Metadata Practices for Digital Collections within Cultural Heritage Institutions: An Investigation of Metadata Strategies Implemented for Enabling Online Content Discovery

A study submitted in partial fulfilment of the requirements for the degree of M. A. in Librarianship

at

THE UNIVERSITY OF SHEFFIELD

by

Muriel Munguia

September 2016
## Contents

List of Tables .................................................................................................................. 4  
Abstract .......................................................................................................................... 5  
Acknowledgements ......................................................................................................... 7  
Chapter One: Introduction & Context ........................................................................... 8  
  1.1 Background ................................................................................................................ 8  
  1.2 Research Aims & Objectives .................................................................................... 11  
  1.3 Structure .................................................................................................................. 12  
Chapter Two: Literature Review .................................................................................... 14  
  2.1 Library, Archive, & Museum User Models and Metadata Practices ...................... 14  
  2.2 Metadata Schemas, Controlled Vocabularies & Their Selection Criteria .............. 16  
  2.3 Metadata Sharing & Interoperability ....................................................................... 19  
  2.4 Collaborations between Cultural Heritage Institutions .......................................... 21  
Chapter Three: Methodology ........................................................................................ 23  
  3.1 Pilot Study ................................................................................................................ 23  
  3.2 Detailed Methodology ............................................................................................... 23  
    3.2.1 Theory ................................................................................................................. 23  
    3.2.2 Sampling ............................................................................................................. 24  
    3.2.3 Collection Method ............................................................................................... 24  
    3.2.4 Coding and Analysis ......................................................................................... 26  
  3.3 Practicalities ............................................................................................................. 27  
  3.4 Ethical Aspects .......................................................................................................... 28  
Chapter Four: Results ..................................................................................................... 29  
  4.1 British Library .......................................................................................................... 29  
    4.1.1 Metadata Standards & their Selection Criteria .................................................... 29  
    4.1.2 Digital Collection Users & Accessibility ............................................................. 32  
    4.1.3 Metadata Sharing & Interoperability .................................................................. 33  
    4.1.4 Collaborating with Other Organisations ............................................................ 34  
    4.1.5 Universal Metadata Schemas and Controlled Vocabularies ............................. 35  
  4.2 The National Archives .............................................................................................. 35  
    4.2.1 Metadata Standards & their Selection Criteria .................................................... 36  
    4.2.2 Digital Collection Users & Accessibility ............................................................. 37  
    4.2.3 Metadata Sharing & Interoperability .................................................................. 39  
    4.2.4 Collaborating with Other Organisations ............................................................ 39
List of Tables

Table 1: Metadata standards used by The British Library.

Table 2: TNA Metadata Fields for Born Digital Collections
Abstract

Background: The public has become more reliant on finding information online and libraries, archives, and museums have striven to enable online accessibility of their digital collections to as many users as possible. Due to the various materials each of these cultural heritage institutions holds they have different item description needs which has led to the development of various metadata schemas and controlled vocabularies.

Aims: This project’s objectives include: discovering the criteria used by each institution to select suitable metadata standards, determining if the selected standards enable users to easily access online collections, examining whether the standards selected can be successfully interchanged between each type of institution, and identifying if an institution’s metadata practices successfully enable metadata sharing and interoperability between them and other cultural heritage institutions.

Method: A sample of metadata professionals from The British Library, The National Archives, and Museums Sheffield were recruited for focus group interviews about the metadata practices within each institution. The qualitative data gathered from the interviews was then used to create case studies for each of the institutions examined.

Discussion: Although the study sample was small there is a good deal of external validity to the general results. Metadata practices used in libraries, archives, and museums vary greatly, but the reasoning behind their implementation is fairly similar. Additionally, all three types of cultural heritage institutions face similar issues when enabling suitable user access, metadata sharing and interoperability, and collaborating with other organisations.
**Conclusion:** It is recommended that museums be examined in more detail and that additional research could also be conducted using alternative data collection methods and recruiting a larger sample from each type of cultural heritage institution. Linked Data and the possible flexibility it could provide should also be examined further.
Acknowledgements

Thank you to my supervisor, Paula Goodale for her guidance and support throughout this project.

Thank you to those from the British Library, The National Archives, and Museums Sheffield who took the time to participate in interviews for this project.

Thank you to my parents, Janel and Gustavo Munguia, and the rest of my family for always supporting me in the pursuit of my goals.
Chapter One: Introduction & Context

This chapter discusses the background of the research topic, and presents the project’s aims and objectives and the structure of this report.

1.1 Background

As use of the internet has grown over the years, cultural heritage institutions such as libraries, archives, and museums, have begun creating digital collections to make resources more accessible to users unable to visit the institution in person. As digital collections continue to grow it has become essential that these institutions implement structures to organise the resources that make up their collections. This can be done by organising resource metadata, a term which Ma shared in her article that the Association for Library Collections and Technical Services (ALCTS) Task Force on Metadata defined as, “structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities” (2006). Due to the various types of resources such as: books, photos, art, motion pictures, audio, public records, and so on, that can make up the collections in various cultural heritage institutions, different digital management standards have been developed to organise digital collections in ways that make their resources easily discoverable for end users.

Metadata schemas and controlled vocabularies have become the common practice for organising and managing the metadata connected to resources in digital collections. A metadata schema is a set of elements and the rules used to determine the type of information associated with each specific element, while a controlled vocabulary is a set of allowable values to be used within particular metadata elements and that have been predetermined based on a collection’s nature and content (Ma, 2006). Implementing controlled vocabularies within specific elements
of an institution’s selected metadata schema ensures that only the approved term is used to describe items within a collection, as opposed to describing items using several different variations of a term. For instance, while working on metadata entry for a digital collection of photographs at a research library, the approved subject heading entered within the ‘Subject’ metadata element for photographs with Native Americans was *Indians of North America*, and then the specific tribe of area of origin if known. Other terms such as Native Americans, or simply Indians, was not permitted. Implementing this strategy ensured that all photographs in a collection which fell into the same subject was made easily assessable to users in one place under the same term.

As previously mentioned, there are a variety of resources collections within libraries, archives and museums might contain, not to mention that some are analogue materials which have been digitised while others are born digital, and should not be described in the same way. As a result, elements within generalised metadata schemas and controlled vocabularies have been modified to more accurately describe certain resources within specific types of collections. For example, a large portion of a library’s collection is comprised of books which are typically organised and described individually, while the archives that make up collections within archival institutions, such as The National Archives, are made up of multiple individual pieces which are described individually and within the context of the archive they are a part of. Similarly, collections in some libraries, particularly large national libraries such as the British Library, are made up of additional objects which cannot be accurately described with the same method used to describe books. Due to the differences between these materials and their description needs, the elements within generalised metadata schemas have not only been modified, but
separate metadata schemas, as well as various controlled vocabularies, have also been developed.

As metadata schemas and controlled vocabularies continue to be modified and redeveloped to best suit the needs of their institutions, the differences in their functionality can sometimes make it difficult for users to find what they are looking for. While some institutions use general metadata schemas and controlled vocabularies recognisable to other cataloguing and metadata professionals to organise and manage their digital collections, others are developing metadata elements and standards that are specific to their collections, which can hinder discoverability. The variety of metadata schemas and controlled vocabularies used across all types of cultural heritage institutions creates issues, not only with user discoverability, but interoperability and metadata sharing as well. In the past some institutions within the libraries, archives, and museums (LAM) community have attempted to create schemas and vocabularies that can be used across all types of cultural heritage institutions, but without much success (Mitchell, 2013). The wide variety of resources within collections and different user needs, create complex metadata requirements within the various types of cultural heritage institutions, which can lead to challenges for institutions when attempting to select appropriate metadata schemas and controlled vocabularies. These issues are why, when decisions are being made regarding metadata strategies, “the users and uses of the resulting digital materials” need to be carefully considered, and the resulting strategy must be communicated to not only the intended user community, but all users, so they are aware of the reliability of the particular digital collection (Trant, 2009).
1.2 Research Aims & Objectives

The aim of the research to be conducted for this study is to identify what criteria libraries, archives, and museums use to determine which metadata schemas and controlled vocabularies to use when creating their digital collections and then investigate whether each selected schema can be used interchangeably between the different types of collections at each institution, or if it is only suitable for the type of collection for which it was implemented. After an exploration of digital collections created by various cultural heritage institutions within the United Kingdom, the British Library, The National Archives, and Museums Sheffield were selected as the institutions to be examined due to the accessibility of their digital collections and the availability of appropriate interview participants.

The objectives of this project are:

- To discover the criteria used by each institution to determine which metadata schema and controlled vocabulary to use.
- To determine if the selected schema and vocabulary enable novice and/or expert users to easily find desired information or specific digital objects.
- To examine whether the schemas and vocabularies selected could be successfully used interchangeably between collections within each of the selected institutions.
- To identify if the metadata practices within each institution successfully enables metadata sharing and interoperability between them and other cultural heritage institutions.
1.3 Structure

This dissertation is organised into the following chapters:

- Chapter 1: Introduction, provides context for the research conducted and presents the aims and objectives of the project, as well as the arrangement of this dissertation.

- Chapter 2: Literature Review, discusses and analyses the available literature relevant to the area of study. Includes topics such as: the differences between libraries, archives, museums and their metadata practices, types of metadata schemas and controlled vocabularies and the types of collections they are generally associated with, and the best practices and advantages of metadata sharing and interoperability among cultural heritage institutions.

- Chapter 3: Methodology, summarises data collection methods. Participant recruitment is discussed and the design and implementation of the qualitative data collection method and focus group interviews is explained.

- Chapter 4: Results, presents the library, archive, and museum case study results from interviews with The British Library, The National Archives, and Museums Sheffield.

- Chapter 5: Discussion, discusses the qualitative research results presented within each case study and the themes which have emerged from the analysis of these results.

- Chapter 6: Conclusion, sums up the findings from the research for this dissertation. The study’s limitations are discussed, recommendations regarding the research results are made, and areas for further research are identified.
The remaining sections are References, which lists all cited works, and Appendices, which includes the ethics application form, participant information sheet and consent form, the ethics approval letter, the participant codes, the interview questions used for data collection, and a table of results constructed as part of the data analysis.
Chapter Two: Literature Review

There is a great deal of literature about the use of metadata in various types of institutions, but for the purpose of this study the literature examined discusses the types of metadata schemas and controlled vocabularies implemented within the digital collections of different types of cultural heritage institutions, as well as how or why they were selected, and whether or not they may be compatible with collections at other institutions. Additionally, the literature examined also discusses various methods of metadata sharing and interoperability between institutions. When searching the literature very little was found discussing the compatible use of particular metadata schemas or controlled vocabularies between libraries, archives, and museums; as a matter of fact a great deal of the literature mainly focused on metadata usage within libraries, but a good portion did discuss the practice of metadata sharing and interoperability between different cultural heritage institutions. Furthermore, the majority of the literature found consisted of research conducted within and about libraries and other cultural heritage institutions specifically within the United States. While some did talk about a few large projects taking place outside the US, such as Europeana, not a large amount of relevant literature was found about studies focused solely in the UK, or other parts of Europe.

2.1 Library, Archive, & Museum User Models and Metadata Practices

In order to understand the challenges behind the development of a metadata schema and controlled vocabulary universally used in libraries, archives, and museums, one must first understand the differences between these institutions, as well as the differences in their metadata use and development practices. As Trant (2009) points out, different types of libraries have different audiences. While research libraries are predominately used by students and scholars, the public
library’s audience is significantly broader. As a result, “the librarian is an enabler in the discovery phase of the research process,” and “[as] repositories of published material that exists in multiple copies … libraries have developed sophisticated systems for helping users find specific resources that correspond to their interests” (Trant, 2009, pg. 370). Alternatively, archive collections represent the organisations within which they are created. “Archival materials provide unique evidence of the transactions of organisations and institutions; collections do not circulate,” and because collections are often organised in whatever order they’ve been given by their creators, the archival staff’s job is to aid researchers in navigating collection finding aids and identifying records relevant to their inquiry (Trant, 2009, p.370-371).

Museums, on the other hand can be a completely different experience. Whereas libraries and archives tend to be used by individuals for a specific purpose, museums are often a social activity users participate in for entertainment or general education on a topic of interest, and rather than examining specific individual pieces of a collection and receiving personal guidance from curators, users visit museums to experience entire exhibits and are guided through exhibits by labels that identify and describe each piece (Trant, 2009).

Eulalia Roel (2005) concisely describes the metadata entry practices within each of these three types of institutions, as well as the reasons behind them, and the strengths and weaknesses behind these practices. For example, she explains that libraries traditionally assign general metadata at the collection item level, which allows metadata to be assigned to a higher percentage of items, but only a very minimal amount for each resource (Roel, 2005). This means, when searching a library’s online catalogue, users are usually only provided with the general publishing and classification information for each item, in addition to its location.
within the library. On the other hand, due to the complexity of archives, detailed item descriptions are highly valued and provide users with a contextualized background for each item, all the way down through each sub-collection, as well as how each item is related to the other pieces within each individual collection. While this is a significant benefit of archival metadata practices, it also permits metadata to be entered inconsistently and at different levels of quality (Roel, 2005). These types of discrepancies can not only cause access and interoperability issues across the three discussed types of cultural heritage institutions, but even just among the archive community. According to Roel (2005), while libraries and archives provide metadata about their collections for the benefit of their users, in museums, “[i]t is common... for curatorial staff to view metadata as intellectual property to which they serve as gatekeepers, reflecting a professional value placed upon contextualizing materials for users.” She reasons that museums don’t typically make their collection metadata as openly available, not just because they aren’t funded to provide access to that information, but because they see it as something they can use to capitalize on, only making it accessible to users who visit their exhibits. The main flaw with this assumption is that there are a significant number of museums, specifically national and local museums within the UK, which provide free admission.

### 2.2 Metadata Schemas, Controlled Vocabularies & Their Selection Criteria

When planning the construction of a digital collection there are various metadata schemas and controlled vocabularies institutions can choose from to best suit their needs. The variety of schema elements and vocabularies were developed because of differences in each institution’s “culture of description” which creates different needs to consider when selecting appropriate metadata schema elements and a controlled vocabulary (Palmer, Zavalina, and Mustafoff, 2007). In addition,
some institutions decide to use more than one type metadata schema, mixing and matching schema elements, or they develop their own local schema or controlled vocabulary. In Ma’s study investigating metadata implementation within Association or Research Libraries (ARL) libraries, MARC was found to be the most (91%) used metadata schema among her respondents, with Encoded Archival Description (EAD) being the second (84%) and Dublin Core (DC) being the third (78%) most used schemas (2009). In a similar study by Park and Tosaka (2010), which examined metadata creation practices for digitized and born digital resources within digital repositories, collections, and libraries, responses from 84.2% of participating institutions again showed MARC as the most widely used schema, with Dublin Core (25.4% Unqualified DC, 40.6% Qualified DC) and EAD (31.7%) again being the next most popular choices. Additionally, Park and Tosaka’s study showed that while Library of Congress Subject Headings (LCSH) was the most frequently used controlled vocabulary across all types of collection resources, the domain-specific terminology of Art and Architecture Thesaurus (AAT) was also highly used within digital collections containing non-print resources such as images, cultural objects, and archival materials (2010). A few things that were noted in the examination of these studies were that Ma’s study focused solely on research libraries and that while Park and Tosaka’s study was generally aimed at cataloguing and metadata professionals working with a variety of resources, their survey was distributed to mailing lists largely directed towards libraries and at an American Library Association (ALA) meeting. If both studies had been designed to also include archives and museums, it may be possible that the results would have been more varied and shown higher usage of metadata schemas and controlled vocabularies that had been designed with archive or museum collections specifically in mind.
According to Park and Tosaka’s study, the criteria used within cultural heritage institutions for selecting metadata schemas and controlled vocabularies follow similar patterns. For example, the types of resources that make up an institution’s digital collection, the nature of the collection, and the target audience were shown to be the top criteria when selecting an appropriate metadata schema and controlled vocabulary; while the institution’s technical infrastructure, resource constraints, and staff expertise were also important factors, they were not as highly considered (Park and Tosaka, 2010). The fact that these factors were the top three criteria in selecting an appropriate metadata schema or controlled vocabulary illustrates the various resource description needs within libraries, archives, and museums, as well as why a variety of schemas and vocabularies have been developed to describe those resources. Additionally, the reason these were considered the top three criteria could arguably be because each collection is unique and, as Skekel explained, “someone viewing a collection of art objects will search differently than someone exploring historical documents,” which supports the assertion that a collection should be organized in a way that is easily searchable for users (2008). At the same time, while different institutions having different metadata and controlled vocabulary needs is perfectly reasonable, it could also be helpful for cataloguers and other metadata entry professionals, as well as users, if the same data format was used across all types of cultural heritage institutions. To use an example by Park and Tosaka, using different standards when inputting a person’s name (e.g., last name, first name, middle initial or last name, first and middle name initials) creates authority control consistency issues, making metadata interoperability between institutions more difficult (2010).
2.3 Metadata Sharing & Interoperability

In a study, Ma stated that “metadata interoperability is, of essence, to facilitate the exchange and sharing of information and to enable cross-domain searching” (2009). At the time of their study a year later though Park and Tosaka revealed that metadata interoperability was still a challenge, even with the growing awareness of its importance. Half of their survey respondents indicated that sharing metadata with OAI (Open Archives Initiative) harvesters, search engines, union catalogues, etc. was not a priority due to financial, personnel, and technical constraints (2010). One of the issues, as Palmer et al. point out, is that in order for institutions to successfully take part in the Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) their metadata would need to be formatted according to the Dublin Core schema, which may not meet their collection’s or institution’s needs (2007). If OAI-PMH were able to work with additional types of schemas, or if the Dublin Core schema was further developed to include description elements and standards that could meet the needs of more collections than it currently does, metadata sharing could potentially become more successful.

Another issue was that institutions with digital collections containing locally created metadata elements developed to meet their descriptive and user needs, created additional difficulty with interoperability between institutions due to there not being sharable mechanisms available for the specialized elements they created (Park and Tosaka, 2010). As explained by Shreeves, Riley, and Milewicz (2006), even if an institution’s metadata is considered to be of the highest quality, once it has been removed from its local context, or if it is too complex or “hand-crafted,” it won’t support interoperability with other institutions. It is for this reason Shreeves et al. (2006) have identified six criteria: content, consistency, coherence, context,
communication, and conformance to standards, which they recommend institutions meet in order to enable metadata sharing and interoperability. That being said, while interoperability challenges are a concern, cultural heritage institutions need to keep in mind that they must strike a balance between functional interoperability with other institutions and easy accessibility for their own users (Palmer et al., 2007). Although it may be challenging, institutions cannot sacrifice the development of one for the benefit of the other.

Mitchell explains that, “while part of the LAM (libraries, archives, and museums) profession focuses on literacy and information engagement issues, the metadata community typically focuses on exploring how information systems and structures support these needs and enable cross-community and cross-repository access and aggregation,” and this is because as cultural heritage institutions continue to digitize more of their collections the needs for appropriate cataloguing and archival processing frameworks and management systems has grown (2013). Mitchell also discusses how institutions within the LAM community have explored using MARC in archival and museum settings in the past with standards such as MARC Format for Archival and Manuscript Control (MARC AMC), MARC Archives, Personal Papers, and Manuscripts (APPM), MARC for Visual Materials (MARC VM), all to mixed success while other cataloguing models and dissemination methods such as Cataloguing Cultural Objects (CCO), Encoded Archival Description (EAD), Metadata Encoding and Transmission Standard (METS), and Describing Archives, A Content Standard (DACS), were developed at the same time and ultimately gained more popularity.

While Mitchell’s book does discuss the afore mentioned points, its main focus is the implementation and use of Linked Open Data (LOD) as a metadata
model that will provide cross-community and cross-repository access to digital information resources (2013). Mitchell explains that with LOD, data open and free from legal and copyright restrictions would be readily able to connect with related information when published on the web, thus making it as accessible to computers as it is to humans (2013). Furthermore, Mitchell describes Linked Open Vocabularies (LOV), with which vocabularies such as term lists, metadata schemas, taxonomies, and ontologies would be published instead of data, would then enable the sharing of data using well-known and copyright-free structures and concepts within the cultural heritage institution community (2013). Although LOD and LOV are relatively new developments within metadata sharing and interoperability, they should not be overlooked as viable options.

### 2.4 Collaborations between Cultural Heritage Institutions

When making decisions regarding metadata practices, libraries, archives, and museums must also consider the types of institutions they are participating in collaborative projects with, or that they might collaborate with in the future. As previously discussed, due to the varying metadata authoring practices within cultural heritage institutions, the discoverability and usefulness of a resource’s metadata is highly impacted (Shreeves, Kaczmarek, and Cole 2003, p.162). Waibel and Erway (2009) point out that libraries, archives, and museums emerged from the cabinets of curiosities put together by seventeenth and eighteenth century scholars who did not separate their collections into library books, archival records, or museum objects, and that by unifying access to these institution’s materials, an interconnected knowledge base can once again be recreated within today’s networked environment. They also mention that, in contrast to coordination, “[c]ollaboration changes behaviours, processes, and organizational structures, and leads to a fundamental
interconnectedness and interdependence among partners…” (Waibel and Erway, 2009). This is a significant observation because if the metadata practices which have been negotiated, agreed upon, and implemented for collaborative projects between differing cultural heritage institutions are evaluated, they could help inform the possible development of a universally suitable metadata schema or controlled vocabulary. Furthermore, Waibel and Erway’s article also explains that developing and implementing metadata practices which help cultivate lasting partnerships rather than temporary ones between different types of cultural heritage institutions can result in a state of ‘convergence’ where “collaboration has become so extensive, engrained, and assumed that it is no longer recognized as a collaborative undertaking” (2009).

Conclusion

As indicated by the literature above there has been a great deal of study regarding the use of metadata schemas and controlled vocabularies in the ever growing digital collections at cultural heritage institutions. Even as new developments emerge, metadata and cataloguing professionals continue struggling to find a controlled vocabulary or metadata schema with element standards appropriate for the variety of resources held within different institutions in order to provide cataloguing and metadata professionals a single reference source and easier accessibility for users. Although no single schema or vocabulary has been found, the studies examined do discuss other viable options that have been explored and continue to develop, such as metadata interoperability between institutions, metadata harvesting and sharing tools, and linked open data and vocabularies, as well as possibilities which could emerge from collaborative projects between institutions.
Chapter Three: Methodology

This study consisted of qualitative research which included semi-structured focus group interviews that were used to construct case studies for three cultural heritage institutions: The British Library, The National Archives, and Museums Sheffield.

3.1 Pilot Study

A pilot interview was conducted with students in the Information School at the University of Sheffield to determine whether the questions created for the interview facilitated responses relevant to the research aims and objectives. As addressed by Bryman (2012) piloting the interview questions also provides the interviewer with some practice, which can result in higher confidence when conducting the real interviews, as well as enable the interviewer to determine if the order of the questions provide a good flow to the interview. Although the subjects of the pilot interview were not especially familiar with the content of the study, they possessed some basic knowledge on the subject matter and provided helpful feedback. According to the feedback received from the pilot subjects, the questions regarding user needs, interoperability, and the development of a universal metadata schema and controlled vocabulary were adjusted to more easily encourage discussion and elaboration, and to avoid leading the participants into providing particular answers.

3.2 Detailed Methodology

3.2.1 Theory

This study consisted of qualitative research, which utilized the grounded theory approach, with the intention of conducting case studies
within three cultural heritage institutions, a library, archive, and museum. Using grounded theory allowed the collected and analysed data to then develop ideas which addressed the research aims and objectives (Bryman, 2012, p. 387). Furthermore, the study also utilized the “publications-as-theory strategy” by which used the results of previous studies to help determine the study’s focus and then used the research results to address and expand on the issues identified within the literature examined as part of the study (Bryman, 2012, p. 22).

3.2.2 Sampling

To ensure appropriate participants such as professionals involved in the metadata development, implementations, and/or management process within their respective institution were recruited for the study, nonprobability purposive sampling was used. This sampling strategy ensured the individuals interviewed were experts on the issue being explored and would be able to accurately answer the questions asked (Flick, 2011). Furthermore, the hope in using this strategy was if anyone initially contacted was unable to take part they would refer other available professionals so the small focus group within each institution might then snowball into similar sizes of two to four people.

3.2.3 Collection Method

According to Morgan (1997) one of the ways focus groups are useful is the way they enable the evaluation of different research sites or study populations (p.11?). This makes it a suitable method of data collection for the case studies conducted within each type of institution and will provide suitable data with which to compare their metadata practices, as well as compare this study’s results to those from previous studies. Using this data
collection method to evaluate the metadata practices within the British Library, The National Archives, and Museums Sheffield will allow a single group to be taken as a unit and then compared to other groups (Flick, 2014, p. 252). As Flick goes on the explain, “[c]omparison then focuses on the topics mentioned, the variety of attitudes towards these topics among the members in the group, the stages the discussion ran through, and the results of the discussion in each group” (2014, p. 252). This leads to the other reason for selecting focus group interviews as the method of data collection: to generate conversation between participants about their views on the available metadata schemas and controlled vocabularies used within their home institutions and other cultural heritage institutions, and how they may or may not suit those institutions and their online collections.

Once a few basic interview questions were developed based on questions Park and Tosaka (2010) created for their study, I had my dissertation supervisor look them over, fine-tuned them according to the feedback she provided, and then developed a couple additional probing questions. The study was then piloted, as previously described, and the questions were again adjusted according to the feedback received. After confirming who would be participating in the focus groups at each institution each participant was provided with a combined information sheet and consent form to ensure consent was given on whether or not their names, job titles, or general area of work would be allowed to be used in the results of the study. Once all questions were finalized, and participants were confirmed, a semi-structured focus group interview was conducted at each institution. While two participants from The British Library and three from The National Archives
were successfully recruited, after countless emails to museums of various sizes, recruiting professionals for the museum case study became a challenge. Eventually one participant was recruited from Museums Sheffield, and to supplement the interview conducted with that person, a brief analysis of Museums Sheffield’s website was also carried out.

### 3.2.4 Coding and Analysis

The interviews were recorded by iPhone, and after each interview was transcribed, thematic analysis was used to extract data which addressed the research aims and objectives. As suggested by Bryman (2012) the thematic analysis was carried out by constructing a matrix made up of central themes and organising the data collected within those themes (p.579). The central themes used within the matrix were determined based on the research aims and objectives, and the matrix was created using Excel. Any additional information which corresponded to data within the matrix was organised in a Word document and colour coded by institution. Additionally, interviewees were also coded according to the institution they represented. These codes can be found in Appendix 2.

The results of the analysis have been presented in the form of a case study for each institution. More specifically, a descriptive case study was constructed for each institution examined, which resulted in the comparative case method also being used to analyse results in order to draw cross-case conclusions. The decision was made to present results as case studies due to their ability to illuminate the reasoning behind a decision, or decisions, and describe them within the case’s “real world context” (Yin, 2014, p.15). Finally, the results of each case study are discussed and recommendations for
improvements and further research in the field have been presented in this dissertation, and while the results of this study apply only to the participating institutions, they do present some external validity, and a number of analytical generalizations have been made.

3.3 Practicalities

The main success factor for this study was the willingness and availability of potential interviewees to participate. The intention was to conduct three interviews with two to three professionals from three different cultural heritage institutions. The most difficult part was attracting participants from a museum to participate in an interview. Numerous museums were contacted, some of which responded that they were unable to participate, and many from which no response was received. As addressed above, an interview was eventually organised at Museums Sheffield, although it was with a single participant rather than two or three, as desired.

Contact with potential interviewees at the British Library, The National Archives, and various national museums began at the end of May, and after receiving ethics approval on the 3rd of June, interviews were arranged with those who had responded. The interviews with the British Library and The National Archives were scheduled for the first week of July, and the remainder of June was spent continuing to contact museums, developing appropriate interview questions and working on the first three dissertation chapters, which included the introduction, aims and objectives, literature review, and methodology. Between the interviews that took place in early July, and the one which eventually took place at the end of July, the remainder of the month was spent transcribing and coding the interviews and then analysing the data before spending August writing the case studies, discussion, and conclusion sections.
3.4 Ethical Aspects

This study was classified as low risk due to the involvement of adult human participants and because of the personal data collected. Participants were asked to take part in an interview, lasting no more than an hour, about the metadata schema and controlled vocabulary practices for digital collections within their institutions, and the reasons behind them. The resulting data from the interviews was stored on the Information School’s research data drive which was only accessible by me, my supervisor, the school’s Examinations Officer, and ICT staff operating the facility. A password protected back up copy was also stored on my personal laptop.

The only identifying information used in the project is the names of the institutions where the participants are employed. Participants have been coded according to their institution of employment. The names and job titles identifying participants who granted consent to do so may be found in Appendix 2. That being said, due to the small size of the pool of professionals suitable to participate in this study, there is no guarantee that the participants are unidentifiable, even without granting permission to use their name or job title. Participants who granted permission to use their name or job title were provided with a transcript of the interview and allowed to retract any statements they did not wish to be published. All data will be deleted upon completion of the dissertation, but data may also be used in any presentations or publications arising from this dissertation.
Chapter Four: Results

Here the results from the interviews with the British Library, The National Archives, and Museums Sheffield, along with a detailed content analysis of each of their websites and digital collections, are presented as case individual studies for each institution.

4.1 British Library

The British Library is the national library of the United Kingdom and the largest library in the world with collections consisting of print and e-books, e-journals and articles, maps, scores, audio recordings, images, data, and moving images. The British Library is also one of the six legal deposit libraries where one copy of every book or article published within the UK must be deposited. As a national and a legal deposit library it has been essential for the British Library to become a leader in digital preservation, accessibility, and management and remain up to speed with the technological advancements available for making their collections available online.

4.1.1 Metadata Standards & their Selection Criteria

Due to the vast array of material which make up the British Library’s collections, a variety of metadata schemas and controlled vocabularies are used to appropriately organise, describe, and manage each type of material as seen in Table 1. Some of these schemas, vocabularies, and standards can be used with different parts of the library’s collections, which might have various types of materials, while others have been adopted for use within specific collections.
<table>
<thead>
<tr>
<th>Metadata Schemas</th>
<th>Controlled Vocabularies</th>
<th>Other Descriptive Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARC (MAchine-Readable Cataloguing)</td>
<td>LOC (Library of Congress) name authority files</td>
<td>RDA (Resource Description and Access)</td>
</tr>
<tr>
<td>SAMI (Sound and Moving Image) MARC</td>
<td>PCC (Program for Cooperative Cataloguing) name authority files</td>
<td>ISAD(G) (General International Standard Archival Description)</td>
</tr>
<tr>
<td>MODS (Metadata Object Description Schema)</td>
<td>NACO (Name Authority Cooperative Program) name authority files</td>
<td>DCRM (Descriptive Cataloguing of Rare Materials)</td>
</tr>
<tr>
<td>METS (Metadata Encoding and Transmission Standards)</td>
<td>LCSH (Library of Congress Subject Headings)</td>
<td>TEI (Text Encoding Initiative)</td>
</tr>
<tr>
<td>Dublin Core</td>
<td>Dewey Classification</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UKAT (UK Archival Thesaurus)</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Metadata standards used by The British Library.

For example, RDA is used for published material, while ISAD(G) is for unpublished material, and while DCRM is used for rare books it is used as a supplement to RDA. Additionally, as new standards are developed that may better suit the needs of the library they may be adopted and replace standards currently being used such as LCSH and UKAT, which according to BL2, may eventually be superseded by FAST (Faceted Application of Subject Terminology). Another important point is that although these are the standards used within the British Library, other organisations from which they might receive some metadata might use different standards, which then need to be translated in order to be compatible with the library’s systems. For instance, metadata from outside suppliers might be in NLM (National Library of Medicine), DTD (Document Type Definition), or JATS (Journal Publishing Tag Suite), which is then generally converted into MODS for the article level material, or from ONIX (ONline Information eXchange) into MARC. Similarly, when the library sends metadata to PRIMO for discovery
the metadata is converted to PNX (PRIMO’s Normalised XML) which is essentially Qualified Dublin Core.

The aforementioned standards were selected by the British Library for various reasons. According to BL2 the continuity and workflows already in place are significant factors in the selection process because “use of ALEPH Library Management System determines that MARC will be used, mostly for published material. Unpublished items go into IAMS (Integrated Archives and Manuscripts System), which can cope with EAD (Encoded Archival Description) EAC (Encoded Archival Context), but uses an internal schema slightly richer than EAD.” BL1 explains further that, “if certain communities [we] want to work with use particular standards then [we] have to be able to produce data in those formats or convert data from those external formats to ones used internally, for example, ONIX to MARC, as Alan mentioned, and developing a way to work with the DCAT (Data Catalogue Vocabulary) format.” Additionally, the value and interoperability of items in terms of collaboration with other parties and efficiency are also important. When deciding whether or not to adopt RDA, a driver in the decision was whether or not Library of Congress would adopt it, and the use of NACO authority file was adopted because of its use in North America. Adopting standards also being used by other institutions not only enables more efficient workflows, but allows interoperability between institutions, thus allowing further accessibility for users.
4.1.2 Digital Collection Users & Accessibility

As BL1 explains:

“… the cluster of users we’ve got … really can be broken down to the kind of group type users who are people like the … commercial organisations who want to take metadata from us, … the individual researchers who might want to use it, … in the case of linked data there are possibly staff developers who might want to consume the viable data that we acquire … there are people who might want to take the data dumps that we provide as well, metadata dumps which are either available from … the BL labs team who also offer content dumps as well … in terms of user groups we have all kinds of different types of users who come to us from you know traditional … library and archive domain users to, who are wanting differing formats again that they can consume in their systems, to individual researchers for whom we’ve got things like the new CSV style format that we come up with so you can open a new file with Excel, do visualisations or just simply research. Other people want simply to have the metadata in formats they can use in a citation manager so like an internet website, or something like that … I mean obviously there are people who have been using our catalogues and things as well …”

A few specific examples of internal and external users of the library’s collection metadata include: the library’s curatorial, web, and licensing teams, reading room users, funders, partners, other national and international library and archives communities, the UK government, and of course the general public.

In order to make online collections and the metadata connected to them accessible to all types of users, as much communication as possible takes place with users to ensure data is as easy to consume as possible.
Efforts are made to ensure metadata for print and e-books have similar formats with identical characteristics despite having different elements, and licensing such as creative commons is being implemented to eliminate boundaries and allow users to use the available data in various ways. Furthermore, indexes are also available with which users can search both authorised and non-authorised formats of authors, titles, etc., which can be a better alternative to conducting a search because browsing within PRIMO is limited to the list of search results. That being said the availability of indexes is determined by search interface capabilities as much as the metadata so while excellent metadata might be entered, it isn’t always taken advantage of.

4.1.3 Metadata Sharing & Interoperability

The British Library enables interoperability between their collections and those at other institutions by using standards according to the communities they’re working with such as MARC with other libraries, EAD with archives, or CSV formats for individual researchers. Additionally, the library ensures that their internal systems are capable of translating external metadata formats to internal ones. The library also shares their metadata with OCLC’s Worldcat, UK COPAC, etc. so that it can be made available to be harvested by Google and other search engines and lead users to the content on the library’s website.

As with all institutional procedures within constantly evolving fields, interoperability and metadata sharing practices have their own issues as well. For instance, when working with other organisations the British Library has to ensure that the agreed upon terms for use of the resulting product and data allow them to perform the needed functions, in other words, the licensing
must enable user access to the material. Another issue is that as new standards continue to be developed, it may become necessary to replace old ones. In this instance the library is currently looking into the possibility of replacing LCSH, which is difficult to train people to apply and only benefits from the strings if a left-anchored browse index is used, with FAST which is faceted and easier to train people to apply because there aren’t a lot of syntax rules. Finally, another big issue is the incompatibility of the library’s internal standards or systems with those used at other institutions, mostly those with collections that only consist of unique items such as galleries and museums.

4.1.4 Collaborating with Other Organisations

The British Library takes part in collaborations with various organisations such as Google’s 19th century book project, Europeana’s World War I project and various projects with commercial organisations who want pictures and illustration from books to be digitised so they can put them on merchandise such as skateboards, t-shirts, ties, totes, etc. These collaborations are important because not only do they enable accessibility, but they allow the library to identify and address possible issues with standards or systems used as part of their collection metadata procedures. In their collaboration with Google Books the British Library supplies them with mostly non-English 19th century publications, which are out of copyright, along with the metadata attached to them. That material is then returned to the British Library with the digitised copies and any additional information provided by Google and the new data is entered into ALEPH where a new record is generated for the digital surrogate of each item, even if the data matches the record for the printed version of an item. This process makes it possible for
more material to be catalogued that would otherwise take much longer due to the library not having enough resources to get through it quickly on their own.

As beneficial as collaborative projects might be there are a variety of issues that can arise if not addressed from the outset. Some of these issues might be related to licensing and copyright, efficiency, inconsistencies with historical metadata, or whether or not the metadata meets the needs of the external companies involved. Having been through various projects, the British Library has been able to become aware of these possible pitfalls and developed processes around them.

4.1.5 Universal Metadata Schemas and Controlled Vocabularies

BL1 and BL2 said that a universal metadata schema or controlled vocabulary wouldn’t be practical or successful due to different representations of metadata being needed within different domains and because there would be no way universal metadata control could be enforced. They then went on to explain that a good alternative would be the creation of triples, codified statements about semantic data, which allow information to be disassembled and then reassembled in forms required for different institutions. Additionally, they also described the possibilities behind linked data registries which hold an infinite number of labels for items and in a variety of languages.

4.2 The National Archives

The National Archives is the UK government’s official archive, publisher, and guardian of over 1,000 years of well-known national documents, and as a leading cultural heritage institution and leader in the archive sector, the archive works to secure the future of physical and digital records, in addition to acting as advisors in
information and records management (The National Archives, 2016). The archive collects and preserves government records, “from Shakespeare’s will to tweets from Downing Street”, in order to make them as accessible and available as possible for current users and future generations (The National Archives, 2016). In addition, the archive also manages Crown Copyright on behalf of the government and assists various communities with research.

The archives online catalogue, Discovery, currently holds more than 32 million descriptions of records held not only by The National Archives, but by more than 2,500 other UK archives as well (The National Archives, 2016). Over 9 million records are available for download, research guides are provided for various subjects, and users are given the ability to tag records within the collections. Popular searches include: World War I army service records, escape and evasion reports, passenger lists, royal and navy service records, wills, RAF combat reports, and much more.

4.2.1 Metadata Standards & their Selection Criteria

The National Archives has developed an in-house cataloguing system which uses the NCA (National Council on Archive Rules), and a core cell of Dublin Core fields in addition to TNA specific ones. These fields are then mapped to ISAD(G) fields when exported to their online catalogue Discovery. The in-house system was
<table>
<thead>
<tr>
<th>Required (mapped field in Discovery)</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Identifier (Arrangement)</td>
</tr>
<tr>
<td>✓ File Name (Title)</td>
</tr>
<tr>
<td>✓ Folder* - referenced in Arrangement field</td>
</tr>
<tr>
<td>✓ Date Last Modified (Date)</td>
</tr>
<tr>
<td>✓ Checksum* - used only for preservation/provenance purposes</td>
</tr>
<tr>
<td>✓ Legal Status (Legal Status)</td>
</tr>
<tr>
<td>✓ Rights/Copyright* - sometimes displayed in Restrictions on Use</td>
</tr>
<tr>
<td>✓ Held By (Held By)</td>
</tr>
<tr>
<td>✓ Closure Type (Access Condition/Closure Status)</td>
</tr>
<tr>
<td>✓ Closure Period (Access Conditions/Record Opening Date)</td>
</tr>
<tr>
<td>✓ FOI Exemption Code* - users only see a record is closed &amp; its opening date</td>
</tr>
<tr>
<td>✓ FOI Exemption Asserted (FOI Decision Date)</td>
</tr>
<tr>
<td>✓ Description Public (Description/Closure Status)</td>
</tr>
<tr>
<td>✓ Description Alternate (Description, closed/sensitive folder names can alternatively be mapped to Arrangement)</td>
</tr>
</tbody>
</table>

*not displayed in Discovery

<table>
<thead>
<tr>
<th>Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ Description (Description)</td>
</tr>
<tr>
<td>✓ Start Date (Date)</td>
</tr>
<tr>
<td>✓ End Date (Date)</td>
</tr>
</tbody>
</table>

Table 2 TNA Metadata Fields for Born Digital Collections

developed to require the bare minimum while still enabling preservation and providing context and provenance and this was done because while some countries such as Norway, Finland, and Sweden have government metadata and export standards, there are no such standards in the UK. This means that a certain level of flexibility is needed when importing records from government departments. Metadata is imported from government departments in CSV files and in addition to the requested fields which provide the minimal amount of information necessary, additional information is accepted for additional fields that might add context and help to make records more findable.

### 4.2.2 Digital Collection Users & Accessibility

Not a great deal is known about the user base other than it is split into users who access the collections just once, and those, such as historians or
students, who are more knowledgeable about what is accessible in the collection and how to find it. Due to so little being known about the user base, little is also known about what they need or want. Furthermore, while the archive does have access to statistics about which records are looked at and how often, there is no way to know if the record was what the user wanted or was looking for. When the digital transfer process first began users were asked what they wanted to see but the majority didn’t know what they wanted. According to TNA3:

“our aim [at the moment] is to get as much stuff in and out onto Discovery so that … we have a pool that’s big enough of digital records … for people to start noticing, and start using, and then generating some feedback because at the moment it’s mostly from internal staff that we get feedback … but that’s not really what we’re looking for. We’re really looking for users to say, and not just about the metadata, but the presentation formats that we’re using, how easy is it for them to download stuff, you know is this what they’re expecting, or are they expecting to be able to just view it, are they expecting to be able to stream video, like we don’t know. We would just be guessing. So we’re kind of, we’re working within … the paper worldview in terms of Discovery and how it was designed, …that’s fine for now, so we can get some stuff through and you know we can just see and learn…”

An example of feedback they have received from users is that when they started closure information wasn’t displayed until users let them know they wanted to see it, they wanted to know why a record might be closed for x amount of time.
4.2.3 Metadata Sharing & Interoperability

Metadata sharing and interoperability practices within The National Archives are made available through the APIs (Application Program Interfaces). The archives also allow the export of Discovery search results into XML or CSV formats. In the past users were only able to export 1,000 files, but that was very recently increased to 10,000 to allow users to examine more of the material available. Additionally, Discovery acts as one portal for digitised, born digital, paper, and other materials from the catalogues of hundreds of other archives in the UK.

4.2.4 Collaborating with Other Organisations

The National Archives has been working with central government departments for years, and as the deposit for government records obtaining government records and making them accessible to the public is their main priority. Working with the government can create some challenges because the archive isn’t in a position to be demanding information from them, and this can lead to accuracy issues with the metadata they are provided.

Although the archive has taken part in some digital preservation projects, such as PLANETS (Preservation and Long-term Access through Networked Services), due to its partnership with government departments it isn’t considered practical for the archive to buy into the various digital preservation standards available. They do still pay attention and try to contribute and participate where they can, but because they aren’t in a position to demand certain types of information for the records they obtain, adopting those standards or cooperating with PREMIS or METS isn’t useful.
Aside from that, other collaboration issues, particularly regarding government records, include ensuring the dates given are accurate and that the core set of metadata elements are accurate for their designated collections. Although ‘date last modified’ may be the least corrupted date, it’s not always indicative of the date of closure, or the date of the record which is what they set closure against. They need a date that’s as accurate as possible because it can affect other metadata fields being collected so sometimes date ranges, or another date that more accurately reflects the record, is used instead. Then, with special collections the core set of metadata elements sometimes need to be evaluated. For example, the collection of UK Supreme Court videos was one of the first large AV collections, and the core metadata used for born digital records didn’t work for the collection. So two people on the digital preservation team examined the core metadata needs and discussed which metadata elements were necessary with people at the UK Supreme Court to determine if they would be possible to export. They also needed to ensure that the systems used by government departments are capable of exporting the data need; for instance, audit trails are huge pieces of metadata which aren’t always feasible to extract from a system.

4.2.5 Universal Metadata Schemas and Controlled Vocabularies

Interviewees at The National Archives agreed that a one-fits-all model metadata schema or controlled vocabulary to be used in libraries, archives, and museums wouldn’t work. They argue that standardisation fails to acknowledge restrictions some institutions might face and that whoever controls the conversation, controls the development of the standards, and if one institution is in control then the standard won’t be meeting the needs of
all institutions. If universal standards were to be developed it would need to be by a super committee where each institution’s view has equal representation. Additionally, if such a standard were developed and it was too restrictive it just wouldn’t get used, and on the other hand, if it is too open to interpretation then it would be pointless. It would also be pointless if it doesn’t make sense to users, or meet their needs or expectations. When developing new standards, you can’t get wrapped up in what data is needed and where it goes and forget about the users. As TNA3 points out, “most people just want to Google and get the results they were looking for, they don’t care what an institution might call it, or if it’s called the same thing at every cultural heritage institution, as long as they get the desired results/needs are met.”

4.3 Museums Sheffield

In 1998 Museums Sheffield was created to take over running the city’s non-industrial museums and galleries from the Sheffield City Council and has been entrusted with the safekeeping and conservation of the city’s treasures, which includes natural history specimens, art, and antiques. Additionally, as well as housing one of the finest metal working collections on the globe, Museums Sheffield’s strengths include their collection of 20th Century British Art, and the recently re-displayed Ruskin Collection (Museums Sheffield, 2016).

4.3.1 Metadata Standards & their Selection Criteria

For the natural sciences Catalogue of Life, a standard data dictionary is used to manage scientific terminology to describe species, and allows accurate tracking of each species’ complicated structure, as well as ensuring correct spelling is being used. SHIC (Social History Collections schema) is
used to classify the social history collections by theme rather than individual object name. An in-house geography thesaurus is also used to classify collections, as is the Museum Archaeological Objects thesaurus; other lists generated in-house for other collections within the museum, such as materials, mediums, art works, etc. are used with other collections across Museums Sheffield. The most important piece of criteria taken into consideration when deciding which standards are best for their collections is detail. Some guidelines, such as TGN (Thesaurus of Geographic Names), don’t go into enough detail which is why so many in-house lists and thesauri have been developed within the museum.

4.3.2 Digital Collection Users & Accessibility

There is no means of tracking who accesses the online collections so it’s hard to identify types of users. The museum’s Facebook and Twitter pages can be examined and give some idea, but nothing can be said for sure. To make online content as accessible as possible vocabulary and terminology are aimed at reading age of about twelve years old, which is standard in the museum world due to it being the average reading age of people in general, according to MS.

4.3.3 Metadata Sharing & Interoperability

Catalogue of Life allows data from the natural sciences collection to be migrated elsewhere and is a way in which Museums Sheffield enables metadata sharing and interoperability. Although not possible at the moment, in the future the museum hopes to implement data mining techniques to further enable metadata sharing and interoperability. The implementation of data mining techniques could enable the creation of a centralised data pool
which knows the fields of each individual institution and would be able to pull them all in rather than institutions just using the same data map.

The major issue in enabling interoperability and metadata sharing within Museums Sheffield, and other museums, is that because they each have their own way of working, use different software packages, and have collections made up of various different materials, museums are renowned for not working together. That being said, when the same software packages are being used within similar or related collections they do communicate with each other and try to do things the same way.

4.3.4 Collaborating with Other Organisations

Although they are not currently participating in any collaborations with other cultural heritage institutions, they have worked with the city archives unit on a digital mapping project in recent years, and have worked with other external organisations such as the city ecology unit and the National Biodiversity Network. Currently, the museum sends meteorological data to the Met Office which they collect from a weather station they run in neighbouring Weston Park. Sheffield museums has been using the weather station to gather meteorological data since about 1882 and when the data is gathered it is input into spreadsheets in a format required by the Met Office.

4.3.5 Universal Metadata Schemas and Controlled Vocabularies

MS said that having a universally used metadata schema and controlled vocabulary would be great if possible; it could help save a great deal of time, but unfortunately coming up with a one-fits-all solution for all cultural heritage institutions would be a struggle. He explains that such a thing would be more possible than in others. For example, some thesauri
make it possible to tweak the set formats, and those decisions are made based on personal preferences that not all institutions share. On the other hand, a mineral index, central repository, or something like Catalogue of Life, which allows people to pull information, would be amazing.
Chapter Five: Discussion

This chapter will discuss the qualitative research results presented within each case study and the themes which have emerged from the analysis of these results. The results from each institution will be compared to identify similarities and differences between each of them, and with the literature examined for this study.

5.1 Metadata Schema and Controlled Vocabulary Use and Selection Criteria

As presented in the case studies within the previous chapter, the British Library, National Archives, and Museums Sheffield each utilize a variety of metadata schemas and controlled vocabularies within their respective institutions. Despite the fact that each institution examined uses multiple standards, with the exception of Dublin Core and ISAD(G), the standards are not used by more than one of the institutions. Dublin Core and ISAD(G) standards are the only exception to this generalisation, as both are used by The British Library and The National Archives, but even then both institutions use the standards in different ways. At the British Library, Dublin Core is only used within particular systems, such as when metadata goes to PRIMO for discovery it is converted into PNX (PRIMOs Normalised XML), which is basically Qualified Dublin Core, and at The National Archives the core fields of Dublin Core are used in addition to an extension of fields specific to the archive. Then, ISAD(G) is only used at the library as the descriptive standard for the unpublished material within their collection, while the archive uses it to map metadata from their in-house cataloguing system to their Discovery catalogue. The overlap in the standards used at the British Library and The National Archives can be seen as a common occurrence in the metadata practices of libraries and archives in general. While different collections are held within these two types of institutions, and for different purposes, sometimes the same metadata standards can be used to
appropriately organise, describe, and manage the material held in both types of institutions, and often is. For example, although developed for libraries, there are archives that have adapted MARC standards to suit their needs, and vice versa with libraries and EAD.

Another observation, which has resulted from this study is that while none of the standards or vocabularies used at The British Library or Museums Sheffield overlapped, due to the differences in their digital collections, both institutions have implemented a larger range of standards and vocabularies as part of their metadata practices. With the variety of materials held at each type of institution, this is not an uncommon practice. According to Palmer et al. (2007) 34% of digital projects in 2003, and 39% in 2006, were using three or more schemas (pg. 391). What’s more, not only is the use of multiple standards common, but it is fairly equally divided among collaborative and non-collaborative projects (Palmer et al., 2007, p. 391). So, while the adoption of multiple standards is largely due to the types of materials in an institution’s collection, meeting the needs of collaborating institutions is an additional reason for this practice. When questioned on the selection criteria for the metadata standards they’ve adopted, interviewees from The British Library explained standards used by external organisations can be a factor in the decision making process because organisations who share their metadata sometimes use other standards, which then need to be translated in order to be compatible with the library’s systems. For instance, standards such as NLM, DTD, or JATS, are generally converted into MODS for the article level material, and ONIX is converted into MARC. Adopting standards or enabling their systems to convert the standards being used by other organisations enables more efficient workflows, allows interoperability.
between institutions, and thus allows further accessibility for users of cultural heritage institutions.

Other selection criteria for the metadata standards used within the institutions examined include: continuity with institutional workflows already in place, the value and interoperability of the standards with other institutions, efficiency, whether the minimum requirements provide relevant context and provenance to enable accurate preservation, and whether the level of detail needed for the item being described is permitted. The variety of criteria which needs to be met by these institutions, in addition to the items within their digital collections, are not only why places like The British Library and Museums Sheffield adopt a variety of standards, but also why The National Archives and Museums Sheffield implement their own locally developed standards. A few reasons an institution might develop their own standards include: a level of customization needed for capturing unique information, information already entered in a database or other local information system needs to be imported, or the existing standards don’t meet a project’s needs (Palmer et al., 2007, p. 392). The first and third reasons presented are why The National Archives and Museums Sheffield developed in-house standards rather than using ones that have already been established. As the interviewees at the National Archives explained, because they don’t feel they are in a position to demand certain information from the government departments they work with, they had to locally develop a standard which enables preservation and provides context and provenance while only requiring the bare minimum information necessary from the government. Though that isn’t to say the archives won’t accept other information that might be offered, on the contrary in addition to the minimal amount of required information, any additional information which might add context and make records more findable
is also accepted. At Museums Sheffield multiple in-house lists and thesauri have been developed, as previously explained, because some guidelines, such as TGN, don’t provide the level of detail needed by the museums.

5.2 Enabling User Accessibility for Digital Collections

As Trant (2009) states, “[t]he image of collections of all libraries, archives, and museums available on-line for consultation, by anyone, anytime, anywhere is a powerful one,” and while this statement rings true, an institution’s success in enabling user accessibility is highly dependent on how knowledgeable they are about those users and their needs (p. 374). Without being able to identify who uses the online collection of a library, archive, or museum, or what their needs or expectations might be, the institution in question will be unable to determine what kind of information should be made accessible or how to present that information in a format that is both easily understood by novice users, and provides the contextualised depth needed by specialist users.

Of the three institutions examined, the British Library is the most knowledgeable about their users and those user’s needs. As previously described in section 4.1.2, in addition to individual researchers using the British Library’s collections for their various scholarly pursuits, users also include commercial organisations wanting to use metadata provided by the library, and staff developers who want to use the data acquired by the British Library, or the metadata dumps that the library provides. Due to their knowledge regarding the various user groups of their digital collections, the British Library is in a better position to meet user’s needs than the others, and because they know their users they are able to have conversations with them and determine what their needs are and the best way to allow access to them. For example, The British Library is able to determine ways in
which users intend to use the information they find within their collections and then figure out ways to allow them to do so, such as removing boundaries on use by implementing forms of creative commons licensing when possible, and this is in addition to the effort taken to ensure the quality and scope of the metadata created is consistent so that users are provided with better recall and more comprehensive search results.

While The National Archives know their users are made up of people who might use the archive once, then go away, and people who are knowledgeable about the collection, how to access the records, and return to the archives catalogue repeatedly, it is still difficult to pin down who exactly is doing what. They do have access to statistics which tell them which records have been looked at, how often, and for how long, but there is no way of identifying if the accessed record was what the user wanted or not. Additionally, when The National Archives began making records available digitally, they did ask users what they wanted to see when looking at the collections online, but the vast majority don’t know, which in turn has made identifying user needs into a bit of a guessing game. As explained in the previous chapter, because of the very general knowledge they possess about their users, The National Archive’s current strategy is to provide as much information as possible through Discovery, and then adjust that information based on any user feedback they receive.

At Museums Sheffield, their inability to track who accesses their online collections makes it extremely difficult to identify the main user groups of their online collections, which in turn makes it difficult to assess those user’s needs. Thus, the museums’ attempt to make their online content as accessible as possible is limited to presenting the relevant information using vocabulary and terminology
aimed at a 12-year-old reading level. Although this practice is apparently standard in the museum world, due to it being the average reading age of people in general, it is a very small step in attempting to make online content accessible to users. However, when considering the issue of online user accessibility, it is important to remember the differences museums have from libraries and archives. While museums are used by individual researchers wanting to find the answer to a specific question, the bulk of museum users are groups of people who are interested in entire exhibits rather than individual items. Additionally, the way material in an exhibit is organised and presented, as well as the format in which information for individual pieces is presented to users, is significantly different than in libraries and archives. These differences can make it more difficult for museums to create digital collections which meet user needs, or create a user experience similar to visiting a physical exhibit.

Furthermore, being knowledgeable about their user base not only enables libraries, archives, and museums to identify user needs and expectations, it enables them to determine ways in which they can make their digital collections more accessible to those users. As other researchers (Kalfatovic, Kapsalis, Spiess, Van Camp, and Edson, 2008; Waibel and Erway, 2009) have described, in order for the digital content created by the cultural heritage institution community to be as accessible as possible, it needs to be moved from the information silos produced within each of these institutions and shared in locations users access on a regular basis, such as search engines or social spaces such as Facebook, Twitter, Flickr, etc. Waibel and Erway (2009) support this by citing an OCLC Perceptions Report from 2005 which showed that, “… 89% of college students claim that they start their quest for information with a search engine, as opposed to the 2% who start with a library
website” (p.325). Additionally, one of the drivers in the Smithsonian’s decision to join Flickr Commons was “… choosing to go where visitors are and not requiring them to come to us…” (Kalfatovic, et al., 2008, p. 270). This demonstrates that being able to accurately identify who their users are would not only allow libraries, archives, and museums to identify their needs and expectations regarding accessibility, but also identify the best methods for meeting those needs and making their content accessible online.

5.3 Facilitating Metadata Sharing & Interoperability

Due to the increasing reliance on the internet for information discovery, users have developed expectations about how easily that information should be found, as well as the format in which it should be presented to them online. As mentioned in section 5.2, in order for cultural heritage institutions to make their content as accessible to users as possible they should not only facilitate accessibility through their own websites and catalogues, but they need to enable metadata sharing and interoperability so their collections are discoverable through each other where appropriate, as well as search engines. Shreeves et al., (2006) share that, “… shareable metadata should be useful and usable to services outside of its local context given the resource described. … institutions need to think carefully about the uses and services they would like to support through their metadata.” Despite the common challenges that can be encountered, enabling interoperability and metadata sharing between each other would help make collections held within libraries, archives, and museums more discoverable for their users.

Each of the three institutions examined for this study have made, and continue to make, an effort to enable metadata sharing and support interoperability in differing ways. Due to the British Library’s level of knowledge about their users and
the communication they have with those users, as well as collaborating institutions, they have a variety of methods for enabling interoperability and metadata sharing including sharing their metadata with lone aggregators. By doing so the library achieves results similar to those attained with the implementation of the OAI-PMH tool, which can be used to bridge the widely varying metadata standards across libraries, archives, and museums (Roel, 2005). According to BL1, sharing their metadata with OCLC (Online Computer Library Centre)’s OPAC (Open Public Access Catalogue) system, Worldcat, or the UK’s COPAC system, makes it available to search engines like Google where it is then harvested and made visible to users. This is important because, as he explains, “… you could have the best metadata in the world, but if you couldn’t make it available outside of your reading room … then effectively to the rest of the world it’s invisible …” The issue of interoperability is also important when negotiating agreements with commercial organisations in regards to licensing. For instance, the library must ensure that the agreed upon terms of use for the resulting product and data allow them to perform the needed functions. In other words, the licensing must enable user access to the material. In situations such as this an embargo period is usually instated which limits how the project’s resulting product, and the metadata attached to it, can be used for a set period of time.

Additionally, The British Library’s communication with their users and collaborators enables them to facilitate metadata sharing and interoperability by implementing standards according to the communities they’re working with such as MARC with other libraries, EAD with archives, or CSV formats for individual researchers, as well as ensure that their internal systems are capable of translating external metadata formats to internal ones. However, this practice is not possible for
most cultural heritage institutions, and it is only possible for The British Library due to their position and remit, not only as a national library, but as one of the largest libraries in the world with a vast scope of items within their physical and digital collections. For example, while The National Archives is also a large national organisation due to their responsibility of collecting and preserving government records for future generations, adopting additional standards which don’t serve those purposes isn’t practical. Furthermore, while The British Library’s practice of adopting standards used by partner institutions is generally successful, it does still encounter incompatibility issues with their internal systems and some of the standards and systems used within other institutions, mostly galleries and museums whose standards have been developed to describe unique items such as paintings or historical artefacts, rather than mass produced items like books and journals.

Another challenge which cultural heritage institutions can come up against is that as new standards continue to be developed, it can become necessary to replace old ones. For this reason, The British Library is currently looking into the possibility of replacing LCSH with FAST because the belief is that FAST will enable the application of subject information more broadly than in the past. The conversion has yet to take place though because, since FAST is just a project within a research program, it is currently unknown whether OCLC is committed to keeping it going. This is an issue because efficiency and sustainability are important considerations when making decisions about implementing new standards to replace old ones, and as BL2 expresses in the case of The British Library, “… I don’t think we could reasonably replace LCSH unless we knew that FAST was sustainable because we need something we could use for the long term and we need partners that are for long term too.”
While The National Archives enables interoperability with other institutions through their APIs, their online catalogue, Discovery, is arguably their best asset in terms of metadata sharing and interoperability. Discovery not only allows access to the catalogues of more than 2,500 other archives in the UK, it also allows users to export search results into XML or CSV formats. Although this method of interoperability doesn’t make their own data or collections more widely available through other platforms, it does make it possible for users to access material they might not have found otherwise if they aren’t aware of the smaller archive whose collection it is a part of.

Just as the British Library and The National Archives provide good examples of how libraries and archives can implement metadata sharing and support interoperability, as well as the types of challenges they face, despite its smaller size Museums Sheffield provides similar examples of these practices within other museums. One way the museums currently support interoperability is by implementing the use of indexes and taxonomies from databases such as Catalogue of Life. As SM explains, as an international standard data dictionary, Catalogue of Life enables interoperability by allowing them, in theory, to migrate data elsewhere.

As pointed out in section 2.2, different types of institutions have different metadata cultures and practices which can create various types of barriers when attempting to implement any shared system for distributing metadata across those different institutions (Roel, 2005). Well, this can also be an issue among similar institutions, such as museums. Due to the various types of items held within their collections, museums have unfortunately become renowned for not working together due to each developing their own practices and procedures to accommodate those different collections, in addition to using different types of software packages. That
being said, according to MS, museums do communicate with each other and try to do things the same way when the same software packages are being used within similar or related collections. In another effort to implement metadata sharing and interoperability in the future, Museums Sheffield hopes to implement data mining techniques which, rather than having institutions use the same data map, could potentially enable the creation of a central pool with knowledge about the fields of the individual institutions involved and be able to pull them all in.

5.4 Collaborative Projects and their Challenges

Initially this study focused solely on collaborations with other cultural heritage institutions, but as the interviews progressed that focus expanded to include collaborations with other types of external organisations as well. This was due to the discovery that a great deal of the larger collaborative projects that the British Library, The National Archives, and Museums Sheffield participate in are actually with various external organisations such as Google, government departments, and the Met Office.

Due to the mass production of books and journals, which tend to make up the bulk of a library’s collection, and the fact that many libraries own a copy of the same item, collaboration has become firmly rooted into the fabric of libraries through the practice of copy cataloguing (Waibel and Erway, 2009, p. 331). While this is a form of collaboration The British Library participates in, along with digital preservation projects with additional cultural heritage institutions, such as Europeana’s World War I project, when questioned about current collaborative projects the library is taking part in BL1 and BL2 spoke mainly about their collaboration with Google and various unnamed commercial organisations. Similarly, The National Archives has participated in some collaborative digital preservation projects, such as PLANETS,
whose primary goal was, “to build practical services and tools to help ensure long-term access to [our?] digital cultural and scientific assets”, but because the archive serves as the deposit for government records their priority is to obtain, preserve and make those records accessible to the public, which results in most of their collaborative efforts being associated with various government departments (PLANETS, 2007). Due to its smaller stature, Museums Sheffield currently only works with the city ecology unit and the Met Office and has worked with the city archives and the National Biodiversity Network in the past, all of which, with the exception of the city archives, are not considered cultural heritage institutions.

In their collaboration with Google Books, the texts which the library supplies Google with are digitised and then the material is returned to the library, along with the metadata attached to them, and any additional information provided by Google. That new data is then entered into ALEPH, where a new record is generated for the digital surrogate of each item even if the data matches the record for the printed version of an item. BL2 explained that this process makes it possible for more of The British Library’s material to be catalogued that would otherwise take much longer due to the library not having enough resources to get through it quickly on their own. Although The National Archives has taken part in some digital preservation projects, due to its partnership with government departments it isn’t practical for them to adopt the various digital preservation standards available. While they do pay attention to developments and changes within the digital preservation sector, cooperating with PREMIS or METS just isn’t useful for the archive because they are not in a position to demand certain types of information for the records they obtain from the government departments they work with. This is not an uncommon situation in regards to tools such as PREMIS. A respondent in Alemneh and Hastings’s (2010)
study explained, “While we can be informed by PREMIS and what worked elsewhere in terms of adopting preservation metadata, we must take account of our own local specific conditions before implementing change” (p. 3). Additionally, because The National Archives is not in a position to demand information from the government departments with which they work, they are confronted with various challenges such as accuracy issues with the metadata they are provided, ensuring the core set of metadata elements used for special collections allows them to appropriately organise and describe the material, and ensuring that the systems used by government departments are capable of exporting the data needed.

Collaborations between cultural heritage institutions and other external organisations are important because not only do they enable accessibility, but they allow the institutions to identify and address possible issues with standards or systems used as part of their collection metadata procedures. On the other hand, as beneficial as collaborative projects might be there are a variety of issues such as licensing and copyright, efficiency, inconsistencies with historical metadata, and whether or not the metadata meets the needs of the external companies involved, which can arise if not addressed from the outset. Additionally, as Waibel and Erway (2009) state, “[t]he groundwork for success on a network level is laid in local collaborations, such as creating cohesiveness between [libraries, archives, and museums] belonging to the same organisation” (p325). Although the institutions examined are not made up of more than one type of cultural heritage institution belonging to the same organisation, such as the Victoria and Albert Museum, which also consists of the National Art Library, archives, and various other collection units, they are each made up of different collections and departments. For example, the British Library’s digital collection is made up of several collections, which are
maintained by different departments. Similarly, Museums Sheffield’s online collection is made up of objects that are held in separate museums across the city.

5.5 Universal Metadata Schemas and Controlled Vocabularies

In their 2015-2018 Collection Metadata Strategy, the British Library explains that collection metadata does more today than it did when it was included on catalogue cards. Descriptions now include content licensing, preservation, and access information and new specialist skills and systems are needed to support the various standards and formats used to make digital materials and their metadata as accessible as possible, and with that in mind, competing cataloguing standards used within different institutions creates additional complications and costs when attempting to harmonise new standards with old ones, or internally used standards with external ones (British Library, 2015). Due to these complications the idea for the development of a universal metadata schema and controlled vocabulary that may be used across all cultural heritage institutions has been brought up and discussed by professionals in the field, but has yet to be executed. This is largely because, as Alemneh and Hastings (2010) explain, “[c]onsidering the complex set of digital preservation challenges, many researchers … agree that there are no effective preservation methods or tools that work for all communities or type of resources” (p. 1)

When the interviewees for this study were questioned on their thoughts regarding the development of a universally used metadata standard, their replies were similar in that yes they think it would be a good idea, but that it just isn’t possible for a myriad of reasons.

The participants from The British Library, The National Archives, and Museums Sheffield, gave similar responses. Interviewees from The British Library and Museums Sheffield don’t believe it would be practical or successful due to the
differences in metadata needs within each of the different institutions and there being no way to enforce such a standard. Those at The National Archives agreed, and further argued that, because whoever controls the conversation about the creation of such a standard, also controls its development, a representative from each type of institution must be involved in the discussion, otherwise the standard won’t be meeting all of their needs. Furthermore, a one-fits-all standard also risks being too restrictive or too open to interpretation, which would result in it never being used or simply being pointless. Another important consideration for the development of a universal schema would be whether or not it makes sense to users. When developing new standards it is important not to get wrapped up in what data is needed and where it goes and forget about the users. A universal standard would also be pointless if it doesn’t make sense to users, or meet their needs or expectations.

Although all of the interviewees agreed that a universal metadata standard wouldn’t be successful or practical, a few alternatives were suggested. BL1 and BL2 explained that the creation of triples would be a good alternative because they would allow information to be disassembled and then reassembled in forms required for different institutions. Additionally, they also described the possibilities behind linked data registries which hold an infinite number of labels for items and in a variety of languages. Similarly, MS explained that while some thesauri make it possible to tweak the set formats, those decisions are made based on personal preferences that not all institutions share. He did suggest however that, a mineral index, central repository, or something like Catalogue of Life, which allows people to pull information, would be amazing.
Chapter Six: Conclusion

In regards to the aims of this study, libraries, archives, and museums typically use the same criteria when selecting appropriate metadata schemas and controlled vocabularies for their digital collections, although the level of importance of each piece of criteria differs from institution to institution. Some types of criteria include the nature of the collection and types of collection resources, user groups, an institution’s IT capabilities, staff experience, and budget and resource limitations. Additionally, each type of institution examined not only implements different schemas and vocabularies, but the results indicate that due to the variety of materials many cultural heritage institutions hold in their collections, they often use more than one schema or vocabulary, or sometimes use established schemas to create their own in-house ones. Many of the established standards, which might have been created to suit a specific type of institution, such as MARC for libraries or EAD for archives, are also often used by institutions other than those they were created for. Alternatively, the standards which are created in-house by some institutions do not meet the needs of others, even if they are the same type, and are very rarely adoptable by anyone other than who it was developed for.

Arguably, the most important consideration when creating, or preparing to implement, new metadata standards, practices, and procedures is how user accessibility will be affected. As indicated repeatedly in the results and discussion chapters of this study, if a cultural heritage institution’s digital collections and the resources they contain are too difficult for users to access then the implemented standards, systems, or other metadata practices or procedures are not appropriate and become pointless. However, sometimes just implementing appropriate metadata schemas or controlled vocabularies is not enough. Online users have become
accustomed to search engines such as Google and the format in which they provide results and because of this, if the search capabilities of a cultural heritage institution’s website are significantly different or a challenge to navigate, users are less likely to return. The most common ways institutions have been combating this issue is by enabling metadata sharing and supporting interoperability with other institutions. Cultural heritage institutions have come a long way in providing online access of their content to users, and although new issues will always need to be confronted, libraries, archives, and museums have all significantly contributed to the many advancements in the online discoverability of the digital collections and resources they hold.

6.1 Limitations of the Study

The most significant limitation to this study was the difficulty in recruiting focus group interview participants, particularly from the museum sector. Due to the lack of participation from the museum sector the results are not as comprehensive as they could have been. Additionally, although the desired amount of participants were recruited from the library and archives sectors, in hindsight more substantial results could have been achieved with the participation of one or two more professionals from each of the institutions. Having interview participants from a slightly wider variety of professional expertise could have resulted in more discussion and a wider breadth of results for analysis.

6.2 Recommendations for Further Research

A great deal of further research can be conducted in this area. Conducting this study with a different data collection method, such as the distribution of questionnaires to a number of libraries, archives, and museums, would provide a broader pool of data to examine. For example, as the results of this study
show, not only is there a wide variety of metadata schemas and controlled
vocabularies for cultural heritage institutions to choose from, but many of those
standards enable different institutions to use their elements in different ways. If more
than one library, archive, and museum, were to be examined an analysis of the
different ways they use those standards and why could be carried out. A similar
procedure could also be utilized to examine user accessibility, metadata sharing and
interoperability, and the thoughts of metadata professional on the future of metadata
practices and procedures.

In addition, while the consensus among those interviewed for this study was
that creating a ‘one-fits-all’ metadata schema or controlled vocabulary would not be
successful, let alone practical, other possibilities did come up which should be
explored further, most specifically, Linked Data. As BL2 explained, while a
universal metadata standard may not be practical, technology does allow for the
possibility of more flexibility, of which linked data registries are an example. Having
a central registry or repository which holds identifiers, labels, terms, and other
information that has been translated, or otherwise modified, and which can be pulled
and used by libraries, archives, and museums, would be useful and could help save
time for all involved.
References


http://www.museums-sheffield.org.uk/about/

http://www.museums-sheffield.org.uk/about/our-story-so-far/

http://discovery.nationalarchives.gov.uk/

http://www.nationalarchives.gov.uk/about/our-role/

http://www.nationalarchives.gov.uk/about/our-role/what-we-do/


http://www.planets-project.eu/

Roel, E. (2005). The MOSC Project: Using the PAI_PMH to bridge metadata


Appendices

Appendix 1: Ethics Application Form

Application 009209

Section A: Applicant details

- Created: Wed 25 May 2016 at 11:59
- First name: Muriel
- Last name: Mangula
- Email: mmangula1@sheffield.ac.uk
- Programme name: MA Librarianship
- Module name: INF6000
- Last updated: 26/05/2016
- Department: Information School
- Date application started: Wed 25 May 2016 at 11:59
- Applying as: Undergraduate / Postgraduate taught

Research project title:
Selection of metadata for digital collections: An investigation of the criteria and strategies used when choosing appropriate metadata schemas and controlled vocabularies

Section B: Basic information

<table>
<thead>
<tr>
<th>1. Supervisor(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Paula Goodale</td>
</tr>
<tr>
<td>2: Proposed project duration</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>Proposed start date: Sun 5 June 2016</td>
</tr>
<tr>
<td>Proposed end date: Wed 31 August 2016</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>3: URMS number (where applicable)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URMS number</td>
</tr>
<tr>
<td>- not entered -</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4: Suitability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Takes place outside UK?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Involves NHS?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Healthcare research?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>ESRC funded?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Involves adults who lack the capacity to consent?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Led by another UK institution?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Involves human tissue?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Clinical trial?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Social care research?</td>
</tr>
<tr>
<td>No</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>5: Vulnerabilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Involves potentially vulnerable participants?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>Involves potentially highly sensitive topics?</td>
</tr>
<tr>
<td>No</td>
</tr>
</tbody>
</table>
Section C: Summary of research

1. Aims & Objectives

The aim of the research to be conducted for this study is to identify what criteria libraries, archives, and museums use to determine which metadata schemas and controlled vocabularies to use when creating their digital collections and then investigate whether each selected schema can be used interchangeably between the different types of collections at each institution, or if it is only suitable for the type of collection it was implemented for. The institutions to be examined are the British Library, The National Archives, and the Victoria & Albert Museum.

The objectives of this project are:
- To discover the criteria used by each institution to determine which metadata schema and controlled vocabulary to use.
- To determine if the selected schema and vocabulary enable novice and/or expert users to easily find desired information or specific digital objects.
- To examine whether the schemas and vocabularies selected could be successfully used interchangeably between collections within each of the selected institutions.
- To identify if the metadata practices within each institution successfully enables metadata sharing and interoperability between them and other cultural heritage and memory institutions.

2. Methodology

The primary method of data collection for this qualitative study will consist of semi-structured interviews with focus groups which include professionals such as: digital curators, cataloguers, and other metadata experts, within the desired cultural heritage and memory institutions. The decision to conduct focus group interviews was made with the hope of generating discussion between participants about differing opinions they might have on the topic. I will begin by reaching out to professionals working at each institution by email, and developing a set of interview questions. I will be using nonprobability purposive sampling in order to be sure I recruit appropriate participants, and my hope is that anyone unable to take part will put me in contact with other available professionals so my focus group within each institution might snowball into similar sizes. I will develop my interview script based on previous studies about metadata selection practices. Once my interview questions have been assembled, and reviewed and approved by a second party, a few additional probing questions will be added. As I confirm who will participate in my study I will be sure to provide them with the required informed consent forms and make sure that I receive their consent to use either their name, job title, or general area of work expertise, in addition to the name of the institution they are employed within, when putting together my research. Once my script has been finalized and I have found a sufficient number of participants from each institution I will conduct my interviews. The interviews will be recorded and transcribed to enable data analysis and determine the study’s results to be discussed in my report.

3. Personal Safety

Raises personal safety issues? No

Personal safety management
- not entered -
Section D: About the participants

1. Potential Participants

Potential participants will be professionals such as: digital curators, cataloguers, metadata experts, and others involved in the development of digital collections within The British Library, The National Archives, and Victoria & Albert Museum.

2. Recruiting Potential Participants

Potential participants will be approached via email. In addition to requesting their participation, I will provide them with an overview about my study’s focus, reasons for requesting their involvement, and what that involvement would entail. If they are unable to participate I will request their assistance in referring me to others within their institution who may be interested and available to participate in an interview.

2.1 Advertising methods

Will the study be advertised using the volunteer lists for staff or students maintained by CICS? No
- not entered -

3. Consent

Will informed consent be obtained from the participants? (i.e. the proposed process) Yes

I will obtain informed consent by providing each potential participant with a consent form via email. I will also take copies of the consent to the interviews and will not interview anyone who has not signed a consent form. I will also allow potential participants to ask any questions about my study that may have previously been unanswered.

4. Payment

Will financial/in kind payments be offered to participants? No
- not entered -

5. Potential Harm to Participants

What is the potential for physical and/or psychological harm/distress to the participants?

I believe there is low risk of physical or psychological harm or distress to participants of my study. My interviews will take place on business premises during working hours. Participating in my study might be considered inconvenient, due to it taking time out of the work day, but effort will be made to work around their schedules to provide as little inconvenience as possible. Additionally, although the topic of research isn’t sensitive, each participant will be allowed to decide their level of anonymity.

How will this be managed to ensure appropriate protection and well-being of the participants?

I intend to conduct the interviews within work place settings and participants will be informed
about the estimated duration of each interview.
As previously addressed, each participant will determine their level of anonymity within the study by choosing to be identifiable by name, job title, area of work, or not at all.

Section E: About the data

1. Data Confidentiality Measures

The only identifying information retained for each participant will be their institution of employment. Depending on the level of anonymity desired by each participant, their name, job title, or general area of expertise/work may also be used. Participants not wanting their names or job titles revealed will be referred to as Participant 1, Participant 2, etc.

2. Data Storage

Data collected will be stored on the Information School’s research data drive, accessible only by the researcher and select members of university staff. NVivo will be used for data analysis, and files will be saved on the researcher’s personal drive on University of Sheffield computers, and password protected backup copies on my personal laptop which will be deleted when the study has been completed and dissertation marked.

Section F: Supporting documentation

Information & Consent

Participant information sheets relevant to project?
No

Consent forms relevant to project?
Yes

Consent Forms

- Consent_Form_MM.docx.pdf
  (Document 022449)

Additional Documentation

None

External Documentation

Note: Both information sheets and consent forms are relevant to my project and, although an information sheet has not been uploaded separately, a combined information sheet/consent form
has been uploaded under 'Consent form(s)'.

<table>
<thead>
<tr>
<th>Official notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>- not entered -</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section G: Declaration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signed by:</td>
</tr>
<tr>
<td>Muriel Mungula</td>
</tr>
<tr>
<td>Date signed:</td>
</tr>
<tr>
<td>Thu 26 May 2016 at 13:24</td>
</tr>
</tbody>
</table>
Appendix 2: Participant Information Sheet and Consent Form

The University of Sheffield Information School

Selection of metadata for digital collections: An investigation of the criteria and strategies used when choosing appropriate metadata schemes and controlled vocabularies

Researchers
Muriel Mungula (mrmungula@sheffield.ac.uk)

Purpose of the research
This research project aims to investigate the selection criteria and procedures for metadata schemes and controlled vocabularies used in the development of digital collections within libraries, archives, and museums; and whether they can be used interchangeably, or enable metadata sharing and interoperability between institutions.

Who will be participating?
I have asked professionals involved in the development of digital collections within leading UK cultural heritage institutions to take part in this study.

What will you be asked to do?
You will be asked to participate in an interview, lasting no more than an hour, about the metadata scheme and controlled vocabulary selection practices for digital collections within your institution, and the reasons behind them.

What are the potential risks of participating?
The risks of participating are the same as those experienced in everyday life. Although your identity will not be revealed without your permission, you may still be identifiable as I will be identifying the institution you work for and the pool of potential participants is small. However, you may refuse to answer any question(s) you are uncomfortable with.

What data will we collect?
I will be recording audio of the interview, to be transcribed later.

What will we do with the data?
I will be analysing the data to be used in my master’s dissertation. The data will be stored on the Information School’s research data drive which can be accessed only by me, my supervisor, the School’s Examinations Officer, and ICT staff operating the facility. I will also store a password protected backup copy on my personal laptop. All data will be deleted after the dissertation has been completed and marked. Data may also be used in any presentations or publications arising from this dissertation.

Will my participation be confidential?
The only identifying information that will be used is the name of the institution that you work for. Your name, job title, or general area of work will only be used if you have granted permission to do so. As the pool of professionals suitable to participate in this study is rather small, I cannot guarantee that you will be unidentifiable, even without granting permission to use your name or job title. My reporting of the results will not include any identifying information about you without your consent. If you grant permission to use your name or job title you will be provided with a transcript of the interview and allowed to retract any statements you do not wish to be published.
What will happen to the results of the research project?

The results of this study will be used in my master's dissertation which will be publicly available. Please contact the Information School at the University of Sheffield for details.

- I confirm that I have read and understand the description of the research project, and that I have had an opportunity to ask questions about the project.
- I understand that my participation is voluntary and that I am free to withdraw at any time without any negative consequences.
- I understand that if I withdraw I can request for the data I have already provided to be deleted, however this might not be possible if the data has already been anonymised or findings published.
- I understand that I may decline to answer any particular question or questions.
- I understand that my responses will be kept confidential, that my name or identity will not be linked to any research materials unless I have agreed otherwise. (Please tick one)
  - I grant permission to be identified by name in the results of this study.
  - I grant permission to be identified by job title in the results of this study.
  - I grant permission to be identified by my area of work in the results of this study.
  - I wish for my name and identity NOT to be linked to any research materials.
- I give permission for all the research team members to have access to my responses.
- I give permission for the research team to re-use my data for future research as specified above.
- I agree to take part in the research project as described above.

Please provide information to be used:

<table>
<thead>
<tr>
<th>Participant Name (Please print)</th>
<th>Job title/Area of work</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Participant Signature ___________________________ Date __________

Researcher Signature ___________________________ Date __________

Note: If you have any difficulties with, or wish to voice concern about, any aspect of your participation in this study, please contact Dr Jo Bates, Research Ethics Coordinator, Information School, The University of Sheffield (school.ethics@sheffield.ac.uk), or the University Registrar and Secretary.
Appendix 3: Ethics Approval Letter

Muriel Munguia
Registration number: 150112662
Information School
Programme: MA Librarianship

Dear Muriel,

PROJECT TITLE: Selection of metadata for digital collections: An investigation of the criteria and strategies used when choosing appropriate metadata schemas and controlled vocabularies
APPLICATION: Reference Number 009209

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 03/06/2016 the above-named project was approved on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 009209 (dated 26/05/2016).
- Participant consent form 1018791 version 1 (26/05/2016).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Yours sincerely,

Matt Jones
Ethics Administrator
Information School
### Appendix 4: Participant Codes

<table>
<thead>
<tr>
<th>British Library</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>BL1</td>
<td>Neil Wilson, Head of Collection Metadata</td>
</tr>
<tr>
<td>BL2</td>
<td>Alan Danskin, Collection Metadata Standards Manager</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The National Archives</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TNA1</td>
<td>Senior Archivist, Catalogue &amp; Taxonomy Team</td>
</tr>
<tr>
<td>TNA2</td>
<td>Digital records and Transfer Manager</td>
</tr>
<tr>
<td>TNA3</td>
<td>Senior Digital Archivist</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Museums Sheffield</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MS</td>
<td>Alistair McLean, Curator of Natural Science</td>
</tr>
</tbody>
</table>
Appendix 5: Interview Questions

**Introduction:**
Can you tell me a bit about your role(s) with regard to digital collections?

**Metadata schema & controlled vocabulary selection criteria:**
What types of materials/resources make up your institution's digital collections?

Which metadata schemas and controlled vocabularies are used at (BL/TNA/STJ)?
- Which are used in relation to which types of materials/resources?
- What criteria are applied when selecting metadata schemas/controlled vocabularies?
- What are your thoughts on the selected schemas/vocabularies?

Do the selected schema & vocabulary enable users to easily search digital collections?
- What types of users take advantage of your digital collections?
  - What types of different needs do they have?

When creating metadata what user needs are considered?
- What types of searching capabilities do your schemas/vocabularies enable?

**Metadata sharing & interoperability with other cultural heritage institutions:**
What are your thoughts on metadata sharing and interoperability?

How does your institution enable metadata sharing or interoperability with other types of institutions?
- What types of issues have you been faced with when enabling metadata sharing/interoperability?

**Interchangeability of schemas & vocabularies between collections within different types of institutions:**
Can you tell me about any collaborative digital projects (past/current) you've done with other institutions?
- How were decisions about metadata made in these types of projects?

What are your thoughts on the possibility of developing a single metadata schema and controlled vocabulary that could be used across libraries, archives, and museums?
- What types of challenges would need to be addressed?
- Would such a schema or vocabulary be desirable?

**Bold = General question**

**Italic = Probing question**
# Appendix 6: Results Table

<table>
<thead>
<tr>
<th>Themes to be Identified</th>
<th>British Library</th>
<th>National Archives</th>
<th>Sheffield Museums</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata Schemas &amp; Controlled Vocabularies Used</td>
<td>Metadata Schemas: MARC21, SAMI (Sound And Moving Image) MARC, MODS, use METS as a wrapper for digital objects, Dublin Core*, Controlled Vocabularies: LOC &amp; PCC/NAACO name authority files, LCSH**, Dewey Classification, UKAT (UK Archival Thesaurus)<strong>. Descriptive standards: PDA for published material, ISAD(G) for unpublished material, DCRM for rarebooks</strong>*, TEI for some manuscripts (mostly oriental manuscripts). [Q4/01]</td>
<td>NCA rules used, in house cataloguing system used with core cell of Dublin Core fields, in addition to TNA specific ones*, which is then mapped to ISAD(G) fields when exported to Discovery** [Q3/01]</td>
<td>Catalogue of Life*, SHIC (Social History Collections scheme), a custom (in house) geograph thesaurus, and other lists generated inhouse i.e. materials, mediums, art works, etc. [Q2/01]</td>
</tr>
<tr>
<td>Metadata Schema &amp; Controlled Vocabulary Selection Criteria</td>
<td>Continuity &amp; workhorse already in place*, value &amp; interoperability of items in terms of collaboration with other parties**, efficiency***. [Q5/01]</td>
<td>Simplicity; in house system developed to require bare minimum while still enabling the preservation context, provenance, etc.* [Q4/01]</td>
<td>Detail; due to some guidelines, such as TNG, not going into as much detail as needed in house lists/thesauri have been developed. [Q4/01]</td>
</tr>
<tr>
<td>Users of Digital Collections</td>
<td>Group type users i.e. commercial organisations wanting to take metadata, individual researchers, and staff developers using the limited data and who might want to use the viable data the BL acquires or want to utilise the metadata dumps made available by it. [Q5/02]</td>
<td>Not a lot is known about the specific user base other than it’s split into users who only access collections once, and those (such as historians and students) who are more knowledge able what is accessible in the collection and how to find it. There is no typical user. [Q5/02]</td>
<td>Difficult to say who accesses online collections due to not having a means of tracking. Looking at those who access the Facebook and Twitter pages, can give some idea, but can’t say for sure. [Email him about documents he said he’d ask for.] [Q5/02]</td>
</tr>
<tr>
<td>How Metadata Schemas and Controlled Vocabularies Enable User Access</td>
<td>Effort is made for print and ebook metadata to have similar formats with identical characteristics, despite having different elements. As much communication as possible takes place with users to ensure data is as easy for them to consume as possible. Attempting to eliminate boundaries for potential use of items within the collection, as well as the data connected to it*. [Q5/02] Indexes are available** with which users can search each both authorized and non-authorized formats of authors, for example. With PRIMO*** browsing is limited to the list of search results. Google/Amazon like search experience vs. research oriented interfaces. [Q6/02]</td>
<td>Due to little being known about the userbase, little is known about what they need/want* so current strategy is to provide as much information as possible &amp; use their user feedback to make adjustments to accessibility**. [Q5/02]</td>
<td>Vocabulary and terminology is aimed at a reading age of about 12.* [Q6/02]</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td>Source</td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>Interoperability &amp; Metadata Sharing</strong></td>
<td>Use of MARC and library standard protocols for libraries, &quot;interoperable versions of things like EAD&quot; for archives, and use other standards according to the community they’re working with. Digitise items for commerical organisations &amp; receive copy of the material in exchange. Sharing metadata with DDC's Worldcat, UK CDG, etc makes available to be harvested by Google &amp; other search engines. Ensure BI's internal systems are capable of translating external metadata formats to internal ones.</td>
<td>[Q8/04]</td>
<td></td>
</tr>
<tr>
<td><strong>APIs (Application Program Interfaces)</strong></td>
<td>APIs (Application Program Interfaces) are made available. [Q7/04] Discovery allows the export of search results into XML or CSV formats. Expert Contributions allow access to the catalogues of other archives from NLA's Discovery catalogue. [Q8/04]</td>
<td>[Q8/04]</td>
<td></td>
</tr>
<tr>
<td><strong>Issues with Interoperability &amp; Metadata Sharing</strong></td>
<td>Licenses that enable accessibility. Vocabulary, looking to replace LCSH w/ FAST. In compatibility with standards/systems used at other institutions.</td>
<td>Museums are renowned for not working together much due to each having their own way of working and the different software packages used, as well as the various types of collections they each hold. [Q8/03/04]</td>
<td></td>
</tr>
<tr>
<td><strong>Collaborations w/ Other Institutions</strong></td>
<td>Google's 18th century books. Digitisation of material for various external organisations. European WII project. Have been working with central government departments for years. Have taken part in PLANETS, [Q9/03] Preservation and Long-term Access through Networked Services, and other types of digital preservation projects.</td>
<td>No collaboration with other cultural heritage institutions, but have had partnerships with other external organisations i.e. work with any ecology unit, have worked with National Biodiversity Network in the past, and meteorological data is also collected and sent to the met office. [Q9/03]</td>
<td></td>
</tr>
<tr>
<td><strong>Metadata Issues in Collaborative Projects</strong></td>
<td>Does the metadata meet the needs of the external companies? Licensing/copyright issues, efficiency issues, inconsistencies with historical metadata. Ensuring the dates given are accurate. Ensuring the core set of metadata elements are accurate for their designated collection, system capability. Due to their specific remit it's not practical to &quot;buy into&quot; the various digital preservation standards available.</td>
<td>No current issues due to minimal amount of collaborative projects [Q7/03]</td>
<td></td>
</tr>
<tr>
<td><strong>Thoughts on Universal Metadata Schemas &amp; Controlled Vocabularies</strong></td>
<td>Different representation of metadata are needed within different domains and universal metadata control can’t be enforced. Standardization fails to acknowledge restrictions some institutions might face. Who controls the conversation/develops the standards? How do you standardise the description of different projects? Is it understandable to users? Too restrictive or not used, too open to interpretation = pointless. A one fits all model wont work.</td>
<td>Would be great if possible, but coming up with a one fits all solution would be a struggle. [Q8/03]</td>
<td></td>
</tr>
</tbody>
</table>
Access to Dissertation

A Dissertation submitted to the University may be held by the Department (or School) within which the Dissertation was undertaken and made available for borrowing or consultation in accordance with University Regulations.

Requests for the loan of dissertations may be received from libraries in the UK and overseas. The Department may also receive requests from other organisations, as well as individuals. The conservation of the original dissertation is better assured if the Department and/or Library can fulfill such requests by sending a copy. The Department may also make your dissertation available via its web pages.

In certain cases where confidentiality of information is concerned, if either the author or the supervisor so requests, the Department will withhold the dissertation from loan or consultation for the period specified below. Where no such restriction is in force, the Department may also deposit the Dissertation in the University of Sheffield Library.

To be completed by the Author – Select (a) or (b) by placing a tick in the appropriate box

If you are willing to give permission for the Information School to make your dissertation available in these ways, please complete the following:

X (a) Subject to the General Regulation on Intellectual Property, I, the author, agree to this dissertation being made immediately available through the Department and/or University Library for consultation, and for the Department and/or Library to reproduce this dissertation in whole or part in order to supply single copies for the purpose of research or private study.

(b) Subject to the General Regulation on Intellectual Property, I, the author, request that this dissertation be withheld from loan, consultation or reproduction for a period of [ ] years from the date of its submission. Subsequent to this period, I agree to this dissertation being made available through the Department and/or University Library for consultation, and for the Department and/or Library to reproduce this dissertation in whole or part in order to supply single copies for the purpose of research or private study.

Name: Muriel Munguia
Department: Information School, MA Librarianship
Signed: Muriel Munguia
Date: 31 August, 2016

To be completed by the Supervisor – Select (a) or (b) by placing a tick in the appropriate box

(a) I, the supervisor, agree to this dissertation being made immediately available through the Department and/or University Library for loan or consultation, subject to any special restrictions (*) agreed with external organisations as part of a collaborative project.

*Special restrictions

(b) I, the supervisor, request that this dissertation be withheld from loan, consultation or reproduction for a period of [ ] years from the date of its submission. Subsequent to this period, I agree to this dissertation being made available through the Department and/or University Library for loan or consultation, subject to any special restrictions (*) agreed with external organisations as part of a collaborative project.

Name:
Department:
Signed:
Date:

THIS SHEET MUST BE SUBMITTED WITH DISSERTATIONS BY DEPARTMENTAL REQUIREMENTS.