

## **OSSArcFlow: Researching Archival Workflows for Born-Digital Content**

*Educopia Institute, in partnership with UNC-CH 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 born-digital content curation in libraries and archives. Developed with our partner institutions, these curation workflows will incorporate at least three leading open source software (OSS) platforms - BitCurator, Archivematica, and ArchivesSpace<sup>1</sup> - and the project will be structured to ensure generalizability of findings to settings that use other platforms/applications. Our 10 partners, including Kansas Historical Society, MIT, Mount Holyoke College, and Odum Institute, will implement these platforms in their workflows; some will also use additional tools (e.g. Hydra, Preservica), and where relevant, these additional tools will be reflected in our research discussions and guidance documents.<sup>2</sup>*

Libraries and archives increasingly bear responsibility for stewardship of born-digital materials. Implementing workflows to support this stewardship is a pressing and persistent challenge. Curating these assets involves a variety of functions (acquisition, description, preservation, access provision) that no single system can manage. Institutions must work within their existing processes, policies, institutional constraints, and technical platforms to develop workflows that combine tools to support local needs. A single turnkey system does not exist and is unlikely to in the future. As Trevor Owens pointed out in IMLS Focus 2015, “we cannot plan to have ‘one mega system.’” Instead, we will adopt and integrate modular systems, with each system using distinct tools and generating its own forms of metadata.

The complications in connecting and integrating tools can be major barriers to implementation. Several recent “proof of concept” projects have investigated OSS tool integration, but have done so within single institutional contexts only. These projects successfully raised awareness of possibilities for integrating OSS tools. Realizing those possibilities requires researching, devising, testing and documenting various strategies for implementing workflows *within institutions of multiple sizes and types*. The research in this project will support digital curation in a broadened range of institutional environments to ensure the availability of born-digital materials for future researchers.

Despite each tool’s popularity, most institutions have not yet implemented more than one simultaneously. In particular, few small-to-midsized libraries have undertaken multiple simultaneous implementations. Our research will address the concerns of these institutions, including public libraries, HBUC libraries, and liberal arts colleges. An advisory board will be formed at the beginning of the project with five members representing these additional communities and institution types. We will collaboratively test, evaluate, and share the research findings with these advisors, and with the robust user communities of each of the three selected OSS environments (BitCurator, Archivematica, ArchivesSpace).

OSSArcFlow addresss 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

<sup>1</sup> See <https://educopia.org/research/incubating-ideas> for criteria used to select the platforms and partners of this project.

<sup>2</sup> We strove to maximize diversity in the research partners we selected for this project. Note that few institutions have implemented more than one these tools, and in particular, very few smaller and under-resourced libraries are able to commit to multiple simultaneous implementations until the type of research and documentation that we propose here exists. The research will be conducted with these additional institutions in mind—including public libraries, HBUC libraries, and liberal arts colleges. An advisory board will be formed at the beginning of the project with five members representing these additional communities and institutional types. We will also test, evaluate, and spread the research findings within a much broader set of players through the user communities of each of the three selected OSS environments (BitCurator, Archivematica, ArchivesSpace).

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 use cases, methods and scripts that can serve as models for other institutions; and (3) an “Integration Guide” on selecting and combining 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 needs?** We will help partners incorporate the OSS systems into curation workflows, describe and compare partner workflows to understand how they align/differ, and document and report results to inform decision making field-wide. This research will identify (1) aspects of institutional workflows that are easily shared across institutional boundaries, and (2) specific sociotechnical factors that define distinct institutional needs. This research will include development, analysis/comparison, and public dissemination of workflow models.

**2. How can institutions implement “handoffs” between different function-based systems?** As there is no single monolithic system to carry out all digital curation functions, OSSArcFlow will investigate, with our partners, various types of “handoffs” between systems that can then be adapted for other institutions. Specific characteristics will be determined as we analyze partners’ needs and practices. Handoffs will include (but not be limited to) filtering/manipulating metadata, identifying decision points that require human judgment, and using methods to invoke other methods. We will research these issues in real-life implementations, developing and disseminating documentation/scripts associated with these handoffs.

In year 1, we will 1) plan, schedule meetings, and hire staff; 2) document partners’ requirements, including local test implementations of the OSS tools; 3) help partners set up local software instances; 4) hypothesize/document system administration and maintenance strategies; 5) draft/refine/disseminate workflow diagrams for public review; and 6) analyze/document feasibility of use cases in different environments. In year 2, we will 7) develop/test methods and scripts for use in partner workflows; 8) refine workflows and documentation based on active engagement with both partners and non-partners to determine how workflows may need to be adapted; and 9) produce/disseminate the “Integration Guide,” scripts, workflow diagrams and a white paper with project findings.

This project will significantly impact curation practices by increasing our understanding of how institutions of different sizes and types choose to integrate digital curation tools in different 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. We expect the empirical findings about institutional needs, as well as formal workflow models, to contribute substantially to digital curation research and practice.

*Estimated Budget:* 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; b) salary and benefits for personnel, including two FTE (Project Manager [to be named] and Kam Woods, Co-PI and Research Scientist) and percentages of the three Co-PIs’ time (Christopher Lee, Sam Meister, Katherine Skinner); and c) federally negotiated indirect costs. Additional personnel committing effort to this project include Sarah Romkey, Artefactual; Brad Westbrook, ArchivesSpace; and Arcot Rajasekar, Gary Marchionini, Paul Jones, and Cristóbal Palmer, University of North Carolina; as well as representatives from our 10 partner institutions.