Virginia Polytechnic Institute and State University

Continuing Education to Advance Web Archiving

Virginia Tech Libraries and the Department of Computer Science, in collaboration with Los Alamos National Laboratory Research Library, Old Dominion University Department of Computer Science, the University of Waterloo's Department of History, the Internet Archive, and George Washington University Libraries, request $249,292 from IMLS for a 2-year *project grant* in the *National Digital Platform* category. Starting from May 1, 2018, we will develop a *continuing education* curriculum and teach library and archive professionals advanced web archiving and analysis techniques. The deliverables include a collection of educational resources, cyberinfrastructure for deploying tools to support the curriculum (including source code), and other related resources.

Statement of Broad Need

Most librarians or archivists do not consider themselves data scientists. But when it comes to web archiving, they are often expected to perform identical tasks. In addition to collecting, archiving, indexing, and curating web content, they also need to clean, compare, arrange, weave, extract, reconstruct, visualize, and more. Since the web is by nature of high volume, velocity, variety, and veracity, big data analytics skills are often implicitly or explicitly assumed. Very few librarians or archivists are trained for these skills, which offers a convincing explanation why so few libraries and archives are providing web archiving and analysis services proportional to the needs of their patrons, even though petabytes of web content has already been archived and a suite of open-source tools, many developed as part of the National Digital Platform, are available to analyze them. Riding on the success of three consecutive Web Archiving and Digital Libraries (WADL) workshops, NSF funded Wikiversity curriculum on Digital Libraries, and four international Archives Unleashed hackathon/datathons, we mobilize and organize a group of expert web archiving technologists, librarians, and researchers to produce a continuing education curriculum to bridge the skill gap.

Project Design

Advancing web archiving requires constructing a well-organized curriculum that brings together the scattered knowledge in the field, constructing related course materials, integrating these with a suitable packaging of key tools, developing effective pedagogy that employs project and problem based learning, field testing in diverse sites, and broad dissemination.

The project will hold one planning meeting at the beginning of the project and one implementation meeting one year later, both as part of the annual WADL workshops. These two meetings will involve tool builders, invited library and archive experts practicing web archiving, and invited web archive researchers. The project team will hold a face-to-face tutorial at a major library conference, and manage up to 15 travel grants to interested librarians and archivists. Feedback from these will be used to further improve the educational resources and curriculum design. We will then hold up to 6 online webinars, possibly through ASIS&T, NISO, NFAIS, or other library and archive organizations. Evaluations will be conducted at all these events. By the end of the project, all course materials will be released online and free of charge, possibly through a major MOOC platform. These open education resources will empower those teaching courses and those completing the courses to continue using the online MOOC platform, and/or related resources, to further propagate the educational opportunities to other librarians, archivists, archive users, and learners.

The proposed project primarily targets library and archive professionals interested in serving the web archiving and knowledge extraction needs of their constituents, and is tailored to their needs and skill levels. The open source tools, protocols, and research areas covered in the curriculum include but are not limited to: Memento and the Time Travel service; Social Media Archiving and Analysis; ArchiveSpark based web archive segmentation, extraction, indexing, and analysis; Warcbase and web archive analysis; Storytelling; and Focused Crawling for events. The curriculum will emphasize 1) a deep understanding of web archiving technologies, web architecture, large scale data analytics and their enabling cyberinfrastructure and technologies and 2) hands-on experience operating these tools to tackle real-world research problems. Data sets and research questions solicited from web archive users will be used throughout the curriculum. We will also invite tool builders to 1) participate in the curriculum design, course materials preparation, and teaching; 2) lower the technology barriers by automating tool deployments using DevOps and Infrastructure as Code techniques; and 3) build class projects and grading methods. The course materials will be evaluated by iSchool experts before, during, and after being delivered to trainees through one face-to-face workshop and multiple online webinars.

Diversity Plan

Many academic libraries are exploring establishing web archiving as a new library service, opening up new career opportunities to library and archive professionals. The project will recruit trainees through various library and archive organizations and include specific language to encourage members of historically underserved ethnic and racial groups to attend the on-site and online training. The call will be distributed to Historically Black Colleges and Universities Library Alliance, Association of Rural and Small Libraries, the Association of Tribal Archives, Libraries, and Museums, and many ALA Affiliate organizations with diverse backgrounds.

Broad Impact

The proposed curriculum emphasizes practical skills essential to conduct research in web science, data intensive science, and social media. Equipped with these skills, library and archive professionals will be able to go beyond their traditional role as information pointers, and form deeper alliances with researchers. This in turn will help accelerate transforming the libraries and archives from information repositories to knowledge producers. Highly skilled professionals will also increase the awareness and value of existing web archives. Through the collaboration of tool builders, library and archive professionals, and end users, this project will also showcase a path to propagate and transfer cutting-edge technologies to library and archive services. The eventual hosting of the curriculum at an open MOOC platform will extend our efforts beyond the project period and ignite further interests in this realm.

Budget Summary

We request from IMLS a total budget of $249,292 over the 2-year period without cost sharing. This includes $206,259 direct cost and $43,033 indirect cost, calculated at Virginia Tech's negotiated rate. The direct cost is further broken down into $51,259 VT portion of salaries and benefits, $40,000 travel expenses, $10,000 consultation fee, $10,000 Amazon Web Services fee, $15,000 travel grants to trainees, and four $20,000 subcontracts to Old Dominion University, University of Waterloo, Internet Archive, and George Washington University.