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 Museum of Indian Arts and Culture (MIAC), a library, archives, and museum located in Santa Fe, under the fiscal sponsorship of the Museum of New Mexico Foundation, the applicant, requests a three year National Leadership Grant (NLG), Project Grant category, in the amount of $248,550, with an additional $84,000 as cost share, though none is required. We will look at ways to increase access to archival material online by amplifying methods that meet people's needs but are very time intensive, such as indexing. Natural Language Processing (NLP), an automated indexing tool, forms a tantalizing option for increasing access to content according to a person's interest. We fill the need for studies on how NLP can be most effectively trained to work with archival material, including documents that pre-date the fairly recent texts previously used to train NLP neural networks, and test ways to combine results with archival description. We'll use recursive usability testing, including accessibility testing for people having different disabilities, to see how to best incorporate appropriately scoped and trained NLP results into online repository user interfaces. We build on our earlier experiments using NLP automated tagging within scanned pages to build a body of starting tags to support a browsing interface in a software toolkit based on the Open Annotation standard and interoperability (Indigenous Digital Archive LG-70-16-0047-16). To assure applicability of our results to various material and software systems, we incorporate repository stakeholders from a range of institutions including the Harvard Law Library and Library Innovation Lab; the American Philosophical Society Library; the US National Archives; and Truman Technologies, consultant to Stanford University Libraries, California State Library, and the New York Art Resources Consortium (the libraries of the Frick, Brooklyn Museum, and MoMa). Continuing as partner organizations are the Indian Pueblo Cultural Center, an organization run by all 19 Pueblo tribes, and the New Mexico State Library Tribal Libraries Program. We'll disseminate results with workshops, usability testing outreach, social media, and a white paper.

National Need: There is great interest in approaches to increasing access to archival materials online in ways that amplify the kind of review and description that can be used as access points for archival material but without the overhead of people doing time intensive detailed indexing. Automated Natural Language Processing (NLP), able to index People, Places, Dates, and Organizations on a collection of texts, is becoming a viable option for increasing access to the content of archival documents according to a person's interest. 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*), our project addresses the lack of studies on how NLP can be most effectively trained to work with archival material, particularly with documents that pre-date the test texts that have previously been used to train the NLP neural net models, such as the documents of the 19th and early to mid 20th century which form a large part of paper archives in repositories nationally, and how those results are best presented in repository interfaces. There is need to benchmark the utility of NLP results, test possible refinements for its use with archival material, and develop empirical information about how to most effectively combine and present this information to people, including non specialists and people with disabilities. This project will also be the first to employ the Indigenous Evaluation Framework developed by Native scholars (NSF grant REC-0438720) on tools of the National Digital Platform.

Proposed Work Plan: 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 test for usability and accessibility the IDA toolkit integrated with Omeka-S (with International Image Interoperability Format, [IIIF] and Open Annotation) developed under LG-70-16-0047-16 using stakeholder advising and creation of user profiles. (Currently 99% complete, pending scheduling in 3 last days with UX designer George Oates.) 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. We'll do usability testing in person in conjunction with conference workshops and presentations each fall (SAA and DLF for researchers and practitioners, ATALM for Native practitioners interested who might work with the software or use case) and online throughout development using a UX testing platform.

Key Personnel: Our advisory panel of Native scholars, practitioners, and community leaders is 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. Dr. Anna Naruta-Moya continues as project director, and continuing technical advising are Dr. Rob Sanderson (Semantic Architect, Getty Institute), Glen Robson (IIIF Consortium), and UX designer George Oates (British Museum, Smithsonian, others). Digirati provides software and NLP testing experiments, and usability and accessibility testing specialists. Evaluator Dr. Shelly Valdez (Laguna Pueblo) conducts evaluations for institutions nationwide including the Smithsonian.

Projected 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 these results.

Performance Goals and Outcomes: We will: a) Produce and share for additional use by others a test corpus of human created or approved topic tags using 500-1000 pages of late 19th and early 20th century archival documents. b) Perform NLP workflow tests per workplan above. c) Conduct usability testing at and conduct workshops and presentations sharing results at 12 conferences or convenings over 3 years. Conduct usability and accessibility testing with at least 160 GLAM practitioners and researchers in person, and at least 60 people via an online testing service platform. d) Build capacity by including developing and experienced Native American professionals throughout the project. e) Gather and analyze automated website analytics, such as Google analytics that can include data on visitors to a website who do not become users. f) Publish code on github and results through a white paper and at least 2 blog postings.

Budget Summary: Grant funds of $248,550 over 3 years support: $99,750 NLP testing, software development, usability testing; $10,500 Native evaluation and user testing personnel; $5k creating tagged test corpus; $800 OCR processing; $11k online usability testing fees; $43,800 project director; $5k tech writer; $13k for assembling tribal librarians etc at IPCC twice; $54k testing & dissemination travel, registrations; $10,800 advisory panel stipends; and $7,600 accounts payable support. An additional $84k in non-required match is server fees, digital preservation, and maintenance subscription ($28k/year).