policy outlined here deals only with the Production platform.

Two main types of electronic data are an intrinsic part of the *v*HMML system: (1) alphanumeric data such as catalog records and user data (registration details and access privileges), and (2) image data and associated image transfer metadata (IIIF JSON manifests). The first is stored in tables in a relational MySQL database; the latter is stored on a VM in a Linux directory structure organized by HMML sub-collection.

Data Type 1: MySQL Database

Backup Procedures

Currently, the *v*HMML MySQL database is backed up in four ways: 1) nightly through a MySQL dump to disk in distinct directories on the VM labeled with the date, as well as an off-campus dump to the data center at our partner institution, the College of Saint Benedict; 2) nightly taped incremental backup kept in the Saint John's University IT Services office; 3) weekly taped full backup (each Tuesday) kept in the Saint John's University IT Services office; 4) monthly taped full backup kept in a fireproof safe in another building on the Saint John's University campus. HMML plans to add a copy of the MySQL database to the backup tapes of our digital image collections sent periodically to off-site storage in Utah (see below). Every two hours throughout the workday, the MySQL database is dumped as a file on the VM, so that at most two hours of user and catalog data could be lost in case of a system failure. These bi-hourly data dumps are overwritten on a daily basis since they are redundant with the nightly taped backups.

Sharing and Dissemination

All catalog metadata is publically exposed in *v*HMML Reading Room and can be viewed without registration. This metadata is available for reuse according to a Creative Commons license (CC BY 4.0). Currently, a *v*HMML Administrator can export catalog metadata as JSON or comma delimited CSV files for specific collections or for the entire database. *v*HMML 3.0 would add tools for export and harvesting of metadata in encoded format, as requested by project partners. Sharing of user data is done under the limited conditions outlined in the Privacy Policy: https://www.vhmml.org/privacy.

Intellectual Property

All metadata creators agree in writing to make their work available according to the CC BY 4.0 license.

Data Type 2: Digital Images

Image Formats and Backups

The cameras used by HMML produce two types of image files simultaneously—a RAW digital image and a JPEG image of extremely high quality. Both types of files are retained after capture. A complete set of both RAW and JPEG images is recorded to a hard disk drive for the use of and retention by the owning library. Another hard drive is prepared for shipment to HMML.

Once the hard drive arrives at HMML, all of the JPEG images are copied to an on-site file server operating on an internal local area network (LAN). This storage area network (SAN) system makes use of redundant hard disk arrays to safeguard the data. The JPEGs on the SAN are backed up by CSB/SJU IT Services using LTO-5 tape cartridges stored in a vault on campus. The RAW images are copied to LTO-5 backup tapes that are shipped to a secure storage facility in Utah (Perpetual Storage: http://perpetualstorage.com/). The original hard drive is retained and stored in the secure, climate-controlled, and fire-protected HMML microfilm vault.

An additional set of JPEG images is produced for use in *v*HMML Reading Room. These high-resolution JPEG images, slightly more compressed than the "maximum quality" JPEG images produced by HMML's cameras, are far more usable for internet delivery but still have excellent viewing qualities. Images are rotated for the proper display orientation and JSON manifests are generated to allow them to be viewed in a IIIF environment. After preparation at HMML, these files are copied via secure FTP to the *v*HMML Reading Room Production image server. The original set of derivative images is retained on hard drives kept at HMML. All of the files on the Production server are backed up on tape by CSB/SJU IT Services as described above for the MySQL database. These images will also be taped for shipment to Perpetual Storage in Utah.

Throughout the process of making and storing backups, careful recordkeeping ensures that the status and location of all digital assets is known at all times. These records are likewise backed up and stored in redundant copies.

Sharing and Dissemination

Image files are being made available in *v*HMML Reading Room. Viewing images for most of the collections requires a one-time, no-cost registration that allows HMML to monitor usage and address violations of the terms of access.

Period of Data Retention/ Future Compatibility

HMML is a permanent institution with an endowment sufficient to guarantee ongoing operations at a level that would ensure continuing availability of data. The images and metadata in *v*HMML Reading Room will be available permanently even if in the future they are migrated to new systems. HMML staff are in continuous communication with CSB/SJU IT Services to ensure that hardware and software are available to retrieve data as needed. As storage formats or media used by HMML become deprecated or superseded, data will be copied to the preferred newer system, while in some cases retaining the data in older formats or media.

9. Data management plan

Expected data to be generated by the project, with **location of storage and access**:

1. Immersive virtual environment. Generated with photogrammetry, onsite measurements, mapped coordinates, using Unity Game Engine + media file assets.

> **To be hosted and preserved for open access at USF's AVC (using the resources of the USF High Performance Computing Center). Mirrored in archival copy at Busa Archive. Downloadable via project website and GitHub.**

2. Website for the project (using Drupal install): with a simple "flyover" version of the immersive environment + data types listed below (map + links to emulations, media files, archival documents, oral histories [mirrored locally at USF].)

> **To be hosted and preserved for open access at USF's AVC. Mirrored in its entirety, once completed in early 2019, in an archival copy at Busa Archive.**

3. Simple pinned, zoomable GIS map (based on Google Maps) of the town of Gallarate, near Milan, with historical layer ca. 1961, showing the location of the center, the Aloisianum Jesuit college, and related light-industry sites in Gallarate.

> **To be hosted at USF's AVC as part of the project website.**

4. Software emulations. Javascript/HTML + images of machinery, configurable in combination.

> **To be hosted at Alberta and McGill by Rockwell and Sinclair, mirrored in archival copies at USF's AVC and the Busa Archive.**

5. Oral histories. MP3 recordings, edited from raw files + transcripts (translated from Italian) in plaintext and basic XML files (from which HTML text will be generated via XSLT).

> **To be hosted at USF AVC, mirrored in archival copies at Busa Archive.**

6. Textual documents in the Busa Archive. Sample correspondence, reports, segments of draft book MSS on data processing, flowcharts, original punched cards. To be scanned as hi-res TIFFs, with derived JPEGS, and transcribed (in XML files linked to page-images). Transcriptions by student workers at the Library (any necessary translations by members of our team) Metadata wrapper = modified Dublin Core tagset already used at Università Cattolica. Mirrored at USF AVC.

> **Copies mirrored at USF's AVC as part of the project website.**

7. Photographs from Busa Archive. 80 images of the center (CAAL) + several representative images illustrating the work of establishing it. Existing medium-high res JPEGs to be prepared for online publication. Metadata wrapper = modified Dublin Core tagset now in use at Università Cattolica.

Copies mirrored at USF AVC as part of the project website.

8. Short film of the opening ceremony of the center in its initial location in the Aloisianum college (before the move to the former factory). Currently on CD-ROM (MPEG-2), to be ripped and compressed for online access.

Original CD-ROM in the Busa Archive; web video to be added there, mirrored at USF's AVC as part of the project website.

9. Co-Authored white paper, as submitted to NEH spring 2019.

Archived at both USF (in the university's digital repository) and at the Università Cattolica in Milan.

10. Presentations and recordings from final panel at USF after the funding period.

Archived by USF (in the university's digital repository) and mirrored at the Università Cattolica in Milan—along with any resulting co-authored journal publication.

The status of data management for the project will be included in both the interim and the final performance reports to the NEH. A complete backup of all data from the entire project at the time of completion, April 2019, will be stored on the Library's servers at the Università Cattolica in Milan, under the supervision of Senna and Passarotti and with dedicated resources for maintenance and preservation via the Department of Special Collections.

All data generated by the project will be made freely available online or via GitHub via a Creative Commons License (CC-BY).

7. Data Management Plan

Curating East Africa will generate two different types of data. First, the project will generate meeting records, proposals, and administrative records associated with the project. Second, the project will generate software code. Each of these materials will be organized, managed, shared, and stored in a different fashion, respecting common practices in each area.

The project will generate administrative records. Administrative Records include emails, correspondence, meeting notes, and other communications. These administrative records will become the basis for a published project white paper. The raw materials from meetings will be printed and retained for a period of five years beyond the finish date of the proposal, in accordance with the Ohio Revised Code, Section 149.33. Eventually, these materials, along with the white paper, will be archived at Cleveland State University, in a publicly accessible archive. The white paper will be archived in the Cleveland State University Library's open-access BePress digital commons and immediately available online upon publication.

Source code for the Curatescape framework, including the original tools for Omeka and the new tools for WordPress, will be open source and made publicly available on GitHub, where interested scholars and developers may track its version history, submit modifications, report issues, and create their own derivative projects. This code will also be preserved on a private server that is backed up daily. After five years, these coding materials will be retained in electronic form, in accordance with the Ohio Revised Code, Section 149.33. At the end of the project period, this code will be archived at Cleveland State University, in its original format, which will be publicly accessible by request.

Responsibility for data management during the project period rests with Dr. Mark Souther, the project director. At the end of the project period, Dr. Souther will pass the materials along to the CSU Library, which will make the appropriate archival arrangements. At that point, the materials will become the property of the CSU Library, which will provide appropriate citation information to users.

The Cleveland State University Library archives materials in multiple places, depending on the particular collection. Print materials will be maintained by the library in long-term storage. Electronic materials and publications will be published within the University's BePress or ContentDM systems. Software code will be housed on servers, updated daily, and referenced through print and digital catalog. Direct downloading of materials may not be possible for electronic materials, such as software code. The rest of the material will be downloadable.

Finally, the Department of History & Archaeology at Maseno University will retain ownership of all its content generated for the project, including interpretive text, images (except those used with permission from another copyright holder), and audio and video recordings and excerpted clips derived therefrom. Also, Maseno University will place a copy of the white paper and other reports generated by the project into their respective print and digital archives.

9. Data Management Plan

I. Legal Considerations for the Use of Podcast Metadata

There are no intellectual property restrictions that obstruct the proposed project. Whereas most of the audio files saved in the PodcastRE database are still protected by copyright, the basic metadata that describes these files are NOT subject to copyright protection. According to the Digital Public Library's Policy Statement on Metadata, "the vast majority of metadata is not subject to copyright protection because it either expresses only objective facts (which are not original) or constitutes expression so limited by the number of ways the underlying ideas can be expressed that such expression has merged with those ideas. To be protectable, a work must be original, which means that it must contain at least a 'modicum' of creativity in its creation, selection, or arrangement. Facts and ideas may not be copyrighted" (DPLA, 2013). Moreover, PodcastRE Analytics uses the metadata in a transformative manner; we either transform existing metadata fields to allow them to be analyzed or visualized in new ways, or we generate new metadata fields entirely through forms of sonic analysis. Ultimately, PodcastRE Analytics has two layers of copyright law on its side: we are making a transformative fair use of metadata, which is generally understood not to be copyrightable in the first place. All other data generated as part of the project will be authored by the project team, and the PodcastRE Analytics platform will be built exclusively using open source software and code.

II. Expected Data that Project will Generate

Type of data	**When will they be shared?**	**How and under what conditions?**
Metadata XML files that describe every podcast in the PodcastRE database.	January 2019 (the start of the second dissemination phase of the work plan).	XML files will be indexed in platform's Solr index and freely available for users to query. XML files will also be preserved and made accessible by Minds@UW.
Open source code for PodcastRE Analytics platform, built using Python, Javascript, and Solr.	January 2019 (the start of the second dissemination phase of the work plan).	The code will be installed and operational on a UW-Madison server as a web application, freely available to all users. The code will also be freely available on GitHub.
Two research journal articles, output as PDFs.	Upon their publication (most likely before the end of 2019).	The articles will be published in open access journals and accessible via thes journal websites, PodcastRE website, and Minds@UW.
Blog posts reporting on the project's progress and early research results.	Immediately as they are posted online.	Blog posts will be available via the PodcastRE website and, by April 2019, Minds@UW.
Podcasts discussing project's progress and early research findings.	Immediately upon their completion.	Podcasts will be freely available via Apple iTunes, Stitcher, the PodcastRE website, and, by April 2019, Minds@UW.
White paper.	Immediately upon the completion of writing and editing.	White paper will be available via the PodcastRE website and Minds@UW before the end of 2019.

III. Period of Data Retention

All of the data described in Section II will be publicly available—in most cases, before the end of the funding period, but in some cases (such as the journal articles) up to a year after the funding period ends. The data will be retained during the grant period and for a minimum of 3 years afterwards on a secure Linux-based web server and XSAN local storage network, both of which are managed by the Instructional Media Center in UW-Madison's Department of Communication Arts. After the grant period ends, the data will be migrated to the University of Wisconsin-Madison Libraries and retained in Minds@UW, the Library's Fedora Commons-based Institutional Repository and preservation platform, for a minimum of 20 years.

IV. Data Formats and Dissemination

- Podcast metadata will be saved in XML files and indexed in PodcastRE Analytics's Solr index and for users to freely query. The XML files will also be preserved and made accessible in Minds@UW.
- The Python, Javascript, and Solr code of the PodcastRE Analytics platform will be installed and operational on a UW-Madison server as a free web application. The code will also be freely available for download on GitHub, where it will be maintained by Eric Hoyt.
- Research journal articles will be published open access in PDF format. They will be disseminated via the journal websites and PodcastRE websites and preserved in Minds@UW.
- Blog posts will be published and disseminated on the PodcastRE website, but also output as HTML files for preservation in Minds@UW.
- Podcasts will be output as MP3s and disseminated via Apple iTunes, Stitcher, and the PodcastRE website. They will also be deposited and preserved in Minds@UW.
- White paper will be disseminated on PodcastRE website and preserved in Minds@UW.

V. Data Storage and Preservation of Access

During the grant period, the work on PodcastRE Analytics will take place on the workstations, networks, and facilities managed by the Instructional Media Center, housed in UW-Madison's Department of Communication Arts. The specific Instructional Media Center resources we will be utilizing include the local XSAN storage network (16 TB), Linux web server (4 TB), and networked Mac Pro workstations (there are 38 different workstations available, any of which the project team can use with authorized network accounts). A letter of commitment from Erik Gunneson, director of the Instructional Media Center, is included as part of this proposal.

During the final three months of the grant period, Eric Hoyt (lead developer and one of the two project directors) will take responsibility for insuring the data generated by the project is migrated from the Instructional Media Center to Minds@UW, the University of Wisconsin-Madison Library's Fedora Commons-based preservation platform. Minds@UW will preserve the project data for a minimum of 20 years.

VI. Contingencies

The participation of two project directors and two established university units provides contingencies in the face of unexpected events that, while unlikely, could potentially befall any project. Eric Hoyt is the project director ultimately responsible for making sure this Data Management Plan is followed. However, if Hoyt were to die, become unable to work, or leave the University of Wisconsin-Madison, then Jeremy Morris will assume his responsibilities related to PodcastRE Analytics and the Data Management Plan. Similarly, if Morris leaves the institution, dies, or suffers serious injury, then Hoyt will take responsibility for the completion of PodcastRE Analytics.

Data Management Plan

People: Brooks Hefner, Edward Timke, Kevin Hegg, and James Madison University and University of California, Berkeley student research assistants (to be determined).

Kevin Hegg, with Libraries and Educational Technologies (LET) at James Madison University, will be responsible for the training of students for access and data entry.

Yasmeen Shorish (Data Services Librarian) at James Madison University will advise on data management protocols and procedures.

Brooks Hefner will be ultimately responsible for the stewardship of the data.

The Center for Open Science's Open Science Framework (OSF) will be responsible for dissemination. James Madison University's IT will maintain a dark archive copy as well.

Expected data

The project will generate tabular data in spreadsheet format (CSV), digitized from print copies of reports submitted to the Audit Bureau of Circulations, and held at the Library of Congress. The data will be transcribed into spreadsheets via the GoogleSheets interface. Spreadsheet data will be transformed to a SQL database on a LET server via Python scripts, undergo quality assurance (QA), and then be outputted to JSON. JSON data will underlie the visualization tools hosted on the project website.

LET servers undergo weekly back-up to tape. QA and JSON will be stored on the OSF. Data stored on the OSF is backed by a $250,000 preservation fund that will provide for persistence of your data, even if the Center for Open Science runs out of funding.

Period of data retention

Upon conclusion of the grant period, all data detailed above (CSV, SQL, JSON, associated scripts and metadata) will be fully available and accessible via the Open Science Framework for an indefinite period of time.

Data formats and dissemination

Data will exist in CSV, SQL, Python, and JSON formats. We will utilize MySQLdump and ReproZip to make SQL data available in a shareable format. Metadata information will be captured in ReadMe files, and a data dictionary will accompany the spreadsheet data. This information will define variable names and units, and will follow spreadsheet best practices (e.g., atomized unit-level information per column).

Data - including scripts, ReadMe files, and procedural information - will be available via the Open Science Framework without embargo or restriction at the conclusion of the grant period. A DOI will be generated for the project space.

Data storage and preservation of access

Data will be available at the Open Science Framework project space. A backup copy will be housed on JMU central servers, where the project data will be stored, backed up, preserved (replicated to one offsite location), and made accessible for no less than ten years. Data stored on the OSF is backed by a $250,000 preservation fund that will provide for persistence of your data, even if the Center for Open Science runs out of funding.

Project Title: Developing *Diviner*, a digital platform
Institution: North Country Public Radio
Project Directors: Ellen Rocco
Grant Program: Digital Humanities Advancement Grant, Level 2

Data Management Plan, Attachment 9

For the Diviner Digital Platform:

Part of our grant covers the development of a more robust data management plan by our Digital Director, Digital Developer, and Advisory Board. They will specifically be designing a plan for updating our WordPress plug-ins and elements as external factors, like WordPress themes, changes.

Right now we envision the Diviner digital platform being housed as a bundle of plug-ins, elements, and themes on the WordPress Plug-In Directory (https://wordpress.org/plugins/). The packaging of this bundle and its uploading into the directory is part of our work plan, executed in the final five months of the project, and made "live" in January of 2019.

North Country Public Radio (through the Digital Developer and Digital Director) will take responsibility for updating and tweaking the elements as needed.

For Individual Projects Utilizing Diviner:

Primary Documents Storage (for all photos, audio, and metadata collected)

We project a maximum of 100 GB of data collected per year (between photographs, audio, and metadata) for every "active start-up" year of our North Country at Work project—going forward, this would be every year we add at least ten public photo collections along with a staff person soliciting and processing private and institutional donations. We imagine this to be between three and five years. For "maintenance years," which come after the start-up period and rely on unsolicited donations, we project a maximum of 20 GB per year.

Therefore, for at least the first decade of the project we will not need more than 1 terabyte of storage space for our primary materials. Our storage system for our collected 'primary documents' (full size photos, WAV audio clips, and background information in Microsoft Word documents) consists of a 1 TB hard-drive which the project computer (owned and maintained by North Country Public Radio) backs up to on a daily basis, as well as a Crashplan cloud account with a second back-up copy of all information.

It will be up to individual organizations that implement Diviner to manage primary materials in their largest forms.

Hosting

The initial test project for Diviner is North Country at Work. The NCAW website will be hosted by the same company that handles the websites North Country Public Radio, NCPR Music, and all other NCPR domains, the R1 Soft Backup Solution for NCPR's private VMware Private Cloud. It provides:

Backup (for data loss):

- SAN (SAN) array and failover disaster recovery SAN (DRSAN) in a RAID 10 configuration.
- Running backups two times per day for the entire server.
- R1Soft Server Backup Manager
- On & Offsite backup

Redundancy (for content availability):

- Fully redundant hardware designed to sustain simultaneous multiple failures at hard drive, network, power, and storage level.
- VMware Servers in secure virtual VLANs with firewalls and fully automated failover and hot migration tools for maximum uptime and resilience to hardware failures ensure near 100% availability.
- Snapshots and on demand backups ensure data integrity. Self-healing High Availability Platform means that there is no single point of failure
- In the unlikely event of a physical failure, the machine simply reboots on another node

The responsibility for choosing and maintaining hosting servers for websites is also up to each individual organization; this is our way of addressing primary document storage and the hosting of the NCAW website.

Go local: Building capacity for public history in York County, Maine

Data management plan

Data for this project will include the needs assessment, proposed projects that result from the planning process, and workshop agendas and evaluations.

We will preserve data via a York County Digital History website that provides links to the white paper for the project, to workshop agendas and evaluations, and a list of proposed public history projects.

We will also link this website to the Digital Maine, the Maine State Library's digital repository, which includes a portal for partner institutions.

Any published articles published related to the project also will linked to the website.