
Large scale digitization initiatives at archival institutions utilize different descriptive methodologies based on the format of material being digitized, the availability of resources, and the descriptive philosophy of the archivists leading each project. This paper gives the results of a study involving in-depth interviews with eight archivists leading and participating in large scale digitization initiatives for manuscript materials. The findings provide information on the current state of description utilized in the field of large-scale manuscript digitization, including the archivists’ perceptions of the effect of their project’s description on resource discoverability. The findings of this study can be used by archivists developing large scale digitization projects to determine the descriptive methodology that would be most effective for their project and user community.

Headings:

Special Collections

Digitization Projects

Archival Materials – Digitization
THE DESCRIPTIVE SWEET SPOT: ARCHIVISTS’ PERCEPTIONS OF
DESCRIPTIVE METHODOLOGIES FOR LARGE SCALE DIGITIZATION
INITIATIVES

by
Helen H. Thomas

A Master’s paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina
April 2014

Approved by

_______________________________________
Dr. Helen Tibbo
# Table of Contents

THE DESCRIPTIVE SWEET SPOT: ARCHIVISTS’ PERCEPTIONS OF DESCRIPTIVE METHODOLOGIES FOR LARGE SCALE DIGITIZATION

- Introduction .............................................................................................................. 2
- Literature Review .................................................................................................. 3
- Study ....................................................................................................................... 8
  1.1 Methodology ...................................................................................................... 8
  1.2 Limitations ......................................................................................................... 11
  1.3 Participants ....................................................................................................... 11
- Findings .................................................................................................................... 11
  1.4 Constraints of Existing Technological and Organizational Structures .......... 12
  1.5 Supplemental Metadata: User Contributed Content and Optical Character Recognition 16
  1.6 Edits to Original Finding Aids ......................................................................... 17
  1.7 Efficacy of Utilized Descriptive Methodologies .............................................. 20
- Discussion ................................................................................................................ 22
- Appendix 1: Interview Questions Sent to Study Respondents ................................ 25
- Bibliography ........................................................................................................... 28
  1.8 Interviews ......................................................................................................... 28
  1.9 Secondary Sources ............................................................................................ 29
Introduction

In recent years, archivists have increasingly endeavored to place manuscript materials online through large scale digitization initiatives (For examples of large scale digitization projects for manuscript materials, see studies described in Aikens, 2009; Boudreau, 2011; DeRidder et. al., 2012; Erway, 2011; and Sutton, 2012). These projects utilize different descriptive methodologies based on the format of material being digitized, the availability of resources, and the descriptive philosophy of the archivists leading each project. This paper describes the results of a study through in-depth interviews with eight archivists leading and participating in large scale digitization initiatives for manuscript materials. The findings provide information on the current state of description utilized in the field of large-scale manuscript digitization in the context of larger archival trends. These findings include archivists’ perceptions on the constraints of existing technological and organizational structures, the inclusion of metadata not assigned by archivists, the editing of original finding aids as part of the digitization process, and the overall efficacy of their chosen descriptive methodologies in promoting discoverability. The findings of this study can be used by archivists developing large scale digitization projects to determine the descriptive methodology that would be most effective for their project and user community.
Literature Review

In their 2005 article, “More Product, Less Process: Revamping Traditional Archival Processing,” Greene and Meissner introduced the idea of minimal processing as a way of reducing the backlog of unprocessed archival materials present in many institutions. They advocated for a “golden minimum,” to “maximize the accessibility of collection materials to users” through minimal description that enables discoverability (Greene and Meissner, 2005, p. 240). The article states that the specific collection should be taken into account, rather than applying a one-size-fits-all approach within an institution, and that describing at the series or file level rather than the item level is sufficient for users to conduct their research. Amid this discussion of the description of analog materials, Greene and Meissner take a moment to consider description and the digitization of manuscript materials. They conclude that “more detailed description […] greatly facilitates the selection of material to be digitized”, and that “[r]etrospectively, the decision to digitize all or part of a collection by definition makes the collection a candidate for improved analog processing” (p. 248-249). Greene returns to this idea in a later paper, noting that MPLP does not hinder the ability to select documents for digitization because “it is simply not that difficult to find items if the description of series or files is done well” and that digitization programs need not be based on individual items (Greene, 2010, p. 193). Drawing an example from his work with visual materials at the Ford Foundation, Greene notes that

“[…] it is not necessary to have cataloged every photograph in a vertical file or record group to find one good one of Henry Ford with Walt Disney. All that’s necessary is a file unit labeled Ford and Disney, or even a series labeled Ford with VIPs; while the latter will take some minutes to discover the “perfect” image, it is far less time than would have been required to process every item.”
These assertions are echoed in Sutton’s case study of the digitization of the University of the Pacific Library’s John Muir Papers. Sutton states that elements of MPLP were successfully utilized in this project, particularly “adopting rapid, minimalist processes when possible and intensive, detailed processes when merited” and acknowledging “that different components of a single collection may necessitate different levels of arrangement, description, or digitization” (Sutton, 2012, p.54).

In their 2010 OCLC report on special collections and archives, Dooley and Luce report findings on the state of digitization projects in the United States and Canada, and compare the scale of projects at different archival institutions. They found that while ninety-seven percent of archives report completion or involvement in at least one digitization project, twenty two percent of archives could undertake digitization projects only with special funding, “suggesting that these libraries have not prioritized digitization of primary sources as an integral element of their programs and services” (Dooley and Luce, 2010, p.54). In addition, they contrast large scale digitization from other digitization projects. Dooley and Luce define large scale digitization as “a systematic effort to consider complete collections - rather than being selective at the item level, as has been the norm for many projects - and using production methods that are as streamlined as possible while also accounting for the needs of special materials” (2010, p.55) and distinguish the term “large scale” from “mass digitization,” which has previously been applied to projects concerning the digitization of books. They state that thirty-eight percent of respondents reported completion of large scale digitization projects, but found “the quantities of material digitized and/or production levels achieved generally were not impressive or scalable” and conclude that a “better overall
understanding of the nature and scope of large-scale digitization of special collections would be valuable” (2010, p.55). This paper will endeavor to contribute to this understanding through investigation of the descriptive practices of these large-scale digitization initiatives.

Referencing trends in the area of mass digitization of books, Conway states that “[w]e are at the end of the era of ‘boutique’ digital scanning projects for which the principal goal is experimentation with new technologies and extraordinary attention to the unique properties of each artifact” (Conway, 2010, p.76). The evidence above would suggest that the same trend may be apparent in the digitization of archival materials. Rather than simply demonstrating collection highlights, digitization appears to be increasingly utilized as a tool for broadening both discovery of and access to their collections. Therefore, it would follow that the description of digital materials would support this function. Reiter argues that in cases of large scale digitization, archivists can best direct their efforts at description by reconsidering “what constitutes essential elements of metadata and how they facilitate discovery and access to materials,” focusing metadata creation rather than applying any and all terms that could potentially be useful (Reiter, 2010, p.16). She also notes the potential value of user-contributed online content (such as tagging and commenting) provided that archival institutions have the technological infrastructure to maintain such a resource (2010, p.16). In 2011, an OCLC survey of capture techniques for high throughput (meaning rapidly digitized) non-book digitization initiatives noted several approaches to initial description and metadata assignment, all of which tend toward minimal description. For example, the University of Minnesota’s Green Revolution Scanning Project noted that scanning staff “are not
creating metadata for the scans, beyond the file names created as part of the process. They are using the folder titles (and in some cases box titles) as the metadata so the existing finding aids link directly to the scanned images” (Erway, 2011, p.13). At the University of North Carolina at Chapel Hill, “metadata is extracted from the finding aid to form folder-level metadata to describe all the scans for that folder. The finding aid provides description and enables discovery and links to the images” (p.16). A pilot digitization initiative at the Mudd Manuscript Library at Princeton University also relies on folder-level finding aid metadata for image description, though not without some hesitation and debate:

“There’s been some debate within the University Library about the quality of the scans and about the reliance on description in finding aids (typically at the folder level) rather than item-level description. While some wish for detailed structural and descriptive metadata to go with the digitized documents, and a more selective approach to digitization, the content is made accessible in accordance with archival standards and practices. Santamaria views the work as an iterative process; as more is learned about how the content is used, decisions can be made about where additional metadata or higher quality imaging might be beneficial” (Erway, 2011, p.18).

Erway’s survey demonstrates that several repositories are utilizing existing description (in the form of file names and folder labels) to describe digitized materials, and that multiple initiatives find this to be sufficient. However, it also introduces the idea that some archivists may prefer a different, less minimal approach to description, as well as the idea that the work being done now will inform future projects (the “iterative process”). It is these perceptions and reflections that I hope to capture in more depth through the research conducted for this paper.

As a corollary to minimizing description created or assigned by archivists, some archives have explored the use of supplemental metadata. This can take the form of user
contributed content or automated transcription through optical character recognition. In theory, both have the potential to increase available metadata and promote discoverability of digitized materials. In the context of a minimally described digital collection, Evans argues the potential advantages of crowdsourcing descriptive efforts as follows:

“Those who object to this minimum metadata model should know that creating and publishing digital images with minimum metadata does more than just make records available for research. It also places these images before thousands of potential volunteers who will use new tools for online metadata collection. It is not just minimum metadata; it is extensible metadata. The data these volunteers collect may include any combination of comments, controlled-or free-text indexing terms, abstracts, or full-text transcriptions. Each archives' institutional policies determine the range of choices, as will the methods for recruiting and managing volunteers” (Evans, 2007, p. 395).

However, the counterargument to this proposal lists inaccurate or offensive user-created content as potential outcomes. In focus groups convened to assess user needs prior to the launch of the Southern Historical Collection’s digitization program, many students and scholars expressed interest in social networking website features, but exhibited “concern about potential problems with anonymously authored content (such as false entries in Wikipedia) and the difficulties in developing an efficient and economical way to monitor, vet, or edit user-generated text or tags” (Southern Historical Collection, 2009, p. 32-33). Weighing potential costs and benefits, the Digital SHC project elected not to include web 2.0 features in their initial website design. Another potential added-value descriptive technique is the application of optical character recognition to textual materials. In 1987, Marie Allen extolled the virtues and possibilities of optical character recognition software for potentially providing searchable transcriptions of archival materials, stating that some programs were even successfully reading cursive handwriting
Currently, Miller claims, “[m]ost OCR software can achieve accuracy rates of 98 to 99.9 percent” (Miller, 2013, p. 531).

In this masters paper, I intend to investigate the current practices and perceptions of archivists working with large scale digitization initiatives, working from Dooley and Luce’s definition of large scale digitization (as noted above). The goal of my research is to determine how descriptive methodologies for digitization projects may have evolved in recent years, such as incorporating elements of minimal description and processing (such as Green and Meissner’s “more product less process”, 2005, p. 240) into their approaches, and to contribute useful research to the archival field focused on the descriptive techniques of manuscript digitization initiatives. I hope to discover more about how the methodology behind the description of digital objects for online publishing may support institutional goals as it relates to potentially expediting digital access to and increasing discoverability of archival manuscript materials.

**Study**

1.1 **Methodology**

This study involved in-depth interviews with archivists at eight institutions, each of which had conducted or is currently conducting a large scale digitization initiative involving manuscript materials. Seven responses came from single archivists, while one institution chose to have two of its employees contribute, returning a collective response. Six interviews were conducted over the phone, and two were returned as written responses to my list of interview questions.

To recruit potential interviewees, I consulted both professional archival literature on the subject of large scale digitization and lists of Institute of Museum and Library
Science grant awardees, looking for descriptions of projects including large scale digitization of manuscript materials. My qualification for this study of what constitutes a large scale digitization project was the digitization of multiple complete collections. I chose not to include projects, of which there were several, that digitized only one complete collection. However large that collection may be, a single collection digitization project would not illustrate the descriptive complications that digitizing multiple collections could pose (such as different styles of finding aids or different subject matter). I chose instead to interview archivists whose techniques had been necessarily refined over the course of the digitization of multiple collections, believing this to be the most valuable and potentially generalizable information for archivists wishing to begin a digitization initiative or program. I also decided to further narrow my approach to the large scale digitization of manuscript materials, excluding collections of published materials (such as projects to digitize books and newspapers) and collections composed entirely of photographic materials. Each of these types of material have a unique set of affordances, attributes, and descriptive standards, and their inclusion would further diversify the responses I would have received. Again, in the interest of making my results generalizable, I chose to limit my focus to one format. I chose manuscript materials due to my familiarity with them and because I believed that they represented a unique set of challenges to descriptive efforts that I wished to explore.

After identifying projects that met the requirements for my study, I used the professional literature, institutional websites (project descriptions and staff contact lists), and referrals from other archivists, to identify archivists who had participated in these projects as a principal investigator, project manager, or other leadership role qualified to
speak about both the descriptive practices utilized and the decisions leading to the implementation of those methodologies. The exception to this was the collaborative written response, to which archivists in a range of project roles contributed. Referrals from other archivists included both snowball sampling (referrals in answer to an interview question) as well as one from an archivist contacted according to previously stated recruitment techniques who referred me to another project member more qualified to answer my questions (the person who had taken over their leadership position when they left). All eight respondents agreed to be part of the study when contacted over email.

The same set of twenty-six interview questions (See Appendix 1: Interview Questions Sent to Study Respondents) was emailed to each of the eight respondents, who then had the option of responding in the form of a phone interview or returning their answers to the questions in written form via email. Respondents 2, 3, 4, 5, 6, and 7 were interviewed by phone, and respondents 1 and 8 responded via email. The phone interviews followed a semi-structured interview protocol and lasted between twenty and eighty-five minutes, with an average of forty-seven minutes. Interview questions were sent to respondents prior to the interview, and phone interviews were recorded and later transcribed. The interviewees who chose to respond in writing were sent, via email, a copy of the same interview questions given to the respondents interviewed by phone. Written responses to the interview questions were returned via email. Important themes and concepts were identified as part of the data analysis. These themes include the impact of working within organizational and technological constraints, supplemental metadata including user-contributed content and transcription via optical character recognition,
digitization resulting in the editing of existing finding aids, and archivists’ satisfaction
with the amount of description their project is able to provide.

1.2 Limitations

While the eight archivists who comprise this study represent a significant portion
of all the large scale manuscript digitization projects which have taken place or are
currently taking place in the United States, their small population may not be considered
generalizable. Though the number of respondents is small, the experience they report on
represents much of the recent history of the developing field of manuscript digitization
initiatives, and can thus contribute to an understanding of common practices and trends in
this work.

1.3 Participants

The archivists interviewed as part of this study had participated in these projects
in a principal investigator, project manager, or other leadership role, and were thus
qualified to speak about both their project’s descriptive practices and the decisions
leading to the implementation of that methodology.

Findings

The findings of this study demonstrate current trends and common practices in
descriptive methodologies for large scale digitization initiatives, including the perception
of the archivists involved as to the effectiveness of those practices.
1.4 Constraints of Existing Technological and Organizational Structures

Each respondent touched in some way on how existing structures, both technological and organizational, impacted their description of resources. Of the eight institutions represented in my interviews, four have both archival and museum components. The difference of approaches between archivists and museum staff was a common theme in the responses I received, with museum staff tending toward a more detailed approach to description than archivists. This was reflected in both philosophy of description and the practical application of that philosophy through digital asset management software. As far as the technological constraints of their content management system, the majority of respondents have utilized or are currently using commercially available content management systems, while two respondents are utilizing systems created by and for their institution. One respondent noted that their decision to create a content management system in-house has allowed them more freedom and ease of description for digitized materials, since they were able to dictate the capabilities of the system.

“We were able, I think, to manage that transfer of the metadata from one system to another probably easier because we could control the actual ingest process and can tweak it if needed. […] We didn’t have to necessarily explore options of what was possible and not possible had we been using either an off-the-shelf or vendor-supplied or even an open source system we had installed” (Respondent 6, personal communication, January 23, 2014).

Another respondent described their experience making their archive’s homegrown content management system comply with the requirements of their parent institution’s digital asset management system, in order to make their digitization initiative’s archival content accessible through the institution’s central search feature.
“It's not great for archival hierarchical description, but we’ve been really lucky to have a strong team of people both on our end and on the [parent institution’s] central DAMS and to develop a way to structure the digital asset management system to shoehorn basically our digital assets to be able to understand and we ingest just enough of the structural metadata for it to be understood in context” (Respondent 5, personal communication, January 23, 2014).

One institution reported that they have been satisfied with descriptive capabilities of a commercially developed content management system not intended for libraries after being able to customize its application.

“I think because we had developer assistance, we were able to really decide what kind of metadata we wanted, and we were able to make it different for whatever object type we had, so different metadata for photographs versus textual folders versus audio recordings. [...] We could do with it what we wanted as long as we had help. We don’t have that help now so we’re a little limited in what we can do in terms of capturing and exporting more metadata, but the system itself can do a lot, so I think that it encouraged flexibility” (Respondent 7, personal communication, January 27, 2014).

Another respondent stated that the off-the-shelf content management system used in their institution had not ultimately affected the description component of the project because “we didn't build our program out of the platform if you will, [...] it did not dictate what we decided to do in terms of description” (Respondent 4, personal communication, January 22, 2014).

Regarding constraints that their content management system’s museum-centric design might have on description, one respondent used their CMS’s controlled vocabulary to illustrate a descriptive challenge.

“[The CMS] is way better on this for the objects side, like, there must be twenty kinds of shovels, like the way you can identify a shovel, but if it’s a photographic print there’s only one way you can identify it, so it’s a little weak on descriptions on the archives side, because it’s more about museum objects” (Respondent 3, personal communication, January 22, 2014).
The same respondent concludes by stating that “except maybe for some usability quirky things it’s not limiting us from including any description or content that we wish we could include, it allows us to put up everything that we want” (Respondent 3, personal communication, January 22, 2014). Another respondent from a museum archive stated that their content management system “has created quite a few issues for us because it is not made to be used for libraries and as a result, there are several issues with the process of MARC coding, the management of authority tables, and how that information is formatted properly to create a MARC record to our standards.” These problems complicate their process of submitting records to WorldCat, in their effort to make their description more widely available (Respondents 1a and 1b, personal communication, January 14, 2014). Another respondent said of their content management system, “It certainly allows for more description, it also has affected the format […] because we’re trying to catalog both 3D and 2D archival materials, archival collections in the same system, so we’re making essentially compromises on certain fields in order to commonize those fields between two descriptive practices. […] Because we are committed to this single collection management system, ultimately this actually should increase the amount of archival description not only available in the collection management system but that is pushed out online” (Respondent 2, personal communication, January 17, 2014).

This same respondent stated that their institution was in the process of moving from item level to folder level description for their digitization project, the former being more closely aligned with museum practice and the latter with contemporary archival practice, saying that, “the reason we’ve tried to incorporate some of the aggregate description is to increase the volume of items that we’re digitizing” (Respondent 2, personal communication, January 17, 2014).

Only two of the institutions (both museums) interviewed as part of this study stated that they were including item level description for all archival materials as part of a
large-scale digitization initiative, citing both organizational and technological reasons for doing so. For the first of these institutions, the digitization project is the first at their institution to address manuscript materials, following the introduction of online exhibits for museum artifacts. Of the institution’s descriptive philosophy, the respondent stated that “[s]ince the images would have little research value, or context, if posted online without a record describing that specific archival item, we decided that cataloging was going to have to be mandatory. Also, cataloging at various levels would also be required to preserve the organizational hierarchies of the collection as well.” This institution also reported that its new content management system “has allowed the archives to describe collections in a much more hierarchical fashion (Collection, Series, File, Item)” (Respondent 1a, personal communication, February 18, 2014). The other respondent using item level metadata described the CMS already in place for museum objects as follows:

“[…] very inexpensive, it’s versatile, it’s kind of a Jack of all trades, and it allows you to work with archival materials and library materials and objects, and it’s just easier for a smaller institution like us not to have multiple databases that are perfectly tailored to each kind of collection we’re dealing with, so we just have the one” (Respondent 3, personal communication, January 22, 2014).

Regarding their decision to use item level metadata for archival materials, the same respondent says,

“[The records] are basically all to the item level and that also goes back to the museum side of things. The museum people are just used to doing everything on the item level because it’s mostly about physical objects and they’ve just continued that through, wanting it that way for the archival collections, so everything is item” (Respondent 3, personal communication, January 22, 2014).
1.5 Supplemental Metadata: User Contributed Content and Optical Character Recognition

Of the eight archivists interviewed, four stated that no user contributed content is currently being used as a supplement to archival description (Respondents 1, 5, 6, and 7). Of these four, two stated that user contributed content was not currently supported by their content management systems (Respondents 5 and 6), one has a transcription program currently in development (Respondent 5), and one stated that if they were to embark on a crowdsourcing project it would be for describing photographs only and they would likely utilize Flickr rather than their own website (Respondent 7). Of the four who do have some capability for user contributed content, two are using it only for the description of photographs. The last two respondents cited limited tagging capabilities but were unsure of how much they were being used. Two respondents specifically cited concerns that the user contributed content could be of poor quality as arguments against its implementation (Respondents 4 and 7). One of those respondents stated:

“The main argument against that has been that we feel like we should be authorities […] We didn't necessarily want to take information from others that may or may not be accurate. The other piece is the time it would take to fact check anything that we got from a crowdsourcing project would be so overwhelming that we wouldn’t be able to do it well enough to trust the data. So that it seems kind of time-consuming, but that said we still sort of toss it around here and there” (Respondent 7, personal communication, January 27, 2014).

Another respondent, whose project does not currently utilize crowdsourcing but whose parent institution is developing a project to do so, also cites limited resources as both a deterrent to and an argument for implementing capabilities for user contributed content.

“[…] just in engaging and enlisting [the] crowd to do work for us actually takes our resources as well to get this platform off the ground, and so it will be a long time before we see how effective it is, but that’s really what we hold out hope for is that the large scale digitization is the perfect place to put our resources into
developing crowdsourcing transcription” (Respondent 5, personal communication, January 23, 2014).

One respondent stated that they would like to use crowdsourcing for transcription, and that “if we had the capability of doing that I think we would get a good return on our investment” (Respondent 4, personal communication, January 22, 2014).

Respondents were evenly split regarding the use of optical character recognition on their digitized materials, with four projects currently utilizing OCR and four not doing so. Of those not using OCR, one stated that their project simply did not have enough textual (non-handwritten) materials to make it worthwhile (Respondent 5, personal communication, January 23, 2014), and another stated that they plan to do so for textual materials slated to be digitized in the future (Respondent 8, personal communication, January 30, 2014). One OCR user summed up their opinion on the practice as follows:

“Through OCR of the digitized images and running them through OCR and allowing that full text to be exposed as well, […] users could actually find things not necessarily solely based on the description but also the actual content of the materials. […] it was highly inaccurate, but there was no effort to clean the OCR, and it was going to be just an as–is. If it leveraged any type of searchability or discoverability, that was a bonus, but again it wasn't our focus for this project” (Respondent 6, personal communication, January 23, 2014).

Another respondent, whose project also did not make any corrections to their OCR, stated that in their opinion, “the rather dirty OCR, as they say, is pretty darn good” (Respondent 4, personal communication, January 22, 2014).

1.6 Edits to Original Finding Aids

Half of the archivists interviewed reported that digitizing archival collections resulted in changes to their existing finding aids in order to facilitate digitization and discovery online (Respondents 1, 3, 4, and 5). These archivists describe how existing
findings aids were inadequate or noncompliant for the purposes of digitization. One respondent described changes that might be made to facilitate description for digitized objects.

“We do take a pretty detailed approach when something is going to be digitized in its entirety, even if that might mean that the archivists goes back to the original finding aid and expands on it and reviews it, perhaps even breaking up the folders into smaller descriptive and physical sections to minimize the number of images per folder, or to make the description more specific, still avoiding what we would call item level description, but trying to apply […] what would be considered the opposite end of minimal level processing, I guess you could say” (Respondent 5, personal communication, January 23, 2014).

Another respondent stated that while the finding aids are not changed for every collection digitized, many do require some changes in order to be compliant with online accessibility, and specifically that findings aids should include container IDs (identification at the container level, e.g. Box 4). Regarding this review process, the respondent states:

“We need to have a finding aid in place, and the finding aid has to be compliant, if you will, with the needs of our program, so that we can extract the metadata from the finding aid and have our system work, basically hooking up containers with the digitized content in our platform” (Respondent 4, personal communication, January 22, 2014).

The same respondent goes on to give an example of a finding aid for a collection including audio recordings that would not be considered compliant with description for digitization.

“Currently the finding aid is such that there are no container IDs whatsoever for any of these various recordings. They are listed individually but […] essentially what we have are unit title and that's it. […] your description is just floating there and it’s not attached to a container and it doesn't have a container ID, so that is noncompliant. The script will not work without the ability to connect the container ID to what we call the hook ID in the content. […] online access cannot happen unless that is there, so that’s a compliance issue” (Respondent 4, personal communication, January 22, 2014).
The other two archivists who reported the addition of description in the process of digitization stated that the existing description was inadequate, described to only the collection level and in one case for internal use only, and had to be completely reformatted as part of the digitization project (Respondents 1 and 3). Of the four who did not report changes to the finding aid, two stated that some description was changed or added only to the digital images placed online rather than editing or adding to the original findings aid (Respondents 2 and 7). One respondent described an effort to increase description for photographs in response to user demand.

“On our Flickr site, we put up something like 40,000 images using just a folder title […] and event […] so the folder title aggregates like 1000 images under the title of the [event] and the year. So, we’ve been questioned previously about why there would be photos of very famous people […] inside that pile of a thousand images, but they’re not identified by, in the title or the subject, but we’ve been asked why those weren’t tagged more specifically or identified more specifically. In a couple cases we’ve gone and added additional description to those” (Respondent 2, personal communication, January 17, 2014).

The final two archivists stated that the finding aids for collections digitized as part of their projects were already developed or existed prior to the beginning of their digitization initiative (Respondents 6 and 8). One respondent described their lack of need for additional or edited description due to previous processing efforts and the scope of the digitization project.

“[…] the collections being digitized had previously been processed under a processing grant, to folder level description. And so the collections that we digitized are approximately 200,000 pages worth of material had all been processed and described in the EAD finding aid at the folder level, and the idea was to repurpose that EAD folder description as the primary description for the digitized resources” (Respondent 6, personal communication, January 23, 2014).
1.7 Efficacy of Utilized Descriptive Methodologies

When asked if they felt that their projects provided adequate description for the purposes of discovery, six out of eight interviewees stated unequivocally that they did (Respondents 2, 3, 5, 6, 7, and 8). One of the other two respondents replied that “we spent a lot of time determining the mandatory fields that we believe are most relevant for our patrons and their research. Our public interface for [the project] is not yet live and public, so it is difficult to say whether or not our patrons require more information, or if they have what they need” (Respondents 1a and 1b, personal communication, January 14, 2014). The last gave a more nuanced response, stating that “academics are going to be able to discover our material just like they always have. I think the really motivated user who does not have the training is going to figure it out. I think the casual person is going to have a really hard time, and I think for some it is simply not worth their while” (Respondent 4, personal communication, January 22, 2014). Only two respondents used the word “minimal” in characterizing their descriptive methodology (Respondent 2, personal communication, January 17, 2014; Respondent 4, personal communication, March 23, 2014), and only one respondent described their methods in terms of MPLP (Respondent 6, personal communication, January 23, 2014).

Of the seven interviewees whose projects have been published online, five reported being satisfied with the amount of use they have received (Respondents 2, 3, 4, 5, and 8), with two of those five (Respondents 2 and 3) stating that while they are satisfied by usage so far they always want more. Two respondents did not feel that they had adequate data to answer the question, but one reported positive anecdotal evidence indicative of use (Respondent 6, personal communication, January 23, 2014) and the
other felt that “we’d like the material to be used more and that’s up to us to just publicize it more and to write more” (Respondent 7, personal communication, January 27, 2014). Of these same seven respondents, all reported employing some usage measurements, though two stated that they have had little impact on the project (Respondents 4 and 7) and two stating that their projects’ history of analytics has been less than desirable (Respondents 2 and 6). Only one project reported utilizing analytics to actively inform and improve usability and description, but stated that the change had yet to be implemented – rather, it has allowed them to “set priorities for web enhancement” when funding may become available (Respondent 5, personal communication, January 23, 2014).

When asked if, given the opportunity and resources, they would add more description to their digitized materials, five out of eight respondents said they would not (Respondents 1, 3, 5, 6, and 7). One characterized their current descriptive methodology as “striking a good balance” (Respondent 7, personal communication, January 27, 2014), while another noted that they had hit the descriptive “sweet spot” (Respondent 5, personal communication, January 23, 2014). Two of the respondents stating that they would add description specifically cited contextualizing information as what they would add (Respondents 2 and 4). One cited identification information for targeted collections of high interest, such as dates, geographic location, personal names, and page numbers for individual documents (for example, to let the user know what page of a letter they were looking at), as having the potential to increase discoverability and help users understand their findings given the limited browsing functionality of the user interface in their chosen CMS (Respondent 4, personal communication, January 22, 2014). Another stated
that placing items in historical context, either by additional narrative description or through arrangement, would be beneficial to the user (Respondent 2, personal communication, January 17, 2014).

**Discussion**

Regarding the effect of technological or organizational constraints on description, it is clear that different approaches to description from archivists and museum staff have the potential to create tension. The majority of archivists in this situation express a preference to describe more minimally despite pressure to conform to a more item-level, object-centered approach. Archivists using content management systems designed for describing artifacts rather than documents report that such a design may not be ideal, but is adaptable for their purposes. Both respondents reporting use of a homegrown CMS were happiest with the description it provided, but one demonstrated the problems of having a homegrown system and attempting to comply with a parent institution’s digital asset management system. The respondents using off the shelf products reported more varied rates of satisfaction, some citing specific ways their description could be better were they not constrained by their CMS, but all respondents were satisfied that the description they were able to provide through each system was adequate.

While the responses contained more variety than I was expecting in terms of granularity of description, the majority of initiatives tended toward a folder-level approach to description, relying on metadata from existing finding aids. Half of the respondents, however, reported making edits to their original finding aids in order to be able to link encoded archival description or otherwise use finding aid metadata for their digitized collections. It would seem that while their overall goal is to use existing
description, it is common for work to precede the publication of materials online in order to make that goal feasible. A majority of respondents demonstrated willingness toward utilizing user-generated content in the form of transcription or description of photographic materials, though only three respondents had such programs in use or in production. Half of the projects reported that the use of optical character recognition for textual materials was enabled on their projects, and that doing so constituted very little effort on the part of the archivist. While there was some dissent on the quality of the searchability it provided, the overall consensus was that it was better to have it than not. The majority of respondents stated that they believed that their description was adequate for the purpose of discovery, with most projects having access to web analytics or other usage measurements. Except for one case, these measurements did not seem to have much impact on the digitization initiatives in terms of directing future policies. Most archivists were satisfied with the quality and amount of description their projects include, with only two expressing a desire for more description, both of which related to context. This wish to enable users to have the closest experience possible to that of a search room environment certainly mirrors the goals of early digitization initiatives, but may or may not be feasible to achieve. The potential difficulties and associated costs with adding such contextual descriptive information (such as pagination for individual documents) is an area where further study would be beneficial. In addition, I would recommend further study in the form of a cost/benefit analysis of custom made (“homegrown”) and commercially available (“off-the-shelf”) content management systems for use by initiatives of this size and nature, as this study revealed a wide variety of systems utilized in the field. Finally, a further examination of the uses of monitored usage measurements
for digital collections would be of interest to those undertaking such projects, in terms of how such statistics are used to the potential advantage of project leaders to demonstrate use and justify costs to institutional administrators, donors, and funding agencies.
Appendix 1: Interview Questions Sent to Study Respondents

1) Tell me about your project’s approach to the description of digitized resources.

2) What led you to choose that approach?

3) [Was/is] the same level of description applied to all series/sections/resources within your project?

4) [Were/are] collections described to the item, folder, series, or collection level, or was some combination of those approaches used?

5) What [was/is] the finest level of description for digital objects you have utilized in your project? How do you feel that that [facilitated/facilitates] discovery and access?

6) [Did you include/are you including] existing metadata (from finding aids, catalog records, etc.) in your description?

7) [Did you use/are you using] or [consider/considering] using tagging or crowdsourcing as part of your description? Why or why not?

8) [Did/do] you use specific metadata standards, either content or structural, in your description, and if so, which ones? Why or why not?

9) [Did/do] you use any specific thesauri to define values in your description, and if so, which ones? Why or why not?

10) If any part of the process [is/was] automated (such as importing existing metadata records), how much subsequent proofing [is/was] there of that description?

11) [Did/does] your project have collaborators outside of the library (community members, professors, etc.)? Did they express opinions or preferences regarding
description? Did their participation influence the description of digital objects at all?

12) Do you feel like the project [had/has] adequate description for the purposes of discoverability?

13) Do you recall any examples of finding description for a resource that you felt was inadequate, either before or after it had been published online?

14) For collections with existing description, [did you add/have you added] more or different description than what already existed in the finding aid or catalog record? [Did you change/have you changed] existing description at all?

15) Do you feel that digitized collections published online have different descriptive needs than the original print materials? If so, how does your project reflect this?

16) Given the opportunity, be that in terms of more time or resources to devote to this project, would you add more description to your digitized collections? Why or why not? Is there a specific kind of additional description that you think would be useful for users of this digital collection? What would it take for you to add more description to these digital objects in terms of resources (volunteer/paid labor, more money, more time)?

17) When did your project begin, and when did it end (if it is not ongoing)?

18) What content management system or digital collection management software was used to publish your digitized collections online? How would you say it [affected/has affected] the format, content, or amount of description your project [was/is] able to provide?
19) How many people [worked/work] on your project? How many of those people [were/are] fulltime, part-time, volunteer, etc.?

20) If you had to guess what percentage of time in the digitization project [was/is] devoted to the description of digital materials, what would that be?

21) What usage measurements, if any, have you employed following the [completion of your project/publishing of digital materials online]? [Do you utilize/have you utilized] web analytics to gauge use of your materials?

22) Have you been satisfied with the amount of use your project has received? Please explain.

23) [Were/are] any resources transcribed or optical character recognition (OCR) used as part of your project?

24) How many (if any) of the resources digitized by your project [have/will have] transcriptions available online?

25) How much [does/did] the need to disclaim/note copyright impact the description of resources?

26) Do you know of any other archivists involved with large scale digitization initiatives with whom it would be useful for me to speak about this project? [If yes]

    What project did [he/she] work on? Do you know the best way to get in touch with [him/her]?
Bibliography

1.8 Interviews
Respondent 1a, and 1b. Interview with authors. Email interview. Chapel Hill, NC, January 14, 2014.


1.9 Secondary Sources


