Grand Valley State University
ScholarWorks@GVSU

Articles University Libraries

1-1-2014

Utilization and Refinement of Standard Curation Models

Max Eckard
Grand Valley State University, eckardm@gvsu.edu

Jonathan P. Leidig
Grand Valley State University, jonathan.leidig@gvsu.edu

Recommended Citation
Eckard, Max and Leidig, Jonathan P., "Utilization and Refinement of Standard Curation Models" (2014). *Articles.* Paper 46.
http://scholarworks.gvsu.edu/library_sp/46

Utilization and Refinement of Standard Curation Models

Max Eckard
University Libraries
Grand Valley State University
max.eckard@gvsu.edu

Jonathan P. Leidig
School of Computing and Information Systems
Grand Valley State University
jonathan.leidig@gvsu.edu

ABSTRACT

The OAIS and Curation Lifecycle Model provide widely accepted models for curation workflows. However, primary and scientific research often produces content in a manner incompatible with the lack of emphasis these models place on integrating curation-supporting activities in early stages within a scientific workflow. Pre-ingest modules are needed in both models to enable curation of complex, domain-specific content during generation processes.

Keywords
Curation, generation, processes, scientific content.

INTRODUCTION

Curation activities performed by information scientists and the systems they develop are often based on a set of standard models that guide these efforts. The models are well suited for large-scale efforts to manage, archive, preserve, and provide access to heterogeneous content from sources across an institution. However, typical research practices conflict with these models. Primary and scientific research does not produce content suitable for ingest into an archive and curation process until the last stages of a scientific workflow, e.g., at publication, despite generation of large quantities of data at earlier stages. Effective curation requires earlier collaborations between researchers and information scientists than is demonstrated by standard practices that have developed in each of these communities. Collaborations between information scientists and researchers have lead to successful management and curation systems when deployed in early stages of the scientific workflow, e.g., archeology, earthquake modeling, network science, public health, and sensors (Leidig, 2012).

Researchers rarely have expertise or training in long-term information management, generation of content in suitable formats, or specification of metadata. Even in data-intensive domains, minimal effort is allocated to the selection, preservation, and manual annotation of scientific content. Curation models and practices have not proven to be effective, due to the human-intensive burden placed on experts for domain-specific data modeling, storage, management, and retrieval. Modifications to the curation models will improve curation of primary research data.

UNIVERSITY LIBRARY CURATION PRACTICES

Libraries and archives have widely accepted two models of curation processes for digital data, i.e., the Open Archival Information System (OAIS) Reference Model (Consultative, 2012) and Digital Curation Centre (DCC) Curation Lifecycle Model (DDC, 2013). These models inform preservation software development, curation curricula, and repository audit processes, risk assessments, and certifications. Librarian efforts are often focused on the curation of collections of born-digital objects, digitized special collections, archival material, electronic institutional records, commercial content outside of the public domain, learning objects, research data, scholarship, and creative works produced by the activities of faculty, staff and students. Longstanding success in these processes demonstrates the suitability of curation models for disseminating collections of simple and complex digital objects, publications, and metadata through digital libraries, institutional repositories, and library catalogs.

Open Archival Information Systems

An OAIS is an "archive, consisting of an organization… of people and systems that has accepted responsibility to preserve information and make it available for a designated community" (Consultative, 2012). Digital objects are ingested or acquired by an OAIS as a Submission Information Package (SIP), archived as an Archival Information Package (AIP), and made available to consumers as a Dissemination Information Package (DIP). As an example, the archives in a university library often curate collections of digital images of university events using the OAIS model. The original SIP may be composed of raw camera files or raster images in various formats and may or may not include a structured description of the people or events they detail. Formal ingest includes copying files to storage media, stripping filenames of special characters, and running a virus scan. Creation of the AIP involves adding descriptive, structural, administrative, and preservation metadata to keep track of information about provenance, authenticity, preservation activity, technical environment, and rights management. AIP production involves normalizing files to preservation formats, e.g., converting proprietary PSD files to the high-confidence, non-proprietary TIFF format. The SIP is archived to allow

ASIST 2013, November 1-6, 2013, Montreal, Quebec, Canada.

for alternative preservation actions in the future, such as emulation. Creation of the DIP may include cropping or editing images, adding a watermark, and normalizing files to access formats, i.e. JPG. The SIP and AIP are kept in secure, geographically redundant archival storage, while online access to the DIP derivatives are provided through digital collection management software.

Digital Curation Centre's Curation Lifecycle Model

The Curation Lifecycle Model provides a "graphical, high-level overview of the stages required for successful creation and preservation of data from initial conceptualization or receipt through the iterative curation cycle" (DDC, 2013). The model also describes sequential activities that process data throughout the curation lifecycle. To use another example, a university library may curate a collection of digitized reel-to-reel interviews chronicling the local history of a particular city in the Midwest. While not involved in the 'conceptualization' phase for the original interviews, the library may be involved in the 'create or receive' phase by asking a vendor to provide high-quality digital masters in a WAV or AIFF file format recorded at a 96,000 Hz sample rate and at 24 bit-depth. 'Appraisal and selection' for digitization may be based on the perceived long-term value of a particular interview, the details of the original consent forms, or state of deterioration of the original tape. Digitized interviews are formally 'ingested' by an archive when returned by the vendor. 'Preservation action' would include the creation of additional descriptive, structural, technical, and preservation metadata. It may include generating transcripts of the audio to aid in keyword search. A university library would then 'store' the digital audio, text of the transcripts, and associated metadata in archival storage, and use digital collection management software to provide 'access, use, and reuse' for derivatives of the audio, i.e., MP3. The digital audio and text may then be go through a stage of 'transformation,' for example, a printed transcript in a local history book, in a video shown during freshman orientation, or by researchers analyzing speech patterns of Midwesterners. Activities also include disposal, reappraisal, and migration as master files become obsolete.

PRIMARY ACADEMIC AND SCIENTIFIC RESEARCH

Relying on standard curation models to support primary research activities is insufficient. Sustainable scientific research requires the capture of selective digital artifacts as produced over the course of data generation activities. Additional curation activities must be added to the curation models to capture workflows and provenance of data as they are produced. Three deficiencies with these models lead to ineffective curation processes and indicate the need to extend the initial stage of the curation models.

Delayed Collaboration

Typically, information scientists will initially engage with researchers at the point of a specific ingest request. This delayed researcher-librarian collaboration is often due to outdated research practices, lack of cross-discipline expertise, and a byproduct of digital repositories modeled after the OAIS model. The OAIS model "seems to assume a minimal level of data fixity, and a single archiving event" (Salo, 2011). Librarians must collectively extend the 'conceptualize' and ingest stages of the Curation Lifecycle Model and OAIS models to include systematic, active identification and management for data, metadata, provenance information, and methodologies. Collaboration should be shifted earlier in the scientific workflow, i.e., concurrent with data generation, instead of *ex post facto*.

Unachievable Preservation

Within an OAIS, preservation is classified as the archive's responsibility. In reality, preservation relies on prior data-management planning, holistic data collecting, validating and verifying, versioning, and cleansing. Researchers perform these activities while conducting research endeavors, long before the ingest stage of a repository. These early activities affect the likelihood that digital objects will be successfully preservation-ready at ingest.

Deferred Curation

Researchers do not utilize librarian expertise or guidance when selecting highly-valued digital objects or developing metadata schemas for domain-specific content. Due to credibility concerns, researchers are generally leery of annotating and archiving incomplete, partial scientific results and datasets. Instead, archival and curation are viewed as activities that take place concurrent with publication and conclusion of a multi-year study. Curation-minded researchers are unable to follow best practices for eventual curation and archival activities beyond the need to manually generate a complete, sufficiently annotated SIP.

CONCLUSIONS

Two models serve as guidelines for the development of curation systems. These models outline the provision of a successful curation process but rely on an ideal ingest request. In scientific domains, these models need to be extended into earlier stages of research workflows to avoid several pitfalls of curation in relation to primary research and researchers. A suggested revision of OAIS and Curation Lifecycle Model entails the addition of data planning, selection, validation, and cleansing to the ingest stage, as demonstrated in (Leidig, 2012).

ACKNOWLEDGMENTS

We thank collaborators at DLRL, NDSSL, and SLIS at Virginia Tech and North Carolina Central Univ.

REFERENCES

Consultative Committee for Space Data Systems. (2012). *Reference Model for an Open Archival Info. System*. http://public.ccsds.org/publications/archive/650x0m2.pdf.

Digital Curation Centre. (2013). *Curation Lifecycle Model*. http://www.dcc.ac.uk/resources/curation-lifecycle-model.

J. Leidig. (2012) *Epidemiology Experimentation and Simulation Management through Scientific Digital Libraries*. Ph.D. Thesis, Virginia Tech, Blacksburg, VA.

D. Salo. (2011). Models and modeling [PowerPoint slides].