Digital Preservation: Tales from the Precipice between theory and practice

Amanda Focke
Woodson Research Center
Fondren Library, Rice University
P.O. Box 1892, MS 44
Houston, TX 77251-1892
afocke@rice.edu

Ying Jin
Digital Scholarship Services
Fondren Library, Rice University
P.O. Box 1892, MS 44
Houston, TX 77251-1892
ying.jin@rice.edu

Monica Rivero
Digital Scholarship Services
Fondren Library, Rice University
P.O. Box 1892, MS 280
Houston, TX 77251-1892
monica.p.rivero@rice.edu

ABSTRACT
Long term access to digital materials relies on active management of these resources. A major challenge for today's data stewards in managing digital collections is how best to apply the plethora of recommended standards and emerging technologies not only to newly created digital content but to legacy digital data. This paper describes our ongoing activities and methods used in applying standards and best practices in support of the preservation and continuous use of our digital assets. This includes: gathering the history of the data in digital curation profiles, analysis of file formats, storage and management of high resolution master files, ensuring critical preservation metadata is maintained and assessing these digital resources in terms of setting preservation priorities.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: collections, standards

General Terms
Documentation, Management, Standardization, Reliability

Keywords
Digital preservation, digital curation, digital archiving, repositories management, standards

1. INTRODUCTION
Digital curation is defined as "maintaining and adding value to a trusted body of digital research data for current and future use" [4]. This concept is a product of both data preservation efforts and the data deluge which is simply the rapid increase in the quantity of digital information. However, the concept of digital curation is more inclusive than digital archiving and preservation and has therefore evolved into a need to understand "how preserving digital materials fits into the broader theme of digital stewardship" and the "aggregation of interconnected services, policies, and stakeholders which together constitute a digital information environment" [4].

This concept of digital curation incorporates both digital preservation and data stewardship into the bigger picture of maintaining digital data throughout its lifecycle. The lifecycle of data includes the creation; selection and ingestion of data into the institutional repository; preservation and assessment efforts; storage and access as well as reuse, repurpose and migration of data into newer formats [1]. As data stewards, we are faced with the challenge to go beyond just preserving data as received but also investigate ways to make this same data continuously accessible to researchers and scholars. This paper will discuss our efforts in supporting data including the creation of collection level digital curation profiles, analysis of file formats and preservation metadata to strategies for archiving master files.

2. BACKGROUND
In 2010, Fondren Library formed a committee to review the challenges and opportunities presented by the task of reliably preserving our locally created digital content. We looked at general best practices and standards in the field of digital preservation (trends, interoperability, preservation metadata, and file format standards), our own storage and back-up system's merits and faults, and migration and auditing issues. We performed an internal TRAC [7] self-assessment as part of our review, and produced a set of recommendations for Fondren's next steps.

Our top recommendation was to compose a Digital Preservation Policy for Fondren's Digital Collections, a one to three page document laying out high level priorities which the Library supported and working out more detailed documentation from there. Areas needing attention included storage and back-up, policy and communication, further data assessment, and improved digitization workflow practices. With so many areas we wanted to address, we chose to begin with a well-defined pilot project where many of these areas could be investigated more deeply given a subset of genuine digital objects. To that end, we began working with our library's digitized Special Collections materials.

3. PILOT COLLECTION DESCRIPTION
Fondren library's Special Collections materials[1] are housed in our institutional repository (IR), a digital archive using the DSpace open source platform. This collection is a body of circa 2,000 digital objects composed primarily of manuscripts, rare books, photographs and illustrations. This group offered a variety of digital file formats (jpg, JPEG2000, TIFF, PDF, plain text, mp3 streaming audio and more), various ages (from 10 year old images

[1] Woodson Research Center
http://scholarship.rice.edu/handle/1911/12341

to the very recently scanned) and metadata using various standards and levels of implementation. Older digital projects and objects in this group tended to have less consistent and complete metadata, especially as relates to preservation. Some PDF files had text recognition (OCR) and some did not. Some printed materials had TEI/XML encoded markup. Some objects had identifiers which would connect logically to a derivative and master filename, and some did not. This variety gave an excellent chance to address what we suspected would be typical problems in the rest of our IR. While handling the problems in the pilot group, we would be preparing our path towards curation in the larger IR.

4. DATA ANALYSIS

4.1 Digital Curation Profile

A digital curation profile is a narrative description of the collection intended to capture the 'story of the data'[2] from the viewpoint of the data creators, drawing on their institutional memory and subject knowledge. We have adopted the concept of a digital curation profile as a technique to help us reexamine and evaluate data and related services. We used a customized template[3], constructed as a series of open ended questions addressed to the creator of the digital resources. This profile acts as a reassessment of the collection, used to confirm existing characteristics of the data, identify changing or emerging needs of users of the collection and identify any "at-risk" data suitable for possible migration treatment.

From the interview with the Woodson Research Center we identified courses of actions to help improve their digitization workflow. Programmers created a customized submission form for manuscript and archival materials that provide metadata fields that are more relevant to these types of documents. This new form makes for a more rapid entry of data into the system. Another outtake from the interview was the need to examine metadata and digital files from older digital collections and reassess quality and determine work necessary to bring these older digital records to current standards. The remainder of this paper discusses some of our efforts in this area.

4.2 Preservation Metadata

Preservation metadata is primarily the technical data necessary for the long term maintenance and potential migration of file formats along with administrative data regarding the intellectual property status of the digital contents. The leading standard in this area is the PREMIS Data Dictionary for Preservation Metadata [6]. However, implementing PREMIS is not a straightforward process. There are a number of redundancies between PREMIS and other well-established metadata schemas and there is no established way to implement PREMIS in our repository. Despite such potential issues, we researched PREMIS to identify which of its parts seemed most important for anticipated future preservation tasks and identified what we were already doing to meet those ends.

4.2.1 Findings

4.2.1.1 PREMIS to Dublin Core mapping

We were able to map a number of PREMIS' preservation-critical elements to our own existing and easily accessible data sources. Wherever possible, it is preferable to use system generated data (such as bit size, fixity checks, uri identifier, etc.) rather than to manually create data. After researching PREMIS and comparing it to the preservation metadata we routinely capture as part of a typical Dublin Core record or collect automatically through the DSpace system, we concluded that that we were already doing a fair job of the end goal of recording basic preservation metadata. We were following the spirit and intention of PREMIS standard. The below table is a short list of the most common elements. A more in-depth semantic mapping with examples and fuller description was prepared during our study.

Table 1. Metadata Crosswalk of PREMIS to Qualified Dublin Core (QDC) Elements

Data	PREMIS	QDc
Size	in size under objectCharacteristics	dc.format .extent
CHECKSUM and CHECKSUMTYPE	in fixity under objectCharacteristics	dc.description .provenance
MIMETYPE	in format under objectCharacteristics	dc.format .mimetype
Unique designation	In IdentifierValue objectCharacteristics	dc.identifier .uri
Format Designation	in format under objectCharacteristics	dc.type.dcmi
Creation of object info	objectDateCreated ByApplication	dc.date.digital
Creation of object info	objectCreating	dc.digitization .specifications

Our IR automatically captures preservation data at the bit-stream level for the following fields: dc.format.extent, dc.description .provenance, dc.format.mimetype. Upon ingest the system assigns a permanent handle or url address (dc.identifier.uri). We manually create or assign values for dc.type.dcmi, dc.date.digital and dc.digitization.specifications (which describes the technical specifications of how an item was digitized, migrated or created. This may include a sundry of information such as method used to create files, types of files and any conditions for accessing files). Our current practice is to also embed copyright and access information directly within the image or provide this statement as an inserted page in PDF files.

4.2.1.2 Completeness and Consistency

We determined that we do need to ensure that all of our collections consistently provide the relevant manually created preservation metadata as listed above, as well as fields for dc.source.collection and dc.rights. As these collections were

[2] Questions were heavily adopted from Witt, M. and Carlson, J. Conducting a Data Interview. in 3rd International Digital Curation Conference, (Washington D.C.,2007), available at: http://docs.lib.purdue.edu/lib_research/81

[3] Digital curation profile Template is available from our library's digital project wiki http://bit.ly/w3mVaC

created at various points in time by various staff members in different standards environments, it is not surprising that some collections have in-depth metadata profiles which include strong preservation elements such as capturing adequate provenance information on a record level, technical specifications (file formats and creation methods) and right/access statements. However, there are a good number of very small collections in our IR that do not have detailed metadata profiles. Accordingly, the metadata for these collections varies significantly in terms of what we now think of as completeness, especially in the area of preservation metadata.

4.2.2 Review Process

Using the curation tools [3] for our IR that will allow us to batch update metadata changes without having to re-ingest the digital files, we have embarked on a project to look at the entire collection focusing on the following fields:

4.2.2.1 dc.type.dcmi

Confirm all records have been assigned a general content type for the resource, using the DCMI Type vocabulary (Collection, Dataset, Event, Image, Interactive Resource, Moving Image, Physical Object, Service, Software, Sound, Still Image, or Text)

4.2.2.2 dc.date.digital

Confirm the digital date is in proper syntax. If not supplied in Dublin Core record may extract from technical metadata embedded in image file.

4.2.2.3 dc.date.issued

Ensure the issue date is the "source" date and not the digital date. Crosscheck information in other qualified Dublin Core fields (e.g. transferring data from other fields like dc.date and dc.date.created).

4.2.2.4 dc.digitization.specifications

Describe scanning equipment used, master and derivative file formats, resolution, color profile and any other relevant information such as TEI markup language version, if applicable.

4.2.2.5 dc.rights

Rights metadata will be critical for future uses of the digital objects. Manually reviewed all items and apply consistent wording as appropriate for our most typical copyright situations.

4.2.2.6 dc.source.collection

Provide reference information to connect back to the physical material. This facilitates patrons' reference requests to view originals and to ensure authenticity of the data source.

4.2.2.7 dc.identifier.digital

Confirm that this identifier matches the filenames for all files associated with the digital object. Assign unique digital identifiers where there is none.

Another finding in our internal study is the importance of digital identifiers as presented in the object level metadata for special collection materials. These are key to preserving the connection between the object and its derivative and master files. In some cases with older objects, our metadata does not include an identifier of any kind. We would rely in that case on accessing the filename of the derivative, using that to look for the master file, and failing that, relying on institutional memory to connect a master to that file. Our more recent digital objects in the pilot group do follow a consistent naming convention where identifiers are always present and always match the derivatives and master filenames. Clearly, this is a case where it makes sense going forward to absolutely require an identifier in the metadata which matches the derivative and any master filenames across the entire IR.

We are looking for completeness of information and consistency of expression, using the appropriate standard or best practice for each field. The results of this manual metadata enhancement, along with the more technical system generated fields, will greatly improve our ability to provide trustworthy access to the materials over time.

4.3 File Formats

The Rice University's digital scholarship archive accepts any type of file. Periodic review of formats helps ensure the integrity of the digital files available from the archive and improves the general trustworthiness and reliability of the archive's content. Through our data analysis we were able to identify the prevalent file formats in the IR as well as any unknown file types submitted to the archive using the curation tool, Bitstream Formats Profile [3]. From a basic count of formats types available from the test collection, we learned that the majority of files were JPEG (60%) and PDF (30%). This inventory confirms quantitatively that text and image media types are of particular importance to this collection and therefore format recommendations should be focused on ways to improve user experience of these media types and ensure their long term access.

A more in-depth analysis is necessary to ensure that files submitted to the archive meet digitization quality recommendations. For example in a sample subgroup of the larger test collection, we exported all digital files and used tools such as DROID [2] and JHOVE [5] to confirm digital specifications. We were able to determine from these tests whether the files were well formed, for example when a file contains an erroneous extent type such as a PDF file with a .jpg extension and made corrections for errors found. However we realize that this process is not practical for larger scaled implementation. We are still exploring exactly how to conduct such analysis systematically for items which are already housed in the IR.

5. PRESERVING MASTER FILES IN THE IR

Our cultural heritage materials are scanned at high resolution, resulting in large bit sized master files. Our historical storage practice has been to back up master files to an independent secure file server outside our IR. Access to this server is heavily restricted; it is cumbersome to transfer files and it can be difficult to track down any particular file as the relationship between the master file and the IR record is not always obvious. Furthermore, we receive a number of patron requests for copies of master level files, making the immediate retrieval of these masters a more urgent priority for supporting user needs. To better manage our master files, we are in the process of transferring them to the IR environment. In implementing this new storage procedure, we will limit masters to admin access only and create a custom ingest script to add the masters to the existing record. Despite adding a very large amount of data to the IR environment, we do not anticipate any impact on IR performance since the masters are hidden from the public and the search function searches metadata and extracted text, not the actual files.

6. CONCLUSION

In assessing collection data in terms of preservation and long-term access goals, we began by conducting data curation interviews to document a narrative account of the data's history, capture current needs and use this information to help guide refinements to our digitization workflows. We found this institutional memory to be critical in auditing legacy records. We also conducted a review of the formats stored in the archive and saw the predominant formats as our preservation priorities. We identified key preservation metadata fields and are updating those fields for completeness and consistency. One recommendation is to write a policy for a core set of mandatory preservation elements for all records. If these elements are not supplied by the submitter, they can be created as part of the IR administrative services.

All of this data analysis work will help to ensure the integrity of our digital collections and improve the overall trustworthiness of the archive's content, as well as preparing data for any future migrations.

Anecdotally, our last recommendation to our fellow data stewards is to start now. This kind of work can begin with a manageable pilot project. Experiment and explore until the right local practical implementation is found. Some of the most important steps as outlined in this paper are very straightforward and starting simply can create critical momentum.

7. ACKNOWLEDGMENTS

Our thanks to Fondren IT specialists, Mang Sun and Marcus Elizondo, for their technical expertise in implementing DROID, JHOVE and EXIFTool software and to DSpace system administrator Sid Byrd for development of custom ingest scripts.

We also wish to recognize Amy Caton, a graduate student in the Library of Information Sciences program at the University of North Texas, for her excellent work on digital curation interviews, transcription and analysis during her practicum at Fondren Library in 2011.

8. REFERENCES

[1] DCC Curation Lifecycle model. Digtial Curation Centre. 2010. Available at: http://www.dcc.ac.uk/resources/curation-lifecycle-model

[2] Digital Record Object Identification (DROID). Computer software. The National Archives of the United Kingdom, PRONOM. Available at: http://droid.sourceforge.net/

[3] Curation Tasks: Bitstream Format Profiler and Batch Metadata Editing Feature. Computer software. DSpace. Available at: https://wiki.duraspace.org/display/DSDOC17/Curation+System#CurationSystem-BitstreamFormatProfiler

[4] Harvey, R. Digital curation: a how-to-do-it manual. Neal-Schuman Publishers, New York, NY, 2010,

[5] JSTOR/Harvard Object Validation Environment (JHOVE). Computer software. Available at: http://hul.harvard.edu/jhove/index.html

[6] PREMIS Data Dictionary for Preservation Metadata, Version 2.0, March 2008. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[7] Trustworthy Repositories Audit & Certification: Criteria and Checklist. OCLC and CRL, version 1.0, February2007. Available at: www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf