LG-71-17-0016-17 Educopia Institute

Abstract

The Educopia Institute, in collaboration with the University of North Carolina at Chapel Hill School of Information and Library Science (UNC SILS), LYRASIS, and Artefactual, Inc., requests $681,178 (with an additional $244,796 as cost share) for a two-year National Leadership Grant (Research Grant category) to investigate, model and test a range of workflows for libraries and archives to curate born-digital content. These archival workflows will incorporate three leading open source software (OSS) platforms - BitCurator, Archivematica and ArchivesSpace - and the project will be designed to generate findings that can be generalizable to settings that are using other platforms and applications.

Our 12 partner institutions – Atlanta University Center Robert W. Woodruff Library, District of Columbia Public Library, Duke University, Emory University, Kansas Historical Society, Massachusetts Institute of Technology, Mount Holyoke College, New York Public Library, New York University, Odum Institute, Rice University, Stanford University – will implement these platforms; many are also working with additional tools (e.g. Hydra, Preservica) that will be reflected in research discussions and guidance documents produced by the project.

OSSArcFlow addresses IMLS priority 2, "Establishing and Refining Tools and Infrastructure," by researching and modeling ways to combine digital curation tools. It also reflects a theme from the 2015 IMLS Focus report: open source software can greatly facilitate "adapting existing tools for easier implementation," but only if interoperability and integration can be addressed.

This project will result in a greater understanding of how the selected tools may be integrated (in combination with each other and with other applications - both open source and proprietary) to support institutional needs. The findings will be made available as (1) project publications, presentations and reports; (2) detailed documentation of partner workflows, including specific uses cases, methods and scripts that can serve as models for other institutions; (3) training materials (including screencasts and webinars); and (4) a guidance document for libraries and archives that describes how to select and combine tools; how to integrate environments in different use cases and scenarios; and how to share institution-specific workflows with the broader digital curation community. We will address the following two research questions:

1. How can institutions combine tools to support workflows that meet local institutional needs?
2. How can institutions implement "handoffs" between different function-based systems?

This project will significantly impact curation practices across the library and archives fields by increasing our understanding of what factors influence the decisions of institutions of different sizes and types as they choose tools and create workflows. Our findings will support a broad range of institutions that are responsible for digital content. The knowledge gained from working with multiple institutions of different types and sizes will also broaden understanding of curation approaches and priorities, and how those impact the use of digital curation tools and capabilities.

OSSArcFlow: Researching Archival Workflows for Born-Digital Content

1. Statement of National Need

Since the late 20th century, the outputs of science, culture, politics, economics, law, and numerous other domains, have been predominantly created in digital form and stored on hard drives, floppy disks, optical disks, tapes, and other media. Libraries and archives increasingly bear responsibility for stewardship of these born-digital materials. Implementing workflows to support these activities is a pressing and persistent challenge. Digital curation involves a variety of functions (including acquisition, description, preservation and access provision) that no single system can manage. Institutions must work within their existing processes, policies, institutional constraints, and technical platforms, and develop workflows that combine tools to support local needs.

A single turnkey system does not exist and is unlikely to exist in the future. As Trevor Owens pointed out in the 2015 IMLS Focus report, "we cannot plan to have 'one mega system.'" Instead, libraries and archives tend to adopt and integrate separate systems for different functions, with each system using distinct tools and generating its own forms of metadata. Modular open-source software (OSS) tools for supporting curation of born-digital content have matured greatly in recent years, and these software applications have growing user communities. However, collecting institutions frequently experience difficulties when attempting to synchronize OSS tools to enable efficient, effective, and scalable curation workflows.

Institutions report that there are both gaps *and* overlaps between different tools and environments that have to be managed. Gaps between tools can make it difficult to push content through a workflow. For example, the output from the first tool/function in a curation workflow may have to be transformed before it is compatible with the next tool. This means that instead of spending time curating the collections, a large portion of time is spent massaging data and metadata so that it can interface with different systems. Overlaps between tools also challenge curators to make decisions about when and where to complete a particular function. An institution may deploy two different OSS tools that contain some of the same modular scripts and functions.

Several recent "proof of concept" projects have investigated the integration of multiple OSS tools within a single institutional context, raising awareness of possibilities for integrating OSS tools. However, institutions typically have highly localized processes and technical infrastructures supporting digital content management, making the implementation of a "one size fits most" workflow a challenging and impractical endeavor. This issue has been discussed in myriad conversations within specific communities (e.g., the BitCurator Consortium, the MetaArchive Cooperative) and in conference settings (e.g., Code4Lib, DLF); it has not yet been addressed.

Achieving OSS combinations and integrations at scale and in diverse institutional environments is a critical, emergent issue; it is also the goal of the OSSArcFlow project. This research team, comprising leaders from each of three leading OSS technologies, will work with 12 partner institutions to research, devise, test and document various strategies for implementing workflows *within institutions of multiple sizes and types*. The partners represent a diverse range of library and archives types; they also represent a range of institutional sizes and technical capacities. Each partner has committed to integrating a set of common OSS technologies - BitCurator, ArchivesSpace, and Archivematica - during the project period. Each partner is also working with a range of other tools and environments, and each institution is grounded by its own specific aims and abilities.

The project will enable these 12 institutions to learn from each other and will enable the research team to study and learn from their collective experiences. The partners will be supported centrally via a project manager and technical lead who assist each institution as it plans and implements workflows based on the OSS systems. The project design will ensure consistent communication across the partners and sharing of important achievements

and scripts and ideas. It will result in clear documentation of the integration pathways—including the decisions made by humans along the way—so that the experiences of these 12 institutions benefit others. We aim to make the daunting task of implementing digital curation tools more achievable for memory institutions nationally.

This two-year effort will help professionals working with born-digital acquisitions in libraries, archives, and museums by enabling more consistent data flows across applications. It will also concretely address IMLS priority 2, "Establishing and Refining Tools and Infrastructure," by enabling system interoperability and by documenting how digital curation tools can be integrated in a manner that informs subsequent implementations in a wide variety of cultural heritage institutions. This project will study and assist project partners in the integration and combination of tools, produce detailed *documentation* of partner institutions' workflows (including specific methods and scripts) to facilitate the flexible synchronization of archival OSS systems by a variety of collecting institutions, *training modules* that will promote the use of the OSS workflow documentation and scripts, and generalizable *guidance documentation* to help institutions of many types as they implement digital curation and preservation tools and workflows in their own environments. These activities will catalyze efforts across the library and archives fields by supporting more efficient and effective digital curation programs that ensure ongoing access to our increasingly born-digital legacy for all Americans.

1.1. Previous Work

For the purposes of this research project, we focus primarily on three OSS environments with strong user bases and communities: BitCurator, ArchivesSpace, and Archivematica.

1.1.1 OSS Environments for Curation of Born-Digital Materials in Libraries and Archives

Archivematica was first developed by Artefactual Systems in 2007. It is an integrated suite of open-source software tools that allows users to process digital objects from ingest to access in compliance with the functional model of the Reference Model for an Open Archival Information System (OAIS). Users monitor and control ingest and preservation micro-services via a web-based dashboard. Archivematica uses METS, PREMIS, Dublin Core, the Library of Congress BagIt specification and other recognized standards to generate trustworthy, authentic, reliable and system-independent Archival Information Packages (AIPs) for storage in your preferred repository. The software is maintained and actively developed by Artefactual, who distribute the software for free, but charge for consulting, support and specific development activities based on a "bounty model," in which customers contract for work that is then incorporated into the code base available to everyone.
ArchivesSpace is a web-based application for managing archival information. Work on ArchivesSpace began in 2009 as an effort to integrate the Archivists' Toolkit and Archon into a single application. The Andrew W. Mellon Foundation provided funding for the first two phases of the ArchivesSpace program, and it is now managed by a membership-based organization administered by Lyrasis. The application is designed to support core functions in archives administration such as accessioning; description and arrangement of processed materials including analog, hybrid, and born-digital content; management of authorities and rights; and reference service. It supports collection management through collection management records, tracking of events, and a growing number of administrative reports. The application also functions as a metadata authoring tool, enabling the generation of EAD, MARCXML, MODS, Dublin Core, and METS formatted data.
BitCurator environment development began in 2011, with support from the Andrew W. Mellon Foundation. It is a Ubuntu-derived Linux distribution geared towards the needs of archivists and librarians. The environment uses software and practices adopted from the digital forensics community to perform a variety of tasks, including the creation of forensic disk images, analysis of files and file systems, extraction of file system metadata, identifying and redacting sensitive information, and locating and removing duplicate files. It has been further developed through a series of follow-on grants from the Mellon Foundation, and the software is now maintained and supported by the membership-driven BitCurator Consortium, administered by Educopia.

1.1.2. Collaborations Between Archivematica, ArchivesSpace, and BitCurator

Since their inception, there has been significant coordination and collaboration between the communities associated with Archivematica, ArchivesSpace and the BitCurator environment. This has included frequent discussions among personnel, as well as more formal relationships, with representatives from each organization serving on advisory boards for multiple research projects.[1] Recent Archivematica/BitCurator collaborations have capitalized on commonalities in the environments (both are based on Ubuntu Linux) to package component tools of the BitCurator environment so that they can be installed and run directly in Archivematica, including the addition of bulk_extractor as a micro-service in version 1.2 of Archivematica.[2] They have also coordinated work around the generation of PREMIS metadata, to ensure consistency across the two systems (Chassanoff, Woods and Lee, 2017). Workflows that connect Archivematica with ArchivesSpace have also been of high interest to their respective communities. As a result of two projects (one sponsored by the Rockefeller Archive Center and the other by the Bentley Historical Library) support for two new types of workflows between Archivematica and ArchivesSpace will be included in the release of Archivematica 1.6. Both have the high-level goal of displaying preservation-object metadata within the ArchivesSpace interface so that archivists and other collections managers can more easily manage their preservation objects within the same tool that they use for analogue collections.[3] The OSSArcFlow project builds on the strong working relationship between these three OSS communities and has the potential to make these types of workflows extendable and adaptable by more institutions.

1.1.3. Forming Collective Agendas around OSS for Libraries and Archives

The integration of OSS tools and environments recently has become an important topic of discussion throughout the library and archives communities. In the past two years, several professional events were designed specifically to advance the dialog between the developers and users of OSS systems used to support digital curation activities in libraries and archives. At the International Conference on Digital Preservation (iPRES) a day-long workshop entitled "Using Open-Source Tools to Fulfill Digital Preservation Requirements" (Chapel Hill, 2015) explored the particular challenges of developing systems and integrating them into local environments and workflows, a conversation later captured in an article that summarizes major themes and outcomes (Gengenbach et al, 2016).[4] This was followed by "OSS4Pres 2.0: Building Bridges and Filling Gaps," which was held at iPRES in Berne, Switzerland in 2016, a second workshop that advanced the discussion with a different mix of international participants. Another set of events under the name Code4Arc has involved pre-conference workshops in association with Code4Lib (2015, Portland, Oregon; 2016, Philadelphia, Pennsylvania). These events provided an opportunity for the archivists and developers who work with them to talk about archives-specific tools, workflows and development, with an emphasis on open-source software. Participants discussed how coding for archives can be different than coding for libraries and interaction with cross-domain tools such as repositories, discovery and access systems. OSSArcFlow project team members were lead organizers of each of the above events, which surfaced key, shared challenges associated with adopting combinations of OSS tools in libraries and archives. This project builds upon the relationships established in these and other venues and will afford the in-depth and systematic investigation of workflows that these participants and practitioners have been seeking.

1.1.4. Institutional Workflows

[1] For example, staff from Artefactual (Archivematica) served as active members of the advisory boards for both phases of the BitCurator project, and the same is true of staff from ArchivesSpace for the BitCurator Access and BitCurator NLP projects.

[2] This connection was further advanced in version 1.6 of Archivematica, in which users can view output from the credit card and PII scans in the Appraisal tab; this work was sponsored by a project funded by the Andrew W. Mellon Foundation at the Bentley Historical Library to implement workflows that incorporate ArchivesSpace, Archivematica and DSpace, among others (Shallcross, 2016).

[3] See http://archival-integration.blogspot.ca/2015/11/digital-objects-and-archivesspace.html http://archival-integration.blogspot.ca/2016/01/archivematica-to-dspace-and-back-again.html for blog posts on some of this integration work.

[4] Topics included current efforts and grant-funded initiatives to integrate different archival OSS tools; the development of workflows involving multiple open source tools for digital preservation, forensics, discovery and access; and the identification of gaps which may need filled by these or other tools.

A review of digital curation literature reveals a steadily growing number of institution-specific workflows released in the last 10 years.[5] The workflows developed by MIT are particularly compelling in their use of clear and consistent conventions for representing the various paths and relationships associated with workflows.[6] While such workflows may have great potential to benefit other institutions, the above efforts to combine the use of OSS systems have been based solely on single-institution scenarios and needs. Additionally, system and tool developers, including the BitCurator environment[7], Archivematica[8], and ArchivesSpace have provided documentation about the workflow elements that they can support.Other research efforts have focused on the documentation of workflows that combine a variety of tools and methods, in order to compare them, including a study conducted of the incorporation of digital forensics tools and methods into libraries and archives workflows (Gengenbach, 2012) and a cross-comparison of multiple institutional born-digital workflows (AIMS, 2012). Additionally, a number of organizations and initiatives are attempting to solicit and share workflows.[9] Documenting and understanding workflows for digital curation can provide decision mappings of use to many institutions. Thus far, workflow documentation has not been created or compiled in standard ways, which limits the degree to which comparisons may be drawn between different workflows. This project will build upon and improve the state of workflow documentation by creating a common mechanism that will be used by 12 institutions to create and share workflows that demonstrate how the same tools and functions may be combined in very different ways depending on institutional needs, priorities, resources, and structures.

1.1.5. Designing for Portability and Reusability of Workflows

Once one has identified an actual or desired workflow, it can beneficial to consider how that workflow could be carried out using different software in the future. For example, in the context of scientific workflows, there has been important work on planning for reusability (De Roure et al, 2011). A primary mechanism to support the movement of functionality from one application to another is modularity (Baldwin and Clark, 2000). Rather than assuming that entire functions will be carried out by a single monolithic system that one treats as a black box, one can define a more discrete set of small tasks that need to be performed, regardless of what software is used. In an increasingly modular environment for library and archives tools, a key purpose of workflow documentation is to describe not just the relationship between existing tools in an environment, but the relationship tools have to particular curatorial functions. In this way, the utility of workflows - including those designed and implemented in this project - is not limited to the specific systems or tools used in one implementation. Instead, they become a sort of roadmap that helps institutions know when and how to "plug" and "unplug" different tools in their workflows. The workflows that we will create in the OSSArcFlow project will have multiple purposes. They will guide the integration and combining of tools during each partner's implementation. They will also enable cross-comparisons of different approaches to such integration in different environments, surfacing factors that may otherwise be invisible. They will provide a set of examples and case studies that other institutions can use as models. They will also continue to be valuable tools in the future as we continue to grapple with new technologies and platforms while trying to perform standard curatorial functions.

2. Project Design

The Educopia Institute, in collaboration with the University of North Carolina at Chapel Hill School of Information and Library Science (UNC SILS), LYRASIS (host of the ArchivesSpace program), and Artefactual, Inc., requests a National Leadership Grant (Research Grant category) to investigate, synchronize,

---

[5] Some of the most detailed examples include those from the Interuniversity Consortium for Political and Social Research (ICPSR) (Whiteman, 2006), Yale/Tufts (Glick and Wilczek, 2006), and the Digital Sustainability Lab at the Massachusetts Institute of Technology (MIT) (Smith, 2015).

[6] Kari Smith, the author of those resources, is an OSSArcFlow partner and has offered this approach as a resource for the project.

7 https://wiki.bitcurator.net/index.php?title=BitCurator_and_Archival_Workflows

8 https://wiki.archivematica.org/Workflows

[9] See the Electronic Records Section of the Society of American Archivists which has encouraged submissions of workflow descriptions through its blog, called BloggERS!, the BitCurator Consortium which solicits and shares workflows that have incorporated digital forensics tools and methods, and the Library Workflow Exchange.

and document born-digital workflows to assist cultural heritage institutions in managing and preserving born-digital content. These workflows will support content transfer and metadata integration among three open source software platforms - BitCurator, Archivematica, and ArchivesSpace. The project will assist professionals working with born-digital acquisitions in libraries, archives, and related institutions by studying and documenting a range of ways to address the gaps and overlaps that currently exist between common OSS tools and environments. Through enabling system interoperability and documenting integration pathways, the project will catalyze efforts across a wide variety of cultural heritage institutions. This project team will accomplish the following:

1. **produce** 12 OSS workflows and integrations between three leading OSS digital curation tools: BitCurator, ArchivesSpace, and Archivematica;
2. **document** partner institutions' workflows (including specific methods and scripts) to facilitate the flexible synchronization of archival OSS systems by a variety of collecting institutions;
3. **design** training modules that will promote the use of the OSS workflow documentation and scripts; and
4. **create** an "Implementation Guide" to help institutions of many types as they implement digital curation and preservation tools and workflows in their own environments.

2.1. Research Questions

Activities will focus on the improving data transfer and metadata sharing between three OSS applications to address the following:

1**. How institutions combine OSS tools to support institution-specific workflows.** We will support 10 partner institutions in incorporating these three OSS systems into local curation workflows, compare partner institution workflows to understand how they both align and differ, and document and report the results to inform decision making by collecting institutions.

2. How to bridge gaps between OSS systems. The project team will develop and implement strategies to facilitate the flexible synchronization of archival OSS systems by a variety of collecting institutions. We will accomplish this by designing metadata integration and handoffs between systems, developing methods and scripts for use in partner institution integrations and workflows, and packaging and documenting methods and scripts that can be reused by additional institutions.

3. How to educate professionals to adapt and adopt existing workflows, methods, scripts and use cases. We will develop and publish guidance documentation and a series of publicly available webinars and videos to educate professionals in collecting institutions on using the workflows, scripts, and novel methods. We will develop three "how to" training modules to complement existing documentation (e.g. BitCurator QuickStart Guide). These guides will focus on how to make decisions about which tools to use and how to combine them, how to integrate the three environments (ArchivesSpace, Archivematica, and BitCurator) in different situations, and how to share institution-specific workflows (including ones adopted from this project) more broadly.

2.2. Partner Institutions - Current and Planned Digital Curation Workflows

The OSSArcflow team will work with partners from 12 institutions selected by the project team to represent a broad range of academic, research, and public libraries and archives: Atlanta University Center Robert W. Woodruff Library, District of Columbia Public Library, Duke University, Emory University, Kansas Historical Society, Massachusetts Institute of Technology, Mount Holyoke College, New York Public Library, New York University, Odum Institute, Rice University, Stanford University. All partners have committed to enhancing their existing workflows and developing new workflows to implement (or in some cases, to more fully implement) the three core OSS systems addressed in the project. In exchange for this commitment, the partners will receive technical support and guidance from the project team, as well as insights and models from the other partners.

The OSSArcFlow team has conducted significant research over the last two years with each of the 12 partner institutions about their current and desired digital curation workflows, including through interviews, group discussions, and a detailed survey completed this December by all 12 partners. Survey responses (synthesized in Supporting Document 2) have affirmed our selection of the partner institutions, highlighting not only their struggles with integration scenarios and commitment to the goals of the project but also the diversity of organizational and technical contexts that they represent.

Like so many of their peers, the project partners report that they are at various stages of adopting the three systems. Some have implemented one of the tools (e.g., nine of the partners have implemented BitCurator). All partners have at least piloted or experimented with one or more of the three systems. All are using a range of additional tools in their local environments (see Supporting Document 2). Some are using entire repository systems (e.g. DataVerse, Preservica), while others are using microservices. The majority of tools used by the project partners are available under OSS licenses, while some others are proprietary. Of the OSS tools, some are already part of the default distributions of Archivematica and the BitCurator environment, while others could be incorporated with relatively little overhead, because they can be installed in Ubuntu Linux.

The partners' stated challenges associated developing digital curation workflows clearly affirm the goals of the OSSArcFlow project. A great deal of their concern relates to metadata generated by and imported into the systems. One partner expressed this quite clearly.

> ...although we can leverage output from one tool as input to another sometimes we have to spend time on manipulating the outputs so that they can be properly formatted inputs. The time and effort that we need to spend in manipulating data so that we can move to the next workflow step using another tool is significant and without the technological expertise on our staff to help, I believe that significant (up to 30%) of our time is spent not on working with the collections material itself, but rather on getting data and metadata from one tool to the next.

Partners cite a need for both guidance and scripts for filtering and transforming metadata output so that it be more easily passed between systems. A related desire is for more consistently and clarity in metadata about the digital curation processes themselves. As one partner has indicated, "With all of this processing, what is a purely manual effort is keeping track of and documenting the processes of using these tools."

The project will afford us the opportunity to work closely with all partners over a two-year period, analyzing, documenting, and helping to resolve the myriad challenges they face in integrating OSS tools to complete digital curation tasks. What we learn from (and with!) the partners will be channeled through a range of documentation forms (case studies, training webinars, script libraries, and an "Implementation Guide") to advance the state of the field in choosing among and successfully integrating the many tools and environments that are available to libraries and archives today.

2.3. Plan of Work

Table 1 provides a summary of project activities, based on a quarterly schedule. The table is designed to identify major milestones and transition points between tasks. It does not reflect several ongoing activities, such as monthly conference call with the project partners, as well as public webinars that we will develop and administer throughout the course of the project. We have not set specific dates for completion or deployment of the "How To" guides, because timing will depend on the scope and focus of those guides, which will be based on ongoing project findings. Also not referenced are outreach, engagement and educational activities in the form of conference presentations, tutorials and workshops. Such activities will represent a substantial time commitment by the team throughout the course of the project. They will be driven by both user demand and the contingencies of conference submissions being accepted, so we will not know the specific dates for such activities until the project is already underway.

Table 1 – Quarterly Activity Overview for OSSArcFlow

Period	Activities
Jul 1 - Sept 30 2017	Initial setup of testbed environment at UNC; elicitation of partner needs and contexts; partners development of initial "as is" workflows; hiring of project staff; hiring of external evaluator; development of first drafts of workflow template; planning for partners meeting; formative project evaluation
Oct 1 - Dec 31 2017	Partners meeting; partners meeting reporting; partners' completion of OSS system test instance installations
Jan 1 - Mar 31 2018	Partners' development of initial aspirational workflows; feedback from project team on workflow feasibility and planning; requirements definition for system handoffs; refinement of workflow template
Apr 1 - Jun 30 2018	Partners' test implementation of aspirational workflows; public dissemination of workflows
Jul 1 - Sept 30 2018	Partners' test implementation of aspirational workflows; refinement of workflows based on partner testing experiences and public input
Oct 1 - Dec 31 2018	Partners' test implementation of aspirational workflows; refinement of workflows based on partner testing experiences and public input; publish and elicit comments on first public draft of "Implementation Guide"
Jan 1 - Mar 31 2019	Partners' test implementation of aspirational workflows; refinement of workflows based on partner testing experiences and public input; conference calls with each of the three user communities (BitCurator environment, Archivematica, ArchivesSpace) to discuss practical implications for future development priorities within those systems
Apr 1 - Jun 30 2018	Final publication of partner workflows; elicitation of example workflows from non-partner institutions; final publication of "Implementation Guide"; final publication of project-developed methods and scripts; summative evaluation

Early in the project, partners will express their institutional needs and goals by completing a detailed questionnaire about their policies, practices and platforms (as explained above, we have already conducted an initial high-level survey of the partners) and helping them to develop "as is" workflows that represent their current activities. Based on the survey responses and "as is" workflows, the project team will develop a summary "context document" associated with each partner institution. The project team will provide the partners with a workflow template, which will then revise if necessary, based on partner feedback. Partners will then create aspiration workflows, based on the desired digital curation activities within their environments. These "context documents" will be useful resources for outside consumers of the later aspirational workflow products, by helping them to understand the situations in which the workflows emerged.

The UNC team will establish and run a testbed environment, which will be customized to reflect (to the best of our ability) the capabilities and constraints at the partner institutions, in order to test strategies for implementing their intended workflows and provide them technical support. This will complement the technical advice that will be provided by the software providers represented on the project team. The project partners will each set up test installations of the OSS systems in their own environments. The project team will provide them support in this process. In fall 2017, we will hold a partners meeting in Chapel Hill, in which we can gain further insights from the partners about their workflow needs, troubleshoot issues they might have experienced with their test installations and gain consensus on overall project goals. There will also be a public event associated with the partners meeting, which will allow us to reach a wider audience.

Partners will test their aspirational workflows, refine them based on testing, and publicly disseminate the revised workflows. This test implementation and refinement will be an iterative process that carries through a substantial portion of the project. In addition to the workflow documents, this activity will yield specific methods and scripts that have been developed either by the project team or the partners themselves. Later in the project, we will ask partners to make their workflows available as completed documents and advertise them widely. We will then encourage other non-partner institutions to share their own through the project site. The project experiences will inform our development of the "Implementation Guide," which will be published in

final form after an initial public comment period. The Guide will reflect comparative analysis of partner institution workflows to identify areas of similarity and difference, as well as distillation of the workflow design and implementation process.

The OSSArcFlow project will also help to complement and Coalition directory tool and the Community Owned digital Preservation Tool Registry (COPTR). Whenever there are entries in COPTR for given tools, the OSSArcFlow workflow documentation will point to them; when there are no entries in COPTR for given tools, the project team will add entries for those tools to COPTR. See the letter of support from William Kilbride regarding this aspect of the project.

The project will hire an external evaluator. As described in Evaluation below, this consultant will provide a formative evaluation, maintain quarterly contact with the team, and then provide a summative evaluation early in the final quarter. The summative evaluation will include conference calls with the representatives of the respective user communities for the three main OSS systems, in order to elicit their feedback about practical implications of findings and deliverables for future development priorities in those systems.

2.4. Financial, Personnel and Other Resources

The project budget requested is $681,178, with an additional $244,796 in cost share. This includes a) travel expenses for partners for an in-person meeting in Chapel Hill, NC, b) salary and benefits for personnel including two FTE and percentages of the three Co-PIs' time (Dr. Christopher (Cal) Lee, Sam Meister, Dr. Katherine Skinner), and c) materials and supplies necessary for the project's success. Additional personnel committing effort to this project include Sarah Romkey, Artefactual; Laney McGlohon, ArchivesSpace; as well as representatives from our 12 partner institutions. Letters of commitment are included from all project partners. The following are brief highlights of project personnel and areas of expertise.

Dr. Katherine Skinner (Executive Director, Educopia Institute and Adjunct Professor) will act as principal investigator (PI) and overall coordinator for the project. Skinner has served as PI for grants and contracts totaling more than $2.5M, including a broad range of cross sector initiatives in digital preservation across libraries, archives, and museums. Skinner is a founder of the MetaArchive Cooperative and the BitCurator Consortium, two leading efforts in digital preservation internationally. She will ensure the project and its deliverables adhere to open access principles and community frameworks, that they are both built and sustained by a range of committed partners, and that all outputs circulate across the memory organization landscape.
Sam Meister (Preservation Communities Manager, Educopia Institute) will serve as co-principal investigator for the project. Meister has served on project teams for multiple grant-funded collaborative research projects, most recently as co-principal investigator for the IMLS-funded Preservation & Curation of ETD Research Data & Complex Digital Objects projects[10]. He also serves as a core instructor in the Digital Preservation Outreach and Education (DPOE) program at the Library of Congress.
Cal Lee (Professor, UNC SILS) will lead the project for UNC. He will oversee the management, progress, and evaluations of all aspects of the UNC team's work. He will take ultimate responsibility for the design, implementation and reporting of findings from the UNC tea. He will interact with members of the Advisory Board, as well as overseeing administrative and financial aspects and reports. Lee is the founder and lead researcher of the BitCurator environment.
Kam Woods (Research Scientist, UNC SILS) will serve as co-PI and Technical Lead. He will be responsible for the technical vision, development and implementation of the BitCurator NLP software. He will work closely with the PI in the management and evaluation of all aspects of the work, as well as contributing to the design, implementation and reporting of findings of the project, including overseeing software development,

[10] https://educopia.org/research/grants/etdplus

configuration management, requirements definition, and engagement with development communities associated with tools upon which the OSSArcFlow workflows are built.

2.5. Outreach and Engagement

The project team is committed to wide dissemination of project findings and outputs. This outreach and dissemination process has already begun through the project planning work our team has engaged in. It will continue to expand and amplify at the start of the project, building throughout the grant period and beyond. The design of the project involves a diverse set of stakeholders, including representatives from different fields and perspectives. It will produce research findings, documentation, training materials (including screencasts and webinars), and an Implementation Guide, all of which will be distributed widely with a particular emphasis on practitioners who are currently working with born-digital materials (or who plan to do so in the near future).

All project outputs will be published with CC-BY or CC-BY-SA or GPL v3.0 licensing and disseminated as freely and widely as possible. Due to our personal relationships throughout the field, our outreach efforts will not be limited to listserv distribution, but will also travel across our (intentionally diverse) networks through a wide variety of webinars, presentations, social media announcements, press releases, media outreach, and invitations targeted to library and archives directors, digital curation and preservation practitioners, trainers, and students to use the research outputs. We will reach out through each of the OSS user communities represented in the project (BitCurator, Archivematica, and ArchivesSpace). We will also encourage the use of outputs by a variety of continuing education and professional development groups throughout the nation and beyond, through such groups as the Coalition to Advance Learning (with more than 20 meta-organization members) and Educopia's project groups and Affiliated Communities (including the Mapping the Landscapes and Nexus LAB efforts, each with more than 35 meta-organizations that offer training to library, archives, and museum practitioners), and with existing trainers in this area to ensure broad uptake and reuse. We will also engage with relevant development groups to help coordinate our efforts and promote the development and documentation of workflows. Members of the project team have active collaborations with numerous initiatives in this space (e.g., ePADD, DPN, APTrust, MetaArchive, etc).

2.6. Risks

The project team recognizes that any project with multiple partners and stakeholders faces an array of potential risks. Specific risks and mitigation strategies we have actively planned include the following (see also Supporting Document 4):

1. Produce findings that lack relevance beyond the project partners - We have selected partners that differ along several dimensions: *Institution type* (large academic libraries, small academic libraries, public libraries, state historical society, social science data repository); *Software installation/hosting configurations* (run directly on a local machine, run locally within a virtual machine, hosted by an external provider); *Storage environments* (on a local machine, cloud storage, shared network drives, centralized repository storage); *IT support arrangements* (provided by curatorial stuff themselves, dedicated IT staff within unit, centralized IT staff within institution, outsourced IT support providers); *Stages of adoption of the three OSS systems*; *IT transition status* (e.g. undergoing ContentDM to Islandora migration, possible Preservica to Archivematica migration)
2. Failure to fully implement workflows at partner institutions - As a research project focused on analyzing and sharing workflow information, core project objectives will be met when partners define and share information about workflows. A large, diverse set of partners reduces dependency on any single workflow in order to meet project objectives.
3. Lack of tools sustainability - Three focal OSS systems have established sustainability structures; Use of both OSS and a modular design approach minimizes future dependency on any specific software; Partners will share information about alternative tools for performing specific functions, providing clear

paths (social support and documentation) for switching between them if necessary.

2.7. Evaluation

The project team will work with an established Evaluator with experience in both formative and summative evaluation. The Evaluator will work with the project team at project kick-off to evaluate the project plan and start up activities against the outcomes and deliverables expected from the project. During this formative evaluation, the Evaluator will make recommendations for adjustments that will strengthen our approach and ensure the success of our work. We will maintain connection with the evaluator, conducting at least quarterly check ins on progress toward all project goals. Using summative evaluation frameworks and methods (e.g., interviews, surveys, logic models and outcomes-based evaluation), the Evaluator will conduct a rigorous evaluation of the project against its stated goals and outcomes. The resulting evaluation will be included in our final report to IMLS.

2.8. Sustainability

The project deliverables will be among the first resources of their kind. As such, it is critical that they are easily accessible and appropriately documented and licensed to ensure that the stakeholder community can readily use and adapt them going forward. The main project website, hosted by Educopia Institute, will promote all review and final versions of project deliverables (archived by the WayBack Machine). All project information - including the narrative, project team, workplan, reports, summary findings, training modules and screencasts, scripts, and the *Implementation Guide* will be released with CC-BY or GPL v3.0 licensing to promote use and reuse with appropriate citation. The three OSS communities partnering on this grant will also maintain and promote use and reuse of all project deliverables. Our project partners and project team commits to maintaining and continuing to build on these documents and tools in the course of future projects as well, as evidenced by the sustained efforts in previous years that this proposal builds upon.

3. National Impact

This project will impact curation practices by increasing our understanding of how institutions of different sizes and types choose to integrate digital curation tools in different workflows. Findings will support a broad range of institutions that are responsible for digital content. The knowledge gained from working with multiple institutions of different types and sizes will also broaden understanding of curation approaches and priorities, and how those impact the use of digital curation tools and capabilities. We expect the empirical findings about institutional needs, as well as formal workflow models, to contribute substantially to digital curation research and practice. We expect the project to achieve the following measurable outcomes:

1. Strengthen relationships between three OSS environments/software developers and model how to think about our work in terms of the national digital capacity, not just our own individual services. **Evidence:** increased collaboration as demonstrated through an alignment map of our activities (drawn at the beginning of the project and at its conclusion).
2. 12 institutions will implement born-digital workflows that support content and metadata transfers/bridges between BitCurator, Archivematica, and ArchivesSpace. **Evidence:** documentation of each partner's work
3. Increase practitioner knowledge of and fluency in workflow modeling and adoption of existing workflows, methods, scripts and use cases to bridge OSS tools and environments. **Evidence:** At least 40 trainees (webinars, screencasts) self-report increase in knowledge and understanding of archival workflows in digital preservation.
4. Increase the number of institutions actively collecting, curating, and preserving born-digital content. **Evidence:** use of OSSArcFlow documentation to inspire or guide at least 40 institutions in their tools integrations in the two years following the grant.

	2017						2018					
Tasks	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
Phase 1												
Initial setup of testbed environment at UNC												
Elicitation of partner needs and contexts												
Partners development of initial “as is” workflows												
Hiring of project staff and external evaluator												
Development of first drafts of workflow template												
planning for partners meeting												
formative project evaluation												
Phase 2												
Partners meeting												
Partners’ completion of OSS system test instance installations												
Phase 3												
Partners’ development of initial aspirational workflows												
Feedback from project team on workflow feasibility and planning												
Requirements definition for system handoffs												
Refinement of workflow template												
Phase 4												
Partners’ test implementation of aspirational workflows												
Public dissemination of workflows												

	2018						2019					
	Jul	**Aug**	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
Phase 5												
Partners’ test implementation of aspirational workflows												
Refinement of workflows based on partner testing experiences and public input												
Publish and elicit comments on first public draft of “Implementation Guide”												
Conference calls with each of the three user communities to discuss practical implications for future development priorities within those systems												
Phase 6												
Final publication of partner workflows												
Elicitation of example workflows from non-partner institutions												
Final publication of “Implementation Guide” and project-developed methods and scripts												
Summative evaluation												

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
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.

Source code contributions will carry the GPL v3.0 license, with the provision that specific contributions identified as candidates for integration or direct distribution with ArchivesSpace, Archivematica, or BitCurator may carry alternate open source licenses (for example, Affero GPL v3.0 or the Educational Community License, v2.0). All other content, resources, and assets produced by the project and intended for distribution to the community will carry the Creative Commons BY-SA 4.0 license. These licenses were selected to maximize community uptake and potential for remix and reuse in future products.

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.

No conditions will be imposed on access or use of the primary products of the project beyond those specified by the licenses CC-BY-SA 4.0 and GPL v3.0.

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.

This project will explore tools and methods that may apply to private and sensitive datasets, but will not collect or otherwise distribute those datasets. Any public demonstrations of the techniques during the project will use only synthetic datasets.

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.

The products of this project will include a report detailing specific workflows for transfer of born-digital materials and translation of metadata associated with those materials when working with at least two of the associated systems under examination (ArchivesSpace, Archivematica, and BitCurator). This report will be supported by software scripts to support transfer of born-digital materials and mapping of metadata between these systems.

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.

The project team at UNC SILS has budgeted for laptop computers and a test server (to run private virtualized instances of ArchivesSpace, Archivematica, and BitCurator) on demand for the term of the project.

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).

Project source code will be distributed on GitHub. Project reports will be distributed in PDF format and comply (at minimum) with PDF/A-1a.

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).

General project tasks and workflow will be managed using Asana (https://asana.com/). The development team at UNC SILS has successfully used this platform to manage multiple prior grant-funded projects. Quality control and feedback for software projects will rely primarily on tools available via GitHub, including Travis CI (continuous integration / automated code testing) and Jenkins.

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).

During the award period, documentation assets will be maintained in both private and public cloud services. These will include Google Docs for contributions shared with project partners and Microsoft OneDrive for those assets internal to the development team at UNC SILS; OneDrive is currently the default cloud institutional asset management system for UNC SILS. All project documentation subject to public release will also be distributed in a format minimally compliant with PDF/A-1a via the BitCurator Consortium.

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).

This project will not produce any technical, descriptive, administrative, or preservation metadata related to collections. It will provide (in the form of descriptive workflows and source code) tools to administer and transform these types of metadata when working with born-digital materials in ArchivesSpace, Archivematica, and BitCurator.

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

This project will not create or maintain structured metadata aside from that metadata associated with software products, which will be stored alongside the code itself on GitHub and in supporting services (see C.3).

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).

This project will use readthedocs (https://readthedocs.org/) to store and publicize relevant documentation related to code and associated APIs.

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).

Project documentation for public release - including workflows, reports, and technical specifications - will be openly available online. To avoid access bottlenecks or single points of failure, the project team will ensure that this documentation is made available (during the funded project period where required and in its entirety on project completion) via existing portals (including https://www.bitcurator.net and https://www.bitcuratorconsortium.org/). Project software tools will be distributed online via GitHub.

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.

URLs are provided below. Note that this is a representative (not complete) accounting of the projects associated with the primary contributors named in this grant:

https://educopia.org
http://metaarchive.org/
https://librarypublishing.org/
https://www.bitcuratorconsortium.org/
https://www.bitcurator.net/
https://wiki.bitcurator.net/
http://bitcurator.github.io/

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.

Using the feedback provided by project partners regarding existing real-world workflows involving two or more of Archivematica, ArchivesSpace, and BitCurator, this project will produce a series of support scripts to help institutions manage specific instances of packaging (for example, in the creation of Submission Information Packages) and metadata transfer and/or mapping (for example, mapping of discrete elements of Digital Forensics XML to technical metadata standards used by archival management software) where those packaging or transfer tools do not already exist, are underspecified, or require modification to manage materials of interest.

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 ongoing efforts in the community underscore the need for improved transfer mechanisms. The Bentley Historical Library recently released a report describing the results of their ArchivesSpace-Archivematica-DSpace Workflow Integration project. An overview of this work can be found at http://archival-integration.blogspot.com/2016/11/the-end-is-just-new-beginning.html, and some of the associated tools can be found at https://github.com/bentley-historical-library.

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 project team will use languages, tools, and distribution platforms appropriate to the needs of the community. Based on our existing experience, we believe these will include (non-exclusively):

Python 2.7/3.5: DFXML handling and export; use of ArchivesSpace RESTful API
(Bash) Shell Scripts: automation, general environment management
Ruby; additional ArchivesSpace integration / plug-in development

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

These scripts will run alongside one or more of the ArchivesSpace, Archivematica, and BitCurator environments, facilitation transfer of relevant archival packaging and/or metadata.

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

We intend that any institution using two or more of ArchivesSpace, Archivematica, and BitCurator will be able to deploy the relevant scripts with minimal overhead.

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

The development team at UNC SILS will operate using an iterative and incremental (agile) development approach based on the needs and features identified by the project partners and extended community. In-development code will be publicly posted to development branches on GitHub. The relevant development repositories will employ Travis CI (continuous integration) for testing and verification of release builds. Basic user documentation and build instructions will be provided alongside the code repository in the form of README.md files. In-depth API documentation (where appropriate) will be provided via readthedocs.

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

A complete list of projects developed by the BitCurator team at UNC SILS can be found at: https://bitcurator.github.io

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.

No additional prohibitions or use will be placed on the software products beyond those specified by the relevant open source licenses, particularly the GNU GPLv3 (for software products) and the Creative Commons BY-SA 4.0 (for derived reports and documentation). The development team expects to work flexibly with partners to provision for alternate or dual-license (for appropriate open source licenses) where such licenses would facilitate additional community uptake.

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

Source code and release packaging will be made available via 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: OSSArcFlow

URL: https://github.com/ossarcflow (this URL will serve as an organizational portal for multiple potential repositories)

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.

N/A - This project does not involve the creation of datasets

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?

N/A - This project does not involve the creation of datasets

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).

N/A - This project does not involve the creation of datasets

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 - This project does not involve the creation of datasets

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).

N/A - This project does not involve the creation of datasets

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 - This project does not involve the creation of datasets

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

N/A - This project does not involve the creation of datasets

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

Name of repository: N/A

URL: N/A

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

N/A - This project does not involve the creation of datasets