Abstract

The Museum of Indian Arts and Culture (MIAC) with partners the New Mexico State Library Tribal Libraries Program and the Indian Pueblo Cultural Center (IPCC), a nonprofit of the 19 Pueblo tribes of New Mexico, and repository stakeholders from Harvard University Law Library, the American Philosophical Society Library, and the US National Archives Office of Innovation, seeks $248,550 to implement "NLP and UX in Online Archives: A Project to Test Natural Language Processing (NLP) Efficacy with Scanned Archival Material and User Experience Validated Approaches to Automated and Human Topic Tagging." The project will build tools to contribute to National Digital Infrastructure by testing when and in what configurations computer automated textual processing is helpful to people seeking information in archives.

This project addresses the need to further bridge the gap between how people seek information and avenues that are open to them for finding information in scanned archival records. NLP and UX in Online Archives follows up on the team's previous experiment initiating computer automated indexing of scanned mass digitized archival records to enable browsable interfaces as an alternative to a straight search box as part of the Indigenous Digital Archive (LG-70-16-0047-16), a project to allow crowdsourced tags within pages of documents in a toolkit overlaying Omeka-S. Clearly, computer assisted indexing through Natural Language Processing, a range of techniques for analyzing texts that identify things like mentions of people, places, dates, and other topics, has a productive role to play in helping people find information of interest. The question remains of how to configure the possible steps in NLP to be best suited for scanned archival material, and what of those results are the most useful for people seeking information, and what arrangements and features of an online repository interface informed by this are most useful and satisfying for people to use?

With a focus on users that is so far unprecedented in NLP work, our three year project will: a) follow established standards for statistical significance to select a range of scanned archival documents and add topic tags by experts to about 1000 pages to create a testing corpus for benchmarking potential NLP techniques; b) run a series of empirical tests on the data to develop a set of NLP and data prep techniques that optimize the information returned; c) add various configurations of NLP results, series information, and different content and conduct evaluations of the online repository interface for usability, including the Indigenous Evaluation Framework (NSF grant REC-0438720) evaluations, until the final iteration achieves at least 90% usability and satisfaction; d) validate the interface against accessibility audits and one on one tests to ensure usability by people with disabilities; e) make the open source software code and testing corpus available on Github; f) disseminate results and receive further shaping feedback through presentations at 6 conferences connecting with diverse audiences of Native and non Native library, archives, and museum practitioners, researchers, and community members.

The results of this project will 1) allow libraries, museums, and archives to make more effective choices in how to provide access to mass digitized or born digital content; 2) increase experience and confidence with online repositories among historically unserved communities; 3) enable other practitioners to test future experiments with our archival testing corpus.

Statement of National Need

Libraries and archives have made increasing amounts of scanned documents available online for access, but it can still be hard to connect people with the information they seek. Reviewing user studies of people seeking information in archival finding aids, Cruikshank et al. (2005) note that finding aids aids "are not transparent to user" and that among other suggestions, archivists need to "simplify options for searching and browsing." People do want to be able to do explorations of content across documents and across collections (Lonie 2018; Cruikshank 2005; Duff and Johnson 2002). This can be particularly important when trying to follow the stories of people, communities, and events across documents in different collections. As it is not always feasible to have the resources for people to create detailed descriptions of archival content, there has been interest and experiments in using computer automated indexing of content.

There is a need to experiment with and evaluate the potentials of Natural Language Processing for helping to connect people to archival content. *Natural Language Processing (NLP)* refers to a range of tools that can be used for computer assisted analysis of texts. One prevalent tool is *topic modeling*, described in the documentation for MALLET, a core topic modeling tool, as providing "a simple way to analyze large volumes of unlabeled text" where a "topic" consists of a cluster of words that frequently occur together in a corpus of text (Mimno 2018). In NLP, a *corpus* is a selection of pieces of electronic texts (reports, correspondence, oral transcriptions, etc) compiled according to goals of representing a variety of language for linguistics research (Caruso, et al., 2014). A *gold standard corpus* is a human evaluated and annotated set of texts that because of the number, size, and variety of texts and accurate annotations by humans can be used for modeling, fine-tuning, and testing NLP (Juckett 2012). Outside of NLP tools, software tools that perform *OCR*, or *Optical Character Recognition*, can digitize the textual content of printed or typescript archival material in paper form into electronic text form usings scans or photographs of the physical pages, thereby allowing the content to be accessed and analyzed as machine searchable and readable text. *Handwritten Text Recognition*, or *Handwriting Recognition (HTR)*, a rapidly developing field, enables automated translation of handwritten material into electronic text (Dunley 2018). *Usability testing* and *usability studies* are a part of the field of studying User Experience (UX), and form a core method of evaluating how well tools and online interfaces are effective for their purpose and are able to meet people's needs.

<u>Trends in NLP Practice:</u> To date, much of the investigation with NLP has focused on characterising the body of records, rather focusing on being being able to get information out of them. These efforts borrow largely from techniques now prevalent in the digital humanities to analyze a large set of texts, or dataset, like finding the emergence and frequency of topics, or analyzing networks as evidenced by named people communicating or collaborating with each

other (Graham, et al, 2012; Posner, et al. 2016).[1] Indeed, manifesting the prevalence of these methods when considering computer automated processing of texts, these very data analyses were the ones chosen by NLP researchers commemorating an anniversary in NLP by studying the literature of the last 50 years (The NLP4NLP Corpus; Mariani, et al., 2019).

Previous Work on NLP in Archives: Archivists have begun experiments using NLP with archival material. In a new toolkit for the BitCurator tools for electronic records, BitCurator NLP (Lee 2018; Chassanoff, et al. 2016) draws on the tools now central in digital humanities, topic modeling and network analysis, as a way of characterizing born digital records. At the University of Illinois Champaign-Urbana, the Cybernetic Thought Collective Team (Anderson 2017) is applying topic modeling and network analysis to highlight emergence of ideas and relationships between the people involved in digitized collections of people central in the field of cybernetics, with the goal of enabling users to browse through network relationship maps, entity relationships maps, other relationships. ArchExtract, a brief experiment starting from an intern project at the Bancroft Library in 2014-15 using digitized items of the John Muir collection, developed a graphical user interface for explorations of topic modeling by MALLET, as a potential tool to assist archivists in arrangement and description (Ellings 2016, 2017). While ArchExtract has not been sustained, explorations of topic modeling derived from NLP are continued also in a 2015-2018 project by ePADD (IMLS LG-70-15-0242), a tool for appraisal, ingest, preservation, curation, and exploration of email archives, where ePADD uses NLP to surface groups of topics that people can browse. At the 2018 Society of American Archivists Research Forum, Stevenson of the Defense Threat Reduction Agency reported on development of automated processing tools that allow a person to request a search for information from scanned archival documents, see what's returned, and then tell the machine whether they found the results of a search useful. However, this technology is developed within the framework of a nuclear safety assurance program, and it appears unlikely that information about techniques and tools will be available to outsiders any time soon.[2]

Recent work highlights the need to benchmark and experiment with NLP techniques. Reviewing projects that have explored NLP for archives, such as BitCurator and ArchExtract, Hutchinson (2017) does some further experiments with topic modeling analysis to evaluate it as a way of characterising records, and investigate its current utility for context based tasks finding documents with personal information that requires restriction. After creating an instance of

[1] Sentiment analysis has been emerging as a technique for determining degrees of positive or negative emotions represented in a body of texts. In an introduction to the topic which uses a case study of the available corpus of Enron emails, Saldaña (2018) explains that "[c]ombined with other NLP methods like topic modeling, sentiment analysis provides a means of characterising the emotions expressed about different topics of conversation. When used in conjunction with network analysis it could shed light on the ways that individuals interact with one another." That is it suggested as useful for enhancing topic modeling and network analysis underscores the current centrality of topic modeling and network analysis and the overall focus on NLP tools for characterizing a body of texts, rather than as tools for finding particular information.

[2] At the same time, this project does highlight the need for recursive feedback from people about whether and how automated processing techniques are helpful in their seeking information, and would likely benefit from usability studies.

ArchExtract and testing it with additional, born digital content, Hutchinson found that he received results that he found useful for the purposes of generating information descriptive of scope and content. The topic modeling produced lists of "topics", or clusters of words, such as this, in a folder of electronic records:

> "Topic 11: management, budget, projects, financial, project, external, including, impact, institutional, activities, reporting, personnel, funds, contract, cost, benefits, resources, implementing, operating, progress".

Undoubtedly, these results require a certain amount of extrapolation by the person viewing them.

Need for Testing NLP Methods, and an Appropriate Gold Standard Corpus related to Typical Archival Holdings: Overall, Hutchinson found topic model had power in identifying general themes, and in identifying records to examine more closely. However, Hutchinson notes that in order to be able to make it of use for finding particular content, that is, to identify particular documents to read, there is need for experimenting with different combinations of pre-processing, training of the NLP, and named entity extraction. Unfortunately, what's currently available and mainly used as testing corpuses are medical dictations, recent magazine articles, and other recent born digital material (Juckett 2012; Caruso 2014). There is a lack of a suitable gold standard corpus that would enable such testing NLP techniques relevant to a wide variety of archival material, especially the later 19th and early to mid 20th century archival documents that comprise a large portion of repository holdings.

Need for Usability Studies: There is a need for usability studies in order to enable a focus on people's needs relative to archival records, beyond looking at the records as a body of data. While the 2015-2018 ePADD project held quarterly meetings with stakeholders to receive feedback about features, formal usability studies of the interface and User Experience (UX) have not yet been a part of the project. Indeed, a keyword search of IMLS awarded grants to libraries over the past decade shows there are just two awarded grants from 2003 to early 2019 that involve usability studies to evaluate and inform a project. One, a software creation project, includes recursive usability testing with stakeholders as part of the 2016-17 National Leadership Grant project to create Cobweb, an open source web archive tool for collaborative identification and preservation of websites (LG-70-16-0093-16). The other was a 2015 small grant ($50k, LG-82-15-0166-15) for Savannah College of Art and Design's UX Design assessment tool for library spaces.

Need to Incorporate Archival Context Information: There is also a need for user validated approaches to effectively combining archival context information, such as series information, to information derived from NLP methods. While the a core topic modeling tool that underlies ArchExtract and other archival explorations, MALLET, explicitly identifies topic modeling as useful to analyze large volumes of unlabeled text (Mimno 2018), most archival collections, in fact, are not truly unlabeled. They have contextual information, such as series information and creator, from the available archival description. In her 1998 *American Archivist*

article evaluating the applicability of NLP to archival material and discoverability, Greenberg noted both potential for NLP to provide a more user friendly approach to searching archival material, and the need to combine NLP information with archival context like series. Twenty years later, there is still need for user validated experimentation to see how to effectively combine archival context information with NLP derived characterizations.

Addressing Needs: As part of "[e]xploring methods, tools, and techniques for sustainably and efficiently providing access to digital content and collections at scale for users of all interests and skill levels" (National Digital Infrastructures and Initiatives, NOFO), we'll fill the need for studies on how NLP can be most effectively trained to work with archival material, including by creating and testing with a gold standard corpus of documents that predate the fairly recent test texts used to train NLP processing systems, and test ways to combine results with archival description. We'll use recursive usability and accessibility testing to see how NLP results can best be incorporated into the user interface.

We understand that this work complements rather than replaces important sources of description such as crowdsourcing, or community sourcing, noted by Yakel (2011; Krause and Yakel 2007) for its conduciveness to sharing the authority for archival descriptions, bringing in diverse voices, and fostering collaboration. In our 2016-2019 IMLS National Leadership Grant, the Indigenous Digital Archive (LG-70-16-0047-16), we created a toolkit layer for Omeka-S to enable people to add International Image Interoperability Framework (IIIF) and Open Annotation compliant tags, annotations, counternarratives, and redactions of sensitive content as needed. We complemented a focus on crowdsourcing, or community sourcing, with an automated initial run of NLP to create tags in order to create to meet the goal of people being able to browse content as they start, and in order to design a browsable, or "generous" interface (Whitelaw 2015; Oates and Whitelaw 2018; EuropeanaTech Community 2019) that provides some insight into content that can be found, we made an initial experiment with incorporating NLP generated tags. This new project draws on the promise of this method seen through community workshops and in our pilot project of Fellows drawn from the 23 tribes of New Mexico plus Hopi, and also the need identified for grounded experiments to benchmark, optimize, and effectively combine the information that can be gained from NLP techniques.

We build on our earlier experiments using NLP to aid browsing and incorporate repository stakeholders from Harvard Law Library, the National Archives Office of Innovation, the American Philosophical Society Library, and Truman Technologies, a frequent consultant to libraries of Stanford, the New York Art Resources Consortium, and others, to increase applicability of our results to various material and software systems.

Project Design

Our project will provide benchmarks on the effectiveness of computer assisted indexing as validated by assessment of the value of the results for users, and the opportunities for accessibility and improved user experience. We make the assumption, indicated by the fields of

usability studies and user centered design, that useful advances in NLP techniques and interface design will be produced through focusing on the end user, not on the data.

Our work will proceed under three basic themes roughly by year:

1. Data usability and UI improvements from expert review; dissemination about project and encouraging people to tag and transcribe on the site.
2. Data engineering NLP improvements; encouraging people to make transcriptions and importing that back into NLP.
3. More raw materials loaded in and transcribed "well" automatically.

Drawing from results from in an article to quantitatively analyze volumes and types of material required for testing corpuses able to produce statistically significant results, or gold standard corpuses (Juckett 2012), we'll develop a test corpus of human created or approved topic tags using 500-1000 pages of late 19th and early 20th century archival documents in the IDA. We'll use this corpus to benchmark against various ways of deriving NLP tags, and for evaluating how useful it is to present even ideal topic tags to a person seeking to access information in a mass of archival material. Changes in the current IDA toolkit workflow we'd test for effectiveness include: 1) Only showing administrator approved tags. 2) Seeding the tagging with large sets of controlled vocabulary, for example, from GeoNames, Library of Congress Subject Headings, Virtual International Authority File (VIAF), and other sources of linked open data. 3) Allowing users to manage individual tags. 4) Experimenting with automatic summary generation, and perhaps subsequent tags, which could be weighted. 5) Re-tagging documents when a transcription is generated. 6) Using machine pre-processing of tags to filter out spurious tags from OCR errors. 7) Making use of handwriting recognition to supplement human created transcription. 8) Reviewing the user experience for how and which tags are shown, and to which audience.

We will conduct baseline usability and accessibility testing of the IDA toolkit for Omeka-S developed under LG-70-16-0047-16 using stakeholder advising and creation of user profiles. (This project is 99% complete, as we enter the final months of the performance period.) We will use our NLP testing results to refine the automated tags that aid browsing and encourage community sourced tags and annotations on our use case of ~250,000 pages of government records from the 1850s to 1960s (bulk 1880s-1930s). We will test to see which aspects of NLP results are useful to display, and useful ways to incorporate them into the user interface. In particular we'll attend to how contextual information from traditional archival description such as series information can be most usefully combined with the automated indexing from NLP.

For most of the duration of the project usability testing will be qualitative, take place through arranged appointment in person when feasible, and otherwise in an online one on one session. We'll follow the Usability.gov (n.d.) guidelines for conducting usability tests, using Concurrent Think Aloud (CTA) method, which helps "Understand participants' thoughts as they

occur and as they attempt to work through issues they encounter [and] [e]licit real-time feedback and emotional responses" (Bergstorm 2013). We will aim for five participants per test, following the optimization recommendations of Neilson (2012, 2000), who graphs results to argue that usability testing with 2-5 people being optimal for many low overhead projects in an agile environment. Each testing session will be comprised of one day of prep, a day of working with people doing the testing, and a day analyzing the results. In preparing for and conducting the tests, we'll use Travis's (2013) one page usability testing plan / dashboard. A one on one session of usability testing can be expected to last 1 to 1.5 hours. Participant compensation is detailed in the Budget Justification.

Analysis of usability tests using NLP variations will be quantitative, and to enable testing multiple user interface variations with larger number of people, we'll use a subscription testing platform, UsabilityHub.com, an affordable service which allows the type of A/B testing (quick checks of which of two variations do people prefer) and analysis that is being found to be so effective in the field of optimization, or shaping websites and other electronic communications for maximum effectiveness (Kohavi and Tomke 2017).

We'll ask one on one usability testing participants to sign a waiver allowing us to use their comments and feedback with project team and possibly in reporting. All comments will be anonymized. For privacy concerns, one of the usability testing information from any of our sources will be considered as a dataset, and it will not be retained once information is gathered from analysis for decision making and reporting.

More detail on the steps to be taken, sequence, and duration in available in the attached Schedule of Completion, Budget Justification, and supplemental Development and Usability Testing work plans.

Two meetings a year with advisory panel members and repository stakeholder advisors help allow for input and consensus building across a wide diversity of repositories. When the project's final UX iteration is approved, our documentation includes a professional technical writer creating user guides and screencasts for common tasks, informed by the usability testing. We will disseminate results and seek additional feedback and input with workshops and conferences, enumerated in the Diversity Plan section below; usability testing outreach; blog posts; and a white paper.

Key Personnel: From the IDA project, we retain collaborating advisors from our partners, our advisory panel and tech advisory panel from the IDA project: Our advisory panel of Native scholars, educators, practitioners, and community leaders, identified in the list of project staff and resume sections, are joined by Dr. Ricky Punzalan (U Maryland) and Dr. Helen Tibbo (UNC Chapel Hill), bringing additional emphasis on inclusion in digital repositories and reference services. Of our collaborating advisors, the newly retired State Library Tribal Libraries Program Coordinator, and founding advisory panel member Alan McGrattan continues on with her successor, former law librarian Faye Hadley, now just three weeks on the job, while Vina Begay, Library Director, continues from the Indian Pueblo Cultural Center. Dr. Anna

Naruta-Moya continues as project director, and continuing on the technical advisory panel are Mildred Walters (Diné, Tribal Librarian and Language Program Consultant), Dr. Rob Sanderson (Semantic Architect, Getty Institute), Glen Robson (IIIF Consortium), and UX designer George Oates (British Museum, Smithsonian, others). Team members from Digirati will provide software and NLP testing experiments, and oversee and participate in usability and accessibility testing and analysis. We're happy to have joining the core IDA team Donovan Pete (Diné), who is serving also a Technical and Research Fellow on the IDA's soon to be announced 2019-2020 sponsored partnership project with the US National Archives, which thanks to an anonymous donor will use the IDA for the repository supporting the creation of a portal for geographic exploration of the Ratified Indian Treaties, now newly conserved and scanned. Mr. Pete brings experience as a tribal library director, graphic and web designer, library usability tester, and linguistics scholar, and will be working closely with the NLP evaluation gold standard corpus creation, and usability testing and analysis. Evaluator Dr. Shelly Valdez (Laguna Pueblo) conducts evaluations for institutions nationwide including the Smithsonian and will be implementing evaluations using the Indigenous Framework for Evaluation.

New and important to this project are our repository stakeholder advisory panel, to provide perspectives from large repositories throughout the nation. They are: Steve Chapman, Manager, Digital Strategies for Collections, Harvard Law Library; Gail Truman of Truman Technologies, a consultant to Stanford University Libraries, California State Library, and the New York Art Resources Consortium (the libraries of the Frick, Brooklyn Museum, and MoMa), and now Cloud Services Product Manager for Oracle; Jason Clingerman, Digital Public Access Branch Chief, National Archives Office of Innovation; and Brian Carpenter, Native American Materials Curator, American Philosophical Society Library.

The financial resources needed are $248,550 in grant funds, with an additional $84,000 as cost share, though none is required.

Measuring Success: The progress of the project and the online interface will be measured by performance indicators in the following areas: evidence of deep engagement with the community of users; evidence that the interface is easy to use; evidence that the toolkits are functionally suitable for the purposes for which it was designed; and evidence that the use of the toolkits and gold corpus can be sustained after the project. Project performance indicators and targets are shown in the table below:

Success Area	Performance Indicators	Project Targets
Deep engagement with the community who would use it	1. Presentations at representative conferences 2. Workshops and feedback sessions with Native librarians; Indigenous Evaluation Framework sessions	1. Present at 6 conferences 2. 2 sessions at IPCC convening for tribal and Native serving librarians and others; 1 session

		with Tribal College Librarians Professional Development Institute
	3. Number of library, archive, or museum practitioners or interested researchers participating in usability tests	3. 100 people participate in one on one usability testing
Easy to use	4. Ability to perform key tasks as measured by usability tests	4. 90% success for last prototype tested
	5. User satisfaction	5. 90% positive satisfaction for the last prototype tested
Functionality and Reusability	6. Gold standard archival corpus meets established statistical significance standards	6. High quality tags added and/or edited on diverse selection of 30 documents (about 1000 pages)
	7. Software tools based on interoperability	7. Tags related to text through IIIF API.
Use can be sustained after the project	8. Results made available	8. White paper and at least 2 blog posts published
	9. Awareness of project and results (citations/references in non-project presentations, etc)	9. 10 mentions
	10. Deployable open source application and documentation	10. Deployed to Github
	11. "Gold corpus" text, tags, and documentation available for use by others	11. Made available on Github

Diversity Plan

The Indigenous Digital Archive project and work to more effectively connect people with historic records related to their communities grows from strategic planning and facilitated community listening sessions that Della Warrior (Otoe-Missouria) undertook after she became MIAC's first ever Native director in 2013. In these sessions, Native American constituents

expressed 1) wanting MIAC to provide online access to documents relating to their history, and 2) for MIAC to provide them opportunities to gain experience with archives, as they have needs gain skills in connecting with archival information as well as develop and operate their own governmental and cultural archives.

The impacts of multiple government agencies and jurisdictions on nearly every aspect of the lives of Native people prior to the Native American Self Determination Act of 1975 means there is a particular need in Native communities for access to especially large quantities of records in order to be able to trace information about a person, family, community, or event. The platform created by the Indigenous Digital Archive provided tools for people to add crowdsourced (or community sourced) tags and annotations help create access points to mass digitized documents and allow collaboration in research. MIAC's experience initiating NLP tools added tags to help provide browsability on the initial interface. Our experience in giving workshops with the interface and in conducting a pilot project of Fellows from the 23 tribes of New Mexico plus Hopi highlighted the appeal and usefulness of the approach, as well as the need for and high potential usefulness of a thorough investigation of what NLP techniques actually optimized information for a researcher, what combinations of archival context made that more effective, and how this information could best be combined in a user interface that's effective and satisfying to use.

Native librarians, leaders, scholars, and community members in our advisory panel and beyond continue to be a central part of the planning process to strategize how to connect people more effectively with archival records. Our usability testing and dissemination plans are designed to further connect with community members and diverse audience to share about the project and to gather feedback. Recruitment for usability testing is a form of dissemination about the project, and we will recruit for people to do usability testing one on one sessions that will take place either in person or online, depending on where they are located. Listservs we will use for recruitment and dissemination include the Society of American Archivists; Digital Library Federation; American Indian Library Association; the Tribal College and University Library Association, to include alumni of the Tribal College Librarians Professional Development Institute; and Native American Studies. We will do presentations and participant recruitment at six conferences of Native and non Native library, archives, and museum practitioners, researchers, and community members. This will include a presentation of findings at DLF Forum to connect with digital library specialists, at the SAA conference to share with a wide audience of archivists, and at the Association of Tribal Archives, Libraries, and Museums (ATALM) to connect with library, archives, and museum practitioners at small and large institutions. Based on initial interest by their constituents, we will do a dissemination and feedback gathering workshop at the Tribal College Librarians Professional Development Institute, an IMLS supported project, at Montana State University in the first week of June 2020. We will have two in person and one web convening of tribal librarians and others who serve Native audiences, at the Indian Pueblo Cultural Center in Albuquerque as a continuation of a multi-year collaborative

training program. At the in person sessions, in addition to some one on one usability testing, we will also do evaluation sessions using the Indigenous Evaluation Framework developed by Native researchers under NSF grant REC-0438720. We will also participate annually to present, recruit people to test usability, and gather feedback from people at Indigenous Pop X, an event near Albuquerque each November highlighting a range of Native intellectual activity.

We chose to participate in Indigenous Pop X three times over the course of the project because a) we are able to dovetail it with the convening at IPCC, nearby, of tribal librarians and others whose service includes tribal populations; b) emphasis on technology makes it a topical fit for an interested audience, and founder is particularly interested in the connection our toolset can enable people to make with documents related to their communities; c) there is no option to connect with this audience by listserv; d) the event has track record of garnering thoughtful and wide reaching national press; e) it gathers extremely diverse crowd, from Native to non Native, and decision makers at libraries even on the East Coast have been spotted there in the past; f) it is an extremely non-threatening environment and so people who might be intimidated by a digital repository may feel more comfortable with engaging; g) with its origins in Indigenous Comicon and emphasis on technology in the service of needs and activities, it draws high numbers of people with computer experience who like to tinker, increasing likelihood that our investment of time will encourage beginning or experienced (or even not-yet) Native software designers and coders to take on further research in this field.

National Impact

This project will produce information and tools that libraries and archives can apply in making decisions about and implementing ways to increase effective access to the content of online archival material through computer assisted indexing (NLP), ability to combine NLP with archival context like series information (a need articulated in Jane Greenberg's 1998 American Archivist article and unaddressed since), and improvements to online interfaces supported by well thought out user interface design (UI/UX) usability testing results. Our institutional stakeholders represent a variety of repositories nationwide, from government to academic to cultural institution, that will be able to draw on and sustain these results. Additionally, complementing the project's developing tools and assessing the possibilities of NLP for fulfilling users' information needs, the project will also contribute a reusable test corpus of 19th to mid 20th century archival material. This deliverable will allow further research and tool building, that previously has focused on recent and narrow bands of content, such as medical and scientific literature, due to the ease of availability. Making cultural heritage content available to computer science researchers enables them turning their research towards this more challenging field. A publicly disseminated white paper and posting of the gold standard corpus dataset and software source code on Github will ensure project deliverables are readily adaptable by other institutions and communities and allow others to build on these results.

MNMF / Museum of Indian Arts and Culture – Natural Language Processing and UX in Online Archives
Schedule of Completion

	2019						2020					
Activity	J	A	S	O	N	D	J	F	M	A	M	J
Administrative activities – review and signing of contracts; project kickoff meeting	■											
Set date for first advisory panel meeting, in September (web conference)	■											
Expert review of web interface with unmodified NLP; first round usability adjustments		■										
Baseline usability testing: Develop testing plan			■									
Advisors' kickoff web meeting; solicit stakeholder desired controlled vocabularies			■									
Internal QA of adjustments made following expert review				■								
DevOps deploys parallel site and content infrastructures for changes and testing				■								
Prep for IPCC meeting and presentation and recruitment at Indigenous Pop X (first week of November)				■								
IPCC joint meeting: in person usability testing; Indigenous Evaluation Framework sessions					■							
Presentation and participant recruitment at Indigenous Pop X					■							
Usability testing Round 1					■	■	■	■				
Advisory panel web meeting								■				
Usability testing, IPCC evaluation, + Indigenous Pop X feedback analysis and action plan								■	■			
Empirical NLP testing & improvements:							■	■	■	■	■	■
Add stakeholder desired controlled vocabularies to toolkit, replace only machine tags							■					
Select texts and develop “Gold corpus” of text well tagged by knowledgable people							■	■	■			
Integrate controlled vocabulary work and 3rd party identifiers (such as GeoNames)										■	■	■
Smart normalization of terms: people. Places, etc										■	■	■
Tag validation implemented to prune automated “junk” tags before ingest										■	■	■
Implement filtering tags on parts of speech (e.g. restrict proper nouns to noun phrases)										■	■	■
Omeka-S Platform Enhancements Phase 1										■	■	■
Updates to Omeka-S to support process of usability testing of new features										■	■	■
Implement use of non Omeka IDs for aggregating topics										■	■	■
Recursive UX improvements and accessibility and usability testing										■	■	■
Dissemination and feedback workshop at Tribal College Librarians Professional Development Institute, first week of June, Montana State University; Indigenous Evaluation Framework session												■

MNMF / Museum of Indian Arts and Culture – Natural Language Processing and UX in Online Archives
Schedule of Completion

	2020						2021					
Activity	J	A	S	O	N	D	J	F	M	A	M	J
Recursive UX improvements and accessibility and usability testing	■	■	■	■	■	■						
OCR Preprocessing and Human Transcription:	■	■	■									
Establish OCR processing rules, using measures of OCR quality to determine usefulness for tagging	■	■	■									
Enable software to automatically identify low quality OCR as candidates for human transcription	■	■	■									
Update software to prefer human transcriptions when available to feed tagging	■	■	■									
Handwriting Recognition: In house test two different available toolsets	■	■	■									
Explore Summary Generation for providing summaries of documents for 1) viewing 2) generating search results	■	■	■									
Advisory panel web meeting			■									
Usability testing participant recruitment via listservs (SAA, DLF, American Indian Library Association, Tribal College and University Library Association, Native American Studies)			■									
Prep for IPCC meeting and presentation and recruitment at Indigenous Pop X (first week of November)				■								
IPCC joint meeting: in person usability testing; Indigenous Evaluation Framework sessions				■								
Presentation and participant recruitment at Indigenous Pop X					■							
Evaluating NLP improvements using testing platform							■	■				
Omeka-S Platform Enhancements Phase 4 – improvements based on actionable recommendations from usability testing and accessibility testing						■	■	■	■			
Data validation and statistics: use lessons learned and usability testing results to identify automated ways of validating data. Generate statistics to inform white paper, blog posts.							■	■	■			
Advisory panel web meeting										■		
Add additional scanned documents to repository, evaluate whether they are being tagged well automatically										■	■	■

MNMF / Museum of Indian Arts and Culture – Natural Language Processing and UX in Online Archives
Schedule of Completion

	2021						2022					
Activity	J	A	S	O	N	D	J	F	M	A	M	J
Scope and write white paper	■	■	■	■								
Technical writer creates user guides as part of documentation of final interface	■	■	■	■								
Presentation at Society of American Archivists conference		■										
Presentation and demos at web conference of tribal librarians and others (virtual IPCC meeting)			■									
Presentation at ATALM, DLF				■								
Presentation at Indigenous Pop X					■							
Advisors receive white paper; contribute feedback					■	■	■					
Advisory panel web meeting							■					
Disseminate white paper								■				
Blog posts on blogs of Indigenous Digital Archive and Digirati or other venue									■			
Complete reporting to IMLS										■	■	

DIGITAL PRODUCT FORM

Introduction

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

Instructions

All applications must include a Digital Product Form.

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

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

Part I: Intellectual Property Rights and Permissions

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

The opensource software tools created or adapted for this project and gold standard text corpus dataset will be assigned the MIT Open Source license. This is to encourage flexibility in uses of the software tools, especially to enable the software to be used as libraries within other systems. The license will be listed in the license section documentation with the code repository at Github.

Tags and annotations that people add to archival content on the website are assigned a Creative Commons license of Attribution-NonCommercial CC BY-NC. This is to encourage people to contribute tags and other information to a library/museum/archive project without worrying that their contributions will be used for commercial purposes. This will be communicated in the Respectful Online Access section of the website and in user guides.

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

See A.1; otherwise not applicable to this project.

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

None of the data from people doing usability testing with us will be retained beyond analysis and generation of reports. We will ask people to sign a consent form allowing us to use their comments and feedback, and will delete potentially personally identifying information and anonymize responses to protect privacy.

The content in our repository is all from open federal records, copies of which can currently be purchased by anyone possessing the required funds. However, to be able to respond to concerns that there could be culturally sensitive material in the documents, our repository allows someone to redact a passage for confirmation by our advisory panel.

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

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

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

We will create a "gold standard corpus" of human tagged text for testing NLP processes. This will be comprised of a selection of approximately 30 scanned typescript archival documents totaling about 1000 pages, from the late 19th and early to mid 20th centuries. These will likely all be government documents, in the public domain. We will add tags to the documents using the International Image Interoperability Framework (IIIF) and Open Annotation international standards.

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

Our opensource software Indigenous Digital Archive toolkit allows tagging within documents with IIIF and Open Annotation compatible tags.

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

Per our repository's IIIF compatible presentation API, the digital images that will be tagged are JP2 images.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan. How will you monitor and evaluate your workflow and products?

Quality control will be monitored by multiple people working on the tagging for the gold standard corpus and through validation through NLP analyses.

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

The JPG images will be maintained in a Amazon Web Services hosting, with automated backup to DuraCloud. The tagged textual dataset forming the gold standard corpus will be deposited on Github.

C. Metadata

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

Descriptive metadata will be transferred from the original archival descriptions of the microfilmed records creating DACScompliant descriptive metadata. As Omeka-S is the content management system, the metadata is entered in Dublin Core format. As needed, we use the convention modeled by the Bentley Library to represent hierarchical arrangement in Dublin Core format, developed for use with Dspace (http://archivalintegration.blogspot.com/2015/09/whatwetalkaboutwhenwetalkabout.html).
User added metadata will be allowed to be entered freeform, but autocomplete will suggest the remaining text of a tag based on previous usage or availability of a semantic tag, and administrators will be able to edit or co

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

(continued) edit or consolidate tags to form an evolving data dictionary. Tags are stored as IIIF compliant metadata in JSON-LD.

Digital objects and metadata will be backed up on the TRAC Certified Trusted Digital Repository when created and ingested. The live server containing user supplied metadata is daily backed up and periodically backed up on the TRAC Certified Trusted Digital Repository.

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

Our use of the IIIF APIs enable harvesting discovery metadata from our repository via IIIF endpoints.

D. Access and Use

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

Openly available online, accessible via standard web browsers.

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

Indigenous Digital Archive
http://IndigenousDigitalArchive.org

Examples of online digital access projects of other organizations executed by our technical contractors:
Digirati:
http://wellcomelibrary.org/ for the Wellcome Trust
Good, Form & Spectacle:
http://whatsinthelibrary.com/ for the Wellcome Trust

Part III. Projects Developing Software

A. General Information

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

We will create improvements to the Indigenous Digital Archive toolkit, a toolkit layer that extends the abilities of the Omeka-S toolkit to create an interface that allows people to tag, annotate, and even redact portions of documents. The IDA toolkit is compliant with IIIF and Open Annotation formats, allowing the content and all added metadata to be exported with the record. In this project, we place an emphasis on developing a Natural Language Processing automated toolkit that aids in browsability of records, and refining the presentation of that information and the user interface through recursive usability testing.

Audiences for this tool include repositories that wish to increase access to mass digitized archival records.

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

A number of digital content management systems such as the opensource Mukurtu and Omeka enable registered users to add tags or comments. Our Indigenous Digital Archive software tool allows users to add tags within pages, rather than at the level of a digital object, and even redaction of a portion of a document for confirmation. This allows making large digitized series of documents available online without individual item cataloging. Also, IIIF and Open Annotation standards aid digital preservation and forward migration of user contributed content. Further, currently no other repository toolkits provide for automated OCR and NLP driven tagging upon ingest . We will further refine these capabilities through user validated testing.

B. Technical Information

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

The IDA uses Omeka S as the content management system, which uses PHP and MySQL. We also make use of Javascript code, with React, to provide client-side annotation capability, and image display. The Digital Library Cloud Service (DLCS) provides the IIIF hosting and also the text enrichment platform which manages the OCR and tagging of digital surrogates. Most DLCS services are either Python 3, .NET or Java, with the bulk of the services using Python with PostgreSQL or Amazon S3 storage for data persistence, with all data backed up using AWS managed services, and regular daily snapshots/dumps of relational database data. We use Python for rapid development, and because wide adoption within (See B.3 for full response.)

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

The IDA tool is accessible from any web browser. Designing annotation and semantic tagging around the International Image Interoperability Format (IIIF) means that this tool can extend software such as Omeka-S. We can interoperate with any IIIF compatible tooling, and the DLCS is designed to be an open source reusable platform which works with all digital surrogates which use the IIIF APIs. We make use of existing tooling from the natural language processing community, such as Spacy.io, and use existing cloud APIs such as Google Vision Document Text Detection for high quality, reusable OCR data.

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

The technology stack proposed in B.1 above allows the software system to be accessible via any modern web browser across all devices. There is no need for special software to be downloaded, or for anything to be insta

(continuation of B.1:) The IDA uses Omeka S as the content management system, which uses PHP and MySQL. We also make use of Javascript code, with React, to provide client-side annotation capability, and image display. The Digital Library Cloud Service (DLCS) provides the IIIF hosting and also the text enrichment platform which manages the OCR and tagging of digital surrogates. Most DLCS services are either Python 3, .NET or Java, with the bulk of the services using Python with PostgreSQL or Amazon S3 storage for data persistence, with all data backed up using AWS managed services, and regular daily snapshots/dumps of relational database data. We use Python for rapid development, and because wide adoption within the cultural heritage IT community means that there are a wide range of existing open source libraries and tools which we can use to integrate with our platform. We use Java and .NET for some core services, because Digirati have a decade or more of supporting enterprise level applications at scale using both of the languages/platforms. Choice of language and datastore is made on an application by application basis, depending on which is best suited to provide: fast development turnaround; high performance; features.

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

Extensive in-line code comments and docstrings; Markdown and restructured text for more general code documentation and usage instructions; Readthedocs for exported documentation; Public blog posts summarising our outputs, which we have typically done using: Medium.com, Github pages, and Digirati's own web pages. We also make extensive use of social media for sharing and promoting access to the outputs of our work. Working documentation is often also provided via Github wiki pages, via Github gists, and via Google Docs and Google sheets for ease of collaborative access. Extensive interlinking between repositories containing code, repositories containing data, and public blog posts and documents ensures that the datasets, code, and documentation are all associated with each other. All repositories will contain a top level README wit

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

Sourcecode for the IDA project is being finalized for the end of the 2016-2019 performance period.
http://IndigenousDigitalArchive.org

Software created includes various toolkit components at https://github.com/digirati-co-uk

C. Access and Use

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

The opensource software tools created or adapted for this project and gold standard text corpus dataset will be assigned the MIT Open Source license. This is to encourage flexibility in uses of the software tools, especially to enable the software to be used as libraries within other systems. The license will be listed in the license section documentation with the code repository at Github.

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

The source code will be available via Github.com.

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

Name of publicly accessible source code repository:

Github

URL:

https://github.com/

Part IV: Projects Creating Datasets

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

We will generate two different kinds of datasets. 1) We will create a "gold standard corpus" of human tagged text for testing NLP processes. This will be comprised of a selection of approximately 30 scanned typescript archival documents totaling about 1000 pages, from the late 19th and early to mid 20th centuries. These will likely all be government documents, in the public domain. We will add tags to the documents using the International Image Interoperability Framework (IIIF) and Open Annotation international standards. 2) We will gather usability testing data from one on one evaluation sessions recursively throughout the project and from large scale A/B testing on a web platform service. We will retain this second dataset only long enough for analysis, decision making, and reporting.

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

Not applicable to this project.

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

In usability testing, we will collect some PII, specifically names, email addresses, and some general information about location. None of the data from people doing usability testing with us will be retained beyond analysis and generation of reports. We will ask people to sign a consent form allowing us to use their comments and feedback, and will delete potentially personally identifying information and anonymize responses to protect privacy.

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

We will not retain data from the usability testing. We will, however, obtain consent agreements for use of anonymized quotes. These consent forms along with copies of the white paper and grant reports will be part of the files of MIAC's Executive Director as the head of a state museum, and thereby be managed as permanent records of the State of New Mexico.

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

W3C Web Annotation Data model and protocol, and the IIIF Presentation and Image APIs to store and disseminate the core data. IIIF Presentation API manifests aggregating IIIF Image API images for digital image surrogates. Elasticsearch for copies of identified tags, which we will use to drive discovery UI, and also to aggregate the data and extract statistical information. The annotations are stored in a PostgreSQL database, hosted and managed on Amazon AWS. OCR data is also stored as JSON objects in Amazon S3. Retrieving this data requires the use of basic HTTP REST, and the W3C Web Annotation protocol. Elasticsearch APIs for accessing the aggregated tag data. We would expect to generate summary data in JSON and/or CSV format, and store and disseminate on Github.

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

Extensive in-line code comments and docstrings; Markdown and restructured text for more general code documentation and usage instructions; Readthedocs for exported documentation; Public blog posts summarising our outputs, which we have typically done using: Medium.com, Github pages, and Digirati's own web pages. We also make extensive use of social media for sharing and promoting access to the outputs of our work. Working documentation is often also provided via Github wiki pages, via Github gists, and via Google Docs and Google sheets for ease of collaborative access. Extensive interlinking between repositories containing code, repositories containing data, and public blog posts and documents ensures that the datasets, code, and documentation are all associated with each other. All repositories will contain a top level README wit

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

(A.6 continued) with Usage instructions, clear licensing and code contribution guidelines, and links on to further documentation, or to relevant datasets.

We will archive the gold standard corpus on Github, including with JSON and/or CSV format. We will publicize its availability in conference presentations, a white paper, and through blog posts and social media.

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

Name of repository:

Github

URL:

https://github.com/

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

Digirati would expect to review this data management plan as part of the release plan, with the data management reviewed at each milestone when we complete a deployment, or a release of a new major version of the software. Digirati are ISO27001 and CyberEssentials certified. Implementation for Digirati will be monitored by our Director of Technology and our Head of Library Solutions, and by the project's advisory panel.