IJDC | *Peer-Reviewed Paper*

Media Digitization and Preservation Initiative: A Case Study

Devan Ray Donaldson
Indiana University

Allison McClanahan
Indiana University

Leif Christiansen
Indiana University

Laura Bell
Indiana University

Mikala Narlock
Indiana University

Shannon Martin
Indiana University

Haley Suby
Indiana University

Abstract

Since its creation nearly a decade ago, the Digital Curation Centre (DCC) Curation Lifecycle Model has become the quintessential framework for understanding digital curation. Organizations and consortia around the world have used the DCC Curation Lifecycle Model as a tool to ensure that all the necessary stages of digital curation are undertaken, to define roles and responsibilities, and to build a framework of standards and technologies for digital curation. Yet, research on the application of the model to large-scale digitization projects as a way of understanding their efforts at digital curation is scant. This paper reports on findings of a qualitative case study analysis of Indiana University Bloomington's multi-million-dollar Media Digitization and Preservation Initiative (MDPI), employing the DCC Curation Lifecycle Model as a lens for examining the scope and effectiveness of its digital curation efforts. Findings underscore the success of MDPI in performing digital curation by illustrating the ways it implements each of the model's components. Implications for the application of the DCC Curation Lifecycle Model in understanding digital curation for mass digitization projects are discussed as well as directions for future research.

Received 9 July 2017 ~ *Revision received* 4 May 2018 ~ *Accepted* 7 May 2018

Correspondence should be addressed to Dr. Devan Ray Donaldson, 700 N. Woodlawn Avenue, Luddy Hall 2126, Department of Information and Library Science, School of Informatics, Computing, and Engineering, Bloomington, IN, USA, 47408. Email: drdonald@indiana.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v13i1.502
DOI: 10.2218/ijdc.v13i1.502

Introduction

Since its creation nearly a decade ago, the Digital Curation Centre (DCC) Curation Lifecycle Model has become the quintessential framework for understanding digital curation. Organizations and consortia around the world have used the DCC Curation Lifecycle Model as a tool to ensure that all the necessary stages of digital curation are undertaken, to define roles and responsibilities, and to build a framework of standards and technologies for digital curation. Recently, researchers have begun exploring the impact of the DCC Curation Lifecycle Model on understanding how digital curation is performed in various contexts with different types of digital data. These include, but are not limited to: video data in social studies of interaction (Whyte, 2009), and brain images in psychiatric research (Whyte, 2008). One unexplored research area involves the utility of the DCC Curation Lifecycle Model to act as a lens for understanding digital curation in mass digitization projects. This type of research is critical given the rise of mass digitization projects over the past decade, which is expected to increase within the cultural heritage and library services domains.

The purpose of this paper is to report findings from a qualitative case study analysis of a mass digitization project using the DCC Curation Lifecycle Model as a means of understanding the scope and effectiveness of its digital curation efforts. Of the mass digitization projects that currently exist worldwide, we selected Indiana University Bloomington’s (IUB’s) multi-million-dollar Media Digitization and Preservation Initiative (MDPI) for our case study because heretofore research has focused primarily on the mass digitization of textual resources (e.g., books). In contrast, MDPI aims to digitize and make accessible a wide variety of time-based, audiovisual media. The main research question this study addresses is: How do the actions of MDPI compare to the actions specified in the DCC Curation Lifecycle Model?

The remainder of this paper is organized as follows. First, the background section explores: 1) the value and importance of mass digitization and recent mass digitization projects; 2) MDPI, including its history and development; and 3) the DCC Curation Lifecycle Model, including previous research on its application in various organizational contexts with different types of digital data. Second, the methods section describes how we applied case study research design and methods to this project. Third, the findings section provides details about how MDPI implements each part of the DCC Curation Lifecycle Model. Fourth, the discussion section addresses the utility of the DCC Curation Lifecycle Model for understanding digital curation in mass digitization projects based on our findings. The discussion section also compares our results with findings of other studies that have used the DCC Curation Lifecycle Model as a framework. The paper concludes with discussion of future directions for research.

Background

Defining Mass Digitization

As early as 1993, mass digitization has been discussed as a way to provide wider access to and preservation for archives and collections. In the past decade, mass digitization has received a considerable amount of attention and scrutiny due to the involvement of large corporations and research institutions such as Yahoo!, Google, Microsoft, Stanford University, the University of California system, and the University of Michigan. Despite several publications on the topic of mass digitization, few definitions for the term exist (Schmitz, 2008). This has led to some vagaries and has, in part, hidden the variety of mass digitization projects. Coyle (2006) defines mass digitization as "the conversion of materials on an industrial scale." For this paper, our working definition of mass digitization builds on Coyle's (2006) by considering six characteristics. Note that they are not mutually exclusive:

- **Aggregation and production** (e.g., whether a project aggregates material that others have digitized, and the extent to which a project performs any digitization in-house);
- **Openness** (e.g., the extent to which materials are open and freely accessible);
- **Business model and cost** (e.g., commercial, non-profit, etc., and who pays for the digitization?);
- **Scope** (e.g., are you digitizing everything, most everything, or specific collections?);
- **Format** (e.g., digitized books, audiovisual materials, 3D Materials, etc.);
- **Time spent digitizing** (e.g., seconds per item, minutes per item, hours per item, etc.).

Based on our scan of recent major mass digitization projects, we argue that these characteristics make the most significant difference in determining whether a digitization project is a mass digitization project. Below we describe each of these characteristics as they relate to specific projects.

Aggregation and production

Arguably the most massive and most publicized digitization project is the Google Books project, formerly Google Print. Officially started in 2004, the Google Books project has amassed a collection of approximately 25 million volumes to create a searchable and meaningful indexed digital library (Heyman, 2015). Google digitized many of these volumes. Libraries may bring books to Google for digitization according to Google's proprietary method.[1] Libraries have also contributed books that they digitized, such as the University of California Davis' contribution of 45,000 volumes.[2]

In contrast to Google Books, the Open Book Alliance was formed in support of open digitization projects.[3] The Open Book Alliance was a consortium of industry, library, and union representatives dedicated to open digital libraries. Members of the Open Book Alliance undertook their own self-titled mass digitization projects. The most successful member of the Open Book Alliance is the Internet Archive, which has over 11 million books and texts.[4] These books are scanned either by the archive or one of its

[1] 5 Things About JSTOR: https://about.jstor.org/5things/

[2] UC Davis Joins Google Book Digitization Project: https://www.cdlib.org/cdlinfo/2013/07/29/uc-davis-joins-google-book-digitization-project/

[3] Open Book Alliance - Mission: http://www.che.ntu.edu.tw/ntuche/safety/upload/browse.php?u=Oi8vd2ViLmFyY2hpdmUub3JnL3dlYi8yMDEzMDYyMDE1NTk0NS9odHRwOi8vd3d3Lm9wZW5ib29rYWxsaWFuY2Uub3JnL21pc3Npb24v&b=13

[4] Internet Archive – About: https://archive.org/about/

450+ partners. The Internet Archive operates 33 scanning centers across the globe.[5] Partners of the Internet Archive include corporate entities such as Yahoo!, which contributed approximately 200,000 texts to the archive.[6] The Internet Archive allows the full download and creative commons use of all 11 million texts with an additional 500,000 texts available for protected viewing.

By aggregating content from other digitization projects, some mass digitization projects have greatly increased their holdings. This is true for the Internet Archive and Google Books. One of the largest digital libraries, HathiTrust, has reached over 15 million volumes by ingesting the digitized works of its partners.[7]

When considering digitization projects that do not involve aggregation, the number of items digitized drops drastically. For example, JSTOR has digitized 50 million pages of data.[8] Assuming an average book size of 300 pages, this is roughly equivalent to 170,000 books. In an effort to effectively utilize its resources, JSTOR staff follow a strict selection process for what they choose to digitize.[9] Similar in size, the American Memory Project has digitized nine million documents; since many of these are only a few pages, the total number of pages digitized is roughly comparable to JSTOR.[10] At the high end of digitization projects that do not involve aggregation, the Million Book Project has digitized 1.5 million volumes (BizResearch, 2007). Yet, this is still an order of magnitude smaller than the huge aggregation/production projects that currently exist.

Openness

Initially, librarians heralded mass digitization initiatives for bound volumes, such as the Internet Archive, Google Books, and HathiTrust. Online libraries yielded new possibilities for collaboration and were capable of addressing issues of access, preservation, and collection management. However, over time, the commercial interests of others, namely Google Books, hampered the optimism motivating these ideas. At present, Google Books serves as an index allowing users to search for words and terms to decide if certain books will suit their needs. It provides minimal contextual information, and restricts access to most of the documents.

As the previous example illustrates, digitization projects differ on how they approach the concept of openness. For example, the Open Book Alliance, HathiTrust, and the Internet Archive all strive to make their materials as open and accessible as possible, while complying with copyright law. They do this by, for example, digitizing works no longer under copyright restrictions. Although openness of digitized files is not a requirement for mass digitization initiatives, it is a characteristic of some of them, depending on institutional goals/beliefs, copyright, and intellectual property issues.

Business model and cost

Institutions involved in mass digitization are either non-profit or for-profit. For example, the Internet Archive is non-profit and Google Books is for-profit. Google funds the Google Books, while the Internet Archive charges its partners to offset its costs.[11] Google is a multi-billion-dollar company, while the Internet Archive receives approximately $10 million in funding annually (Womack, 2003). Google describes the

[5] Internet Archive – Scanning Services: https://archive.org/scanning

[6] Open Content Alliance: https://archive.org/details/opencontentalliance&tab=about

[7] HathiTrust – Statistics and Visualizations: https://www.hathitrust.org/statistics_visualizations; HaithiTrust – Welcome: https://www.hathitrust.org/about#

[8] JSTOR – About: https://about.jstor.org/5things/

[9] Journals: https://about.jstor.org/whats-in-jstor/journals/

[10] American Memory Project – About the Collections: http://memory.loc.gov/ammem/about/about.html

[11] Internet Archive – Scanning Services: https://archive.org/scanning

original motivation behind Google Books as creating a searchable, meaningfully indexed digital library.[12] However, the limited access of the site has led librarians and scholars to question how closely the intentions of Google align with the values of libraries (Hahn, 2011). Furthermore, Google Books has taken a risky, opt-out approach to digitizing copyrighted works, which has led to a series of litigations arguably counterproductive to public digitization projects (Vaidhyanathan, 2007). Regardless of the business model in place, mass digitization requires the mobilization and coordination of significant institutional resources, as can be seen by the prevalence of collaboration in mass digitization projects.

Scope

Similar to Google Books, the Internet Archive does not discriminate in terms of what is digitized. This, according to Coyle (2006), is the goal of mass digitization: "the goal of mass digitization is not to create collections but to digitize everything." However, more recently, it has been argued that such a lack of discrimination should be avoided, lest digital libraries turn into digital "garbage dumps" and "towers of babel" (Gooding, 2013). In contrast, other mass digitization projects specify their scope. For example, in 2000, the China-America Digital Academic Library (CADAL) project scanned 1.43 million: 1) ancient books published before 1911, which are out of copyright; 2) books published since 1911, which may be in the public domain; and 3) dissertations from the 16 partner Chinese libraries (Jihai, 2008).

It appears that scope and size interrelate and are open to interpretation when considering mass digitization. For example, Europeana is a multi-lingual online collection of European heritage institutions.[13] One of its contributors digitized nearly the entire museum collection of the Civic Archaeological Museum in Milan, Italy, producing 500 models (Guidi, Barsanti, Micoli and Russo, 2015). This project has significantly fewer items than even the most modest text digitization project. Yet, researchers who are responsible for this contribution to Europeana claim that they have undertaken mass digitization because of the size of the files that comprise each of the 500 models.

Format

As with most mass digitization projects, Google Books focuses on the digitization of print objects. This may be due to libraries' historical emphasis on print materials, the straightforward nature of text digitization, and/or perhaps this is merely a reflection of the predominance of text as the subject of digitization. More recently, mass digitization projects similar to the one at the National Library of Australia (2017) are continuing to digitize print materials but are also expanding to include music. Other large-scale digitization projects focus specifically on non-textual resources. For example, the Netherlands Institute for Sound and Vision has digitized close to 100,000 hours of audio and video respectively along with 22,000 hours of film.[14] If we are to use some number of items digitized as a criterion for determining mass digitization, the addition of different formats necessitates a change of units. How might we meaningfully compare two formats (e.g., the digitization of books to the digitization of video)?

[12] Google Books – History: https://www.google.com/googlebooks/about/history.html

[13] Europeana: http://www.europeana.eu/portal/en

[14] Netherlands Institute for Sound and Vision: http://www.beeldengeluid.nl/en/about

Time spent digitizing

The Netherlands Institute's self-reporting indicates that for many legacy audiovisual formats the play time of the item is roughly equivalent to the time necessary to digitize. However, with advances in technology and innovation in technique, the time necessary to digitize texts has changed drastically. A direct comparison of formats, whether it be number or digitization time, becomes even more convoluted when one introduces the digitization of 3D objects, such as with the Europeana collection.

Media Digitization and Preservation Initiative (MDPI)

Established in 2010, MDPI is the culmination of incremental infrastructural improvements and practically gained expertise at IUB. The remainder of this section briefly describes its origins and background.

In 1998, the then Vice President and Chief Information Officer Michael A. McRobbie helped to found the Digital Library Program at IUB. It was a collaboration between Indiana University Libraries, University Information and Technology Services (UITS), and the then-School of Library and Information Science[15] to "produce, maintain, deliver, and preserve networked resources for scholars and students at Indiana University and elsewhere to improve the teaching and research of IU faculty, improve the learning and research of IU students, and increase knowledge about the development of digital libraries."[16]

In 1998 IUB implemented the Scholarly Data Archive (SDA), then known as the Massive Data Storage System (MDSS), an extensive storage service that currently has over 40 petabytes of magnetic tape storage. The SDA is a system built with the goal of preserving bit-level data for the long term. When data is ingested into the SDA it is dual written to the IUB and the Indiana University – Purdue University Indianapolis (IUPUI) data centers.

IUB continued to develop infrastructure as well as technical expertise. With units and IU entities such as the IU Libraries Moving Image Archive and the Variations Project, IU continued to be at the forefront of learning and developing techniques and best practices for digitization and preservation of audiovisual materials. IUB was constantly striving to digitize more materials while publishing works on best practice for digitizing audio materials, such as those through the Sound Directions project.[17]

While IUB has had a long-standing commitment to the preservation of audiovisual media, at $15 million dollars and counting, MDPI is one of the most expensive archival projects IUB has ever undertaken. In 2009, an IUB task force published a survey of the media in the IUB collections with a frightening message; there is only 15 years left to preserve over 750,000 valuable recordings in the IUB collection. IUB developed MDPI to address this need for preservation to safeguard the valuable cultural heritage found within its audiovisual collections.

According to the six characteristics defined above, MDPI is a mass digitization project:

- **Aggregation and production** – in partnership with Memnon (Sony), audiovisual engineers and experts digitize materials.

[15] MDPI – History: https://mdpi.iu.edu/about/history.php

[16] History of Digital Preservation at Indiana University: https://wiki.dlib.indiana.edu/display/DIGIPRES/History+of+Digital+Preservation+at+Indiana+University

[17] Sound Directions: http://www.dlib.indiana.edu/projects/sounddirections/

- **Openness** – currently, the digital objects are stored in a dark archive to avoid copyright infringement. However, the ultimate goal is to provide access to these materials, after appropriate metadata has been assigned and copyright review has taken place.
- **Business model and cost** – the business model for MDPI is non-profit. The IUB Office of the President, the Office of the Provost, and the Office of the Vice President for Research provide support for the cost of preparation, digitization, and storage of the materials. Additional support is provided by numerous campus facilities, such as UITS as well as repositories and libraries across the university.
- **Scope** – MDPI digitizes specific items that have significance and are in need of long-term preservation. In total, approximately 280,000 items from IUB and approximately 25,000 items from IU's regional campuses will be digitized.
- **Format** – MDPI is exclusively focusing on audiovisual material. There are additional files being attached to the digital objects, such as images of the original, but these are secondary to the main goal.
- **Time spent digitizing** – For extremely delicate items, MDPI digitizes items one at a time. For less delicate materials, MDPI uses parallel processing for digitization (i.e., it digitizes multiple items at the same time).

The DCC Curation Lifecycle Model

Digital curation "involves maintaining, preserving and adding value to digital research data throughout its lifecycle."[18] Pennock (2007) describes digital curation as a lifecycle process. The DCC Curation Lifecycle Model acknowledges that digital curation and data preservation are ongoing processes (Higgins, 2008). It was designed by the DCC "as a training tool to help curators understand the processes involved in successful curation, and develop curation and preservation methodologies for their organisations" (Higgins, 2008). It includes Full Lifecycle Actions (e.g., *Description and Representation Information*; *Preservation Planning*; *Community Watch and Participation*; and *Curate and Preserve*), Sequential Actions (e.g., *Conceptualise*; *Create or Receive*; *Appraise and Select*; *Ingest*; *Preservation Action*; *Store*; *Access, Use, and Reuse*; and *Transform*) and Occasional Actions (e.g., *Dispose*; *Reappraise*; and *Migrate*). For detailed descriptions of each of these actions see Higgins (2008).

The DCC's research on the Curation Lifecycle Model underscores its applicability to digital archival projects. For example, in a series of seven case studies, the DCC used the Curation Lifecycle Model as a tool to analyse the current practices of digital archives and provide recommendations for future improvements. As a high-level abstraction, the Curation Lifecycle Model served as a meaningful way to organize the actions of the archives studied. However, once these relations were made, it was found that the archives' actions often only covered a subsection of the entire model. The DCC concluded that the Curation Lifecycle Model was most effective when adapted to the specific situations of the archives (Lyon, Rusbridge, Neilson and Whyte, 2010). In two of the seven case studies, the DCC explicitly demonstrated this benefit: Case Study 1 and Case Study 5.

[18] DCC – What is Digital Curation?: http://www.dcc.ac.uk/digital-curation/what-digital-curation

In Case Study 1, the DCC applied the Curation Lifecycle Model to the curation of brain images in psychiatric research. In this case, the Curation Lifecycle Model was used as a tool to identify risks and mitigations present at each phase. This information was then used to prescribe steps to improve data policy in accordance with the Curation Lifecycle Model. Similarly, researchers from the University of Oxford Research Data Management Project and the UK Research Data Service used the Curation Lifecycle Model to identify problematic aspects of their data management practices (Martinez-Uribe, 2008; Sykes, 2008).

In Case Study 5, the DCC applied the Curation Lifecycle Model to the reuse of video data in the social studies of interactions. The DCC specifically examined the video archival practices of two Scottish research groups. The DCC drafted a recommendation for an iterative process targeted at tailoring the Curation Lifecycle Model to digital video management in the social sciences.

In summary, several archives and advisory groups in digital curation education, project planning, and curation strategy development have applied the Curation Lifecycle Model (Higgins, 2009). In these cases, the model has provided valuable insights. However, the types of projects to which the Curation Lifecycle Model has been applied have been primarily concerned with born digital or already digitized objects. To date, the Curation Lifecycle Model has not been applied to any mass digitization projects, such as MDPI. Research on applicability of the Curation Lifecycle Model to mass digitization projects could further demonstrate the validity of the model while increasing our understanding of how digital curation is performed in mass digitization projects.

Methodology

To critically examine the functions of the DCC Curation Lifecycle Model in the context of MDPI, we performed a qualitative case study analysis (Yin, 2014). The Indiana University Human Subjects Office approved this study (IRB Study #1703553601). From August 2016 through February 2017, the research team collected data concerning MDPI's history, organizational structure, facilities, staff, processes, collections, and the project's partnership with the commercial entity Memnon-Sony (Sony Europe Ltd.). Data sources included: 1) participant-observation, 2) a tour of MDPI's physical space, 3) documentation, 4) the MDPI website, 5) guest lectures, and 6) interviews. We received permission from the tour guides, guest speakers, interviewees, and other relevant MDPI employees to use any relevant information that they shared with us during class presentations, tours, emails, or conversations as data for our research project. We also promised not to report anyone's identity in any publications resulting from this research.

To study digital curation in a naturalistic environment, the second author acted as a participant-observer. She was a non-traditional participant-observer because she worked on MDPI in two different capacities for two years prior to joining our research team. Her prior experience brought a wealth of expertise, including tacit knowledge, which was useful for us in understanding MDPI. In her role as a Strategic Media Access and Resource Team (SMARTeam) Member, she beta-tested and provided feedback for the development of the Physical Object Database (POD) and MediaSCORE/MediaRIVERS systems, including helping to assemble and edit user guides for the systems. Additionally, she developed workflows for processing audiovisual materials for digitization, developed training documentation, trained other team members, helped create and return shipments of materials from Memnon and the IUB Digitization team, and diagnosed technical/physical problems and other issues for audiovisual materials. She also coordinated with IUB's Cook Music Library faculty and staff to digitize and process materials for digitization, as well as troubleshoot and address any anomalies.

The research team examined MDPI in its naturalistic environment by taking a tour of the Innovation Center, the physical location for MDPI, which is housed on the IUB campus. During the tour, we took field notes as MDPI staff described how they do their work.

Documentation for MDPI included two reports and a user guide, all of which are publicly available: 'Media Preservation Survey: A Report; Meeting the Challenge of Media Preservation: Strategies and Solutions' and the Media Research and Instructional Value Evaluation and Ranking System (MediaSCORE/MediaRIVERS) user guide. The first document contains key information about the state of digital media at IUB and justified the need for the establishment of MDPI to address preservation of those media (Casey, 2009). The second document contains information about MDPI's background, preservation planning, preservation strategies for film, access, technology infrastructure, structure and personnel, and engagement on IUB's campus (Indiana University Bloomington Media Preservation Initiative Task Force, 2011). The third document is the Media Selection: Condition, Obsolescence, and Risk Evaluation/Media Research and Instructional Value Evaluation and Ranking System (MediaSCORE/MediaRIVERS) user guide that is available on GitHub (Bohm et al., 2015). We compared all this documentation against the DCC Curation Lifecycle Model Checklists.[19]

The MDPI website contained valuable information about MDPI's history, development, and status. For example, as of April 2018, MDPI has preserved 281,882 items.[20]

Guest lectures by MDPI staff took place during Fall 2016 in a graduate-level course taught by the first author in the Department of Information and Library Science in the School of Informatics, Computing, and Engineering at IUB, Z586 – Digital Curation. The guest speakers defined their roles and responsibilities and described the initiative from their perspectives. Members of the research team took field notes during the speakers' guest lectures, which we added to our dataset.

After analysing data from the aforementioned sources, we realized that we had a few gaps in our understanding of how MDPI applied some of the concepts in the DCC Curation Lifecycle Model. Consequently, we conducted interviews with MDPI staff to better understand how the *Description and Representation Information*, *Preservation Action*, and *Store* concepts in the DCC Curation Lifecycle Model applied to MDPI.

The research team compiled all the data into one dataset in a password-protected, encrypted file folder. The research team reviewed each other's field notes, checking for consistency and accuracy. Next, the research team compared these data against descriptions of each component of the DCC Curation Lifecycle Model. Afterwards, the research team wrote reports on how MDPI addresses each component of the DCC Lifecycle Model based on the aggregated dataset. We cited the documentation we analysed for our case study throughout this methodology section, and Donaldson et al. (2018) contains our field notes.

[19] DCC Curation Lifecycle Model: http://www.dcc.ac.uk/resources/curation-lifecycle-model

[20] Media Digitization and Preservation Initiative: https://mdpi.iu.edu/index.php

Findings

We have organized our findings based on the components and subcomponents of the DCC Curation Lifecycle Model. These components and their definitions are provided in Higgins (2008). Our analysis of how MDPI addresses each component is as follows:

Data (Digital Objects or Databases)

Digital objects

MDPI focuses on preserving the content of single digital objects, specifically sound and video recordings. These objects are processed by the SMARTeam using the Physical Object Database (POD), a database specifically designed by IUB Libraries developers to describe, process, manage, inventory, and track physical objects as they go through digitization workflows. In addition to the audiovisual content, Memnon has added a workflow to capture images of 78 rpm records and their packaging (when available). This allows for the capture of descriptive metadata that otherwise may not have been recorded, such as matrix numbers of the disc recordings.

Databases

As described above, MDPI uses the POD to describe, manage, and process materials before, during, and after digitization has occurred. MDPI has identified discovery and access of their audiovisual resources as a key component of their mission, and has adopted three systems to fulfil this requirement: Avalon, HydraDAM2, and Fedora.

Avalon

The public, front-facing instance of Avalon for IUB is Media Collections Online (MCO). Avalon Media System is an open source, online program that allows institutions to manage and provide access to audiovisual materials.[21] This system is integrated with the IUB servers, and can use other resources such as IUCAT, IUB's online library catalog. By submitting a "catalog key" from IUCAT along with the file during ingest, Avalon imports bibliographic data that enables users to discover and access the audiovisual resources of MDPI.

HydraDAM/HydraDAM2

HydraDAM is a digital asset management system designed by WGBH Boston,[22] an American public broadcaster, to support libraries, archives and cultural institutions with digital preservation of audio and video files. PHYDO, formerly HydraDAM2[23], is a joint NEH-funded project between Indiana University Libraries and WGBH to extend the functionality of HydraDAM to meet the needs of more complex storage environments. It will be used as the preservation repository for MDPI content. When an item is processed by PHYDO, the system will create a YAML file to extract all metadata properties deemed to be of importance to long-term preservation activities and discoverability for content managers. This extracted metadata will be stored within the Fedora repository, while the content bitstreams will be put into the Scholarly Data Archive due to their size and the robustness of the storage. An asynchronous storage

[21] Avalon: http://avalonmediasystem.org/

[22] WGBH: http://www.wgbh.org

[23] Phydo: https://wiki.dlib.indiana.edu/display/HD2/PHYDO

gem implemented within PHYDO will allow the Fedora repository to support actions on the content bitstream, so that a user of the system can stage the file for download or for fixity checks, as needed. PHYDO also implements the preservation metadata schema PREMIS in order to log preservation actions undertaken on content within the system.

Fedora

Key features of Fedora that makes it well-suited for the needs of MDPI are: no file restriction, machine- and human-readable metadata, interoperability with other systems to increase search and discovery of resources, advanced storage options, built-in fixity checks, audit trails, versioning control, and backup/restore.

Full Lifecycle Actions

Description and representation information

MDPI captures and maintains descriptive and technical metadata for each object as it goes through the digitization process. MDPI also includes information such as which box or bin an item was placed into for shipment. Prior to the release of materials for digitization, information is captured in a single-cell spreadsheet that functions as a manifest for delivery to the digitization vendor. The metadata in the manifest is limited to what is needed for the digitization stage.

Preservation planning

In numerous publications, MDPI established a solid preservation plan prior to any digitization. For example, ‘Media Preservation Survey: A Report’ generally described the action plan that should be undertaken, such as establishing a media preservation and digitization center (Casey, 2009). ‘Meeting the Challenge of Media Preservation: Strategies and Solutions’ outlines MDPI’s approach to several of the sequential actions in the DCC Curation Lifecycle Model (Indiana University Bloomington Media Preservation Initiative Task Force, 2011).

In 2009, the IUB Media Survey Task Force investigated the media holdings at IUB. After analysing the degradation of the objects in combination with rapid obsolescence of playback machines and expert knowledge, the Task Force consulted with leading experts and specified what digitization should occur within 15 to 20 years. This time frame was considered a maximum. For instance, they determined that some digital objects required preservation action within as little as five years.

After the IUB Media Survey Task Force completed its work, the Media Preservation Initiative Task Force was created to outline a concrete preservation plan, setting forth numerous guiding principles that would influence all preservation decisions (Indiana University Bloomington Media Preservation Initiative Task Force, 2011). These guidelines ascribe to widely-accepted practices, such as considering long-term preservation, as opposed to short-term or temporary, as well as establishing partnerships on- and off-campus, assuming long-term responsibility, and prioritizing the objects to be digitized. The Task Force also stated that items should be digitized once, with the exception of film, to prevent undue stress to those items.

Furthermore, ‘Meeting the Challenge of Media Preservation: Strategies and Solutions’ established the stages media would undergo to ensure long-term preservation and access (Indiana University Bloomington Media Preservation Initiative Task Force, 2011). These include: determining ownership; selecting items for digitization; preparing for digitization and transfer; preservation transfers under the supervision of experienced technicians; quality control; assigning metadata to master files; describing files for discovery and use; storing derivative files; and providing access. MediaSCORE and MediaRIVERS (see ‘Appraise and Select’ below) assist in ranking items for digitization. Curators of the holding institution would then review the items to ensure selection is appropriate. Initially, only minimal metadata is attached, to increase the speed of the mass digitization process so that audiovisual materials that are danger of being lost forever are preserved more quickly. Later on, more complete metadata are added to the digital files. There is also an emphasis on using software to assist with metadata capture and quality control. Finally, the Task Force recommended a combination of 1:1 transfers and parallel transfers (e.g., multiple:1 transfers). This allows engineers to pay more attention to fragile items in 1:1 workflows, while other items, with fewer preservation concerns, are digitized at higher rates (e.g., 2:1, 4:1, 8:1, or even 16:1) to quickly digitize as many objects as possible.

We found that the preservation planning process for MDPI was iterative. In some instances, decisions about preservation planning in ‘Meeting the Challenge of Media Preservation: Strategies and Solutions’ needed to be modified. For example, in the report, it was recommended that IUB digitize exclusively in-house to provide digitization training for students and to establish IUB as a center for digital preservation. However, due to financial and time constraints, IUB partnered with Memnon (Sony) to expedite the process. This example is indicative of how the preservation planning concept in the DCC Curation Lifecycle Model applies in the context of MDPI; not only did MDPI staff thoroughly research the digitization process, they also partnered with Memnon to further establish thorough digitization procedures and enhance MDPI’s long-term preservation plan. It is also important to note that, when faced with obstacles to executing their preservation plan, MDPI staff carefully re-evaluated their preservation plan to decide what other actions could be taken, consulted the relevant literature on what they needed to address, and ensured that any changes they made to the initiative were calculated and methodical.

Community watch and participation

The organizational structure for MDPI includes oversight at many points from the president and provost of IUB to a layer of co-chairs and directors responsible for specific elements of the initiative. Direct management of the initiative is the responsibility of two co-chairs who hold titles as Dean of University Libraries and Vice President for Information Technologies and Chief Information Officer (CIO). A third person who is executive director for the initiative also holds the title of Associate Vice President and Deputy CIO.

The IUB MDPI Director of Technical Operations and the Sony Memnon director of US operations guide day-to-day operations. Various working groups have been convened as needed to address IT issues, library operations, development of workflows, access, etc. Even though these working groups appear to be helpful and effective, they come and go as needed and are not permanent parts of the operating structure.

Curate and preserve

As discussed previously (see ‘Preservation Planning’), ‘Media Preservation Survey: A Report’ (Casey, 2009) and ‘Meeting the Challenge of Media Preservation: Strategies and Solutions’ (Indiana University Bloomington Media Preservation Initiative Task Force, 2011) both represent actions planned to promote curation and preservation of IUB’s time-based media over their curation lifecycles. By partnering with Memnon for most of the digitization work and employing a factory-style preservation workflow, MDPI continues to work toward the efficient and careful preservation of audiovisual materials at IUB. Specifically, MDPI captures content and places it in newer digital formats before the content's original analog formats degrade and become obsolete.

Sequential Actions

Conceptualise

Based on the 15-20 year preservation window[24] defined in the preservation survey (Casey, 2009), and the goal of preserving more digital materials in less time, MDPI staff chose to capture minimal descriptive metadata including: format; the IUB entity that owns or manages the original object, or *unit*; shelf number; collection number or title; recording title or one-phrase description; digitization destination (e.g., IU or Memnon); and storage location. Post-digitization, detailed metadata are attached to files allowing researchers to reference digital files with complex content.

Corresponding metadata are attached to resources with barcodes. Specific information contained within individual resource barcodes facilitates inventory control, quality control, and efficiency as resources are processed through individual life cycle phases and are moved to different physical locations.

MDPI produces lossless compressed FFV1/MKV video preservation master files with a Matroska wrapper, 50 Mbps MPEG 2 video mezzanine files, and 24-bit and 96 kHz high-resolution audio preservation and production masters. The estimated long-term storage for resources digitized in Phase 1 is 6.5PB.

Create or receive

MDPI has technical and digital provenance metadata collection and documentation processes incorporated into its workflows to accurately describe each digital object that is created. The SMARTeam helps to prepare materials for digitization by barcoding items, organizing them, and creating descriptive and technical metadata that can later be enhanced and used for access purposes (Indiana University Bloomington Media Preservation Initiative Task Force, 2011). The initial descriptive metadata includes the item title, format, physical location, unit, etc. MDPI uses the POD as a standards-based database to contain enough metadata to allow for tracking each physical object through the digitization process in one system. This allows the SMARTeam and MDPI staff to describe materials in the digitization queue, and track batches and shipments of materials throughout the process. This initiative cultivates and maintains "the collection of a rich set of administrative metadata about digital media" to support long-term preservation (Indiana University Bloomington Media Preservation Initiative Task Force, 2011). Additional descriptive metadata used for MDPI come from MARC bibliographic records and archival description in XML using the EAD (Encoded Archival Description) schema, which is used by archival institutions for the description of finding aids and metadata. MDPI also leverages existing item-level metadata for IUB's Archives of Traditional Music (ATM) field collections, as well as metadata for collections from all non-library/archives units on campus.

MDPI creates data and digital objects from physical objects. When MDPI begins the action of *Create or Receive*, appraisal and selection has already occurred. The items are transferred from the IUB unit (e.g., department, library, or school) to MDPI. Some metadata are already in the database describing the physical object; as stated earlier,

[24] This has since been modified by the IUB MDPI Director of Technical Operations to 10-15 years.

these metadata exist in IUCAT and are extracted from its database. If metadata for any given digital object do not already exist in IUCAT, the SMARTeam creates enough descriptive and technical metadata for it within the POD to allow for identification, tracking, and digitization. The team processes items either by a generated pick list, which has some metadata already for each object, or they create a new record for the object and create new metadata. When the SMARTeam processes a pick list, they verify that all information and metadata are correct and then include any missing data. When they process an object, and create new descriptive and technical metadata, they also verify that all metadata are correct for the item record. The POD contains these descriptive, technical, and administrative metadata that are created to describe and track the location of the physical materials that are digitized.

Appraise and select

Appraisal and selection of materials to be curated and preserved is necessary, as the resources available to preserve materials should be allocated to those in most need of curation into perpetuity. The DCC provides the following checklist for the *Appraise and Select* sequential action:

1. Begin to appraise and select as early as possible.
2. Plan to keep what will support your findings.
3. Know the audience of the data (who you're keeping it for).
4. Know what you need to dispose of to comply with legal policies.
5. Make sure data complies with minimum quality assurance metrics.
6. Reappraise before long-term storage.
7. Develop policies; identify realistic workflows.
8. Appraise for current needs with a mind towards what will be useful in the future.[25]

We found that MDPI follows all but one of these guidelines (i.e., checklist item two). Regarding the first checklist item, MDPI developed MediaRIVERS and MediaSCORE software in combination with the IUB Media Survey to appraise and select before the initiative began. Later these tools were not necessary because the project received enough funding to digitize all audio and video considered appropriate for long-term preservation. Checklist item two is not entirely relevant to MDPI due to the nature of the materials preserved; MDPI preserves audiovisual recordings as opposed to scientific research datasets. MDPI staff know the audience for the material they are preserving – students, community users, scholars, researchers, faculty, staff, and university administration (checklist items three and eight). They considered copyright and access issues, and have been developing policies that will allow materials to be removed from public access if an issue arises, therefore complying with legal conventions (checklist items four and seven). As part of the digitization process, materials go through quality control (and therefore are reappraised) after digitization (checklist items five and six). Members of IUB Quality Control (QC) staff verify that digital audio files are complete, accurate, and can be stored for the long-term.[26] In

[25] DCC Curation Checklists – Checklist for appraise and select: http://www.dcc.ac.uk/sites/default/files/Select%20and%20Appraise%20Checklist.pdf

[26] This function is also carried out by the post-processing system's automated QC checks.

addition, MDPI's Access Task Force Committee, including metadata and rights subcommittees, is working to make sure that policies and realistic workflows are in place for preserving, curating, and maintaining digitized content through databases and the streaming platform Avalon (checklist item seven).

We found that MDPI staff looked at factors such as importance to research and curricula, access rights and copyright. By considering what will support research and curricula, MDPI staff adhere to appraisal for current needs as well as future research trends by selecting items that are unique and non-commercial. MDPI staff adhere to the DCC guideline regarding knowledge of legal policies by consulting with the Libraries Copyright Librarian, who works with IU Legal Counsel, to determine what can be digitized and streamed online. MDPI also digitizes commercial recordings, but keeps them unpublished in 'Dark Avalon,' with the idea that at some point it will be permissible to provide some level of access to these materials (e.g., behind firewalls of IUB that require user authentication). In addition to collaborating with legal experts to determine rights and legal compliance, MDPI sought the expertise and input from librarians and collection managers who had a better understanding of what the collections held to determine current and future use.

Developed specifically for MDPI and IUB in partnership with AVPreserve, MediaRIVERS ranks the importance of materials based on their research and instructional value, and it provides units and selectors of objects for digitization based on a numerical score of importance for selection and triage (Bohm et al., 2015). Values are assigned to *asset groups*, or materials grouped together based on similar characteristics (e.g., not necessarily a collection in the archival sense, but rather things of the same nature or subject, creating entity, etc.), regarding subject of interest, content quality, rareness, documentation, technical quality, generation, and intellectual property (Bohm et al., 2015). By having collection managers assign values to objects based on research value within asset groups, whole collections of materials can be appraised or selected at once, speeding up the process.

In addition to selecting and triaging based on subject characteristics and research importance, there is also the factor of obsolescence when dealing with audiovisual materials. To address this issue, MDPI partnered with AVPreserve to develop software that could rank and score asset groups of materials based on risk factors related to obsolescence and degradation, or "degralescence" (Casey, 2009). By including metadata for groups of materials about tape stock and duration as well as preservation problems such as sticky shed or vinegar syndrome for open reel tapes, delaminating lacquer discs, or damaged 78 rpm discs, the software returns a score for groups' risk. Table 2 describes the range.

Table 2. Range of Risk of Obsolescence for Audiovisual Materials (Bohm et al., 2015).

Score (Range from 0-5)	Risk	Description of Risk
4.5-5	Extreme danger	Digitize immediately (Action statement: "I'm begging my director for money tomorrow").
3.5-4.49	Danger ahead	Digitize near-term (Action statement: "I'm writing a grant in the next year").
2.5-3.49	Caution	Moderate risk—digitize soon (Action statement: "I'm actively planning for digitization").
1.5-2.49	Lower priority	Digitize medium term.
0.6-1.49	Low priority	Digitize within 10-15 years.
0-0.5	Safe	Not in need of digitization.

Once materials are appraised and selected based on their scholarly importance, and triaged based on their physical condition, they are categorized into two separate workflow paths: items at low risk go through a cheaper, faster preservation workflow, and items at high risk or of exceptional value go through a more expensive, careful workflow.

Ingest

After digitization, master copies of the files are stored as preservation copies, while access files are made available to users via the online access system MCO. With the files from Memnon, the items are transferred to IUB with an MD5 checksum. These data objects, alongside those created by the IUB preservation team, are then ingested into the SDA.

Preservation action

Since the preservation planning phase, MDPI staff anticipated the need for constant quality control. Throughout the digitization process, staff perform quality assurance and quality control actions, ensuring that all digital objects and metadata records are valid. Staff, including engineers, perform quality control actions. They also perform quality control when derivative files are created, an issue they addressed by partnering with Memnon, the third-party mass-digitization partner (Indiana University Bloomington Media Preservation Initiative Task Force, 2011). MDPI ensured that the necessary infrastructure and support was in place for the long-term preservation of these digital files and their associated metadata files. SDA storage at both IUB at IUPUI is used as a part of this long-term preservation plan.

According to staff, MDPI relies mainly on open source file formats that have a long-term future, and IUB Libraries monitor current trends so that they can quickly migrate file formats if needed. At MDPI, audio files are currently being created in the Broadcast WAVE (BWF) non-proprietary format, which is a 24-bit, 96 kiloHertz (kHz) long standing, high-quality format currently preferred among audio preservationists (Behl, 2015). In comparison with this widely accepted format, video files at MDPI are being stored in an FFV1 codec in a Matroska wrapper. The FFV1 is a mathematically lossless program for encoding digital streams, currently being developed and used for archiving by repositories around the world, while Matroska is an audiovisual container and format, with numerous flexible features, designed to store descriptive metadata in the file, thus mitigating the risk of metadata software failure (Murray, Rice and Blewer, 2015). Even though this format is not as established as others, MDPI staff are currently involved in the maintenance and development of FFV1 and Matroska to ensure that the current and future needs of MDPI continue to be served (Media Area, 2016).

Store

After the digital objects have been ingested, the data are stored in a viable location and file format that will persist for at least the mid-future. MDPI writes their digitized files to two data storage centers: the Bloomington Data Center and the Indianapolis Data Center. These centers are separated by approximately 50 miles and are built to withstand F5 tornadoes. The second digital copy reduces the risk of human error and general file corruption while the physical separation mitigates the potential threat of a natural disaster critically affecting both data centers. However, as explained by MDPI staff, a third copy, preferably housed in a location outside of the state of Indiana, would further reduce these risks.

Access, use, and reuse

IUB Libraries has created a Task Force and subcommittees to address the issues of metadata and rights as they pertain to access to MDPI digitized material. As materials are digitized and pass quality control, they are stored on the SDA, with compressed streaming derivatives stored on disk. The front-facing platform for collection managers to preview these derivatives is 'Dark Avalon.' The current idea is to have a function that will request materials to be moved from 'Dark Avalon' into MCO, making the materials publicly accessible. To facilitate discoverability and use of the materials, collection managers for each collection unit use MCO to create metadata, context, description, and structure for the materials.

Through MCO, users can listen to audio files and view video files that have been digitized by MDPI. Users can jump between sections of a recording based on the structural metadata provided by a collection manager, view contextual information either imported through IUB's OPAC or manually entered by a collection manager, and can add materials to personal playlists associated with their accounts. The functionality of MCO allows for materials to have different levels of access. It can either be open and available to all, available only to IU affiliated individuals, collection staff only, or a specified user and/or group. There is also the option to leave materials "unpublished" so that only the collection manager can view the record for the object. The unit that manages the content can also apply usage restrictions into the catalog record/bibliographic data for the object, although it is not required.

Collection managers can determine how long specific users have access to materials via special permissions settings for objects and collections. Collection managers can choose to provide information about restrictions to users or restrict access to materials without explanation. Collection managers also provide context for audio and video files by adding bibliographic information and structural metadata.

Transform

MDPI currently creates several derivative files during the preservation process including when files are required for access and use. When MDPI digitizes an original audiovisual object, the content within the object is saved in three separate files: a preservation master file, a production file (called a mezzanine file for video), and an access file. All three files are stored at the SDA at IUB and IUPUI. Another set of three derivative files is automatically created by scripts for ingest into Dark Avalon. Currently, collection managers upload material into MCO manually. All derivatives are made from the production master files while the preservation file remains untouched and acts as a "pure" copy of the original audiovisual object. In the future, MDPI staff plans to develop their systems to enable movement of files en masse into MCO.

Occasional Actions

Dispose

MDPI is not transferring materials to other repositories or disposing of any media materials at this point.

Reappraise

It is common for MDPI staff to reappraise resources that are found in poor physical condition that will be rejected by Memnon. Transferring the resources to IUB's workflow prior to rejection saves time that can be used for other resources that will not be rejected.

Migrate

Migrating data to different formats is a concern MDPI has attempted to pre-emptively address by their choice of file formats. With regards to audio files, the BWF file format is non-proprietary, ubiquitous and widely accepted, suggesting the format will remain viable for an extended period, and will require minimal preservation actions. Moreover, MDPI leaves these files uncompressed, which may mitigate the effects file corruption or 'bit rot.'

As for the FFV1/Matroska format, although currently not supported by a specific institution, this format is currently being standardized. Moreover, as a non-proprietary format with a strong community of developers, FFV1/Matroska should require minimal preservation actions (e.g., migration).

Discussion

This study adds to the body of research literature aimed at performing empirical tests of the DCC Curation Lifecycle Model. Specifically, it tests the ability of the model to apply to different domains as its authors suggest. In contrast to prior studies where only parts of the model were the primary focus (e.g., Whyte, 2008; Whyte, 2009), we applied the entire model in our study. We found that the model is a useful framework for understanding how digital curation is performed in a mass digitization project, in this case, MDPI. We gathered empirical support for each of the components of the DCC Curation Lifecycle Model. These findings underscore the validity of the model.

To a certain extent, the DCC Curation Lifecycle Model is intentionally generic so that it can apply across different disciplines and digital data. Therefore, case studies are important for elucidating the ways in which the concepts in the model can apply in different contexts. We argue that our study provides key insights regarding how concepts in the DCC Curation Lifecycle Model can apply to mass digitization projects for audiovisual materials. In particular, mass digitization projects that are just beginning or are at the planning stages could benefit from reviewing the findings of this study as a way of understanding what they should consider as they move forward and prepare for digital curation. For example, staff dedicated to a new mass digitization project for audiovisual materials may learn key insights from MDPI based on their approach to the *Access, Use, and Reuse* concept in the model, particularly if they too are responsible for materials that have varying levels of access restrictions. For instance, staff for the new mass digitization project may want to adopt MDPI's approach of offering different levels of access to digitized materials based on whether copyright restrictions apply. Or staff may want to use the MediaSCORE and MediaRIVERS systems to help them apply the *Appraise and Select* concept in the model by rank ordering the importance of materials and digitizing them based on their priority and risk.

The DCC Curation Lifecycle Model's relevance to new digitization efforts does not decrease its applicability to initiatives that have already begun or have ended. This is because our findings suggest that certain actions that are necessary for digital curation are iterative (e.g., *Preservation Planning*). We found the digital curation process to be iterative, even within individual components and subcomponents of the model. We argue that institutions can use the DCC Curation Lifecycle Model to refine procedures and re-evaluate practices periodically to provide the best preservation for their digital objects. The cyclical nature of the DCC Curation Lifecycle Model is a useful reminder that, while digitization initiatives may end when the last item has been converted, the curation process for those materials is never-ending.

This study also advances the concept of mass digitization by providing a more nuanced definition for the term. In addition to defining mass digitization as "conversion of materials on an industrial scale" as Coyle (2006) suggests, we recommend consideration of six characteristics when determining whether a digitization project is a mass digitization project: 1) Aggregation and production, 2) Openness, 3) Business model and cost, 4) Scope, 5) Format, and 6) Time spent digitizing. More research on and comparison of mass digitization projects could validate our proposed definition of mass digitization, or help to refine it.

Conclusion

Mass digitization projects require a significant amount of effort and millions of dollars in resources. Without proper digital curation, mass digitization is all for naught. That is, without proper data (digital objects or databases); description and representation information; preservation planning; community watch and participation; curation and preservation; conceptualization; creation or receipt; appraisal and selection; ingest; preservation action; storage; access, use, and reuse; transformation; disposal; reappraisal; and migration, mass digitization projects could waste valuable resources. We need ways to measure the effectiveness of mass digitization projects' digital curation efforts. One way to do this is to compare their activities to all the actions specified in the DCC Curation Lifecycle Model. Future studies could compare the results of this study to the results of studies that apply the model to different types of mass digitization projects to better understand how the model can apply to mass digitization projects.

Beyond these types of studies, we need studies focused on transforming the DCC Curation Lifecycle Model into a measurement model. This would allow us to more meaningfully compare any given organization or project's adherence to or compliance with each action that the model specifies. To facilitate this type of research, we need formalized mechanisms for turning the DCC Curation Lifecycle Model into a measurement model. For example, researchers could develop survey questions designed to capture information about each of the Full Lifecycle Actions, Sequential Actions, and

Occasional Actions in the model. The surveys could be administered to staff who are responsible for digital curation. Then survey data could be analysed to more meaningfully compare digital curation effort across organizations and projects. This would bring us closer to knowing if, when a staff member at an organization says “yes, we perform digital curation activities effectively at our organization,” this means the same thing as when someone else at another institution says it.

References

Behl, H. (2015). Audio formats: Characteristics and deterioration. In S. Brylawski, M. Lerman, R. Pike, K. Smith (eds.) ARSE Guide to Audio Preservation (pp. 14-36). Washington, DC: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub164/pub164.pdf

Bohm, A., Casey, M., Feaster, P., Lyon, J., Moore, A., Reynolds, C., …, Shelby, J. (2015). MediaSCORE and MediaRIVERS media preservation prioritization software user guide. Retrieved from http://www.avpreserve.com/wp-content/uploads/2015/03/IU_mscore_mrivers_guide.pdf

Casey, M. (2009). Media preservation survey: A report. Indiana University Bloomington. Retrieved from https://mdpi.iu.edu/doc/survey.pdf

Coyle, K. (2006). Mass digitization of books. *The Journal of Academic Librarianship 32*(6): 641-645.

Donaldson, D., McClanahan, A., Christiansen, L., Bell, L., Narlock, M., Martin, S., & Suby, H. (2018). *Field notes for media digitization and preservation initiative: A case study* [Data set]. Bloomington, IN: IU ScholarWorks. Retrieved from http://hdl.handle.net/2022/22101

Gooding, P. (2013). Mass digitization and the garbage dump: The conflicting needs of quantitative and qualitative methods. *Literary and Linguistic Computing*, *28*(3), 425-431.

Guidi, G., Barsanti, S.G., Micoli, L.L., & Russo, M. (2015). Massive 3D digitization of museum contents. In L. Toniolo, M. Boriani, & G. Guidi (Eds.), *Built Heritage: Monitoring Conservation Management* (pp. 335-346). Springer International Publishing.

Indiana University Bloomington Media Preservation Initiative Task Force. (2011). Meeting the challenge of media preservation: Strategies and solutions. Indiana University Bloomington. Retrieved from https://mdpi.iu.edu/doc/strategies-solutions.pdf

Hahn, T.B. (2011). Mass digitization. *Library Resources & Technical Services*, *52*(1), 18-26.

Heyman, S. (2015, October 28). Google books: A complex and controversial experiment. *The New York Times*, Retrieved from https://www.nytimes.com/2015/10/29/arts/international/google-books-a-complex-and-controversial-experiment.html?_r=2

Higgins, S. (2009, May). Applying the DCC curation lifecycle model. Paper presented at the 35th Conference of the International Association for Social Science Information Services and Technology (IASSIST 2009), Tampere, Finland.

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation 2*(1), 134-140. Retrieved from doi:10.2218/ijdc.v3i1.48

Jihai, Z. (2008, October). Mass digitization of the collections of the academic libraries in China. Paper presented at Pacific Rim Research Libraries Alliance Annual Meeting (PRRLA'2008), Singapore. Retrieved from http://pr-rla.org/2008/10/mass-digitization-of-the-collections-of-the-academic-libraries-in-china/

Lyon, L., Rusbridge, C., Neilson, C., & Whyte, A. (2010). Disciplinary approaches to sharing, curation, reuse and preservation: Final report. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

Martinez-Uribe, L. (2008). Using the data audit framework: An Oxford case study. Retrieved from http://www.disc-uk.org/docs/DAF-Oxford.pdf

Media Area. (2016). No time to wait: Standardizing FFV1 and matroska for preservation. Retrieved from https://mediaarea.net/MediaConch/2016/07/26/No-Time-To-Wait-Preservation-FFV1-Matroska-Symposium/

Murray, K., Rice, D., & Blewer, A. (2015). Improving technical options for audiovisual collections through the PREFORMA project [Interview transcript]. Library of Congress: The Signal. Retrieved from http://blogs.loc.gov/thesignal/2015/09/improving-technical-options-for-audiovisual-collections-through-the-preforma-project/

National Library of Australia. (2017). Digitisation and digital activities at the National Library of Australia. Conference of Directors of National Libraries in Asia and Oceania Newsletter, 90. Retrieved from http://www.ndl.go.jp/en/cdnlao/newsletter/090/901.html

Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. *Library and Archives Journal*, 1. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf

Schmitz, D. (2008). *The seamless cyberinfrastructure: The challenges of studying users of mass digitization and institutional repositories*. Digital Library Federation, Council on Library and Information Resources. Retrieved from https://www.clir.org/wp-content/uploads/sites/6/schmitz.pdf

Sykes, J. (2008). Managing the UK's research data: Towards a UK research data service. *New Review of Information Networking*, *14*(1), 21-36.

BizResearch. (2007). The million book project – 1.5 million scanned! Retrieved from https://web.archive.org/web/20080614093014/http://lbslibrary.typepad.com/bizresearch/2007/11/the-million-boo.html

Vaidhyanathan, S. (2007). The googlization of everything and the future of copyright. *UC Davis Law Review*, *40*(3), 1207-1231.

Whyte, A. (2008), Curating brain images in a psychiatric research group: Infrastructure and preservation issues - SCARP case study 1. Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/scarp

Whyte, A. (2009), Roles and reusability of video data in social studies of interaction - SCARP case study 5. Digital Curation Centre. Retrieved from http://www.dcc.ac.uk/scarp

Womack, D. (2003). Who owns history? *Cabinet*. Retrieved from http://www.cabinetmagazine.org/issues/10/womack.php

Yin, R.K. (2014). *Case study research: Design and methods.* Fifth edition. Los Angeles: SAGE.