OSSArcFlow Learning Module 1: Common Steps in OSS Born-Digital Archival Workflows
Script

In this video, we will discuss 13 common workflow steps in born-digital archiving identified by OSSArcFlow, or the Open Source Software Archival Workflows project. We will also cover examples of these steps being used in different workflows, and showcase how some common open-source tools bundle together multiple steps.

Since the late 20th century, records and other knowledge objects have been created predominantly in digital form and stored on a mix of hard drives, floppy disks, optical disks, tapes, and other media containers. Archives, libraries, museums, and other collecting institutions increasingly serve as the stewards of these born-digital materials and must implement digital curation workflows to support their acquisition and care. The OSSArcFlow project looked at the lifecycle of these materials from acquisition by the collecting institution through processing.

The term “open source” commonly refers to software that uses an open development process. One of the primary benefits of open source software is that its code base is transparent, so you are not tied to any particular vendor, unlike proprietary software. The transparent, collaborative, and often cost-effective nature of open source software has contributed to its widespread usage in archives, libraries, and museums.

Modular open source tools supporting the curation of born-digital content have matured greatly in recent years, and many open source applications now have solid user communities, stable code bases, and documentation. However, there is currently no single “end-to-end” solution for born-digital archiving. Most institutions must adopt separate systems for different functions in the born-digital archiving process. Although we dream of automated acquisitions-to-access digital archives workflows, the reality is that collecting institutions frequently experience difficulties when attempting to synchronize these open source tools to enable efficient, effective, and scalable curation workflows. This issue is discussed in greater detail in the *Guide to Documenting Born-Digital Archival Workflows.*

The 13 steps outlined here were identified through the workflow assessment and documentation process done in collaboration with the 12 OSSArcFlow partners, listed here. These steps were the most common processes undertaken by all partners when archiving born-digital content, though not all of these steps are undertaken by each institution. Each institution may also perform these steps in a different order, so they are not necessarily listed here in order of operation.

The steps are:

- Gather information before acquisition about the material and the nature of its creation
- Transfer materials to the collecting institution using packaging and transfer tools
- Create a disk image, or a sector by sector copy of the data from physical storage media

- Run virus check(s) and quarantine any infected items
- Identify files and characterize their formats
- Check file integrity to verify that there have been no undocumented changes to digital objects
- Create accession records by logging key information about physical media or digital objects
- Analyze and identify sensitive content in order to prevent unintentional disclosure
- Analyze forensic or technical metadata to extract a range of important information about the disk and its content that may be "hidden"
- Create or extract digital object metadata in order to facilitate discovery, access, preservation and use for born-digital materials
- Assemble the Archival Information Package, which bundles together digital materials for long-term preservation
- Assemble the Dissemination Information Package, which converts all or a subset of the Archival Information Package into an access copy
- Transfer the Archival Information Package to a dedicated preservation environment for long-term storage and maintenance

It is important to emphasize that this list of steps is not prescriptive or linear. As evidenced by our OSSArcFlow project partners, these steps are often assembled in different ways for different collections and units even within one institution.

For example, in this workflow for born-digital accessions at New York University, forensic disk images are not created for all materials. Whereas heterogenous born-digital materials get the "forensic disk treatment", staff simply create a logical copy of the files when they are donor-appraised, curated digital files, or large structured transfers from partners. This demonstrates that even within the same institution, not all material moves through the workflow in the same way.

In this example, also from New York University, filetypes are identified at the beginning, when the archivist is conducting a field visit to review a potential accession. In this way, the file characterization step can be used to help aid in pre-appraisal of files before acquisition. However, this example from Duke University shows that file characterization can also occur later in the workflow, after the material has been transferred to the institution and a forensic disk image has been captured. There is no evidence that one of these workflows is a better practice than the other; the order of born-digital archiving steps are configured based on local needs and practices.

Although many of these steps can be performed with separate tools, there are a few open source software tools that were used by project partners that bundle multiple steps into one integrated environment. The benefit of using one of these tools is to potentially cut down on the number of handoffs that need to be managed between different systems, which may require you to transform the data or metadata into different formats. This is not an exhaustive list of tools

that can be used for born-digital archiving, and there are more tools that are named in the *Guide to Documenting Born-Digital Archival Workflows.* This is merely a representative sample of the kinds of tools that are currently used in this work, each of which has a robust community of users.

The BitCurator Environment includes a suite of open source digital forensics tools that are bundled into one Linux environment. BitCurator includes tools that can create disk images, run virus checks, identify filetypes, check file integrity, identify sensitive content, extract digital object metadata, and conduct forensic analysis.

Archivematica is an open source digital preservation system that packages together multiple open source tools to aid in the preparation of archival information packages and dissemination information packages. Archivematica includes functionality for running virus checks, identifying filetypes, checking file integrity, identifying sensitive content, creating digital object metadata, assembling archival and dissemination information packages, and transferring packages to a preservation environment.

ArchivesSpace is an open source archives information management application for managing and providing web access to archives, manuscripts, and digital objects. ArchivesSpace is commonly used to create accession records and digital object metadata in order to facilitate discovery, access, preservation and use options for born-digital materials.

The next video will cover the process for documenting born-digital archiving workflows, including how to assess your current practices, how to describe your workflows, and how to create a visual workflow diagram.