When compared to the policies and procedures that have been developed for their non-digital counterparts, the field of digital preservation is still in its infancy. This lack of established procedures and standards does not detract from the importance of ensuring the preservation of digital collections. One aspect that lacks sufficient research is the role that preservation metadata plays. Preservation metadata is necessary for effective preservation of digital collections. Of equal importance is the ability for this metadata to be interoperable between institutions. If preservation metadata is present, but it cannot be understood by other institutions, its value is greatly diminished. This research seeks to determine the level of use of preservation metadata for digital collections for the 123 members of the Association of Research Libraries. From this determination, this research will also briefly address the role of interoperability for preservation metadata-collecting libraries, and the type of workflows currently being used.
EXAMINING THE USE OF PRESERVATION METADATA
FOR THE LONG-TERM PRESERVATION OF DIGITAL COLLECTIONS

by

Brody Selleck

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
July, 2009

Approved by

_______________________________________
Jeffrey Pomerantz
Acknowledgements

I would first like to thank the survey participants for their time and involvement in this study. Special thanks goes to Maggie Dickson, who introduced this project to me and offered guidance and support throughout its duration; Diane Kelly, who assisted in the development and implementation of the survey used for this project; and Priscilla Caplan, who reassured me that this project would be beneficial to the field of digital preservation. I would also like to thank Jeffrey Pomerantz for advising this project.
Table of Contents

Acknowledgements ...........................................................................1

List of Figures ..................................................................................3

1. Introduction ..................................................................................4

2. Literature Review .........................................................................9
   2.1 What Metadata? .................................................................9
   2.2 The Need for Agreement ...............................................14
   2.3 Digital Preservation and Metadata, in General .............17
   2.4 Summary .................................................................21

3. Method ......................................................................................22
   3.1 Method Selection ..............................................................22
   3.2 Sample Selection .............................................................24
   3.3 Data Collection Instruments and Materials ...................25
   3.4 Study Procedures ..........................................................28

4. Results and Analysis ..............................................................29
   4.1 Level of Response ...........................................................29
   4.2 Overview of Digital Collections and Preservation Metadata .........................................................29
   4.3 Cooperation, Interoperability, and the Collecting Process .................................................................35

5. Summary and Conclusion .......................................................41

References ........................................................................................47

Appendices ......................................................................................49
   A. Survey Consent and Survey ..............................................49
   B. Survey Email ...............................................................55
   C. Reminder Email .............................................................57
List of Figures

Figure 1: Age of Digital Library/Collections ........................................30
Figure 2: Institutions Collecting Preservation Metadata ......................31
Figure 3: Age of Digital Collections vs. Creation of Preservation Metadata .................................................................32
Figure 4: Plans to Develop Preservation Metadata Procedures ..............33
Figure 5: Institutional Cooperation vs. Use of Standard Schemas ..........36
Figure 6: Metadata Standards Referenced by Respondents ..................37
Figure 7: Responsibilities for Collecting/Creating Preservation Metadata .................................................................38
Figure 8: Methods for Creating Preservation Metadata ....................39
1. Introduction

As digital objects and digital collections have become increasingly integrated into the core materials for libraries across the globe, librarians have been confronted with a set of challenges that they have never had to deal with before. While these digital objects have undoubtedly changed how information and materials are accessed, presented, and disseminated, they have also changed how librarians must maintain and preserve information. With traditional, paper materials librarians’ and archivists’ main concern was how to best preserve the physical object itself. While this task can be a challenge depending on the type of material, the preservation process has become relatively routine. Currently there is no such routine for preserving digital objects. What makes the preservation of digital material unique is that not only does one have to preserve the object (the digital files) via conversion to new file formats and backing up the objects with exact copies, but librarians also must address the issue of preserving the metadata that is associated with the original file.

One of the largest issues that comes with preserving a digital object’s metadata is trying to determine what can, and should be, considered necessary for preservation and what can be considered as excess and not essential. For example, in order to effectively preserve a digital image, one would likely record the metadata for things like the image’s dimensions, color map reference, and compression scheme. For a digital audio file, one would likely record duration, resolution, and bit rate. For a text file it would be useful to preserve the metadata for things like the character set and markup language.
(Knight, 2005). Although most librarians in the field of digital preservation recognize that there is a need to collect/create metadata to be captured and recorded for each and every digital object, there is still a minimal amount of research that has been done on this subject. This lack of research was the initial cause of a rapidly developing interest in investigating the issues surrounding digital preservation and metadata. This interest was further developed through an interaction that I had with Maggie Dickson, the Watson-Brown Project Librarian in the Carolina Digital Library and Archives, when she said in an email to myself that librarians “all talk a pretty big game about the importance of digital preservation and metadata standards, but not everyone is doing it [...] and everyone seems to agree that it is a really good idea, but I don’t think many institutions are actually employing it” (M. Dickson, personal communication, January 14, 2009).

In terms of the problem that this research will address, it is to expand upon what Maggie Dickson conveyed to me by looking at the efforts being taken to create and maintain interoperable preservation metadata for digital collections. Preservation metadata is defined as metadata that “supports activities intended to ensure the long-term usability of a digital resource” (Caplan 2009, p. 3). As the definition suggests, this type of metadata is not directly concerned with discovery and access or description of the digital materials, but instead exists to assist with the preservation of these functions. This research is necessary because there has been little or no previous research done that has surveyed practices across institutions. To be sure, there have been a number of articles written that detail the practices of a single institution. One example of this is Steve Knight’s article explaining the digital preservation practices of the National Library of New Zealand. While his piece is a good display of what one library is doing, it does not
go beyond that in scope. With that said, the purpose of this research is to perform a survey of institutions to determine the level of use of preservation metadata in the digital preservation of their objects and collections. This research will allow librarians to see and understand what the current level of use is for digital collections, whether the surveyed libraries are striving to create interoperable preservation metadata, and if the use of preservation metadata will increase in the future.

When one begins to look at the role preservation metadata plays in the conservation of digital collections, it becomes increasingly apparent that although its role is an important one, this is an area of library and information science that has yet to accumulate a significant amount of research. According to Priscilla Caplan, director for Digital Library Services at the Florida Center for Library Automation, in a recent issue of *American Libraries* magazine, “digital preservation is a young field, heavily dependent on research and experimentation. It is a fast-moving area that advances rapidly but as yet has few exemplars” (Caplan 2008, p. 38). Based upon what Caplan states and through a thorough investigation of the literature on preservation metadata and digital preservation, it would seem that scholars have not yet been able to catch up with, and make sense of, the rapid advancements in the realm of digital preservation. To be sure, there is still a substantial amount of literature on the subjects; it is simply that much of it is focused on description and recommendations of various plans and projects, and less on empirical research. This need for empirical research is precisely why the research questions proposed for this study have been developed.

As noted above, this study’s research questions are focused on preservation metadata and specifically if librarians are collecting it for the digital collections that their
libraries are creating. However, prior to addressing the research questions, it is necessary to know what is meant by the term: “digital collections.” As Linda Cantara writes, “a key issue for digital libraries and the repositories designed to preserve them […] is how a digital collection is defined: is it a collection of images, a collection of PDF files, a collection of audio files?” (Cantara 2006, p. 40). For the purposes of this study, the definition of a digital collection will have few restrictions. The definition will depend entirely upon what each respondent to the survey chooses it to be. However, institutional repositories (digital collections that preserve and provide access to the intellectual output of an institutional community) will be excluded from this research because they are much more complex than traditional digital collections (Crow 2002, p. 5). With that said, the research questions for this study are:

What is the level of use of preservation metadata for the purpose of long-term preservation of digital collections?

What steps are libraries taking to ensure the interoperability of the preservation metadata they create?

These questions came to fruition after a number of conversations with professionals in the field and an investigation of the related literature led to the recognition of the aforementioned lack of hard research on the subject matter. The purpose of this study is to learn more about the current state of preservation metadata in digital collections and to fill in a portion of the void that the present literature on preservation metadata has left.

There are three key elements for the importance of this study. These elements are: the need for empirical research on the use of preservation metadata in digital collections, the need to understand the current use and practice of preservation metadata by a population of libraries actively involved in creating digital collections, and to make
librarians aware of the current state of interoperability, or lack of, for a very important aspect of digital preservation. By asking, answering, and analyzing the survey questions it is hoped that the importance of the research questions that have been detailed throughout this paper will be affirmed. Conversely, if this research reveals that there is little or no use of preservation metadata or that there is minimal uniformity in the metadata being collected, then perhaps the results could be used by librarians to advocate for the inclusion of preservation metadata in the management of digital collections and an increased amount of interoperability across libraries to more effectively preserve their digitized materials for the future.

The results of this study are expected to be important to librarians working with digital collections, metadata librarians, and librarians simply interested in the preservation of intellectual content. Similarly, this study contributes to research in each of the fields that the aforementioned librarians work in. The fields of digital libraries, metadata, and preservation should all be interested in the results of this research, as it is expected to create a unifying bond between these fields. Additionally, this research will also be beneficial to advocates of interoperability. If each institution is applying their own standard to preservation metadata, it will show that libraries have not yet grasped the importance of interoperability for ensuring access to digital collections in the distant future. And if the results show that a large number of libraries are using preservation metadata and working together in doing so, it will show that librarians have recognized the importance of interoperability.
2. Literature Review

The literature review for this study is divided into three main sections: What Metadata?, The Need for Agreement, and Digital Preservation and Metadata, in general. The first section examines two articles that address the type of metadata elements that should be collected for long-term preservation. The second portion of the review examines two studies whose results advocate for uniformity and practicality in the preservation metadata that is being collected by institutions. Lastly, the review scrutinizes a number of articles that give additional information related to digital preservation and metadata in order to fully develop the need and context of this study.

2.1 What Metadata?

Before delving into the details of this research, one needs to have a better understanding of what exactly preservation metadata encompasses. One of the more effective, and simpler, explanations comes from Diane Hillmann, et al (2008). Hillmann writes that preservation metadata is “designed to ensure access to information resources remains over a long period and records details about format migration and data refreshment. This is typically not done in traditional cataloging as most traditional resources are static and unchanging; the digital world is conversely more dynamic, and metadata must accommodate these changes and updates” (p. 9). While all types of metadata are important for ensuring effective access and management of a digital collection, preservation metadata has a very unique and equally important role in ensuring the long-term access and management of the digital collections they support.

Although the overall importance of preservation metadata has become a widely accepted notion amongst librarians working in the field, there is still a great amount of
consternation when it comes to determining what metadata elements should actually be generated and populated. Over the past decade a number of generalized guidelines have been established by varying groups. The standard that most digital preservation repositories use as a model is the Open Archival Information System (OAIS). The OAIS “defines a vocabulary for preservation-related concepts, describes a model for preservation information, and itemizes at a high level the functions an archive should perform” (Caplan 2008, p. 38). But what OAIS, and other schemas like it, does not do is explicitly state the metadata elements that should be collected for long-term preservation. Because of this shortcoming these models are only intended to act as guidelines for developing a digital preservation plan/structure. This is not to say, however, that no one has attempted to suggest a set of elements for preservation metadata. There have actually been a large number of articles published that explain the preservation metadata that was collected for a specific collection. For the purposes of this literature review, two models for preservation metadata will be looked at in greater detail. The first of the two, published in 2000, is based largely upon the guidelines laid out by the OAIS, whereas the second model seeks to “develop a conceptual model for preservation metadata that complies with a standard ontology for cultural documentation” (Constantopoulos 2007, p. 1).

Even though much of the technology that was used for digital collections in the year 2000 could almost be considered antique by today’s standards, the work done by Catherine Lupovici and Julien Masanès to develop a core set of preservation metadata elements is still very influential. Working for the Networked European Deposit Library (NEDLIB), Lupovici and Masanès recognized the need to develop an appropriate schema
to deal with the long-term digital preservation of the large amount of material that NEDLIB received. While much of their paper, entitled: “Metadata for long term-preservation”, is dedicated to explaining the inherently complex structure of OAIS and NEDLIB’s own digital preservation structure, the article’s contribution to preservation metadata lies within the development of 8 metadata elements and 38 sub-elements based upon the OAIS model.

Lupovici and Masanès’ preservation metadata elements were some of the first to be suggested for inclusion in digital repository records. The main elements included hardware, microprocessor, multimedia, and peripheral requirements, operating system, interpreter and compiler, object format, and application (Lupovici 2000, p. 18-20). The sub-elements that were detailed in the article, expanded upon a number of the main elements in order to provide greater specification when necessary. These elements are a great source of context when one is trying to grasp the fluidity of the digital world. They were created less than ten years ago, but one can already begin to see how a number of the main elements have already lost some of their importance. Operating systems and microprocessor requirements, although still somewhat necessary, are nowhere near what they were when Lupovici and Masanès were writing. They deemed these two elements to be a core element to preservation metadata, while in today’s world it could be suggested that they are almost not necessary. To be sure, there are undoubtedly digital collections that contain items which would require such special treatment, but with the near universal functionality and constantly improving processor speeds these cases are likely to be in a very small minority.
The preservation metadata elements created for NEDLIB less than ten years ago prove to be very useful towards understanding the importance of, and difficulty maintaining, an effectively preserved digital collection. The authors write that “we can be sure that the modality of data processing will be different in 20 or 100 years. It is thus our task to collect key information about today’s data processing to ensure future access to these documents” (Lupovici 2000, p. 2). This modality has even changed substantially since that statement was written. In relationship to the research questions at hand, it would seem that this development reaffirms the questions’ necessity. Librarians and information professionals need to be aware of the importance of preservation metadata in order to ensure that their digital collections remain accessible in the future, and hopefully this research brings this need to light or at least shows that digital collection managers have already begun to take the appropriate steps towards digital preservation.

The second article that addresses the metadata elements that should be generated and collected for digital preservation is much more recent (2007) than Lupovici’s, and takes a much different approach to quantifying the necessity of metadata. “An ontological model for digital preservation” by Panos Constantopoulos and Vicky Dritsou presents a model for determining preservation metadata that “is compatible with CIDOC CRM, the ISO standard ontology for cultural documentation” (Constantopoulos 2007, p. 2). The authors feel that preservation metadata would make the most sense, and be most effective, if it models the same ontology used to preserve non-digital material. This is not to say that the authors chose to ignore previous guidelines for preservation metadata, as they actually did the exact opposite. To develop the model, the five most influential (as determined by the authors) models for metadata sets (OAIS, Dublin Core Metadata
Initiative, Curl Exemplars Digital Archives (CEDARS), the Pittsburgh Project, and the National Library of Australia proposal) were comparatively studied in order to extract a set of common elements. These 14 core elements were then paired with parent concepts from the CIDOC CRM (Constantopoulos 2007, p. 2). By associating the common elements of the most influential preservation guidelines with the ontology for cultural documents, this model would seem to effectively bridge the gap between the digital and non-digital world. Whether or not this suggestion is accurate will have to be determined through future application and research of this ontology-based preservation metadata model.

Although it seems that the authors’ ideas are clearly headed in the right direction, by associating their metadata elements with non-digital preservation methods, their article does not properly explain how and why this linkage exists. Additionally, they do not address how they paired the common elements with CIDOC CRM parents when no direct link between the elements existed. Constantopoulos and Dritsou believe that their model’s “merit lies in the inference capability stemming from the explicit semantic structure, as well as in the integration with the domain of cultural documentation” (Constantopoulos 2007, p. 6). This assertion would indeed be accurate if their model is to be successful. However, such success is by no means guaranteed.

As was the case with the first model, this preservation metadata model’s relationship to the research questions lies in its ability to show the effectiveness/necessity of creating such elements. By affixing this model to the preservation ontology of cultural documents, it would appear that this set of elements is more likely to endure the changing preservation environment of the future. Non-digital documents have a much more
concrete preservation system than digital collections have, and even though the conversion from CIDOC CRM to preservation metadata elements is imperfect, it certainly provides a very strong base on which to manage a digital collection. Both of the articles discussed in this section were reviewed in order to show that at least some efforts have been made to develop coherent sets of preservation metadata. While these models are imperfect, they still do well in providing a greater understanding for what can and should be expected of preservation metadata for digital collections.

2.2 The Need for Agreement

In this section of the literature review, an attempt has been made to show why it is important for institutions to agree upon, or at least come close to agreeing upon, the type of preservation metadata that they create for their digital collections. If every institution was collecting their own version of preservation metadata, it would be as useless as if they had not collected any preservation metadata in the first place. The two articles examined here have been reviewed in order to show the need to establish a standard for preservation metadata. This is done because the research questions that this literature review is based on seek to show the same need. If an analysis of the question shows that institutions are not collecting any preservation metadata, or are collecting their own personalized version of the necessary elements, then it is hoped that the research will deliver the same message the following two articles do.

In 2003, Michael Day performed a content-analysis of 14 preservation metadata initiatives in order to identify problems related to the practicality, sustainability, and interoperability of the varying perspectives. His research also sought to categorize the initiatives based on how conceptual or practical they were considered to be. The
initiatives that Day examined originated from the OAIS, national and research library-based programs (e.g. NEDLIB), the archives and records domain (e.g. the Pittsburgh Project), digitization projects, and others (Day 2003, p. 1). While examining the programs, Day characterized the various initiatives on a spectrum ranging from conceptual to practical. What he found from doing this was that “one may be able to detect a gradual move from the conceptual to the practical. […] The current focus is on developing metadata schemas that can be implemented” (Day 2003, p. 6). The trending that Day discusses would seem to suggest that more and more institutions are developing models for generating and populating preservation metadata fields; however, this would also suggest that these same institutions are not concerning themselves with issues related to sustainability or interoperability. These are two very important issues when dealing with long-term digital preservation, which Day also addresses.

The remaining sections of Day’s article highlight the problems that he sees with having practical, but unsustainable and poorly interoperable preservation metadata initiatives, along with advocating the creation of metadata registries to help alleviate these problems. Day believes that the creation of a registry for preservation metadata could support three separate functions. A metadata registry is intended to provide guidelines and assistance for creating metadata for a digital collection. Firstly, the registry would act as an authoritative source for information in order to ensure proper interpretation of the data. Secondly, it could help to automatically generate metadata once partially populated. Lastly, it could help support the exporting of metadata to other repositories (Day 2003, p. 10). Coupled with a predicted increase in the automatic capture of metadata, Day thinks that the creation of a preservation metadata registry
could go a long way in creating a more interoperable world of preservation metadata.

Although Day’s research showed that preservation metadata initiatives are slowly trending toward practicality, thus suggesting that there is an increased use of preservation metadata, the research also indicates that these initiatives are not much closer to obtaining sustainability and interoperability. It is from these points that Day’s research reflects the need for an investigation of the preservation metadata practices of specific institutions. There is a significant need to find out if these institutions are implementing one of the larger initiatives or if they have developed their own model from OAIS.

“Trends in metadata practices: a longitudinal study of collection federation” by Carole Palmer, Oksana Zavalina, and Megan Mustafoff is another study whose results indicate a need for uniformity of metadata, and specifically preservation metadata, across institutional boundaries. This research, done via surveys, interviews, and case studies, was performed with the intended goal of understanding how local digital collections were collecting metadata and how effectively collections were being integrated into federated collections. The authors performed their study over a number of years in order to account for a potential increase in interoperability and to determine whether local collections were making a concerted effort to adhere to more universal metadata schemas (Palmer 2007, p. 388). This trending is important in terms of preservation metadata for the obvious reason that if they are using more traditional/universal metadata schemas, then the preservation metadata they are collecting (if they actually are collecting any) should also be trending in the direction of interoperability. One shortcoming of this study was that it does not directly address preservation metadata, so there is no way of knowing if the surveyed collections were indeed collecting metadata with an eye towards digital preservation.
While the authors did not address any implications regarding preservation metadata and successful inclusion of local collections into federated collections, the results of their study were still promising. They found that the local collections, on a generalized level, were developing more a uniform metadata practice, while still collecting specific metadata to be kept at the local level (Palmer 2007, p. 393). These findings are important for the future of preservation metadata because, as noted at the beginning of this section, if every institution is using their own set of metadata elements to support their digital collections, then thirty to eighty years into the future no one will be able to make much sense of the records regardless of if they took what they felt were the appropriate means to digitally preserve their material. While digital preservation is important in its own right, without at least some level of interoperability between institutions it suffers greatly.

2.3 Digital Preservation and Metadata, in General

This final section is dedicated to four loosely related articles that find unity under the subject headings of digital preservation and metadata. These articles, while not directly relating to preservation metadata, highlight some of the more important aspects of digital preservation. Collecting preservation metadata is just one aspect of preserving digital collections. Collection managers have to deal with issues like preservation format, object storage, monetary costs, along with simply choosing the preservation process that best suits their collections. All preservation initiatives deal with “three main strategies: emulation, migration and metadata – or some amalgam of these which relies on the encapsulation of the digital object with detailed preservation metadata” (Hunter 2003, p.
1). This exploration of interrelated material allows one to have a better understanding of digital preservation, in general.

One aspect of digital preservation that is often overlooked concerns what elements of an object are necessary to preserve as viewed by the users. The research completed by Margaret Hedstrom, Christopher Lee, Judith Olson, and Clifford Lampe, entitled “Digital preservation from a user’s perspective,” tested the reactions of subjects to varying forms of digitally preserved material. According to the authors, “most criteria for evaluating digital preservation strategies rely on requirements from the archival perspective that emphasize ease of accessioning, simplicity of long-term maintenance, and authenticity. Users’ needs and preferences are rarely considered when evaluating digital preservation strategies or when choosing which methods to apply” (Hedstrom 2006, p. 158). The needs of the users should be taken into consideration when preserving material because they are the ones who are going to be using it again in the future. By understanding what the users prefer, then perhaps the methods used for digital preservation may change. For their study, the authors chose to test users’ reactions to three common preservation formats: original, migrated, and emulated (Hedstrom 2006, p. 158). This allowed the researchers to understand what specific elements the users preferred as well as their preference of format in general.

Unfortunately, this study did not produce the results that the authors expected. They attributed a number of the study’s shortcomings to varying quality of emulators, questionable methods, a small number of subjects, and only testing two digital formats (Hedstrom 2006, p. 187). Even though the study did not live up to expectations, it certainly opened the door for any number of follow-up and related studies. Changing the
type of digital format, increasing the number of subjects, and changing the method of the research could do wonders for a topic that needs more attention. Margaret Hedstrom and her colleagues may not have been satisfied with their own findings, but by simply bringing awareness to a digital preservation issue that is often overlooked, they have made their article an important source on digital preservation.

In the introduction of this paper it was mentioned that the idea of a digital collection could take on any number of meanings. It could refer to a collection of digital texts, HTML documents, images, video, or audio among other things. One digital collection of audio files is vastly different than one of HTML documents. The same differences that plague digital collections also affect the digital preservation programs that are designed to manage them. Jane Hunter and Sharmin Choudhury recognized these variances and sought to create a digital preservation strategy that addresses one of the more uncommon digital collections, complex multimedia objects (Hunter 2003).

Although much of what Hunter and Choudhury cover in their article is not important to understanding some of the larger digital preservation issues, they still raise a number of key points that cannot be overlooked. Interoperability of digital preservation initiatives may be the goal of many individuals in the field, however, the authors “believe that no single approach is the right one. Each object or class of objects needs to be evaluated on a case-by-case basis” (Hunter 2003, p. 5). Universal interoperability is not something that can realistically be obtained. There will always be special circumstances that make one digital collection different from another in some way. What is important, however, is that these digital preservation projects take advantage of interoperability when they are able to. Another significant item that arises from Hunter and Choudhury’s
work is their example of a preservation metadata input form. They provide a number of screenshots that detailed how a user would input preservation metadata for a complex multimedia object (Hunter 2003, p. 10-12). This gives one a visual understanding of how an input process would possibly work for a digital collection or repository that they could not acquire from simply reading about the process. Hunter and Choudhury’s overview of how to implement a digital preservation process gives a much more realistic view of how the process actually works that would be difficult to acquire from another source.

The previous articles dealt with issues that surrounded the digital preservation process. Although digital preservation is at the core of this study’s research questions, the metadata aspect of the question should not ignored. With that said, this final article details an assessment of one digital collection’s metadata. This work, done by Eun G. Park in 2006, addresses the metadata quality in a Canadian architectural collection. What is important about this study is to understand the need to have quality metadata. Any collection, big or small, that does not have quality metadata will not perform well individually, nor will it have any interoperability with other institutions or in federated collections. The discussion of this research aims to reiterate the importance of developing a specified set of preservation metadata elements and to highlight one of the many problems that need to be dealt with in order to achieve interoperability.

In doing this research, Park found that the 11 architectural databases examined had a number of inconsistencies in the metadata schemes used. For example, one database called an element: “box no,” while another called the same element: “box ID” (Park 2006, p. 216). While adherence to a predetermined metadata scheme such as the Dublin Core Metadata Initiative would most likely do away with such schema errors,
there is no guarantee that a particular collection will adhere to such a scheme. Librarians and those who create metadata for digital collections are free to create schemas and elements as they choose. Even though this practice is not wise, as it creates the potential for decreased interoperability, the inconsistencies in the relatively small databases that Park examined show that errors can arise almost anywhere. In order for preservation metadata and metadata in general to be effective in practice, particular attention needs to be paid to the quality of the metadata elements that have been created. Errors in the metadata will cause problems in the short term, and there is no telling what effect small errors in the metadata would have in two or three decades from now.

2.4 Summary

The purpose of this literature review was to provide a better understanding of the core concepts surrounding preservation metadata and digital preservation, to delve deeper into actual preservation metadata elements, to recognize the need for interoperability of digital preservation systems, and to experience a number of unique, yet equally important subjects that are related to digital preservation and metadata. By exploring proposed metadata element sets from both the early years of digital preservation, as well as a more recent and relatively unique set, it was hoped that one would begin to understand the ever-changing semantics of the digital world. This dynamic is evidenced by the recognition that preservation metadata elements that were considered core pieces to the set in the year 2000 were not even mentioned in the proposed set that was released just seven years later. In looking at the research done by Michael Day and Carole Palmer one could begin to see the significance of the need to develop practical, sustainable, and
interoperable digital preservation initiatives. All three aspects are equally important towards ensuring the long-term preservation projects actually are long-term.

The other purpose of this literature review was to show the need for more research on this subject matter, and to especially show the necessity of exploring the research questions that were presented in the introduction. It is not essential to know what specific preservation metadata is being created because it changes on a collection-by-collection basis. It is, however, very essential to know if institutions are using the digital preservation initiatives that Michael Day analyzed in order to know the current and future state of interoperability of digital collections. To return to the quote from Priscilla Caplan that was used to open this paper “Digital preservation is a young field, heavily dependent on research and experimentation. It is a fast-moving area that advances rapidly but as yet has few exemplars” (Caplan 2008, p. 38).

3. Method

3.1 Method Selection

As has been discussed at length in the previous sections of this paper, the research questions being addressed are exploratory in nature. There has been little work done on the use of preservation metadata, in general, and virtually none for its use in digital collection metadata. With that said, it is felt that the most useful method to use to investigate these questions would be to perform a survey. According to Earl Babbie, “survey research is probably the best method available to the social researcher who is interested in collecting original data for describing a population too large to observe directly” (Babbie 2007, p. 244). Because this research intends to investigate the use of preservation metadata by a large group of institutions, a survey will be the most effective
way of contacting these institutions and acquiring the necessary data needed for the
analysis of the research question. Although Babbie notes that surveys most often have
individuals as the unit of analysis, they can also measure groups, or in the case of this
study: digital libraries at institutions. While the survey will be used to determine the
practices of digital libraries, the respondent for the survey will only be one individual at
each institution, thus maintaining the necessary qualifications for using the survey
method (Babbie 2007, p. 244). An additional benefit to using a survey for this research is
that it allows for multiple types of questions to be asked to the respondents. For example,
this research plans on asking closed and open-ended questions, as well as contingency
questions. The variety of information that can be retrieved from a survey make it the
most effective method to get answers for the research questions at hand.

While the benefits of a survey made the decision to use this method relatively
easy, the decisions that were made to choose the sample size and population to survey
required more consideration. As the research questions state, the goal of this study is to
determine the use of preservation metadata for the long-term preservation of digital
collections. When one begins to think of the variety of digital collections that have been
created, let alone the sheer number of digital collections, this study would appear rather
daunting. How would it be possible to determine what type of metadata is being
collected for all digital collections across the world? Even with established standards like
OAIS and Dublin Core, it would be incredibly difficult, if not altogether impossible, to
effectively survey the entire digital collection population. Upon recognizing the
difficulty of sampling such a large and indefinable population, it was felt that finding a
smaller, finite population to survey would still achieve good results and allow for more effective sampling.

3.2 Sample Selection

Although there are a variety of populations that this study could focus on, such as digital collections created by museums or government institutions, it was decided that restricting the sample to the 123 libraries who are members of the Association of Research Libraries (ARL) would be the most sensible option. By surveying the ARL libraries it is felt that this study will still address a large population, and a sample that is distinctly different from other institutions. Additionally, it is believed that it is highly likely that a very large percentage of these 123 libraries will have some sort of digital library department or at least some set of digital collections that have been created. The selection of a clearly defined, finite population eliminates any need to create a sample, and limits the need to identify and recruit subjects for the survey.

The only thing that will need to be done in terms of subject determination will be to seek out the most appropriate individual to contact at each institution to complete the survey. To be sure, this will still take a great deal of time and consideration to find the proper individuals. Discovery of these individuals will be accomplished by searching the websites of the ARL libraries to find the contact information for their digital library director or creator of their digital collections. Upon finding this contact information, a significant amount of subjectivity will have to be applied to determine the most appropriate respondent. However, the intention of this study is to contact the director of each digital library. These individuals will likely know enough about their institution’s digital collections to answer the survey questions or they will know who to direct the
survey to on their staff. If there is no director, or no actual digital library at an institution, the survey will be directed towards the individual responsible for the management of the institution’s digital collections. If a library has no digital collections, additional efforts will need to be taken to find the appropriate respondent that would have knowledge of any plans for digital collections in the future. As the subject matters for this study are digital collections and their interoperability, it makes the most sense to contact the subjects and administer the survey via the internet. Therefore, when the appropriate respondent is determined at each library, they will be contacted via email with information about the study and a link to the web survey. By completing the necessary communication online as well as the survey, it will allow for the best response rate and results from the survey. This feeling is supported by Babbie as he writes “that some populations are ideally suited to online surveys.” (Babbie 2007, p. 273) It would seem that creators and directors of digital collections are one of those populations.

3.3 Data Collection Instruments and Materials

As this study is intended to be an exploratory introduction into the relationship between preservation metadata and digital collections, the survey will consist of only eight brief questions. It is felt that the content of these questions will be sufficient to determine whether institutions are collecting/creating preservation metadata, as well as some additional contextual information about an institution’s procedures. Of these eight questions, only the first two are directed at all respondents. There are then two contingency questions for respondents whose libraries do not collect preservation metadata. The four remaining questions are directed at libraries that collect preservation
metadata to provide context for the policies, procedures, and standards. A sample of this survey can be seen in Appendix A.

The first question of the survey asks the participants how long their library has had a digital library, or how long they have been creating digital collections. The respondents were given six options: less than two years, 3-4, 5-6, 7-8, 9-10, and more than 10. This ordinal measurement of institutions’ experience with digital collections is asked in order to provide a potential interpretation for a library’s use of preservation metadata. For example, it could be expected that a library that has 9-10 years of experience is more likely than a library that has less than 2 years of experience to create preservation metadata. By asking this question, it is hoped that the remaining questions will be able to be interpreted more directly. The second question that will be asked to all respondents inquires whether or not the subject’s library collects/creates preservation metadata for their digital collections. This is intended to weed out any libraries that have no involvement with preservation metadata. If a respondent states that their library does not collect any preservation metadata they are directed to answer a contingency question. Any respondents that answer “Yes” are directed to skip to the next relevant question.

In order to better understand why some libraries may not collect/create preservation metadata the survey has two contingency questions built in. When they respond with a “No” to the above question, they are then asked if their library has any plans to collect preservation metadata in the next five years. This Yes or No question is asked in order to see if these libraries are aware of the importance of preservation metadata for digital preservation, and that they have a plan in place to eventually collect this metadata. For example, one of the libraries that has little digital collection
experience may simply not be at the preservation stage of creating their collection. The five year time limit was included because most libraries would not be able to plan to a point much further in the future than five years. If a respondent also answers “No” to this question, they are then directed to a second contingency question. This question asks the respondent to briefly explain why their institution has no plans for preservation metadata. This is asked of the respondent in order to better understand the position their library is in. As this study does not expect many libraries to be in this position, any that are should be documented. After answering one or both of the contingency questions the respondents are asked to submit their survey, as the remaining questions pertain to only libraries that are collecting preservation metadata.

The remaining four questions pertain to the practices that institutions employ to develop, create, and collect their preservation metadata. The first two of these questions ask whether or not the library is involved in any efforts to improve interoperability of their preservation metadata. They question whether or not a library is working with other libraries to develop standard schemas, and if they are using previously developed digital preservation standards or are creating their schemas and standards based on local needs and decisions. These two questions are necessary to the study because they will potentially show a library’s awareness, or lack thereof, for the need for interoperability in their metadata. These questions will possibly provide some intriguing results. For example, a library that is creating their own local standards and schema is aware of the need for preservation metadata, but has yet to consider the need for interoperability. It will be very interesting to see where the responses fall for these questions. The final two questions address a library’s procedure for creating/collecting preservation metadata.
They ask who is responsible for the metadata creation (i.e. librarians, student assistants, someone else) and how it is created (i.e. automatically, manually, or some other way). These questions are intended to better understand the overall collection procedure. By understanding who and how the metadata is dealt with may yield additional details about how a library approaches the process. Additionally, the final questions have comment boxes to allow for further explanation of procedures if the respondents choose to elaborate. The possible responses here also makes for some potentially intriguing results.

3.4 Study Procedures

As was noted in the section, 3.2 Sample Selection, the procedure for completing this study is fairly straightforward. The websites for the 123 ARL libraries were be searched for contact information of the most relevant subject to issue the survey to at each institution. A significant amount of time was devoted to this portion of the method as it was felt that in order to achieve a high level of response, the survey should be distributed to the individuals who would be most qualified to answer the questions. There was no specific method that was followed to determine these individuals; however, as the search progressed, certain job titles began to be sought out at each institution. A small sampling of the positions titles that the survey was sent to include: Metadata Management Librarian, Digital Projects Librarian, Digital Preservation Strategist, Digital Initiatives Librarian, and Digital Projects Coordinator. As will become apparent in the Results and Analysis sections of this paper, the selection of individuals with these job titles, and similar titles, proved to be well worth the effort it took to conduct the search for them. Upon determining these 123 respondents an email was sent to them explaining the background and importance of the study, and also provided a link to a web survey to
be completed. The respondents were given a two week window to complete the survey, and a follow-up email was sent one week after the initial message to all respondents who had not yet completed the survey.

4. Results and Analysis

4.1 Level of Response

One of the largest concerns before distributing the survey for this research was that there was a very large chance that the response rate by the participants would be much too low to make any confident declarations regarding the results of the data. Fortunately, the response rate for this study proved to be more than satisfactory for drawing concrete conclusions as a total of 52 of the 123 respondents completed the entire survey for a response rate of 42%. Additionally, 65 of the 123 respondents completed at least a portion of the survey, yielding a partial participation rate of 53%. It is felt that with a level of response of this degree that much of the results that will be presented below can be interpreted with a sufficient degree of confidence. To be sure, this is not to say that all of the members of the ARL function in a distinctly similar fashion. In actuality it is felt that although there are indeed similarities between these institutions, the results of this research will indicate that there is a rather large amount of variance between the practices that each are employing in their creation, management, and preservation of their digital materials.

4.2 Overview of Digital Collections and Preservation Metadata

As was explained in the section, 3.3 Data Collection Instruments and Materials, the first portion of the survey addressed the age of an institution’s digital library/collections and whether or not their institution collected preservation metadata.
Figure 1 shows the length of time that the institutions have had a digital library or have been creating digital collections.

![Age of Digital Library/Collections](image)

46% of all ARL institutions have been creating and processing digital collections for more than ten years, and over 75% of these institutions have been creating collections for over seven years. This level of collection creation was partially expected when this population of libraries was selected for survey distribution, and it is useful to recognize that the large majority of these institutions have a very significant amount of experience in working with digital objects. While this data is informative on its own, it was solicited in order to determine whether or not there was any statistical significance between the creation of preservation metadata and the length of time an institution has been creating digital collections.
With it being established that over 75% of the ARL libraries have had digital collections for over seven years, it is now necessary to turn to the most important question pertaining to this research, whether or not a library collects/creates preservation metadata for their digital collections. The results of this question can be seen in Figure 2.

As the primary point of this research was to determine the extent to which the members of the ARL collect preservation metadata, there were no assumptions that were made regarding this. However, as can be seen above, over two thirds (68%) of these institutions do indeed collect preservation metadata for their digital collections. This figure does not take into account any special circumstances that may come with certain institutions’ metadata practices, but a variety of these issues will be addressed below in section 5.3 Preservation Metadata Practices. Returning to the level of preservation metadata creation that was revealed through these results, it is important to note that while this figure (like the age of digital collections) is interesting on its own, the true
importance of these results can be found when determining if there is any statistical significance between the two figures.

In order to determine if there is any statistical significance between the length of time an institution has been creating digital collections and whether or not they collect/create preservation metadata, a cross-tabulation and Chi Square test of the data was performed. The results of these tests can be seen in Figure 3.

As can be seen in the above figure, the Chi Square test revealed that there was indeed statistical significance between the two figures at the .05 level. This would then suggest that the longer the length of time that an institution has created digital collections is, the more likely that the institution is to create preservation for their digital collections. Similarly, this would also tend to suggest that as a digital library gains more experience in dealing with digital objects they are also more likely to recognize the importance of preserving and maintaining these collections. However, what this relationship does not
show is the amount of resources (financial and “man-power”) that each institution may or may not have. It could be expected, but not statistically proven, that the libraries who have had their digital collections for a significant amount of time not only have a great deal of experience, but also have a greater amount of resources to operate with than a library that is just beginning a variety of digital initiatives. Although this suggestion does not take into account the role of grant-funding for fledgling digital library projects, the expectation of greater resources coming with age stretches beyond the role that money has in the development process. Nevertheless, there is still a clear relationship between the two figures discussed above, and from this one may be able to infer that as the remaining ARL members continue to gain experience and knowledge that they too will begin to recognize the importance of collecting preservation metadata.

In an effort to find out if the libraries that are not presently collecting preservation metadata due indeed plan to do so in the near future, these respondents were asked to indicate whether or not their institutions had plans for collecting in the next five years. These results proved to be a bit more mixed than the other questions that were posed to respondents, and are presented in Figure 4.

<table>
<thead>
<tr>
<th>Does your library plan on developing a procedure for collecting preservation metadata within the next 5 years?</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>10</td>
<td>56%</td>
</tr>
<tr>
<td>No</td>
<td>2</td>
<td>11%</td>
</tr>
<tr>
<td>I don't know</td>
<td>6</td>
<td>33%</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>100%</td>
</tr>
</tbody>
</table>

As Figure 4 indicates, more than half of those institutions who are not presently collecting metadata have plans to do so within the next five years. Of these institutions
seven of the ten have had digital collections for more than seven years, which again suggests that the length of time a digital library has existed has an effect on how conscientious they are of preservation issues. Those respondents that answered “No” or “I don’t know” to the above question were presented with the opportunity to explain why they do not plan on collecting preservation metadata in the next five years. One of the most relevant responses that came from this portion of the survey highlights a number of concerns that need to be addressed when dealing with the preservation of digital objects.

The respondent wrote:

I am concerned about the bigger picture of preserving digital objects themselves, as well as preserving their metadata, preservation or access. Having the preservation metadata doesn't help if the digital objects themselves are not preserved. [We] have been urging our institution and our consortium to take a strong look at cooperative efforts such as MetaArchive. From what I know to date, digital objects need to be stored redundantly, be geographically dispersed, and perhaps also be saved in different storage architectures, to maximize our chances of preservation. Preservation metadata being included in what is preserved in this way might help with corollary issues, such as format migration, but I think is largely putting the icing before the cake, if you will.

The main point that the respondent makes is that the collection/creation of preservation metadata is, as they put it: “putting the icing before the cake.” In order for preservation metadata to achieve its goal, the objects that the metadata is referencing also need to be properly digitally preserved. This lack of preserving the digital objects is an element of the preservation process that clearly cannot be overlooked. Additionally, with regards to the respondent’s comments, it is important to note that individuals at this institution are at least aware of the idea behind preservation metadata. The only thing that is preventing them from developing a preservation metadata process is the fact that their digital collections are simply not established enough to begin addressing additional preservation
issues. Another positive aspect that can be extracted from the respondent’s comment is their recognition of the need for interoperability and cooperative efforts between institutions. While the respondent’s motivation to get involved in MetaArchive may be partially driven by a lack of resources at their own institution, the overarching theme of cooperation and interoperability is still apparent.

4.3 Cooperation, Interoperability & the Collecting Process

Where the first portion of the survey data dealt with the creation/collection of preservation metadata in general, this section of the analysis will address the levels of interoperability and cooperation that are being achieved in the development of metadata standards and the creation of the metadata records themselves. Additionally, this section will also briefly address the different creating/collecting workflows that the ARL members are presently using for their digital materials’ preservation metadata. An investigation of these elements of the preservation metadata process, interoperability, and workflow is necessary to better understand the current state of digital preservation at these institutions, as well as to understand how those libraries who are not currently addressing preservation metadata issues may attempt to start their own initiatives.

An examination of the role that interoperability plays in the current practices of the preservation metadata collecting members of the ARL can be seen in Figure 5. In the form of a cross-tabulation of responses to a question regarding whether or not an institution cooperates with other libraries to develop a uniform preservation metadata schema with responses to a question that addresses whether an institution uses a previously developed metadata standard or a locally created standard.
What is immediately clear from this data is that the large majority of respondents (72%) use some form of a standard digital preservation schema when creating the preservation metadata for their collections. Additionally, the largest number of respondents also work cooperatively with other institutions to create uniform preservation metadata schemas. These two pieces of information would seem to indicate that there is a considerable amount of effort being taken by digital libraries to make sure that they are not simply doing things on their own. As was discussed earlier in this paper, even if an institution has gone to great lengths to develop a preservation metadata schema that perfectly suits their digital collections, it will be of little value in the future if it is not interoperable with the schemas of other libraries. While the creation of preservation metadata is important to the process, it alone does not accomplish the goal of digital preservation. Libraries and institutions must work together to ensure that a uniform standard is at least partially being followed, and the results from this data would tend to suggest that some institutions are moving in this direction.
Although a number of libraries are working together to develop uniform schemas, more than half of the respondents are working alone in their preservation metadata efforts. Fortunately, however, a larger number of these libraries are still using standard digital preservation schemas even if they are not directly working with others. In order to better understand the variety of schemas being employed, the survey respondents were also asked to briefly describe the schemas that their institution was using in the metadata-creation process. A breakdown of the responses to this question can be found in Figure 6.

<table>
<thead>
<tr>
<th>Standards referenced by Respondents</th>
<th>Number of References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dublin Core</td>
<td>17</td>
</tr>
<tr>
<td>Local</td>
<td>8</td>
</tr>
<tr>
<td>PREMIS</td>
<td>7</td>
</tr>
<tr>
<td>MARC</td>
<td>5</td>
</tr>
<tr>
<td>MODS</td>
<td>4</td>
</tr>
<tr>
<td>MIX</td>
<td>4</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
</tr>
<tr>
<td>OAIS</td>
<td>3</td>
</tr>
<tr>
<td>METS</td>
<td>2</td>
</tr>
</tbody>
</table>

From this information one can see that the largest number of references were to the Dublin Core Metadata Initiative. With regards to the local standards referenced by respondents, some of these were local adaptations of the Dublin Core standard. For example, one respondent wrote that “local needs are heavily applied / added to DC set.” Conversely, a number of these local standards are totally unique to the institutions that have created them. Two of these institutions have created detailed metadata standards of their own that they use for the digital collections they create. Another respondent wrote that they “use a local standard that adapts to each collection. Collections are unique, and
some have better metadata than others.” While this response does make it entirely clear the involvement, or lack of involvement, this library has with uniform standards, it does still make one wonder how much they are considering interoperability in the creation of their preservation metadata. “Other” standards that were referenced by respondents included a variety of niche schemas that included: textMD, audioMD, AES, and NISO. In later studies it would be beneficial to better understand the use of these standards as well as the local standards that have been developed by libraries.

The last portion of this analysis is devoted to investigating the type of workflows being employed by the preservation metadata-collecting libraries of the ARL. The data for this brief analysis was obtained by asking the survey respondents two questions that dealt with who was responsible for collecting/creating the preservation metadata and how the preservation metadata is created. The results of these questions can be found in Figure 7 and Figure 8.

<table>
<thead>
<tr>
<th>Who is responsible for collecting/creating the preservation metadata for your library’s digital collections?</th>
<th>Response</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>The librarian managing the digital collection</td>
<td>19</td>
<td>29%</td>
</tr>
<tr>
<td>Student Assistants (Graduate or Undergraduate)</td>
<td>14</td>
<td>22%</td>
</tr>
<tr>
<td>A separate Metadata librarian</td>
<td>16</td>
<td>25%</td>
</tr>
<tr>
<td>Other</td>
<td>15</td>
<td>23%</td>
</tr>
<tr>
<td>I don’t know</td>
<td>1</td>
<td>2%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>65</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>

Figure 7
From this data one is not able to infer much with regards to a standard workflow for creating/collecting preservation metadata as nearly every response for both questions was represented at a similar level. In examining the comments made by the respondents for who is responsible for collecting/creating the preservation metadata there was one relatively consistent theme. For many of these libraries it would appear that they use student assistants to collect the metadata that is then vetted by the librarian managing the digital collection or a separate metadata librarian. One respondent wrote that “student assistants create metadata under the supervision of the Digital Archivist. Digital Archivist also performs quality control measures on metadata.” The preservation metadata at other libraries is done by a variety of individuals that include: technical services staff, software developers, project managers, and “staff in the DL department.” It is clear from the variety of responses and the comments made by respondents that there is not yet one particular type of individual who is responsible for the creation of preservation metadata.

Looking next at the question of how the metadata is created (see Figure 8), a similarly low level of confidence can be found in the types of processes currently being used in this portion of the procedure. There are relatively equal levels of automatic and
manual creation being done at these libraries. In looking at the comments made by respondents for this question it would seem that a number of libraries currently use a mixture of automatic and manual creation of their preservation metadata. This mixed process is best described in the comment of one respondent:

Depending upon who creates the collection, all of the above methods are employed. If the "donating department" has skill and time to provide automatically generated metadata, they do. If not, and it's the type of file that can have a batch run to extract automated metadata, then we run it. If there is not time to create metadata at the time of the creation of the collection, it may be created later.

This comment would tend to suggest that the library does what they can, when they can; which is to say that they take advantage of whatever opportunities they have to automate the process, but also recognize that they do not have the resources to automate the entire process. The comments for this question also seem to show that many of these institutions wish that they could automate the process as a means of streamlining the workflow and reducing their personal workloads, but that they realize that total automation is not possible for reasons related to insufficient resources and the simple fact that some of the metadata must simply be manually input, such as the digitization settings. The results from the final two questions of the survey attempted to develop some type of standardized workflow for creating and collecting preservation metadata, and even though the large variety of answers prevented this workflow from being established, some consistencies can be drawn from this data. For instance, many libraries use a metadata librarian or a digital projects librarian to create, or supervise, the creation of the preservation metadata. Secondly, the data and comments also seem to show that many institutions are using a mixture of both automatic and manually created preservation metadata for their digital objects. While no specific workflow can be
extracted from the survey data, the framework for successfully creating and collecting preservation metadata is still present.

5. Summary and Conclusions

Throughout the development, completion, and analysis of this research I repeatedly returned to comments made by two librarians working in the field. As was noted in the introduction of this paper, Priscilla Caplan wrote that the field of digital preservation is young and dependent on research, and that “it is a fast-moving area that advances rapidly but as yet has few exemplars” (Caplan 2008, p. 38). Additionally, in talks with Maggie Dickson, a digital librarian at the University of North Carolina at Chapel Hill, she similarly suggested that there was a realization that the field of preservation metadata is important, but that the empirical support for it was not on the same level. These comments set the stage for the development of this study’s research questions and the methods completed to carry out these questions, which were reaffirmed by the review of related literature and more importantly through the results and analysis of the data acquired throughout this research.

The results of this research revealed a significant amount of information regarding the research questions that this study was built around.

What is the level of use of preservation metadata for the purpose of long-term preservation of digital collections?

What steps are libraries taking to ensure the interoperability of the preservation metadata they create?

The first of these two questions, that sought to determine how many members of the ARL were collecting preservation metadata, was clearly established from the data provided by the survey respondents. To date, approximately 68% of these institutions collect
preservation metadata for their digital collections, and another 10 libraries expect to be collecting this metadata within five years. As the Introduction and Literature Review sections of this paper discussed, there has been little research completed that dealt specifically with preservation metadata for digital collections, and therefore there was no way to effectively predict what level of use the survey results would reveal. A level of 68% is certainly noteworthy and altogether reassuring that a large number of the major libraries in North America are aware of the importance of preservation metadata in the overall goal of digital preservation.

Another aspect in the development of preservation metadata was the statistical significance that was established between the age of a digital library/collection and whether or not they collected preservation metadata. This data would tend to suggest that as the younger digital libraries gain experience and knowledge, that they too will begin to collect this valuable metadata for their own collections. This information provides a significant amount of reassurance that preservation metadata may eventually become a universal element of the digital collection and digital preservation process. More specifically, the results of this data can act as a message to those digital libraries that are not collecting preservation metadata. These non-collecting libraries should recognize that this data shows that the more experienced libraries have acknowledged the importance of preservation metadata, and that they too should make an effort to improve their digital preservation practices. Additionally, the comments that were provided by those respondents whose libraries are not presently collecting preservation metadata also lead one to believe that these institutions are aware of the need for this metadata, but that extenuating circumstances are currently preventing them from developing this metadata.
The future of preservation metadata for digital collections would appear to be rather bright.

The second research question addressed by this study dealt with the level of interoperability that the preservation metadata creating libraries are currently employing in their development and use of standards. The survey data was not as clear as it was for the first half of the research, however, a number of conclusions can be extracted. The first of these conclusions is that a large number of these libraries use some form of a standardized metadata schema when creating the metadata for their own collections; with the Dublin Core Metadata Initiative being the basis for the greatest percentage of these institutions. The use of these standards is likely caused by a combination of the desire to create interoperable metadata and the simple fact that it is easier for a library to use a standard schema as opposed to creating one from scratch. That is not to say that libraries avoid creating their own schemas altogether, as a number of respondents noted that their institution had indeed done so, or had at least made major alterations to the Dublin Core schema. Even though a lack of time and resources may be partially responsible for the use of standardized schemas, the fact that they are being used by a majority of the ARL members reaffirms the presence and importance of interoperability. Another point of discussion that this portion of the data creates is in regards to the large variety of standards being used, the creation of local standards, and the lack of communication by some libraries in creating their metadata standards. As a number of institutions are creating or adapting standards to their own collections, it could reasonably be assumed that there is at least some overlap between these alterations. This overlap, as well as the fact that some libraries may not reference any other standards in the creation of their own,
reflects the need for interoperability that other institutions have recognized. Even though the results of this research cannot confidently claim that there is, or is not, a movement towards interoperable metadata standards, the variety of responses to the survey suggests that a greater level of importance should be placed on interoperability in the future. Libraries should work collectively or at least consult previously developed standards in the creation of preservation metadata standards, and metadata standards in general.

Any continuation of this study or any related future research would be wise to take into consideration a number of issues that did not get the attention that they deserved or were simply overlooked. The first of these issues is in regards to clarity of some of the questions on the survey. For example, the second question asked if a library collected preservation metadata for their digital collections, excluding digital repositories. This exclusion was intended to prevent the inclusion of responses that may have been in reference to institutional repositories of scholarly work. These repositories function differently than most digital collections, and the goal of this study was to only look at digital collections. What was overlooked, however, was that some institutions have established repositories that house both their scholarly material as well as their digital collections. With these institutions, additional clarification was necessary in order for the respondents to better understand the goal of this research. Future work on this topic should contain additional clarification regarding the role of repositories for any survey questions and in the discussion and development of the research questions.

Another element of preservation metadata that deserves greater attention in future work would be the reasons why libraries are not collecting this data for their collections. Although some respondents to the question related to this topic for this research provided
some background as to their decision not to collect, it would be highly beneficial to understand these reasons in greater detail. There are any number of reasons why these libraries may not be collecting, and it would be beneficial to the field of digital preservation to more effectively identify with and reach out to the institutions who are struggling to preserve their digital materials. In a similar manner, a more detailed look at the cooperative efforts being made by ARL members in developing preservation metadata standards would also be worthwhile. As this paper has discussed, interoperability is an important aspect of digital preservation, and knowing why certain libraries choose to work, or not work, with other libraries in creating standards. Are these libraries motivated by a need to share resources or is interoperability one of the driving forces in forging these cooperative efforts?

The goal of this research was to shed more light on the role that preservation metadata currently has in the creation and management of digital collections at the member-institutions of the Association of Research Libraries. Through the survey that was distributed to specifically selected individuals at each of the ARL members, it was revealed that over two-thirds of these libraries presently collect preservation metadata for their digital collections. The research data also showed that there is a statistically significant relationship between the age of a library’s digital collections and their decision to collect preservation metadata. This information would tend to suggest that the younger libraries that are not presently collecting this metadata will begin to do so as the gain knowledge and experience in managing and preserving their collections. Lastly, this research also briefly addressed the varying levels of interoperability that are present with institutions’ metadata processes, and discussed the potential for developing a
standard workflow for creating/collecting preservation metadata. The survey data for both of these topics allowed for some assumptions to be drawn, but would require additional research before any concrete assertions can be made. Overall, the results of this research show that the importance of preservation metadata has been acknowledged by most members of the ARL, but there is still a significant amount of room for this aspect of digital preservation to grow in the future.
References


Appendix A: Survey Consent and Survey

University of North Carolina-Chapel Hill
Consent to Participate in a Research Study
Adult Participants
Social Behavioral Form

IRB Study # 09-1084  
Consent Form Version Date: 5/20/2009

Title of Study: Use of Preservation Metadata in Digital Collections

Principal Investigator: Brody Selleck  
UNC-Chapel Hill Department: School of Information and Library Science  
UNC-Chapel Hill Phone number: (607) 857-7669  
Email Address: bselleck@email.unc.edu  
Faculty Advisor: Jeffrey Pomerantz  
Faculty Advisor Email Address: jpom@ils.unc.edu  
Funding Source and/or Sponsor: None

Study Contact telephone number: (607) 857-7669  
Study Contact email: bselleck@email.unc.edu

What are some general things you should know about research studies?
You are being asked to take part in a research study. To join the study is voluntary. You may refuse to join, or you may withdraw your consent to be in the study, for any reason, without penalty.

Research studies are designed to obtain new knowledge. This new information may help people in the future. You may not receive any direct benefit from being in the research study. There also may be risks to being in research studies.

Details about this study are discussed below. It is important that you understand this information so that you can make an informed choice about being in this research study. You will be given a copy of this consent form. You should ask the researchers named above, or staff members who may assist them, any questions you have about this study at any time.

What is the purpose of this study?
The purpose of this research study is to learn about the level of use of preservation metadata for digital collections for the 123 members of the Association of Research Libraries. From this determination, this research will also briefly address the role of interoperability for preservation metadata-collecting libraries, and the future plans for libraries not yet collecting/creating preservation metadata.
You are being asked to be in the study because you possess the required knowledge of metadata procedures and your library’s digital collections to be an effective representative of your institution.

**How many people will take part in this study?**
If you decide to be in this study, you will be one of approximately 123 people in this research study.

**How long will your part in this study last?**
Your part in this will be amount of time it takes you to complete the survey. The survey is expected to take you no longer than 10-15 minutes. There will be no follow-up to the survey.

**What will happen if you take part in the study?**
If you take part in this study, you will be asked to complete a brief web survey. This survey will ask you a number of questions related to your institution’s digital collections and your use, or lack of use, of preservation metadata. The only requirement of this study is the completion of the survey to the best of your abilities.

**What are the possible benefits from being in this study?**
Research is designed to benefit society by gaining new knowledge. You may not benefit personally from being in this research study.

**What are the possible risks or discomforts involved from being in this study?**
There may be uncommon or previously unknown risks. You should report any problems to the researcher.

**How will your privacy be protected?**
Your privacy and confidentiality will be protected in a number of ways. All records will be kept in password-protected files, only the principal investigator and faculty advisor will have access to your individually identifiable data, and upon completion of this study, all individually identifiable data will be destroyed.

Participants *will not* be identified in any report or publication about this study. Although every effort will be made to keep research records private, there may be times when federal or state law requires the disclosure of such records, including personal information. This is very unlikely, but if disclosure is ever required, UNC-Chapel Hill will take steps allowable by law to protect the privacy of personal information. In some cases, your information in this research study could be reviewed by representatives of the University, research sponsors, or government agencies for purposes such as quality control or safety.

**Will you receive anything for being in this study?**
You will not receive anything for taking part in this study.
Will it cost you anything to be in this study?
There will be no costs for being in the study.

What if you have questions about this study?
You have the right to ask, and have answered, any questions you may have about this research. If you have questions, or concerns, you should contact the researchers listed on the first page of this form.

What if you have questions about your rights as a research participant?
All research on human volunteers is reviewed by a committee that works to protect your rights and welfare. If you have questions or concerns about your rights as a research subject you may contact, anonymously if you wish, the Institutional Review Board at 919-966-3113 or by email to IRB_subjects@unc.edu.
Section 1: Age of Digital Library/ Digital Collections

1. For how many years has your library had a digital library, or been creating digital collections?
   ○ Less than 2 years
   ○ 3-4
   ○ 5-6
   ○ 7-8
   ○ 9-10
   ○ More than 10 years

Section 2: Collection of Preservation Metadata

2. Does your library collect/create preservation metadata for its digital collections, excluding digital repositories?
   ○ Yes (Please proceed to question 5 of this section.)
   ○ No

3. If you answered no to the above question, does your library plan on developing a procedure for collecting preservation metadata within the next 5 years?
   ○ Yes (The remaining questions pertain only to those libraries that collect preservation metadata. Please submit your survey now.)
   ○ No (Please proceed to the next question.)

4. If you answered no to the following 2 questions, please briefly explain why your library has no plans to begin collecting preservation metadata.

   After your explanation you may submit your survey.
5. Does your library work with libraries at other institutions to create a uniform preservation metadata schema?

○ Yes
○ No

6. When collecting/creating preservation metadata does your library use previously developed standard digital preservation schemas (ie. OAIS or Dublin Core) as a basis, or does your library create the metadata schema based on local needs and decisions?

○ Uses a standard digital preservation schema
○ Uses a locally created standard and schema

Please provide the name of the schema used and any other relevant information:


7. Who is responsible for collecting/creating the preservation metadata for your library’s digital collections?

If you select more than one option, please provide an explanation below.

○ The librarian managing the digital collection
○ Student assistants (Graduate or Undergraduate)
○ A separate metadata librarian
○ Other (Please specify below)

Please specify:


8. How is your preservation metadata created?

If you select more than one option, please provide an explanation below.

- Automatically generated at time creation of digital collection
- Automatically generated at a later time
- Manually created at time of creation of digital collection
- Manually created at a later time
- Other (Please specify below)

Please specify:
Appendix B: Survey Email

Dear [Name of Subject],

I am inviting you to participate in a research project to study the use of preservation metadata in digital collections by the members of the Association of Research Libraries. At the bottom of this email is a link to a short questionnaire that asks a variety of questions about your institution’s digital collections and use of preservation metadata. I am asking you to look over the questionnaire and, if you choose to do so, complete it. It should take you about 10 minutes to complete.

Preservation metadata is defined as metadata that “supports activities intended to ensure the long-term usability of a digital resource” (Caplan 2009, p. 3). The results of this project will help to determine the level of use of preservation metadata in the digital collections of your institution, and the membership of the Association of Research Libraries as a whole. You were selected to represent [Name of Institution] because it was felt that you were the individual who could most effectively answer the questions in the web survey. I hope that the results of the survey will be useful for librarians working in the fields of digital libraries, metadata, and preservation; and I hope to share my results by presenting them in my Master’s Paper.

I do not know of any risks to you if you decide to participate in this survey and I guarantee that your responses will not be identified with you personally. I promise not to share any information that identifies you with anyone outside my research group which consists of me and my faculty advisor, Jeffrey Pomerantz.

The survey should take you about 10 minutes to complete. I hope you will take the time to complete this questionnaire and return it. Your participation is voluntary [and there is no penalty if you do not participate]. Regardless of whether you choose to participate, please let me know if you would like a summary of my findings.

If you have any questions or concerns about completing the questionnaire or about being in this study, you may contact me at bselleck@email.unc.edu.

Follow this link to the Survey:
[Link to Survey]

Or copy and paste the url below into your internet browser:

[Link to Survey]

Sincerely,
Brody Selleck,
Master’s of Library Science Candidate ‘09
University of North Carolina – Chapel Hill
Link to Survey Fact Sheet:
http://unc.edu/~bselleck/SurveyFactSheet.doc

Reference:

Appendix C: Reminder Email

Dear [Name of Subject],

This email is being set to you as a reminder to please take the time to complete my survey on the use of preservation metadata for digital collections. The survey is very brief, and should not take you more than 5-10 minutes. Your participation is greatly appreciated, and I thank you in advance for taking the time out of your busy day to help me complete my Library Science degree.

I have copied my original email to you below, as well as a link to the survey.

Thank you,

Brody Selleck

[Link to Survey]

Brody Selleck
Master's of Library Science Candidate '09
University of North Carolina--Chapel Hill

--------------------------------------

Original Email:

Dear [Name of Subject],

I am inviting you participate in a research project to study the use of preservation metadata in digital collections by the members of the Association of Research Libraries. At the bottom of this email is a link to a short questionnaire that asks a variety of questions about your institution’s digital collections and use of preservation metadata. I am asking you to look over the questionnaire and, if you choose to do so, complete it. It should take you about 10 minutes to complete.

Preservation metadata is defined as metadata that “supports activities intended to ensure the long-term usability of a digital resource” (Caplan 2009, p. 3). The results of this project will help to determine the level of use of preservation metadata in the digital collections of your institution, and the membership of the Association of Research Libraries as a whole. You were selected to represent the [Name of Institution ] because it was felt that you were the individual who could most effectively answer the questions in the web survey. I hope that the results of the survey will be useful for librarians working in the fields of digital libraries, metadata, and preservation; and I hope to share my results by presenting them in my Master’s Paper.
I do not know of any risks to you if you decide to participate in this survey and I guarantee that your responses will not be identified with you personally. I promise not to share any information that identifies you with anyone outside my research group which consists of me and my faculty advisor, Jeffrey Pomerantz.

The survey should take you about 10 minutes to complete. I hope you will take the time to complete this questionnaire and return it. Your participation is voluntary [and there is no penalty if you do not participate]. Regardless of whether you choose to participate, please let me know if you would like a summary of my findings.

If you have any questions or concerns about completing the questionnaire or about being in this study, you may contact me at bselleck@email.unc.edu.

Follow this link to the Survey:
[Link to Survey]

Or copy and paste the url below into your internet browser:

[Link to Survey]

Sincerely,
Brody Selleck,
Master’s of Library Science Candidate ‘09
University of North Carolina – Chapel Hill

Link to Survey Fact Sheet:
http://unc.edu/~bselleck/SurveyFactSheet.doc

Reference: