Modular Curation for ETD Repositories

Matt Schultz
Educopia Institute
Atlanta
GA 30309, USA
+1 (616) 566-3204
Matt.Schultz@metaarchive.org

Stephen Eisenhauer
University of North Texas
Denton
TX 76205, USA
+1 (817) 458-8585
Stephen.Eisenhauer@unt.edu

Nick Krabbenhoeft
Educopia institute
Atlanta
GA 30309, USA
+1 (917) 409-7907
Nick@metaarchive.org

ABSTRACT

Led by the University of North Texas and funded by the Institute of Museum & Library Services (IMLS), the *Lifecycle Management of ETDs Project* (2011-2013) is documenting lifecycle curation practices for electronic theses & dissertations (ETDs) and improving implementations for curation tools such as Clam AV (for virus checking), and JHOVE/JHOVE2, DROID, and FITS (for format identification and validation), to better facilitate the curation and management of ETD collections. The project has evaluated several open source institutional repository (IR) systems and related submission systems currently being used for ETDs to determine their provision and extensible support for such curation tools. The IRs and submission systems evaluated included E-Prints, OpenETD, ETD-db, DSpace, and Vireo. ETD programs are primarily implementing these repository software systems for quality control and workflow management to enable authors to submit ETDs (often via ProQuest), and deposit them for on-going access and long-term archival management. Some of these software systems already provide modular support for virus-checking, format identification, and format validation whereas others do not. The following paper will explain the research methodology the project took to evaluate these various tools, IRs and related submission systems; share findings; and discuss how these findings have solidified implementation improvements for the above mentioned curation technologies (Clam AV, JHOVE/JHOVE2, DROID & FITS).

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Collection, Systems issues.*

General Terms

Documentation, Management, Reliability, Security, Verification

Keywords

Curation, Digital Preservation, ETDs, File Formats, Institutional Repositories, PREMIS, Virus Checking

1. INTRODUCTION

The *Lifecycle Management of ETDs Project* (2011-2013) being led by the University of North Texas and funded by the Institute of Museum & Library Services (IMLS) is performing research toward several deliverables that will have bearing on institutional repositories (IRs) and the workflows that feed into them for the archiving of electronic theses & dissertations (ETDs). As submission of student theses moves from print-based to digital, the project is evaluating, first and foremost, how institutions will begin to address the entire lifecycle of ETDs, ensuring that the ETDs they acquire from students today will be available to future researchers? The answer to this question is shaping up to be one of both practices and technologies. Good guidance documents are needed for ETD curators (and these are under their own research and development within the project). But institutions also need easy-to-implement modular curation tools that can work with and/or alongside of the repository systems that they are deploying for their ETDs. As a first step, institutions' IR workflows and systems should accommodate enhanced curation functions such as virus checking, file format identification and file format validation so that they can verify that ETD submissions are clean, stable, and what they purport to be as newly ingested digital objects. How easy are the implementations of current modular curation tools? What barriers exist to their implementation at the IR level or within an ETD program's workflows?

The following paper will begin by describing the typical submission workflows employed by ETD programs, and identifying several open source repository systems with current or potential support for enhanced curation functions. The case for enhanced curation of ETDs in a modular fashion will then be made clear, followed by a thorough overview of the evaluations undertaken by the *Lifecycle Management of ETDs Project* on several modular open source curation tools (Clam AV, JHOVE/JHOVE2, DROID & FITS). These evaluations were carried out to determine their ease of implementation within existing ETD submission workflows and open source IR systems. Based on these evaluations, it is clear that even though minor structural and programmatic barriers do exist to the proper implementation of such tools, the tools themselves are adequately developed for the current curation needs of ETDs. As will be made clear, ETD program managers and stakeholders will benefit from the project's forthcoming simplified documentation and functional requirements for requesting similar such curation features from external submission/archiving service providers.

2. Current ETD Submission Workflows

Most U.S. universities currently have a small variety of fairly standardized, higher-level workflows for managing the submission of ETDs. In many cases ETDs enter these workflows from a student author via ProQuest's UMI ETD Administrator interface, where representatives from the graduate schools can review the submission prior to deposit with ProQuest. Once ProQuest has received an ETD for archiving and disseminating, the university's ETD program may in turn receive (via FTP or some other agreed upon transfer method) a copy of the ETD, supplemental files, and metadata from ProQuest for deposit in

their own institutional repository. Work is then often undertaken by the unit operating the university's IR to catalog the ETD, re-process the metadata in-line with local schemas, and then deposit the ETD and any supplemental files in the repository for long-term archiving and access purposes. In many cases, either the library handles these activities, or in some cases the graduate school and library collaborate. Occasionally student submission begins at the graduate school and is then stewarded on to ProQuest for deposit. In some cases the final submission approved by ProQuest may also be the version that gets returned to the university's ETD program for subsequent deposit into their IR. [1]

3. Open Source Repository Implementations

Universities and libraries are employing a number of software solutions—proprietary, non-proprietary, hosted and non-hosted—to facilitate the workflows described above. For the purposes of this project, research has focused on evaluating the predominant open source software systems currently in use for ETD submission and archival deposit, in an effort to determine their current support and extensibility for modular curation enhancement (more on this below). The primary open source ETD-related software systems include EPrints [2], OpenETD [3], ETD-db [4], DSpace [5] and Vireo [6].

EPrints, an open source repository software developed by the University of Southampton, was selected for evaluation because of its add-on/extension architecture that allows for the incorporation of other modular services that can be run over top of submitted content.

OpenETD is a web-based ETD submission system developed by Rutgers that can be locally hosted and maintained—it has no add-on/extension architecture but its support service invites feature requests and its open code base and license can support minor extensible scripting.

ETD-db, developed by Virginia Tech, is also an open source web-based ETD submission system that is currently undergoing significant re-architecting and currently has no support for add-ons or extensions.

Finally, though some universities are currently re-thinking open source approaches to their institutional repositories—considering such things as Hydra [7], Islandora [8], and Archivematica [9] among others—DSpace has frequently been the repository software of choice for depositing ETDs for long-term archival management and access. DSpace is freely available and supports a robust add-on/extension platform, including their Curation System [10], which provides for a series of tasks such as virus scanning, format identification, and fixity creation.

Vireo, also a web-based ETD submission tool, was originally architected to layer over top of DSpace to serve as a front-end for the overall ETD submission workflow into that popular institutional repository software system.

There are several other ETD submission and repository systems in existence, many of which however are not available for download and testing due to their licensing or because of their strictly hosted status—these include OhioLink (and their Digital Resource Commons) [11], ExLibris DigiTool [12], bepress Digital Commons [13], and ProQuest's UMI [14].

4. Modular Curation for ETDs

As mentioned above, some of the systems evaluated in the *Lifecycle Management of ETDs Project* are already equipped with curation features, usually in the form of modular add-ons or extensions. Typical curation functions include virus-checking, format identification and format validation, as well as fixity creation. None of these curation functions are typically geared to a specific content genre, but they can be used to ensure clean and complete deposits of content for long-term archival management, and can generate useful information for quarantining or making on-going preservation decisions such as format migration and normalization.

Though ETDs are often produced in fairly non-proprietary formats (typically PDF), this is certainly changing as more proprietary or less standard multimedia and research datasets increasingly come to comprise the supplemental files that accompany an ETD submission. In addition, ETD files are created and passed between multiple platforms and servers before they arrive in an IR and therefore run an increased risk of acquiring viruses.

Which is all to say that ETDs warrant the application of enhanced curation features at various stages in the workflow of their processing for long-term archival management. The *Lifecycle Management of ETDs Project* has therefore sought to advance the ease with which ETD programs can implement such curation services as modular standalone features within their current ETD workflows and/or in coordination with their ETD submission and IR software systems.

5. Modular Open Source Curation Tools

A number of such modular open source curation tools are already in existence and have received thorough use case testing and even production implementation in various repository settings.

For virus checking, Clam AV continues to be the tool of preference, primarily for its ease of use and extensibility—there are a number of specialized implementations and the API is quite robust [15].

JHOVE (JSTOR/Harvard Object Validation Environment) [16], and JHOVE2 [17] are command-line tools in active deployment by curators for performing file format identification and validation, as well as metadata extraction.

DROID (Digital Record and Object Identification), a GUI-based tool, is also in usage for format identification with links to the PRONOM technical registry. [18]

FITS (File Information Tool Set), which encapsulates both JHOVE and DROID (among other utilities), is yet another command-line driven tool that is gaining popularity. [19]

Each of these tools take different approaches to analyzing digital file formats and have very different outputs. They are also in various stages of development and refinement. Each of these open source tools is intended to work primarily as standalone utilities but are also often tool-chained into larger digital preservation workflows and systems. E-Prints, for example, has a Preservation Toolkit add-on that invokes DROID, and DSpace makes use of Clam AV as part of its Curation System task set.

The tools mentioned above were chosen for the project evaluation because of their broad application to analyzing digital files of *most* format types (as opposed to limited and specialized formats). There are other similar modular open source curation tools such as ExifTool [20] and the New Zealand Metadata Extractor [21] (both encapsulated by FITS) that are geared primarily towards analyzing and extracting metadata from files as opposed to simply validating their cleanliness and integrity. As will be mentioned below, though use cases for enriched technical metadata are important, for the immediate purposes of the project deliverables the goal is to explore ease of implementation for modular curation tools that can provide simple identification and pass/fail outputs to fulfill lightweight preservation metadata (e.g., PREMIS Events).

6. Evaluation Methodologies

To properly evaluate the current ease of implementation and any needed development for these modular open source curation tools, a number of research activities were undertaken in the project. As a first step, it was important for our project team to become familiar with installing and implementing the tools based on their current documentation. In addition, each of the open source IR systems in widespread usage was installed by the project team and investigated for their degree of openness or extensibility for the incorporation of such modular curation tools. Once both the tools and the IR systems had been analyzed, the project team then turned its attention to performing a needs analysis across the project partners' ETD programs to determine the best fit for inserting the curation tools into their existing ETD workflows and IR systems. Partners interviewed included Boston College, Indiana State University, Penn State, Rice University, University of Arizona, and Virginia Tech. Finally, because Vireo has proven itself to be a helpful open source and downloadable submission system that layers over top of IRs such as DSpace (and soon to be other systems as well), the project team also took some time out to speak with the lead developers to better understand its extensibility for incorporating the curation tools under investigation.

7. Findings

The modular open source curation tools that we investigated proved to be adequately documented, developed, and easy to install and begin directing toward their intended usage. During the course of our partner interviews the project team explicitly asked whether there were any resource or skill barriers to making use of the tools in their current state. All sites reported that their library and repository technical staff would have no problems with implementation of these tools on behalf of their ETDs, once the proper workflows were addressed.

There are currently two major open source IR systems (E-Prints and DSpace), which thanks to their add-on/extension platforms, provide out-of-the box support for virus scanning and/or file format identification and validation. In addition, Vireo also holds forth some future potential, thanks to its APIs, for the insertion of modular curation tools. Unfortunately, though OpenETD and ETD-db could have their functions modified to handle calls to external tools such as Clam AV, JHOVE/2, DROID and/or FITS, such modifications would undoubtedly introduce software vulnerabilities—these two IR systems do not currently support add-ons or extensions. [22]

Without question, the most important set of findings that emerged from the project team's research were that current workflows for ETD submission create problems for the proper application of the evaluated curation tools. The reasons for this are two-fold. First and foremost, though there exists out-of-the box support for curation tools at the IR level (as mentioned above), by the time an ETD and its supplemental files reach the library/repository host and receive final preparation for submission into the IR, it is likely far too late to remedy and request a re-submission should a virus be detected and/or a file format proves itself to be invalid. Which leads immediately to the second problem, namely that the graduate schools, whom are the agents best suited from a workflow perspective to apply such curation tools at early stages of submission, do not typically have the hosting resources, mission, or technical skills required to perform this role and function.

Two potential solutions have emerged in this research. First of all, university libraries or repository system managers/staff could begin to work more closely with the graduate schools (perhaps in negotiation with submission service providers) to apply such modular curation tools at early stages of submission. Alternatively, the graduate schools and the extended group of ETD program stakeholders could work strategically to advocate for the application of such services and output reports (where they are not already available) from submission/archiving providers like ProQuest, Vireo, or OhioLink's Digital Resource Commons (to name just a few). Particularly if the university is archiving local copies of ETDs and supplemental files that they are receiving secondarily (usually via FTP or some other agreed upon transfer method) from a service provider (e.g., ProQuest).

8. Solutions

Based on the findings covered above, the Lifecycle Management of ETDs Project aims to contribute several practical resources that can help to bridge such solutions.

First of all, though modular curation tools such as Clam AV, JHOVE/2, DROID & FITS are already very well documented for their standalone usage, using them to programmatically fulfill particular curatorial goals (e.g., the curation of a genre of content with shared characteristics like ETDs) is not. The project team is now working to simplify documentation along these lines for an ETD program audience. This can help to increase their uptake and usage on behalf of ETDs.

In addition, the project team will also document functional requirements that could be used by ETD program stakeholders (graduate schools, libraries, IT, etc.) to help to advocate for filling curation gaps where such curation services reach ETDs too late in the submission workflow.

Finally, as mentioned above, though the proper and systematic usage of the enriched technical metadata that is the output of such modular curation tools is out of scope for this project, the project team will document some lightweight measures that can be taken to record the identification and pass/fail information as basic preservation metadata (e.g., PREMIS Events). This basic

preservation metadata can be coupled with open source tools such as the PREMIS Event Tracker (reported on at Open Repositories 2011) [23] to facilitate quarantining and/or migrating/normalizing of ETD submissions.

9. Conclusion

The *Lifecycle Management of ETDs Project* is evaluating the current ease of implementation and any needed development for the increased usage of modularized curation tools on behalf of ETDs. Research and findings have demonstrated that the uptake of modular open source curation tools such as Clam AV, JHOVE/2, DROID & FITS will benefit from simplified implementation documentation for an ETD program audience. ETD programs will also benefit from the project's documentation of functional requirements that can be used to advocate for expanded curation features and output reports from external submission/archiving service providers.

10. References

[1] Based on needs analysis interviews with project partners Boston College, Indiana State University, Penn State, Rice University, University of Arizona, and Virginia Tech

[2] http://www.eprints.org/

[3] http://rucore.libraries.rutgers.edu/open/projects/openetd/

[4] http://scholar.lib.vt.edu/ETD-db/index.shtml

[5] http://www.dspace.org/

[6] https://www.tdl.org/etds/

[7] http://projecthydra.org/

[8] http://islandora.ca/

[9] https://www.archivematica.org/wiki/Main_Page

[10] http://www.dspace.org/1_7_1Documentation/Curation%20System.html

[11] http://drc.ohiolink.edu/

[12] http://www.exlibrisgroup.com/category/DigiToolOverview

[13] http://digitalcommons.bepress.com/

[14] http://www.proquest.com/en-US/products/dissertations/

[15] http://www.clamav.net/lang/en/

[16] http://jhove.sourceforge.net/

[17] https://bitbucket.org/jhove2/main/wiki/Home

[18] http://digital-preservation.github.com/droid/

[19] http://code.google.com/p/fits/

[20] http://www.sno.phy.queensu.ca/~phil/exiftool/

[21] http://meta-extractor.sourceforge.net/

[22] Based on software analysis of OpenETD and ETD-db conducted by the project team.

[23] Mark Phillips, Matt Schultz, Kurt Nordstrom, "PREMIS Event Service," Open Repositories 2011, Austin, TX, June 9, 2011. Available at: https://conferences.tdl.org/or/OR2011/OR2011main/schedConf/presentations