Continuing Education to Advance Web Archiving
Virginia Tech, et al.
Jan 16, 2018

Abstract

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, University of Waterloo's Department of History, the Internet Archive, and George Washington University Libraries, request funding from the IMLS Laura Bush 21st Century Librarian Program for a 2-year project grant in the National Digital Platform (NDP) project category. Starting from May 1, 2018, we will develop a continuing education curriculum and teach library and archive professionals advanced web archiving and analytics techniques.

In the past few decades, tens of petabytes of web content have been collected and archived by the Internet Archive, national libraries and archives, and a growing number of university, public, and special libraries and archives. Since the web is by nature of high volume, velocity, variety, and veracity, web archives are increasingly used in ways beyond traditional searching, browsing and close reading. Suites of open-source tools have been developed, many supported by IMLS in the NDP project category, to assist researchers conducting analyses and extracting knowledge. Very few librarians or archivists have been trained to understand, utilize, maintain, and manage these tools. This at least partially explains why so few libraries and archives are providing web archiving and analytics services that satisfactorily address the needs of their patrons. In this project, 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.

This project directly addresses the IMLS agency-level goal: Learning, with the main emphasis on "Train and develop museum and library professionals", but also approaches aspects on "Support communities of practice" and "Develop and provide inclusive and accessible learning opportunities". The goals of this project are to 1) train library and archive professionals to effectively use innovative web archiving tools to answer research questions and as a result, 2) enable new web archiving services based on these tools. Through a planning meeting and follow-up collaborations, the project team will produce a project- and problem-based curriculum for advanced web archiving training. It consists of approximately 6 course modules and associated training plans, covering a wide range of web archiving tools. To lower the learning and service implementation barrier, we will also produce open source software to automate the deployment of the tools. We will then recruit trainees and deliver the course through in-person and online workshops. Through this training, participants will gain better understanding of web archiving, and be able to effectively use advanced tools to answer research questions.

Continuing Education to Advance Web Archiving

Virginia Tech and its 5 partner organizations request $248,451 from the IMLS Laura Bush 21st Century Librarian Program for a *2-year project grant* in the *National Digital Platform* (NDP) project category. Starting from May 1, 2018, we will develop a *continuing education* curriculum and teach library and archive professionals advanced web archiving and analytics techniques. This project directly addresses the IMLS agency-level goal: *Learning*, with the main emphasis on "Train and develop museum and library professionals", but also approaches aspects on "Support communities of practice" and "Develop and provide inclusive and accessible learning opportunities". The deliverables include a collection of educational resources, a series of in-person and online training workshops, and cyberinfrastructure for deploying tools to support the curriculum and workshops (including source code).

Statement of Broad Need

This project addresses library and archive professionals' needs for advanced training in web archiving, particularly its associated analytical skills. Since the nation's innovation and economic growth are increasingly driven by data and data-intensive sciences and technologies, data literacy is becoming more critical to knowledge workers. Library and archive professionals are no exceptions. With libraries and archives around the world amassing more data, we need to put more emphasis on extracting knowledge and value from them. This has been made abundantly clear by the participation of JISC and IMLS in multiple rounds of the Digging Into Data Challenge[1]. Although many library and information science education programs are adding data and analytics components, working professionals still lack on-the-job training, especially advanced training that is customized to their knowledge and skill levels, and can be integrated into their job responsibilities, as well as prepare them for future library service developments.

Web archiving is a promising growth area for library and archive services. In the past few decades, tens of petabytes of web content have been collected and archived by memory institutions. These include the Internet Archive, national libraries and archives, and a growing number of university, public, and special libraries and archives. Since the web is by nature of high volume, velocity, variety, and veracity, web archives are increasingly used in ways beyond traditional searching, browsing and close reading (Graham, Milligan & Weingart, 2015, Brugger & Schroeder, 2017, Weber, Ognyanova & Kosterich, 2017). Suites of open-source tools have been developed, many supported by IMLS in the NDP project category, to assist researchers conducting analyses and extracting knowledge. These tools usually assume a high level of data literacy, sometimes even proficiency in big data processing and analysis. Yet, it is unreasonable to require patrons to be highly tech-savvy in order to use web archives. Neither is it realistic to perpetually fixate tool builders' time and efforts on customer support. Accordingly, very few librarians or archivists have been trained to understand, utilize, maintain, and manage these tools. This at least partially explains why so few libraries and archives are providing web archiving and analytics services that satisfactorily address the needs of their patrons.

[1] https://diggingintodata.org

This project aims to bridge this skill gap by training library and archive professionals to work on real-life web archive research questions using the cutting-edge tools developed for these purposes. They will be exposed to perspectives of researchers interested in archived content, archive patrons, and tool builders. The training will equip them with a deeper understanding of the patrons' needs, the web archives used as data sources, the tools developed to process the data, and the potential library services that can be offered based on the above.

This project will be beneficial to a wide spectrum of library and archive stakeholders. Library and archive professionals will acquire essential skills to enable or expand their career prospects in new, data service areas. Equipped with a better-trained workforce, libraries and archives can then expand their services from mere repositories to knowledge producers, and more specifically, offer web archiving and analytics services. Supported by trained web archive professionals, researchers can focus more on their domain research rather than on keeping abreast of technology and system changes. In addition, web archives will see more insightful explorations of their resources. Information Schools can also leverage the training modules developed in this project as capstone projects on top of the more general data science courses in their MLS/MLIS programs. Web archiving tool builders, through teaching and directly engaging prospective service providers, will gain deeper insights on the potential skill obstacles for deployments, so they can improve existing tools, expand their scope, and design better tools in the future.

This project builds upon the project team's large body of prior research, development, and outreach work, including:

1. Los Alamos National Lab Research Library and Old Dominion University: the Memento project[2] and related work, partially supported by the Library of Congress and the Andrew W. Mellon Foundation.
2. Internet Archive: IMLS NDP funded "Systems Interoperability and Collaborative Development for Web Archiving" (WASAPI)[3] and related work, e.g., ArchiveSpark.
3. Old Dominion University: IMLS NDP funded "Combining Social Media Storytelling with Web Archives" project[4], NSF funded "Web Archive Cooperative" project[5] and "Increasing the Value of Existing Web Archives" project[6], and related work.
4. Virginia Tech Libraries: IMLS NDP funded "Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse" (LCI) project[7] and related work.
5. Virginia Tech Department of Computer Science: NSF funded "Integrated Digital Event Archiving and Library" (IDEAL) project[8], "Global Event and Trend Archive Research" (GETAR) project[9], "Curriculum Development: Digital Libraries"[10], and related work.

[2] http://mementoweb.org/
[3] https://www.imls.gov/grants/awarded/lg-71-15-0174-15
[4] https://www.imls.gov/grants/awarded/lg-71-15-0077-15
[5] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1009392
[6] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1526700
[7] https://www.imls.gov/grants/awarded/lg-71-16-0037-16
[8] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1319578
[9] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1619028
[10] https://www.nsf.gov/awardsearch/showAward?AWD_ID=0535057

6. University of Waterloo: the “Archives Unleashed” project[11] and related work, supported by the Andrew W. Mellon Foundation and NSF
7. George Washington University Library: the “Social Feed Manager” project[12], partially funded by IMLS and the National Historical Publications & Records Commission.

This project consolidates and extends the team’s prior outreach and training activities, including but not limited to:

1. Wikiversity curriculum on Digital Libraries[13]
2. Three consecutive Web Archiving and Digital Libraries (WADL) workshops[14, 15, 16]
3. Four international Archives Unleashed hackathon/datathons[17]
4. Nine Archive-It Partner Meetings[18]

Specifically, the Wikiversity curriculum on Digital Libraries offers a blueprint for project- and problem-based curriculum development (Gruss et al, 2015, Oh et al, 2016). The WADL workshops (Fox & Xie, 2015; Fox, Xie & Klein, 2016; Fox, Xie & Klein, 2017) provide a cross-pollinating platform, where the synergy among researchers, librarians and archivists, educators, and tool builders helped to develop priorities and modules for web archiving training and service development, which ultimately led to this project. The Archives Unleashed hackathon/datathons, especially its NSF funded travel grants[19, 20], informed the in-person training portion of this project and also provided valuable insights on training librarians and archivists on big data analytics. Through the WASAPI project and the Archive-It Partner Meetings, the Internet Archive has established working relations with a lion’s share of the nation’s libraries and archives interested in providing web archiving services. This is the community this training project is targeting, though we expect a much broader community will be impacted too.

We will also leverage experience and tools developed through many other related training programs, most significantly the training delivery approach pioneered by Software Carpentry[21], Data Carpentry[22], and Library Carpentry[23] (Baker et al., 2016). These programs, along with the programming, data science, and big data analytics courses offered at many iSchools and MOOCs

[11] http://archivesunleashed.com/
[12] https://library.gwu.edu/scholarly-technology-group/social-feed-manager
[13] http://en.wikiversity.org/wiki/Digital_Libraries
[14] http://www.dlib.vt.edu/WADL2015/
[15] http://fox.cs.vt.edu/wadl2016.html
[16] http://fox.cs.vt.edu/wadl2017.html
[17] http://archivesunleashed.com/archives-unleashed-events/
[18] https://support.archive-it.org/hc/en-us/sections/115000428706
[19] https://nsf.gov/awardsearch/showAward?AWD_ID=1723430
[20] https://nsf.gov/awardsearch/showAward?AWD_ID=1624067
[21] https://software-carpentry.org
[22] http://www.datacarpentry.org
[23] https://librarycarpentry.github.io

like Coursera[24], edX[25], and Udacity[26], form the basis of fundamental knowledge, based on which our more specialized, advanced training program can be developed.

This project will also coordinate closely with the International Internet Preservation Consortium (IIPC) Training Working Group. IIPC has identified staff training as a top priority, and has compiled an inventory of web archiving training resources. The vast majority of existing training focuses on basic web archiving skills, e.g., Heritrix and Wayback Machine style web crawling, indexing, searching, and browsing. Although our curriculum will also include similar topics in the "Web Archiving Fundamentals" module, we will go several steps further, emphasizing how to innovatively use web archives to answer specific research questions. Our delivery approach is also different. We will combine lectures and presentations with project-based problem solving, as developed in the Wikiversity curriculum on Digital Libraries. Although this approach has also been explored in the Archives Unleashed hackathon/datathons, our curriculum will cover more than one set of tools. We will reveal relations between various tools and use cases to form a more holistic view of the field.

Another significant difference between this project and other training program is the inclusion of the tool deployment component. We realize that although many open source web archive analytics tools have been developed, deploying them typically requires significant technical know-how not readily available at the libraries and archives interested in offering these services. In order for the training to produce tangible results of value to end users, we must make explicit efforts to lower the deployment barrier, especially for those big data processing tools.

Project Design

The goals of this project are to 1) train library and archive professionals to effectively use innovative web archiving tools to answer research questions and as a result, 2) enable new web archiving services based on these tools.

To achieve these project goals, we will develop a curriculum that serves as a guided tour to show how innovative tools can be used to creatively solve research questions. Course modules will be developed using a project- and problem-based learning approach. The training will focus on how to install, configure, and use the tools, but will not delve into the inner workings of the tools and why they are built that way. We will produce a collection of educational resources to support the curriculum, and will also deliver the training, both in-person and online, to selected librarians and archivists. To ease the technical burden to install, configure, and maintain the cyberinfrastructure for tools deployment, we will also develop software for automatically deploying these tools in the cloud, and offer such cloud deployments for free to trainees throughout the training.

[24] https://www.coursera.org
[25] https://www.edx.org
[26] https://www.udacity.com

The proposed training targets two broad categories of library and archive professionals: 1) who holds a master's degree in information or library information science, has some experience in digital preservation and digital libraries; or 2) who is an IT professional with some digital library or archive development experience. We do not assume extensive programming experience or expertise, but expect that participants are not afraid of reading source code under guidance and typing command lines. Some modules assume the participants have some knowledge in basic programming, data science and/or big data analytics. If they do not already have these skills, we will recommend MOOC courses as a remedial measure.

The project team will first build a "Web Archiving Fundamentals" module and then build on this module to introduce 5 selected tool sets representative of cutting-edge web archiving research and development, schematically shown in Figure 1. Six solid blocks constitute the core modules to be developed in this project, although it is unnecessary to require every trainee to complete all six modules. Blocks in dotted outline denote essential skills not included in this program. As mentioned above, we will recommend openly available courses or training, e.g., MOOCs and/or Software/Data/Library Carpentry events, that may be taken to acquire these skills. Arrows in solid line denote hard dependencies. For example, no other course module should be taken without successfully completing the "Web Archiving Fundamentals" module. On the other hand, arrows in dotted line denote soft dependencies, such that understanding "Social Feed Manager" will help learning "Event Archiving" but is not required.

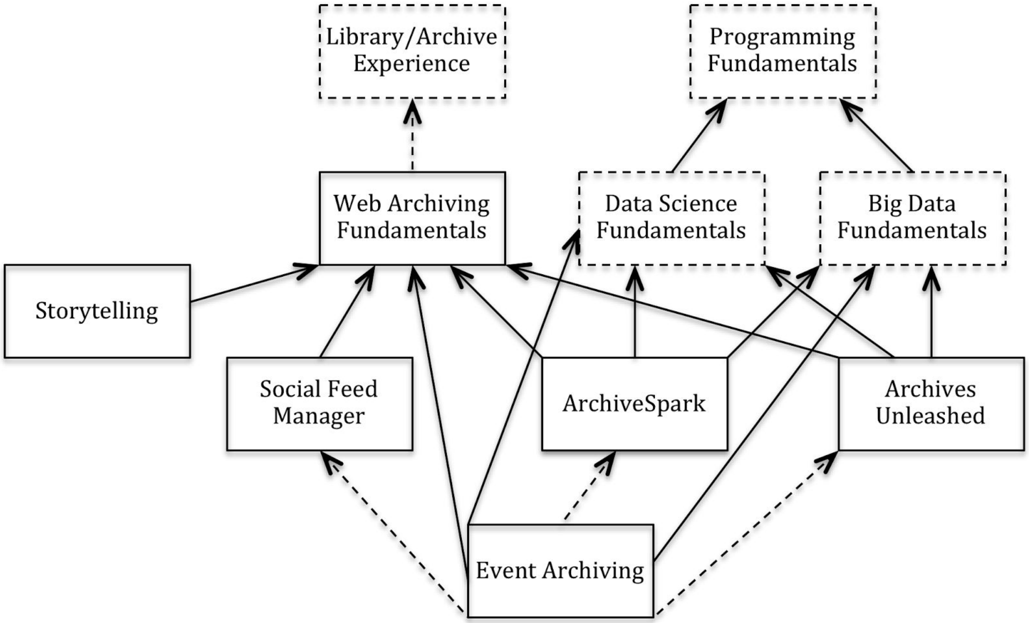


Figure 1. Curriculum Design

Module Descriptions

The following module descriptions are subject to change upon further discussion during the first project planning meeting.

The *Web Archiving Fundamentals* module will teach the web architecture and explain why web archiving is necessary and how it is done, e.g., using crawlers Heritrix or using Archive-It. This module will also introduce the Memento protocol, the TimeTravel service, and related work (Van de Sompel, Nelson, Sanderson, Balakireva, Ainsworth & Shankar, 2009; Van de Sompel, Nelson & Sanderson, 2013; Klein, Van de Sompel, Sanderson, Shankar, Balakireva, Zhou & Tobin, 2014; Ainsworth, Nelson & Van de Sompel, 2015; Jones, Nelson, Van de Sompel, 2016; Jones, Van de Sompel, Shankar, Klein, Tobin & Grover, 2016; Klein & Van de Sompel, 2017), as well as research resulting in better understanding of current web archives, their coverage and characteristics (Klein & Nelson, 2011; Klein & Nelson, 2014; AlNoamany, Alsum, Weigle & Nelson, 2013; Ainsworth & Nelson, 2015; Brunelle, Kelly, SalahEldeen, Weigle & Nelson, 2015; Brunelle, Weigle, & Nelson, 2017; Alkwai, Nelson & Weigle, 2017). This module will also include discussion on the ethics and limitations of web archiving.

Social media constitutes a specialized subset of the web. With the increased usage of social media, its archiving requires special attention. *Social Feed Manager* or SFM (Littman et al, 2016) is open source software for libraries, archives, cultural heritage institutions and research organizations. It empowers those communities' researchers, faculty, students, and archivists to define and create collections of data from social media platforms. SFM will harvest from Twitter, Tumblr, Flickr, and Sina Weibo, and is extensible to other platforms. In addition to collecting data from those platforms' APIs, it will collect linked web pages and media. The SFM team has also developed TweetSets, a service for sharing and reusing Twitter datasets and f(b)arc, a command line tool and library for collecting public Facebook data. The SFM team proactively collects social media datasets on a wide variety of topics, including government, news organizations, and elections. This module will introduce 1) social media APIs and the Twitter API, in depth; 2) tools for collecting social media data, specifically SFM; 3) tools for sharing existing datasets, including TweetSets; 4) ethical and privacy considerations in collecting social media data; 5) social media collection development strategies; 6) social media consultation as a library service; 7) overview of other sources for acquiring social media data; and 8) approaches to sharing social media datasets.

The *storytelling* module is derived from the outcome of the IMLS NDP funded "Combining Social Media Storytelling with Web Archives" project (AlNoamany, Weigle & Nelson, 2016a; AlNoamany, Weigle & Nelson, 2016b; AlNoamany, Weigle & Nelson, 2017). When a web archive grows larger, it becomes more difficult to understand its content. This module will demonstrate how to automatically extract summary stories from Archive-It collections. Events in these stories are summarized by sampling web pages from the Archive-It collections, arranged in a narrative structure ordered by time, and replayed through storytelling social media interfaces such as Storify. This module showcases how web archiving tools can be used to complement and alleviate the limitations of close reading.

Today's web archives can easily grow beyond traditional computation and storage limits. Doing research on such archives requires a totally different approach. The *Archives Unleashed* module is derived from a project with the same name, and the Archives Unleashed Toolkit, formerly known as Warcbase (Milligan, 2016; Ruest & Milligan, 2016; Milligan, Ruest & Lin, 2016; Milligan, Ruest & St.Onge, 2016; Maemura, Becker & Milligan, 2016; Jackson, Lin, Milligan & Ruest, 2016; Lin, Milligan, Wiebe & Zhou, 2017). The project aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past. The toolkit is an open-source platform built on Hadoop. The platform provides a flexible data model for storing and managing raw content as well as metadata, and for extracted knowledge. Tight integration with Hadoop provides powerful tools for analytics and data processing via Spark. The toolkit will be deployed in a cloud-based environment that will provide a one-stop portal for scholars to ingest their collections and execute a number of analyses with the click of a mouse. This particular training module will focus on analyzing archived web data using the Archives Unleashed Toolkit. We will demonstrate how web archives can be used to examine social science phenomena that have evolved over time, such as how news media is presented on the web or how individuals interact in online discussions. Participants are encouraged to explore the toolkit, the data, and learn how to design appropriate research studies that fully utilize large web archives.

As of Oct. 23, 2016, the Internet Archive alone has archived 273 billion webpages, taking up 15 petabytes of storage. The WASAPI project will enable interoperability among many large archives over APIs. Web archives typically do not offer in situ data analytics capabilities. However, even if moving petabytes of data becomes realistic, typical research cyberinfrastructure such as a 20-node Hadoop cluster cannot handle this much data. *ArchiveSpark* (Holzmann, Goel & Anand, 2016) allows dicing large web archives into manageable chunks, to be moved from web archives to big data platforms for analysis. This module will introduce the concepts, technologies, and APIs behind web archives interoperability, and demonstrate how to generate a corpus from a large web archive, build a metadata index, and perform queries against the generated files.

The *Event Archiving* module is derived from the outcomes of the NSF IDEAL and GETAR projects and related work (Yang, Fox, Wildemuth, Pomerantz & Oh, 2006; Gruss, Farag, Kanan, English, Zhang & Fox, 2015; Kanan & Fox, 2016; Oh, Yang, Pomerantz, Wildemuth & Fox, 2016; Kavanaugh, Sheetz, Sandoval-Almazan, Tedesco & Fox, 2016; Farag, Lee & Fox, 2017; Farag, Nakate & Fox, 2016; Castro, Chakravarty, Williamson, Pereira & Fox, 2017). It leverages and extends the capabilities of the Internet Archive to develop spontaneous event collections that can be permanently archived as well as searched and accessed. Through a new model-based approach to intelligent focused crawling, it improves the quality (e.g., accuracy, coverage, and elimination of noise) of collections of webpages so as to ensure comprehensiveness, balance, and low bias, as is needed for scholarly study of historically important events by social scientists. Event Archiving connects the processing of tweets and webpages, combining informal and formal media, to automatically detect important events, as well as to support building collections on chosen general or specific topics. Integrated services can then be performed on these collections, such as topic identification, categorization (building upon continuously evolving ontologies), sentiment analysis, and visualization of data, information, and context. This module will demonstrate the capabilities of event archiving.

Automated Deployment

This project will develop automated deployments to be used in training for the Social Feed Manager, ArchiveSpark, Archives Unleashed, and Event Archiving modules. This work is informed by the IMLS NDP funded LCI project and related work (Xie, Van de Sompel, Liu, van Reenen & Jordan, 2013; Xie, Chandrasekar & Fox, 2015; Xie, Chen, Jiang, Speer, Walters, Tarazaga & Kasarda, 2015; Xie, Chen, Speer, Walters, Tarazaga & Kasarda, 2015; Xie & Fox, 2017). The preliminary findings of the LCI project indicate that in comparison to other options, commercial cloud services still provide a better balance of low entry barrier and ease of use. Amazon Web Services is the most widely used cloud provider, therefore is chosen for this project. Virginia Tech Libraries has previously developed automated deployment for a Fedora/Samvera based institutional repository. We leveraged Vagrant and Ansible to deploy systems to either a local laptop, the library servers, or the Amazon or OpenStack clouds[27]. The project team has also developed docker based deployment in the Amazon Clouds for Fedora 4[28] and SFM[29].

The cost disadvantage associated with the commercial cloud usage will be offset by the project funding. During the training and a short period before and after the training sessions, all related cloud costs incurred by trainees will be covered by this project.

Activities and Timeline

At the beginning of the first year, the project team will meet in person for planning and discussion of the curriculum and its implementation. We will then develop course modules, prepare course materials, and develop automated deployment software. Subject to change, the total length of the training program for all six modules will be around 2 days, inclusive of lecture/presentation and problem-solving exercises. In the second year we plan to hold a full-day in-person training workshop and three full-day online training workshops, for a total duration of 4 days. This allows us to teach all modules twice during the project period, and have the chance to compare outcomes of online vs. in-person training, specifically on those modules perceived to be more difficult to deliver online. The project will also offer up to 15 travel grants for interested librarians and archivists to attend the in-person training. We will start recruiting trainees by the end of the first year. More details on the project timeline can be found in the attached document named “Scheduleofcompletion.pdf”.

We will weigh various options of meeting venues. In order to attract more participants in addition to those receiving travel support, we will give preference to co-locate the in-person training with a major library/archive meetings or Software/Data/Library Carpentry events.

[27] https://github.com/VTUL/InstallScripts
[28] https://github.com/fcrepo4-labs/fcrepo4-docker
[29] https://github.com/gwu-libraries/sfm-docker

Examples of the former include JCDL/WADL, Open Repositories, ACRL, and Code4Lib, etc. Online training may also be delivered through ASIS&T, NISO, NFAIS, or other library and archive organizations through their continuing education programs.

Personnel

The project team has a balanced mix of web archiving technologists and tool builders (Xie, Fox, Klein, Nelson, Littman, Goel), educators (Fox and Nelson), librarians (Xie, Klein, Littman), web archive representative (Bailey), library IT personnel (Xie and Littman), and researcher/patron (Milligan). We are advised by representatives from iSchools (Chen and Weber) and libraries (Griffin), who will help us adjust the curriculum to current and future library and archive professionals.

Project director Xie will provide leadership and project management services. He will also provide budgetary oversight, facilitate the trainee selection process, and coordinate the curriculum development and delivery through the in-person and online training workshops. The training module development and delivery work is allocated as following: Xie, Fox, and Klein, possibly also Nelson (Web Archiving Fundamentals); Nelson (Storytelling); Littman (Social Feed Manager); Goel, Praetzellis, and Bailey (ArchiveSpark); Milligan (Archives Unleashed); and Fox (Event Archiving). The automated deployment work is allocated as following: Littman (Social Feed Manager); Milligan (Archives Unleashed); Goel or Xie (ArchiveSpark); and Xie (Event Archiving). All key project staff will participate in the trainee selection, advised by the project advisors. Klein will also act as liaison with the IIPC and iPRES communities, and coordinate the training efforts.

The detailed time and resource commitment is given in the budget justification and the sub-recipient agreements.

Evaluation

The success of the project will be evaluated by 1) the completion and successful delivery of the curriculum and training materials; 2) the number of applications received, travel support awarded, and training completed; 3) what trainees learn; and 4) a feedback survey prepared by the project team for the trainees to complete at the end of the training. In addition, the course materials will be evaluated by iSchool and library experts before, during, and after being delivered to trainees. All surveys will follow Virginia Tech's Institutional Review Board procedures to ensure high standards and ethical behavior. The survey component will specifically include the Performance Measure Statements required by IMLS Learning projects.

Communications

By the end of the project, all course materials and software will be released online and free of charge, through a combination of institutional repositories, Jekyll on GitHub, and possibly also a major MOOC platform. Our past experience indicates that hosting software training materials via

Jekyll on GitHub strikes a decent balance between usability and sustainability. It's versioned, openly hosted so people can fork it or adapt it if they want to, and more importantly, allows rapid, distributed revisions to adapt to the rapidly-changing technology stacks.

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.

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 multiple 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 the 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.

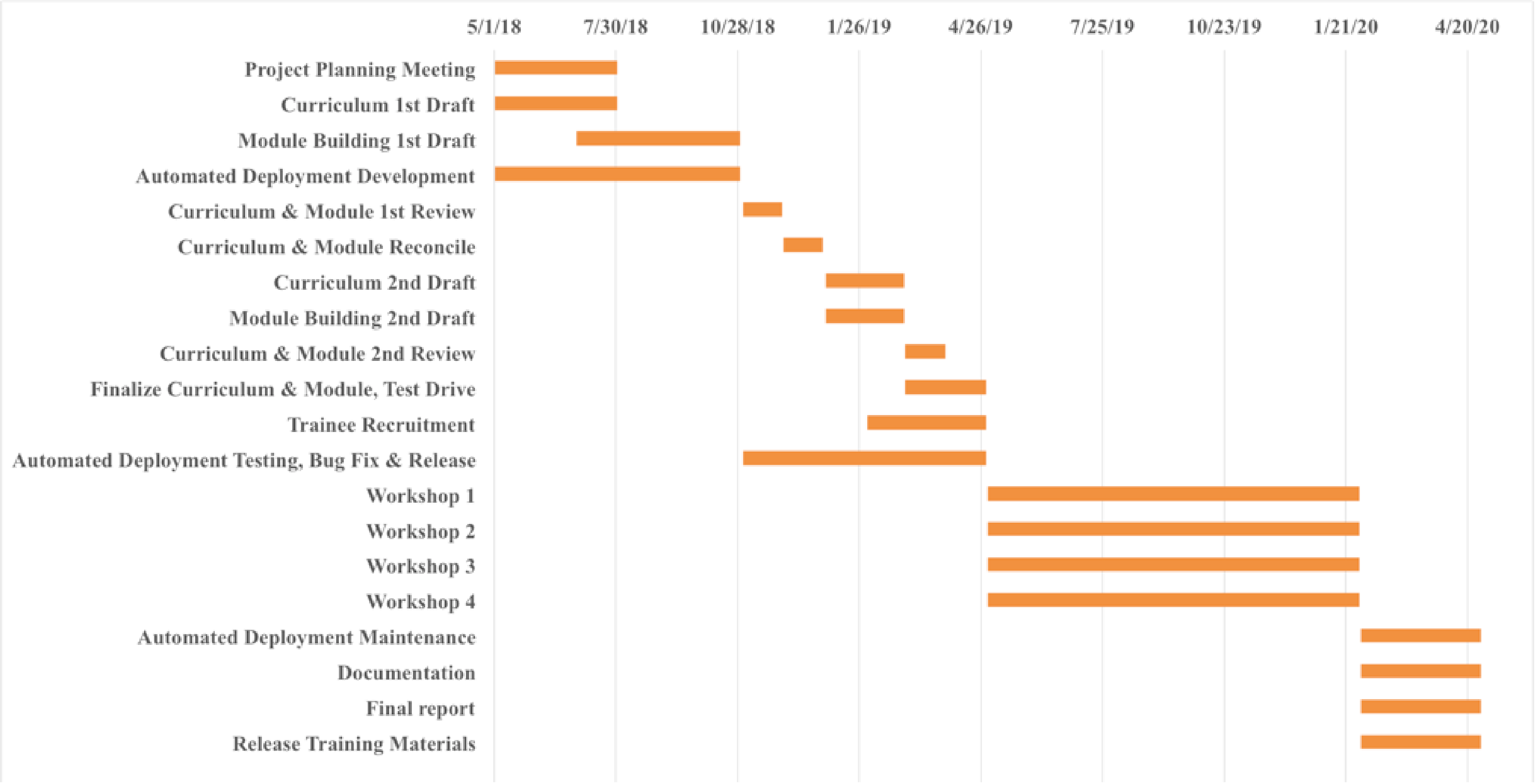
In addition, the project director will appoint a diversity advocate from within the project team. Both the project director and the diversity advocate will be required to complete a Diversity Advocate training comparable to that currently offered at University Organizational & Professional Development, Virginia Tech[30]. Both will advocate for diversity during the curriculum development and implementation, the trainee selection, and the training delivery.

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 providers/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. Making the curriculum and training materials openly available will extend our efforts beyond the project period and ignite further interests in this realm.

[30] http://uopd.vt.edu/other-development/diversity-development/certificate-programs/diversity-advocate.html

Schedule of Completion



DIGITAL PRODUCT FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded digital products (i.e., 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. However, applying these principles to the development and management of digital products can be challenging. 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

□ 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.

All course materials will be published using CC0 (Creative Commons-Zero). Software developed in this project will be released in open source licenses compatible with the upstream open source software. All licenses will be included in the released documents and software.

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.

Virginia Tech will not claim any ownerships of the content, nor would it make any warranties about the work.

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.

Permissions will be obtained before creating any media (e.g., photographs, video, audio, testimonials) product, by requiring related party to sign the standard Virginia Tech Media Release Form (http://www.unirel.vt.edu/photography/media_release_form.pdf).

Evaluation data will not contain any Personally Identifiable Information (PII) data.

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 format you will use.

We will create course curriculum, slides, and other reading materials, released in PDF/A-1a format.

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.

Equipment will include desktop and laptop computers, tablets, and personal mobile devices. Software will include word processors, spreadsheets, presentation software.

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).

Course materials will be primarily generated from Microsoft Office, but final publication will be in PDF/A-1a to ensure broad accessibility.

B. Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

N/A

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 course curriculum, slides, and other reading materials will be stored in openly accessible institutional repositories at Virginia Tech with a Creative Commons Attribution License. Survey data will be stored in Virginia Tech's data repository (VTechData) with protection on files that require confidentiality. Files that do not require confidentiality will be made Open Access with a Creative Commons Attribution License.

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).

Documents and data archived in institutional repositories will use the Qualified Dublin Core.

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

Virginia Tech’s institutional repositories will be leveraged for preservation and maintaining metadata during and after the award period..

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).

The repositories we have selected are OAI-PMH and OAI-ORE compliant.

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).

It will be openly available online in non-proprietary standard formats in repositories that do not require authentication, and in formats that are ADA compliant.

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.

https://vtechworks.lib.vt.edu
https://data.lib.vt.edu

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 develop software to deploy web archiving service platforms and tools on Amazon Web Services. The intended audiences are library and archive staffs wishing to use these tools on Amazon cloud.

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.

N/A

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.

Bash, YAML, Vagrant, Ansible, and Docker will be used to create the deployment software. These are the widely used DevOps tools to deploy complicated applications.

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

We create software that can deploy web archiving tools in the Amazon cloud.

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

The software depends on the Amazon cloud to perform intended functions. In some cases, we also assume Vagrant and/or VirtualBox is installed.

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

We will utilize agile software development process for development, documentation, and for maintaining and updating documentation.

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

Install Script for deploying an institutional repository: https://github.com/VTUL/InstallScripts

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.

We will choose open source licenses to release software developed in this project. The main constraints on licenses are the ones used in the upstream open source software. Our licenses will be chosen to be compatible with the upstream software.

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

Source code will be made publicly available on github.

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

Name of publicly accessible source code repository: Virginia Polytechnic Institute and State University Libraries

URL: https://github.com/VTUL

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.

Survey data will be created to evaluate the usefulness and effectiveness of the curriculum and learning outcome. The data will be collected after each training workshop.

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?

The evaluation survey will be reviewed by Virginia Tech's IRB prior to execution.

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).

No

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.

N/A

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).

Survey data will be collected using Qualtrics software, and stored in Excel.

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?

N/A

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

All data will be deposit in VTechData, Virginia Tech’s institutional data repository.

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

Name of repository: VTechData

URL: http://data.lib.vt.edu

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

We will review the data management plan in Feb 2020 to ensure its satisfactory implementation.