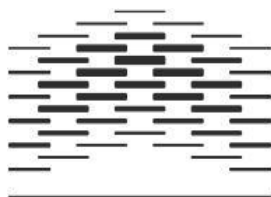




TALLINNA ÜLIKOOL



OSLO AND AKERSHUS
UNIVERSITY COLLEGE
OF APPLIED SCIENCES



UNIVERSITÀ DEGLI STUDI DI PARMA



Education and Culture DG

ERASMUS MUNDUS

Leire Arrula Eugui

**Case studies on digitization and metadata creation
and management**

Supervisor: Michael Preminger

Master thesis

International Master in Digital Library Learning

2012

ABSTRACT

This thesis' initial aim is to study the life-cycle of metadata from its creation, its development and management and the role they play during digitization processes. Three memory institutions and their respective digitization projects are studied in order to know how they carry out the digital object production process, what is the life-cycle of metadata, by who, how and when are generated and to know what is the best way to generate and implement metadata in digitization projects.

The research method is based on a qualitative approach and the research strategy is based in case studies. Data were collected with different techniques: documentation, questionnaires and semi-structured interviews. These questionnaires were sent to 3 key informants (one person for each institution) and subsequently interviews to the same informants were made in order to get deeper and more concrete information.

The collected data suggest that most institutions do not have explicit guidelines but follow leading organizations' standards. Because of the use of standards no institution had to cope with interoperability problems. Also in the opinion of respondents digitizing and creating metadata is not a remarkable challenge, they perceive it as another process among others. Considering the problems or challenges the informants stated to have, the emphasis can be put on budget. That is probably the aspect that most influences on the result, the institution with no budget limitations is the one that best and more complete metadata generates. Also budget has appeared to be a problem for preservation, and the institution with lower budget availability had to delegate the custody and management of its preservation copies to another institution.

AKNOWLEDGEMENTS

I feel very grateful to many people for helping and supporting me during my study period and writing of this thesis.

I would like to express my sincere gratitude first of all to my loving parents, Nekane and Javier, and aunt Alicia, who supported me since the first day and gave me their encouragement throughout all the writing process.

I would also like to thank my supervisor at Høgskolen I Oslo og Akershus, Michael Preminger, for the great assistance, guidance and help. And thanks to my teacher and Master's programme coordinator in the University of Barcelona, Miquel Térmens, for the great directions he gave me before starting my thesis. Thanks also to all staff of both universities which helped me with administrative procedures. Thanks to the European Commission for the funds I received through the Erasmus programme.

Of course, I am also very grateful to Petter Rønningsen, Gunleiv Handland and Lluís Vicente for their patience, for providing me with information and for collaborating with me all this time. Thank you, without you this thesis would not exist.

Big hugs to the incredibly great and beautiful friends whose companionship was an inexhaustible source of happiness and pleasure during my stay in Oslo: Sandra, Eirik, Alfons and all the other awesome people I met (you know who you are)!!! Also big hugs to my classmates in Barcelona, you are awesome!

Many thanks to my loving friends in Barañain for being always there for me no matter the distance. Also thanks to my excellent friends in Barcelona, you have been my family during the last years. I love you all!!!

TABLE OF CONTENTS

LIST OF FIGURES.....	vi
LIST OF TABLES	vii
CHAPTER 1: INTRODUCTION AND OBJECTIVES.....	1
1.1 INTRODUCTION	1
1.2 BACKGROUND OF STUDY.....	1
1.3 STATEMENT OF PROBLEM	2
1.4. RESEARCH QUESTIONS	3
1.5 AIMS AND OBJECTIVE OF RESEARCH	3
1.6 BENEFIT OF RESEARCH	4
1.7 METHODOLOGY OF RESEARCH.....	4
1.8 LIMITATION OF STUDY.....	5
1.9 OUTLINE OF STUDY	5
CHAPTER 2: BACKGROUND AND LITERATURE REVIEW.....	7
2.1 INTRODUCTION	7
2.2 DIGITAL LIBRARIES & MEMORY INSTITUTIONS	7
2.3. DIGITIZING COLLECTIONS	8
2.3.1 Digitization of written documents.....	11
2.3.2 Digitization of images	12
2.4 METADATA	12
2.4.1 Metadata categories	14
Descriptive metadata:	14
Administrative metadata:	14
Structural metadata:	14
2.4.2 Metadata schemes and element sets.....	16
Dublin Core	18
Text Encoding Initiative (TEI).....	19
Metadata Encoding and Transmission Standard (METS)	19
Metadata Object Description Schema (MODS)	20
MARC 21	20
PREMIS Data Dictionary for Preservation Metadata.....	21
2.4.3 Metadata creation.....	21

2.4.4 Metadata quality control	24
CHAPTER 3: METHODOLOGY	25
3.1 INTRODUCTION	25
3.2 THEORETICAL FRAMEWORK.....	25
3.3 RESEARCH DESIGN.....	26
3.4 RESEARCH STRATEGY	28
3.5 COLLECTING DATA:	29
Questionnaire.....	29
Interviews	30
3.6 DATA ANALYSIS:.....	31
3.7 ETHICAL CONSIDERATIONS:.....	31
3.8 LIMITATIONS IN STUDY	32
CHAPTER 4: DATA ANALYSIS AND FINDINGS	33
4.1 INTRODUCTION	33
4.2 BACKGROUND STUDY OF KEY INFORMANTS	33
4.3 CURRENT DIGITAL OBJECT CREATION LINE	34
4.3.1 NATIONAL LIBRARY OF NORWAY	34
4.3.1.1 Some technical issues	35
4.3.1.2 Digital objects' production line:	37
4.3.1.3 Problems/Challenges.....	43
4.3.2 NORWEGIAN PETROLEUM MUSEUM.....	44
4.3.2.1 Some technical issues	45
4.3.2.2 Digital objects' production line	48
4.3.2.3 Problems/Challenges.....	49
4.3.3 ATENEU BARCELONÈS.....	49
4.3.3.1 Some technical issues	50
4.3.3.2 Digital objects' production line	51
4.3.3.3 Problems / Challenges.....	52
CHAPTER 5: CONCLUSIONS AND DISCUSSION.....	53
5.1 INTRODUCTION	53
5.2 FINDINGS.....	53
5.3 RESEARCH QUESTIONS	54

5.3.1 How do the studied institutions carry out the digital object production process?.....	54
5.3.2 What is the life-cycle of metadata during a digitization process? By who, how and when are generated?	55
5.3.3 What are the main differences between institutions?	56
5.3.4 What is the best way to generate and implement metadata during a digitization process?	57
5.4 FURTHER RESEARCH	57
REFERENCES	60
APPENDICES.....	63
APPENDIX 1	63
APPENDIX 2:	66
APPENDIX 3:	75
APPENDIX 4	78
APPENDIX 5	85

LIST OF FIGURES

Figure 1: steps followed by the National Library of Spain in a digitization project.	11
Figure 2: life cycle of an information object (Baca, 2008).	21
Figure 3: the workflow of metadata in a digitization Project (Day et al.,2010).	23
Figure 4: the model proposed by pickard (2007) (adapted from krumar, 1999, and lincoln and guba, 1985).	27
Figure 5: data analysis process (Miles and Huberman, 1994).	31
Figure 6: example of how after applying an OCR it is possible to make searches by content.	36
Figure 7: detail of how the final user sees the digitized book. In this caption the barcode with the physical books id can be seen.	38
Figure 8: inside the red circle the persistent link based on the URN of the current page.	39
Figure 9: NB's digital object production line.	40
Figure 10: Photostation's default tool for editing and adding metadata. These metadata fields can be customized.	41
Figure 11: digital newspapers production line in NB.	42
Figure 12: relationship between the analog and the digital object in NB.	44
Figure 13: PRIMUS XML export from Petroleum Museum. Each photograph's descriptive metadata are displayed in these captures.	47
Figure 14: that is how PRIMUS interface looks like.	48
Figure 15: display view of a digitized document of Ateneu Barcelonés.	52
Figure 16: digital object production line in general terms.	54

LIST OF TABLES

Table 1: different types of metadata and their functions (Baca, 2008).....	154
--	-----

CHAPTER 1: INTRODUCTION AND OBJECTIVES

1.1 INTRODUCTION

This chapter introduces the thesis, background of study and statement of the problem. Aims and objective of research are explained and research questions are set out. Following this, benefits of research are stated and the methodology and limitations of study are described. Finally the outline of study is introduced.

1.2 BACKGROUND OF STUDY

During the last decades most memory institutions have adopted new technological solutions and have started carrying out digitization projects in order to preserve and make more accessible their collections, as well as to adapt to new information needs.

Digitization is the process by which analog (e.g. “paper-based”) materials become a sequence of 0s and 1s ordered using a binary code to be readable for a computer. Digital information has certain characteristics and qualities such as: the content can be linked to other materials and create multimedia digital objects, digital information is not limited by space or time barriers, can be stored and distributed in different ways and unlimited copies can be made without degrading the original. Digital content can be searched, indexed or be collated instantly. At this point, it is important to emphasize that for such tasks to be performed quickly, easily and satisfactory for the users and the managers, it is necessary to have generated quality metadata. The decisions on this issue will influence other components of the electronic object, ultimately influencing the life-cycle of the digital object. For this reason, an important part of the resources when designing the digitization policy should be invested in the optimal design of the metadata schema that will be implemented and the process to develop it.

The result of a digitization process is a digital object. This digital object may consist of diverse content files and metadata. The three main types of metadata that are generated during digitization are: descriptive, administrative / technical and structural. In summary it may be said that the descriptive metadata are used to

provide access to collections and resources. Administrative or technical metadata allow the institution to manage its digital collection. And finally, the structural metadata are the ones that define how to display the digital object to users and the interrelationships with other objects in the collection or with external resources. It is manifest the important role played by metadata, the reason why it is vital to ensure the correct implementation and management of them.

For creating quality metadata which ensure interoperability and consistency it is very important that the creators use standard schemes.

1.3 STATEMENT OF PROBLEM

There are nowadays different and standardized metadata schemes, such as METS (Metadata Encoding Transmission Standard) which uses XML for encoding descriptive, administrative and structural information about digital objects. Because of its great ability to encode information of various kinds, is one of the most widely used schemes. Other scheme used to encode information of a digital object is DCMES (Dubic Core Metadata Element Set) as a basic framework from which other metadata schemes have been developed. Other typical patterns that encode information describing the digital object would be MARC (MACHINE Readable Cataloging), DTD (Document Type Definition), etc..

The literature on the metadata and digital object creation is very extensive, but for the most of them describe the role and importance of metadata, their interoperability and define metadata schemas or standards. There is also very extensive literature on how to carry out a digitization project (theoretical planning and development of it) but do not specify how to manage it in detail in each of the phases. This gap is due to digitization strategies do not follow a single rigid scheme, they are usually based on a series of recommendations that the managers have to adapt to their project. Each project is unique regarding the objects to be scanned, time, budget, human resources and other aspects that differ substantially from one project to another.

But because properly generated and managed metadata are of vital importance to ensure access, management and preservation of digital objects, it is interesting to define some guidelines on the process of generating and managing metadata during a digitization project.

1.4. RESEARCH QUESTIONS

- How do the studied institutions carry out the digital object production process?
- What is the life-cycle of metadata during a digitization process? By who, how and when are generated?
- What are the main differences between institutions?
- What is the best way to generate and implement metadata during a digitization process?

1.5 AIMS AND OBJECTIVE OF RESEARCH

This thesis' aim is to study the life-cycle of metadata from its creation, its development and management and the role they play during the digitization processes in order to:

- Identify good practices
- Identify potential malpractice
- Identify key moments of the life-cycle of metadata
- Propose guidelines or recommendations to be followed during the scanning process and implementation of metadata

As metadata allow users and institutions find, manage and use digital objects, it is considered of great importance to generate quality metadata to ensure further access to digital information. Although metadata schemes are largely standardized, it would be interesting to analyze through different institutions what type of scheme they have chosen and how they have implemented it in their digitization projects.

The objective is to define a group of good practices or guidelines in order to generate the best metadata for a collection of digital objects which have been digitized from a paper-based original. It is not planned to create a rigid or strict template of how metadata should be implemented, because depending on several factors (type of digitized collection, size, use of it, etc.) metadata needs may vary. But it will be possible to draw conclusions about which methods work best in each type of process, what metadata schemes work best with each collection and how to implement them during the digitization process.

In this point it is important to add that the questionnaires and interviews were supposed to obtain very technical information about metadata creation and management. This objective of the research techniques was not fulfilled during the research process. The researcher found obstacles obtaining technical information as the informants were not aware of all these precise details. As time and means are limited, it was not possible to contact all the staff involved in digitization processes and metadata creation and management. This is the reason why during the research the methodology and approach had to be slightly modified. The approach could not be so technical and had to involve a wider range of aspects of a digitization process. Digitization processes had to be approached from a wider view, analyzing the digital object production line, including metadata implementation.

1.6 BENEFIT OF RESEARCH

- We may improve our understanding of digital object creation and metadata generating in current digitization projects and apply this knowledge in further researches.
- Institutions interested in carrying out digitization projects will be able to avoid mistakes done previously by others and follow guidelines and good practices.
- The researcher will also be benefited from the observations and work done and will be able to apply in further works the knowledge obtained during the research.

1.7 METHODOLOGY OF RESEARCH

This research uses a qualitative approach and the research strategy will be based on case studies. Data will be collected with different techniques: documentation, questionnaires and semi-structured interviews.

The studied Norwegian institutions are the National Library of Norway and the Petroleum Museum in Stavanger (Norway). A third institution will be studied, the Ateneu Barcelonès, in Spain. These institutions have been digitizing their collections for the last decade and represent different kind of memory institutions: one is a big national institutuion (Nasjonalbiblioteket) and carries out all the

digitization process itself as a leading institution; while the other ones are smaller and some parts of their collection are not digitized by themselves, at least in Petroleum Museum's case. Some key informants who work in these institutions will be contacted and will be asked to give information in order to carry out the research.

1.8 LIMITATION OF STUDY

Limitations on this research could be:

- The study approach focuses in only three institutions which are supposed to be representative but it does not mean that all the rest of memory institutions work in the same way.
- The informants are one person per institution. These people work in these institutions but it does not mean they know everything about all the processes carried out. Informants' level of knowledge can be a limitation.
- Some of the required information may be confidential.
- Informants, as part of an institution, may not give completely truthful information. The informants may try to disguise problems or obstacles and transmit a too positive idea of the work done at their institution.

1.9 OUTLINE OF STUDY

This thesis consists of five chapters.

The first one is the introductory chapter. It explains the background of study; statement of the problem; research questions; aims and objectives of the research; benefits of the research; limitation of study and the outline of this thesis.

The second chapter describes the studied institutions and the literature review on digitization practices and different aspects and characteristics of metadata describing the many schemes used nowadays.

In chapter three the methodology of research is revealed. It describes research paradigm; research design; data collection techniques; determination of key

informants; how data will be processed and analyzed and few ethical considerations to take in account during the research process.

In chapter four collected data will be exposed, analyzed and discussed.

Finally, in chapter five conclusions will be drawn. The way of extracting conclusions will consist of giving an answer to stated research questions after carrying out the data analysis. Finally suggestions for further research will be exposed.

CHAPTER 2: BACKGROUND AND LITERATURE REVIEW

2.1 INTRODUCTION

This chapter is dedicated to the presentation of the review of some of the most important works concerning digital libraries, digitization projects and metadata strategies. Digital libraries will be described and digitization process of text and images will be defined. Metadata types and functions will be deeply described such as the definition of metadata; different metadata categories and schemes; metadata creation processes; quality of metadata...

2.2 DIGITAL LIBRARIES & MEMORY INSTITUTIONS

Digital libraries, apart from having the same functions and roles as traditional “physical” libraries, expand their influence to a much wider field.

In *How to build a digital library* (I. H. Witten, 2009) they write this about digital libraries:

“Digital libraries are about new ways of dealing with knowledge—preserving, collecting, organizing, propagating, and accessing it—not about deconstructing existing institutions and putting them in an electronic box.”

In that book a digital library is defined as:

“A focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection.”

Ane Landøy in *Aspects of the digital library* (Garnes et al., 2006) states that implementing a digital library, therefore, digitizing collections as well, consists for the most part of technological and cultural changes and rises new technological needs.

In general terms, all authors define digital libraries as collections of digital information resources accessible through a network. Digital libraries emerged at the same time as new technological resources were developed, and the traditional data management adapted to this new technological tools. One of the big advantages of digital libraries is the improvement of interoperability and resource sharing between different libraries or institutions. Also new digital objects' management and preservation is a new challenge in a digital environment. As in traditional libraries, metadata play a key role in these issues.

The library staff is formed by expert professionals of cataloging and the use of metadata. They are concerned about the information management and retrieval and are familiarized with the use of standards such as MARC, the Anglo-American Cataloging Rules, Library of Congress Subject-Headings, Dewey Decimal classification system, etc. In a digital context metadata are the new way of cataloging objects and this community is the one that has to manage with it. As Dorner in *Cataloging in the 21st century* (Dorner, 2000) states:

“metadata is about standardizing information. Standardizing information is what catalogers have done for centuries.”

In this research one of the studied institutions is part of another kind of memory institution: a museum. Lately, as well as libraries, many museums are digitizing its collections and making them accessible online for the interested users. Also in the same way as in libraries, using metadata standards when cataloging digitized objects allows different museums to collaborate between them (Marty et al., 2003).

2.3. DIGITIZING COLLECTIONS

In *Digitizing Collections: strategic issues for the information manager* (Hughes, 2004) the author says:

“Digitization is the process by which analogue content is converted into a sequence of 1s and 0s and put into a binary code to be readable by a computer.”

The main purposes of digitizing collections are to preserve and to expand the access to items stored in memory institutions.

“Libraries have collections, but these are much more accessible (than archives' collections) and can often be borrowed. One can go to a library and get access to a wide range of Information. Important information is often

stored in a redundant manner, so that the library may lend more copies of a medium simultaneously. An important role for libraries is to provide public access to information and knowledge, but users must make an active effort to acquire it.

The museums are even closer to final users and offer their collections in an entirely different way from archives and libraries. In museums visitors can not borrow objects but will however have a high quality experience and the team of experts who prepared the access to collection plays an important role as they pass a particular view of the objects.” (ABM-Utvikling, 2010).

Digitizing collections is expensive and big amounts of money and time must be invested, as well as staff, designing and coordination efforts, etc. As stated in (ABM-Utvikling, 2010):

“A digitization project requires a good coordination process, organizing what is digitized and what should be digitized across the sector. A part of material is in multiple copies in multiple locations, and it is not always appropriate to spend time and money on duplication of already existing digital objects...”

Both text and images can be digitized in very similar ways, using a scanner or a digital camera and capturing an image resulting in a digital image file. From that moment on, this digital object starts its life-cycle in which metadata will be its “passport or ID” containing all the needed information for its management for the rest of its life.

When first digitization projects were developed, one of the most relevant difficulties was to ensure that digital text could be represented on different computer systems and exchanged across these systems without loss of formatting (Hughes, 2004). These difficulties were solved by developing the standard bodies mentioned before. As the same author states metadata importance lies in that:

“... mark-up systems are the 'glue' that links electronic resources, and enables their interoperability. Similarly, the use of metadata standards will enable the long-term management and re-use of electronic textual resources, and make the long-term preservation and migration of electronic content less of a challenge.”

When designing a digitization project shape, size and condition of the primary source material will dictate how faithful to the original the digital surrogate can be (Hughes, 2004). This author in the same work, *Digitizing Collections*, speaks about digitization of text and images, the issues treated in this research.

In Europe nowadays there is a huge digitization project involving cultural and scientific content. This project, which name is *Minerva* (Minerva Project), tries to create a common frame for different digitization projects carried out in the countries of the European Union. To achieve that standardization in processes they have published some guidelines and good practices that institutions should bear in mind when designing this kind of projects. In their *Good Practice Handbook* (Minerva Project, 2003) edited by Minerva Working Group 6, they give practical guidelines for planning digitization projects, selecting source materials, preparation for digitization, manipulation of originals, use of scanners, digital cameras, software applications (OCR), etc. They also provide guidelines for a correct preservation of master copies (file formats, media choices, migration strategies), metadata, publication (image processing, 3D and virtual reality, online publication), IPR and copyright issues, management of digitization projects (team development, staff training...) and standards (for image, audio, video, 3D, metadata and taxonomies).

Europeana is another example of a thematic network which aggregates and publishes digital cultural heritage objects from memory institutions (libraries, archives and museums) from all Europe. In Norway's case the Nasjonalbibliotek participates, parts of its collections are accessible via *Europeana*. Being based on the aggregation of digital objects from different institutions, one of the main challenges *Europeana* had to face was the interoperability problems as not all institutions base their records on the same points.

In Norway, a local version of *Europeana* could be *Digitalt Museum*. This portal provides an overview of collections in Norwegian museums. The collections that are available are based on artifacts, photographs and art (since September 2010). *Digitalt Museum* has content from 25 to 30 institutions and includes more than 1 million digital objects. It is developed by KulturIT, and data in the portal are retrieved from Primus database, a system for management of museum collections. No automated access methods have been yet implemented (*Digitalt Museum*).

The National Libray of Spain created a digital collection named BDH (*Biblioteca Digital Hispánica / Hispanic Digital Library*) (Biblioteca Nacional de España, 2010) which is still growing as new documents are being digitized and added to the collection's catalogue. In a publication of the Spanish National Library in 2011 they made public all the process concerning to that project. They explain in a very graphic and simple way the steps followed in any digitization project of a memory institution. The steps are the following (Biblioteca Nacional de España, 2010):

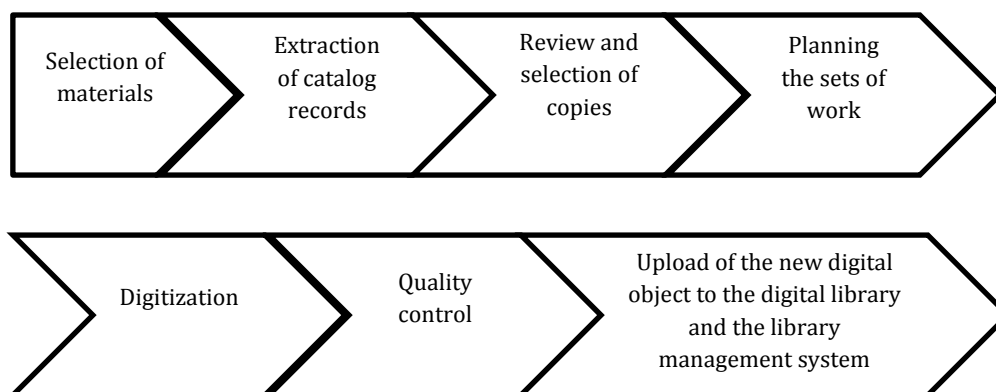


FIGURE 1: steps followed by the National Library of Spain in a digitization project.

They explain in detail how they carry out all the digitization process, including metadata creation and implementation. When they have the TIFF master copy of the digitized image (book page, document, etc.) they create a PREMIS preservation metadata structure. Further, they link dissemination files (PDF, JPEG) with their MARC records, generating a correspondent METS/MARC/COMPLEX/SIMPLE structure. They create descriptive and preservation metadata for each digitized work. Descriptive metadata are in MARC21 (XML) scheme. Preservation metadata follow PREMIS standard and are linked to master copies. See appendix 1 for an example of the PREMIS scheme they use.

Other relevant and interesting initiatives have been developed in France (Louvre Museum, etc.), in the United Kingdom (British National Library, data.gov.uk, BBC open data, etc.) or Germany.

2.3.1 Digitization of written documents

Text digitization can be approached from different angles (depending on the documents and the uses). Hughes (2004) describes two main approaches: the creation of page images on the one hand, and a full-text conversion approach on the other hand.

The first type, creation of images, is not advisable from the point of view of creating searchable content documents. Created documents cannot be processed or edited. One example is Cornell University's digitized historical collections (Cornell Library, 2012). From the point of view of metadata, these kinds of digital objects should have technical, administrative and descriptive metadata, but it's important to mention that descriptive metadata would be limited in some aspects. For example, no METS-ALTO (XML) would be possible, as textual content would not be identified (with OCR or via other resources).

Full-text resources are machine readable and fully searchable (Hughes, 2004). There are two ways of creating full-text documents. One is by using Optical Character Recognition (OCR) software, applied to a page image produced by a scanner or a digital camera. The other way of creating full-text resources is to manually transcribe text into machine-readable form (for handwritten documents). A good example would be *Documenting the American South* (University of North Carolina, 2012), a project in which digitized full texts are accessible from their website both in HTML or in TEI (XML) formats.

2.3.2 Digitization of images

As in the case of written documents, one of the main aims of digitizing photographs is to preserve their content (information) and to make this information more accessible, for example, on the World Wide Web, or via other digital resources in order to make easier for users in other geographical places to reach this information and in order to preserve the original photographs, very vulnerable and easy to destruct.

As Solveig Greve in *Aspects of the digital library* (Garnes et al., 2006) writes:

“Photo-conservation is expensive if done to optimal standards. The photo-archivist is often placed in a difficult position, finding the right balance between allocation of resources to physical preservation and to digitization for preservation of content. Also, there is a continuous pressure from the research community and the general public of making rare and important source information easily accessible through searchable databases and internet publication. All of this makes any digitization project easier to finance than projects of physical conservation.”

2.4 METADATA

Metadata consist of “*information about information*”, as the author pointed out in *Metadata fundamentals for all librarians* (Caplan, 2003), metadata is structured information about an information resource. Any information object should have associated metadata in order to describe its properties. All information objects have basically three features: content, context and structure. Metadata should at least describe these three properties of an information object, regardless its form. In order to provide these three types of information metadata are divided into three categories depending on their functional use: administrative, structural and descriptive.

Metadata are the tools we have to specify the contextual information associated to each document: its content, the history of transformations of each digital object, the formats of each file, programs that allow you to access each record, etc.

The definition of Gail M. Hodge (2004) in *Understanding metadata* for metadata is that metadata consist of structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource. As Hodge says in the same text

“Metadata is key to ensuring that resources will survive and continue to be accessible into the future.”

The same author describes also the main objectives of metadata:

1) Resource discovery. Quality metadata help discovering the information resource in the same way as good cataloging does. It allows the resources to be found by relevant criteria, it helps identifying resources, bringing similar resources together and giving location information.

2) Organizing electronic resources.

3) Interoperability. The description of an object with metadata helps both humans and machines to understand the resource and promotes interoperability. Interoperability means to allow different systems with diverse hardware and software platforms, data structures and interfaces to exchange data without content and functionality loss. Using standardized and defined metadata schemes helps interoperability.

4) Digital identification. Many metadata schemes have defined elements that allow the identification of the digital object and its location (URL, DOI, PURL, etc). In order to avoid objects loss due to location change, it would be interesting to add metadata not only pointing the object, but also as a set of identifying data, differentiating one object from another for validation purposes.

5) Archiving and preservation. Metadata is key for ensuring that resources will survive and continue to be accessible in the future. There are specific metadata elements with archiving and preserving purpose (describing where it comes from and how it has changed over time, its physical description, etc) (Hodge, 2004).

There are different models for metadata management. It is common that metadata are stored independently from the objects they describe, and only maintains a reference that links the original or the digital master file. To ensure that the metadata accompany a digital object in all cases (including the distribution and

copying), it is in some cases desirable to store the metadata in the file itself (ABM-Utvikling, 2010)

2.4.1 Metadata categories

Descriptive metadata: this kind of metadata contribute to the discovery, identification, selection, collocation and access to the digital resource (Caplan, 2003). Their function may also be to indicate evaluation (ratings), linkage and usability. One very well-known descriptive metadata scheme is Dublin Core Metadata Element Set (Dublin Core).

Administrative metadata: the main purpose of administrative metadata is to facilitate to the caretakers the management of the resource (Caplan, 2003). In this group we can include metadata that concern rights management, preservation metadata and technical metadata.

Structural metadata: these kind of metadata hold compound digital objects together and define the relationship between the different components. E.g. how pages are ordered to form chapters.

According to Anne J. Gilliland (Baca, 2008) these are the different types of metadata and their functions:

Type	Definition	Exemples
Administrative	Metadata used in managing and administering collections and information resources	<ul style="list-style-type: none"> · Acquisition information · Rights and reproduction tracking · Documentation of legal access requirements · Location information · Selection criteria for digitization · Version control and differentiation between similar information objects · Audit trails created by record keeping systems
Descriptive	Metadata used to describe or identify collections and related information resources	<ul style="list-style-type: none"> · Cataloging records · Finding aids · Differentiations between versions · Specialized indexes · Hyperlinked relationships between

		<ul style="list-style-type: none"> resources · Annotations by creators and users · Metadata for record keeping systems generated by records creators
Preservation	Metadata related to the preservation management of collections and information resources	<ul style="list-style-type: none"> · Documentation of physical condition of resources · Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration · Documentation of any changes occurring during digitization or preservation
Technical	Metadata related to how a system functions or metadata behaves	<ul style="list-style-type: none"> · Hardware and software documentation · Technical digitization information, e.g., formats, compression ratios, scaling routines · Tracking of system response times · Authentication and security data, e.g., encryption keys, passwords
Use	Metadata related to the level and type of use of collections and information resources	<ul style="list-style-type: none"> · Circulation records · Physical and digital exhibition records · Use and user tracking · Content reuse and multiversioning information · Search logs · Rights metadata

TABLE 1: different types of metadata and their functions (BACA, 2008).

Mind that Gilliland does not talk about the 3 main categories that most authors define (descriptive, administrative and structural). But Gilliland's preservation, technical and use type of metadata could be placed inside the administrative metadata group. Structural metadata should not be forgotten as they are very important in the case of compound digital objects.

2.4.2 Metadata schemes and element sets

Metadata schemes are “sets of metadata elements and rules for their use that have been defined for a particular purpose” (Caplan, 2003).

Hodge (2004) gives a very similar definition to metadata schemes: “Metadata schemes are sets of metadata elements designed for a specific purpose, such as describing a particular type of information resource. The definition or meaning of the elements themselves is known as the semantics of the scheme”.

“Metadata standards have generally been developed in response to the needs of specific resource types, domains or subjects” (Kelly, 2006).

Metadata may have specific semantics, content rules or syntax, but metadata schemes not always define these three aspects. When we say that they specify semantics it means that the scheme defines the meaning of each metadata element and if that element is required, optional or conditionally required. Regarding content rules, metadata schemes define how information or data should be recorded (order of words, use of words or identifiers extracted from an authority file, etc.). Finally, metadata schemes may also define a syntax. It means that they define how elements must be encoded in order to be machine readable. It is important to specify the syntax of metadata in order to provide a common exchange format so that data can be interchanged. Some of the most commonly used syntax in metadata schemes are SGML and XML.

Because of the lack of a completely specified metadata scheme (as pointed above, usually they do not specify all the semantics, content rules and syntax at once so they offer many choices to the user) creators and users of metadata schemes follow previously existing guidelines that help defining a schema. There are informal guidelines, which are quite popular but also there exist (*application*) *profiles* which are formally developed specifications for using certain metadata schemes for a certain use.

There is not any standard for metadata schemes concerning what kind of information they should support and how it should be represented. However, there is an ISO standard for data elements, and as metadata are, as a last resort, a kind of data, this ISO 11179 (ISO/IEC 11179, Information Technology) standard should be applied when metadata creation.

In *A Framework of Guidance for Building Good Digital Collections* (National Information Standards Organization, 2007) listed some of the questions to consider during deciding which metadata standard(s) to adopt:

- What is the purpose of the digital collection?

- What are the goals and objectives for building this collection?
- Who are the targeted users? What information do they need, and what is their typical information-seeking behavior?
- Are the materials to be accessed at the collection level or as individual items, or both?
- Do multiple versions or manifestations of the object need to be distinguished from each other?
- Will the collection or its objects have metadata before the digital collection is built?
- What subject discipline will be involved? What are the metadata standards that are commonly used within this discipline?
- What metadata standards are used by organizations in this domain? Which ones are most appropriate for this particular collection?
- How rich a description is needed, and does the metadata need to convey hierarchical relationships?

Also Anne J. Gilliland (Baca, 2008) specifies some key questions that information professionals will have to have in mind when deciding a metadata strategy:

- Identifying which metadata schemes to use in order to fulfill the needs (it can be one schema or more combined, it depends on the collection and needs)
- Deciding which aspects of metadata are essential for the goal and how granular each type of metadata needs to be.
- Ensuring that metadata schemes and controlled vocabularies (thesauri) and taxonomies (including folksonomies) being applied are the most up-to-date and appropriate for the materials being described.

The *IMPACT Best Practice Guide* (Day et al., 2010) also speaks about metadata standards in digitization projects:

“Naturally, there are many options available for this, but the vast majority of text digitization projects use one of two main standards: the Metadata

Encoding & Transmission Standard (METS) and the Text Encoding Initiative (TEI) guidelines. The syntax of both of these standards is based on XML. The general approach of these two standards differs. METS is a generic means of packaging metadata, content and links together to produce logical objects, and is widely used in a range of digital library contexts. In text digitization projects, METS is typically used as a means of creating logical containers that are able to link all of the content files and metadata that make up a given work, can represent its structure (e.g. page order), and as a means of linking page images with OCR text, e.g. using extension schema like the Analyzed Layout and Text Object (ALTO) standard. The TEI guidelines, by contrast, were primarily designed for the detailed markup of texts, and its use in large-scale text digitization contexts tends not to use all of its features. ”

The most popular markup languages or syntaxes are HTML and XML (cataloguing rules follow AACR) while data structure standards are more numerous: MARC (Machine-Readable Cataloging)... Data value standards such as LCSH (Library of Congress Subject Headings) or AAT (Art & Architecture Thesaurus).

Minerva Group's Guidelines (Minerva Project, 2003) made also a review of the most popular metadata standards. As they state in their good practice guide, it is important in a way to choose an existing metadata scheme that fulfills institution's needs, and the most popular it is the more it will remove search/retrieval problems, exchange problems, etc. In that sense, they mention Dublin Core as it is one of the most popular, among many others:

Following, some of the most popular metadata schemes are introduced (the ones cited in this research, but it is important to remind the reader that many other schemes exist):

Dublin Core

It was created in 1995, originally intended for libraries. The continuing development of this scheme was done by Dublin Core Metadata Initiative (DCMI). It provides a short list of the most commonly used metadata elements and an extension mechanism (Minerva Project, 2003). Nowadays it is widely known and used also out of library contexts (researchers, museum curators, music collectors...), as in the Internet. Dublin Core XML is the required basic XML schema for OAI harvesting. It is simple and concise and because of that, it has been sometimes controversial among specialists and collection managers. Some of them support a minimalist view, emphasizing the need to keep the elements to a minimum and the semantics and syntax simple. Other specialists support a

structuralist view, who defend finer semantic distinctions and more extensibility for particular uses (Hodge, 2004).

The DCMI has gone furthermore from simply maintaining the Dublin Core Metadata Element Set into an organization that describes itself as “*dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for discovery systems.*”

Text Encoding Initiative (TEI)

It is an international initiative to develop guidelines for marking up electronic texts such as novels, plays and poetry, primarily to support research in humanities (Hodge, 2004). TEI also specifies a header embedded in the resource consisting of metadata about the work. This TEI header (and the rest of the TEI) is defined as a SGML DTD (Document Type Definition). It means that the set of tags and rules have a SGML syntax that describes the document's structure and elements.

Metadata Encoding and Transmission Standard (METS)

As Hodge (2004) explains in his text, this standard was created to fulfill the need for a standard data structure for describing complex digital library objects. METS uses XML schema language for creating XML documents with the structural metadata of digital library objects, it also acts as a container for the associated descriptive, administrative and technical metadata and the names and locations of the files that comprise the digital object. “*METS provides a document format for encoding the metadata necessary for management of a digital library objects within a repository and for exchange between repositories*” (Hodge, 2004).

A METS document contains seven major sections very clearly described by Hodge (2004):

- *METS Header: metadata describing the METS document itself, including information such as creator, editor, etc.*
- *Descriptive Metadata: points to descriptive metadata external to the METS document (endorses Dublin Core, MARCXML and MODS descriptive metadata schemes) or to internal descriptive metadata, or both.*
- *Administrative Metadata.*
- *File Section: lists all files that comprise the digital object.*

- *Structural Map: gives a hierarchical structure for the digital library object and links the elements of that structure to content files and metadata pertaining to each element.*
- *Structural Links: allows METS creators to record the nodes in the hierarchy outlined in the Structural Map.*
- *Behaviour: associates executable behaviors with content in the METS object.*

Metadata Object Description Schema (MODS)

This descriptive metadata scheme is derived from MARC21 and compatible with it and it is intended to carry selected data from an existing MARC21 record or to enable the creation of original resource description records. MODS is also expressed using XML schema language. It is a schema that can stand on itself, but usually is used in combination with other metadata formats. It usually works as a METS extension scheme. MODS can give very granular descriptions of constituent parts of an electronic object, that is why it works well with METS' Structural Map for complex digital library objects.

MODS focuses in description of electronic objects and it is richer than other schemes e.g. Dublin Core. Its elements are more compatible with library data than Dublin Core standards, and at the same time they are simpler to apply than the full MARC21 bibliographic format.

MARC 21

It is a long-established standard for exchanging bibliographic records between the library communities (National Information Standards Organization, 2007). This scheme has been enhanced to support descriptive elements for electronic resources.

MARC is not exactly a metadata scheme, but it is part of a multifaceted scheme used in library cataloging (Caplan, 2003) and a bunch of rulesets for cataloging and format specifications that when putting them into practice work as a metadata scheme. These schemes include ISBD, AACR2R and MARC21XML, as well as a MARC Lite scheme. MARC is basically formed by two components: *ANSI/NISO Standard Z39.2*, which provides a machine-readable structure for records; and *MARC21 Format for Bibliographic Data* encoding rules.

PREMIS Data Dictionary for Preservation Metadata

This is a set of core preservation metadata elements developed by Preservation Metadata Implementation Strategies (PREMIS).

2.4.3 Metadata creation

Metadata creation is one of the main activities for memory institutions. As told before, creating quality metadata is essential for the preservation, display and dissemination of information objects.

Metadata schemes may specify a syntax, it means to define a common way of recording data in order to make them machine readable and interchangeable between systems.

As Anne J. Gilliland (Baca, 2008) states:

“Metadata creation and management have become a complex mix of manual and automatic processes and layers created by many different functions and individuals at different points during the life-cycle of an information object.”

As we can see, metadata creation process is complex and many agents take part in it. Here we can see Anne J. Gilliland's graphic representation of the life cycle of an information object:

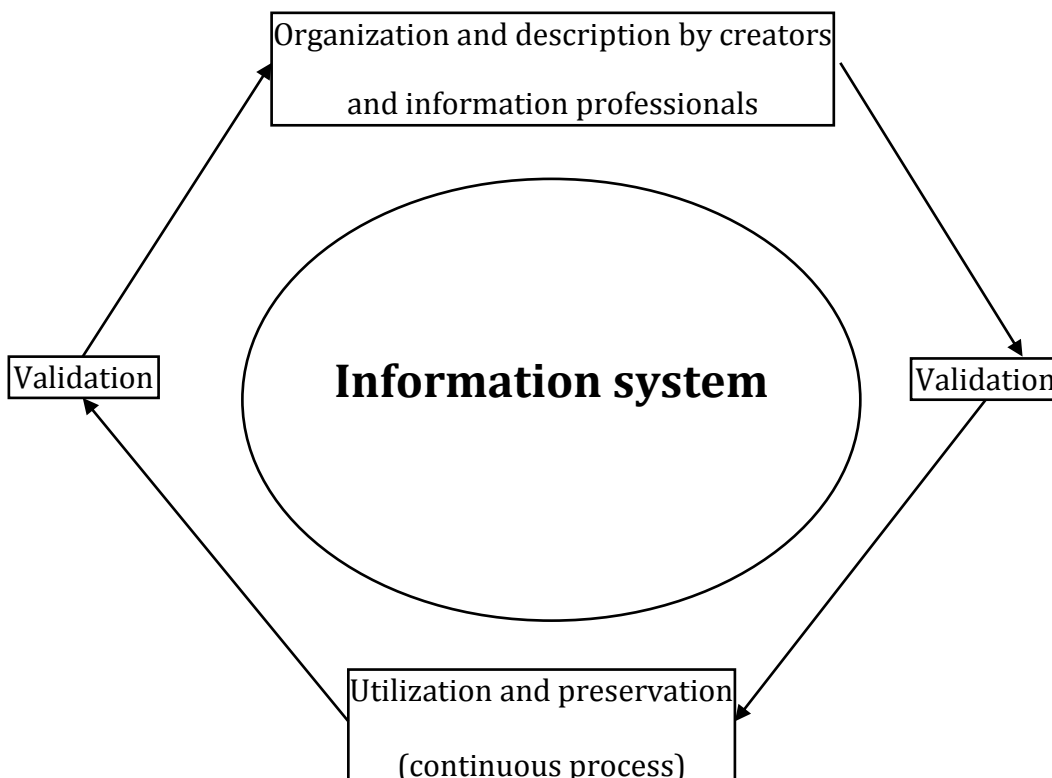


FIGURE 2: life cycle of an information object (Baca, 2008).

In each of these phases different kinds of metadata are added. Metadata can be added inside the information object itself (METS packs, which include structural, administrative and descriptive metadata, with the information object inside a packet and specify the relationship between the different elements). Metadata can also be added as exempt elements (not inside the same pack as information object) and be connected with the information object through hyperlinks. Some authors, as Anne J. Gilliland (Baca, 2008) recommend integrating metadata within the information object, avoiding the storage and hyperlinks elsewhere.

When digitizing a collection some administrative and descriptive metadata should be included by the digitizer. Then, information managers should add more metadata while registration, cataloging and indexing processes. But during further stages of information object's life-cycle also more metadata may be added, even user-created metadata may be generated (i.e. folksonomies) (Avery, 2010).

In libraries when metadata are created by individuals they are usually created in two ways. On the one hand resource description can be exported from open resource catalogs (OCLC and RLIN) and be represented in the most used metadata schemes as MARC21/AAC2 or Dublin Core. In this way, librarians reduce the time invested in metadata creation and ensure that data will be correct and without "human" errors. Another way of creating metadata records is using library's ILS (integrated library system), most of these systems support MARC21 records.

There exist several tools that help information professionals to create and implement metadata. Hodge (2004) lists some of them:

- Templates: metadata values are entered into preset fields that match the element set being used. The template will automatically generate a formatted set of element attributes and their corresponding value.
- Mark-up tools: for structuring metadata attributes and values into the specified schema language. Most of these tools generate XML or SGML DTD (Document Type Definitions).
- Extraction tools: create metadata from an analysis of the digital resource. These tools are limited to textual resources and the quality of the extracted metadata can vary depending on many aspects. These tools should be an aid during the metadata creation phase, but not the main tool for creating them. The resulting metadata should always be manually reviewed.
- Conversion tools: translate one metadata format to another.

NISO (2007) published a chart in which they advise information professionals with guidelines and application of each metadata scheme listing one by one all the

online resources where templates, examples, implementation tools, or any other useful online resource.

As seen before, the Minerva Working Group published very complete guidelines covering the many different aspects of a digitization project. With regard to this thesis, I will focus on what Minerva Group's Guidelines say about metadata. They emphasize the importance of choosing the correct metadata standard, the one that best describes the information object and the goals of its digitization. They recommend using preferably existing metadata standards, and only in specific cases when standards cannot cover the use and goals of a collection, a new metadata model should be created. They encourage the collection managers to spend the needed time to identify and implement the best metadata (key attributes and descriptors) as it means investing in a further efficient search, exchange, etc. of information objects. They also encourage to use controlled vocabularies (if exist for the concerning kind of elements/attributes) to make metadata more standardized and so to increase the success of searches.

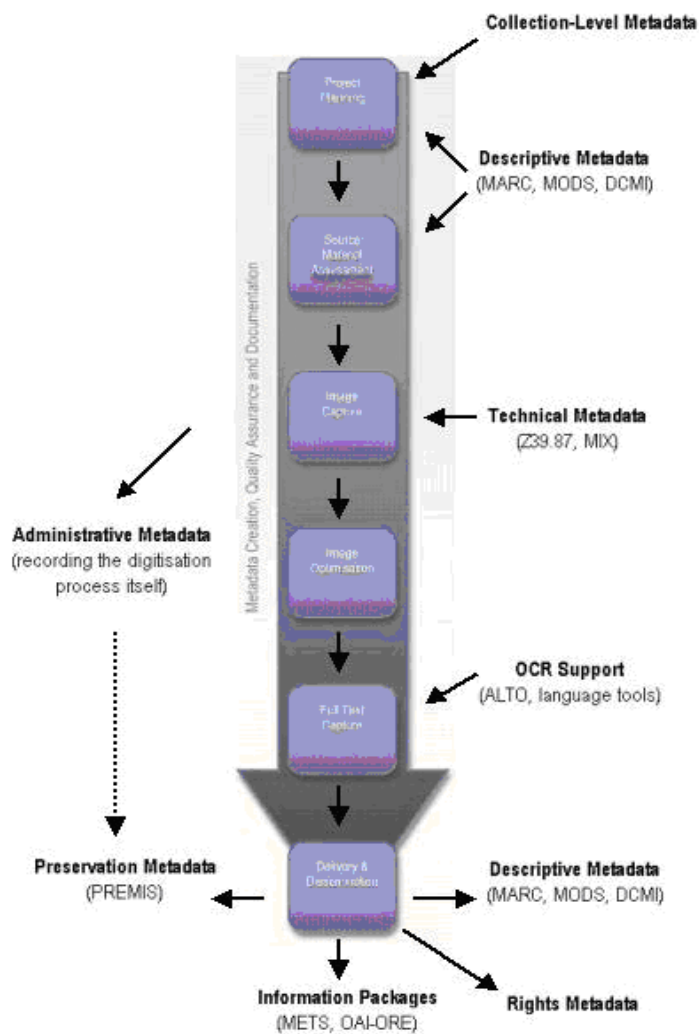


FIGURE 3: the workflow of metadata in a digitization Project (Day et al.,2010).

Day et al. (2010) give some general recommendations for a good practice in digitization projects:

- Use existing standards (wherever possible)
- Reuse legacy metadata (wherever possible)
- Automatically capture metadata (wherever possible)
- Identifiers are important
- More on persistent identifiers
 - The Digital Object Identifier System (DOI)
 - Uniform Resource Names (URN)
 - ARK identifiers

2.4.4 Metadata quality control

During the metadata creation phase there can occur errors or quality problems (Hodge, 2004). In *A Framework of Guidance for Building Good Digital Collections* published by NISO (2007) they define 6 principles to create quality metadata:

1. Metadata Principle 1: Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential future uses of the collection.
2. Metadata Principle 2: Good metadata supports interoperability.
3. Metadata Principle 3: Good metadata uses authority control and content standards to describe objects and collocate related objects.
4. Metadata Principle 4: Good metadata includes a clear statement of the conditions and terms of use for the digital object.
5. Metadata Principle 5: Good metadata supports the long-term curation and preservation of objects in collections.
6. Metadata Principle 6: Good metadata records are object themselves and therefore should have the qualities of good objects, including authority, authenticity, archivability, persistence, and unique identification.

CHAPTER 3: METHODOLOGY

3.1 INTRODUCTION

This chapter defines the methodology used to conduct the research. The research is a complex process which design should take into account various factors. First, it is necessary to define the theoretical framework of the research process and later, according to the nature of the object of study, research strategy will be designed.

The research will be based on a descriptive survey. The extracted and analyzed data will be qualitative and based on real case studies mainly drawn from questionnaires and face to face interviews and if not possible, from online interviews. Documentation will also be consulted as a technique of data collection.

The staff responsible for designing and coordinating various digitization projects will be contacted and will be sent a questionnaire and interviewed in order to collect precise information to analyze the life cycle of metadata during the scanning processes: how metadata were generated, who generated them, when they arise, how they connect and interact with other elements that are involved in the process of digitization ...

Afterwards, in chapter 4 and 5, collected data will be analyzed and conclusions will be drawn.

3.2 THEORETICAL FRAMEWORK

Any research has to define its philosophical approach as a starting point, in order to determine the research's design and development. There are three big questions that help determine the focus of the research (Lincoln and Guba, 1985):

- *“What is the nature of reality? (Ontology)”*
- *“What is the nature of the relationship between the knower and the known? (Epistemology)”*
- *“How to get to know? (Methodology)”*

In the case of social sciences, and in particular, information science, the research

approach that best fits and is most popular among researchers since the mid-twentieth century is the interpretivist research. The first research papers that were made from the empirical approach of interpretivism were framed mainly in the field of ethnography and the study of human groups. Subsequently, researchers tried to give this approach a body of theory that would allow this type of analysis to be applied to any scientific research.

At the ontological level, interpretivism is based on relativism, i.e.: realities are multiple and holistic, or what is, reality is actually conceived more as a whole (a sum of components) different from each of its component parts. The reality is constructed by the subject and cannot be dissociated from the context that surrounds it. At epistemological level, interpretivism is transactional and subjectivist. All knowledge is acquired as a result of the interaction between the researcher and the object of study and the descriptions of reality are limited in time and context. Finally, this philosophical approach to science raises at methodological level an empathic interaction between the researcher and the object of study. The data extracted from an investigation might be in itself product of research. The timing and the context in which data are taken also influence them. The interpretivism seeks to understand the whole context (both micro and macro) and for this purpose, the methodology used is qualitative. To this end, case studies are carried out, as in the present work. The main purpose of this interpretive study is to transfer the findings to other contexts in which they could be applicable.

3.3 RESEARCH DESIGN

The research will follow a qualitative research model. A qualitative methodology assumes that reality is constructed socially (Gorman et al., 2005). Qualitative research cannot be planned in great detail and in part will be designed "on the fly", but there have been created design patterns or guidelines that can be useful in this case.

The model proposed by Pickard (2007) (adapted from Krumar, 1999, and Lincoln and Guba, 1985) is as follows:

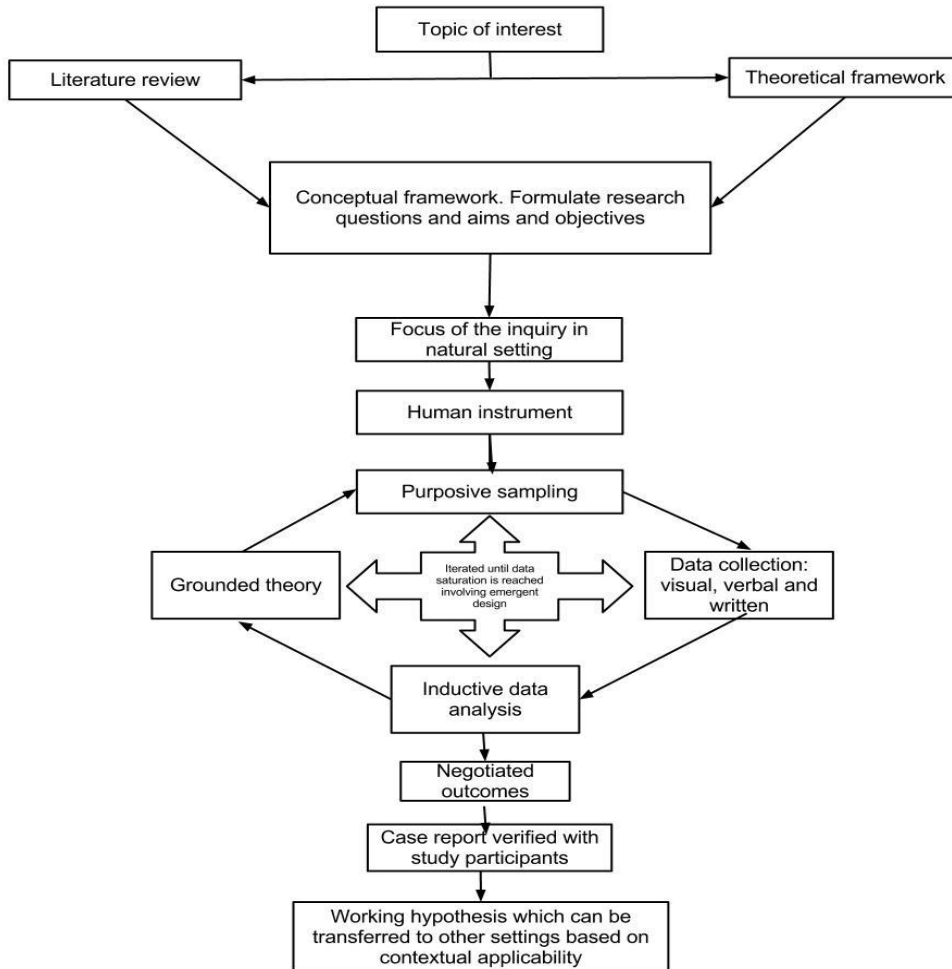


FIGURE 4: the model proposed by pickard (2007) (adapted from kumar, 1999, and lincoln and guba, 1985).

During the research process it will be essential that the human instruments apply a suitable data collection techniques complemented with tacit knowledge of the researcher.

The main phases of qualitative research include: literature review, the definition of the theoretical framework, field work, (real case studies in their natural environment) using a human instrument (researcher), the proper techniques for data collection, inductive analysis, emergent design, etc. and finally propose a hypothesis that allows transfer the results to similar contexts.

In qualitative research the researcher is a human tool, actively involved in research.

3.4 RESEARCH STRATEGY

The research strategy is based in case studies. The methods that are chosen will be those that help obtaining the necessary information to complete the research.

The case study is an empirical approach to a contemporary phenomenon in its real context and when the boundaries between phenomenon and context are not so obvious, and when using different evidential sources (Yin, 2009).

A case study is basically a particular study in context. For this reason, in the current research the most appropriate methodology is this, because it is intended to extract information from individual cases (National Library of Norway, Petroleum Museum and Ateneu Barcelonès). In this case we have instrumental case studies, research focuses on investigating a particular phenomenon: the creation and implementation of metadata during a scanning process, and then extract the knowledge and guidelines on what would be a good practice.

The different phases of this case study are:

Phase 1:

- Focusing: Establishing a research question
- Determination of what kind of case study approach use (single, collective...)
- Site: Nasjonal Bibliotek, Petroleum Museum and Ateneu Barcelonès
- Unite of analysis: creating and implementing metadata in a digitization process
- Determination of the techniques of data collection: questionnaire and interviews. These techniques are the best suited to the research questions previously raised while they are feasible and involve techniques that fit the time and means that the researcher counts with.

Phase 2:

- Data collection: send the questionnaires, record and transcription of interviews, observation notes, etc. In the data collecting the basis will be the previous knowledge of the researcher, who must be open and willing to expand it. At this stage interpretation of any kind is not yet carried out, and there will not be a classification or creation of categories of data at that moment. Data will just be collected and stored.

Phase 4:

- Writing conclusions and new knowledge extracted.

Using case study technique has also some disadvantages. The generalities we can extract from a case study are more vulnerable to criticism.

3.5 COLLECTING DATA:

Following the data collection techniques are introduced. Besides the study techniques used, testing them in advance is essential for the research.

In this research the main data collecting tool is the questionnaire. Afterwards an interview is made to obtain more in depth answers that were not answered in the questionnaire for different reasons. Interviews are a supporting tool to deepen understanding questionnaires' data.

Questionnaire

In this case the questionnaire will be a major data collecting tool of the research. In order to design an efficient questionnaire these steps are followed (Pickard, 2007):

1. Review data requirements of the research questions
2. Write a list of potential questions to provide the needed information
3. Prioritize list of questions
4. Evaluate each potential question using the following criteria:
 - Can potential participants understand the question?
 - Are they in a position to answer the question?
 - Will they be prepared to answer the question?
 - Is there ambiguity, bias or potentially offensive content?
5. Determine form of questions: open-ended, closed or scale items.
6. Construct specific wording of each question.
7. Organize the structure of the questionnaire.

8. Evaluate and pilot the questionnaire.
9. Make necessary amendments and distribute.

Interviews

Interviews are a very useful research technique when we seek in-depth qualitative data. In this case we want to know how a process was performed in the past and how they continue doing it, for this reason, interviewing the library staff will be the most fruitful way of obtaining the data we need. The interviews allow us (along with interviewer-interviewee interaction) to reconstruct events, gather opinions, thoughts, feelings, etc.

Following the guidelines of *Interviews : an introduction to qualitative research interviewing* (Kvale, 1996) an interview process can be divided into 7 phases: thematizing, designing, interviewing, transcribing, verifying and reporting. For designing a good interview these steps will be followed:

- Thematizing: (defining the “why” and “what” of the research). This is linked with the research questions stated before.
- Designing: the interview guide will be designed in this stage. In this case the interview will be semi-structured. The main purpose of this interview is to extract the feelings and opinions of the interviewee, in this case about how they carried out the digitization process and metadata implementation during scanning. The questions will be open-ended so the interviewee can explain his/her own point of view on the experience. The interviewer will prepare a guided interview with a prepared list of subjects that should be covered during the interview.
- Recording: the interviews will be recorded not to lose any important information and to be able to consult over time as many times as necessary this source.
- Transcribing: as soon as possible after the interview it will be transcribed in order to extract the most important information for the research and to make a first “analysis” of the provided information.
- Analyzing: it will be a constant process (not only carried out during the last stage).

3.6 DATA ANALYSIS:

(Miles and Huberman, 1994) described data analysis process as follows:

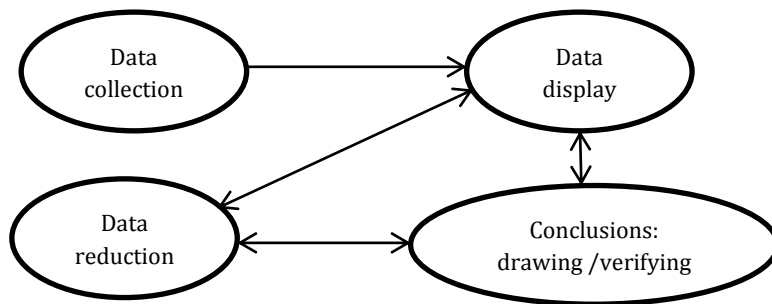


FIGURE 5: data analysis process (Miles and Huberman, 1994).

The qualitative data collected from the questionnaire and interviews must be organized and reduced in order to make easier the analysis. Data will be ordered and transcribed so the researcher can visualize and analyze them.

3.7 ETHICAL CONSIDERATIONS:

Ethical aspects are very important in any research.

All informants of this research will be always informed of the nature of the project, the use of the provided information and the treatment it will receive. The informants will receive a consent letter in which the researcher will emphasize the volunteer nature of the participation and will ensure respondent's confidentiality.

“Integrating ethics into the research process, from selecting the research problem to carrying out research goals and interpretation and reporting research findings, is critical to ensuring that the research process is guided by ethical principles beyond informed consent.” (Hesse-Biber and Leavy, 2011)

3.8 LIMITATIONS IN STUDY

During the research process the researcher may find some obstacles or limitations on the research.

First of all it is important to remind that the study approach focuses in only three institutions which are supposed to be representative, but it does not mean that all the rest of memory institutions work in the same way. With such a limited sample, the results and conclusions may be partial and not easy to extrapolate to other institutions or digitization projects.

The researcher must also take in account that the informants are just a single person per institution. These people work in these institutions but it does not mean they know everything about all the processes carried out. Informants' level of knowledge can be a limitation as they may not have all the requested information.

In the same line of limitations, it is important to know that some institutions may have confidential information for diverse reasons. The researcher may not collect all the needed data because of institutions' confidentiality policies.

Finally, informants, as part of an institution, may not give completely truthful information. The informants may try to disguise problems or obstacles and transmit a too positive idea of the work done at their institution.

CHAPTER 4: DATA ANALYSIS AND FINDINGS

4.1 INTRODUCTION

In this chapter collected data from the questionnaires (see appendix 4) and interviews (see appendix 5) and findings will be described. We will discuss how the Norwegian Petroleum Museum and Norsk Nasjonalbibliotek implemented metadata in their digitized collections.

Key informants were asked to describe their metadata implementation strategy. As mentioned in the methodology chapter, an online questionnaire and an interview were the used tools for collecting the needed information. The obtained data were classified in different subjects.

As stated before in chapter 1.5 it is important to mention that the questionnaires and interviews were supposed to obtain very technical information about metadata creation and management. This objective of the research was not fulfilled during the research process. The researcher found obstacles obtaining technical information as the informants were not aware of all these precise details. Digitization processes had to be approached from a wider view, analyzing the digital object production line, including metadata implementation.

Following the collected information is shown:

4.2 BACKGROUND STUDY OF KEY INFORMANTS

For this research three people were sent the questionnaire and interviewed, one from each of the studied institutions. One of them was the responsible of the technical department of the National Library of Norway (Norsk Nasjonalbibliotek). Another informant was the responsible of the photography collection of the Norwegian Petroleum Museum (Norsk Oljemuseum). Finally the manager of the library of Ateneu Barcelonés was also asked about their digitization projects.

4.3 CURRENT DIGITAL OBJECT CREATION LINE

Thanks to the questionnaire and interviews with the key informants it is possible to extract the needed information to describe the current digital object creation and implementation procedures carried out by these two institutions which are digitizing part of their collections.

4.3.1 NATIONAL LIBRARY OF NORWAY

The National Library of Norway (Nasjonalbibliotek, NB from now on) has been digitizing its collection during the last 16 years and is still doing it. The NB participates in joint digitization projects with other institutions (Petroleum Museum, National Archives, Parliament...). The materials that have been digitized are very eclectic: books, newspapers, periodicals, manuscripts, photographs, films, radio programs, TV programs, music... In this research only textual documents and photographs were taken on account. This means that both the IT development and the scanning activities are decided on a business level and according to plans made for cooperative projects, user demands or strategic reasons.

The main reasons for digitizing their collections were to preserve the originals and to improve the accessibility to its content. For that reason, their digital collections are published online in NBDigital-Bokhylla (NasjonalBibliotek) web site. The access to this digital collection is made in two ways: digitized objects can be read online or users can download the object in PDF format (when copyrights allow it). Bokhylla is a national service under the auspices of the National Library. More archives and museums have registered their book collections in the national library systems and opened up the collections for searches based on the internet. The solution to the Bokhylla is identical to the National Library. It uses both the unique identifier of digital objects (URN: NBN) and a URI that leads directly to the digital object. Metadata are not linked to the objects, but they are based on text strings.

In 2005 the NB made an important decision consisting of digitizing every single object of their collection. That decision gave them a quite open frame to design their digitizing policies and eliminated the need of prioritizing and the problem of selecting materials. They have to invest some efforts in deciding what to digitize first, but they do not need to invest time and efforts in selecting what materials will be digitized and what materials will not.

“The digitalisation of National Library’s collection happens in line with the demands which are made in relation to long-term storage of digital content,

and the National Library establishes, in collaboration with international participants, necessary standards for this. The digital objects are enriched by metadata and lasting identifiers which increase the possibilities for preservation, use and re-use over a 1000-year perspective. The National Library arranges for numerous and varied uses of the content in the collection. The content is made available in attractive formats.” (National Library of Norway)

The main factors that will influence NB’s scanning practices in the near future will be staff commitment in a very important level, the development of policies on scanning and digital collection management, the increasing number of digital objects, technology and users’ needs. More technical issues, as metadata schemes, are not determinant factors at all when designing and carrying out scanning practices.

4.3.1.1 Some technical issues

The metadata schemes used by the NB are MARC21, Dublin Core and METS-ALTO. These schemes were chosen because of their flexibility and due to their interoperability, because they support information sharing. Other reasons were that these schemes are supported by leading organizations in library field and they are widely used, also NB had used them previously. The usability (easy to use) of metadata schemes was not a decisive factor. Anyways, for each type of material there were different reasons to choose its corresponding metadata scheme.

They have a metadata preservation policy consisting of the creation of a “preservation package” containing:

- A lossless copy of all images (in JPEG200 format).
- XML file with descriptive, structural and administrative (including preservation) metadata.
- XML file (METS-ALTO) file with the content (OCR-text and structural information of the content).

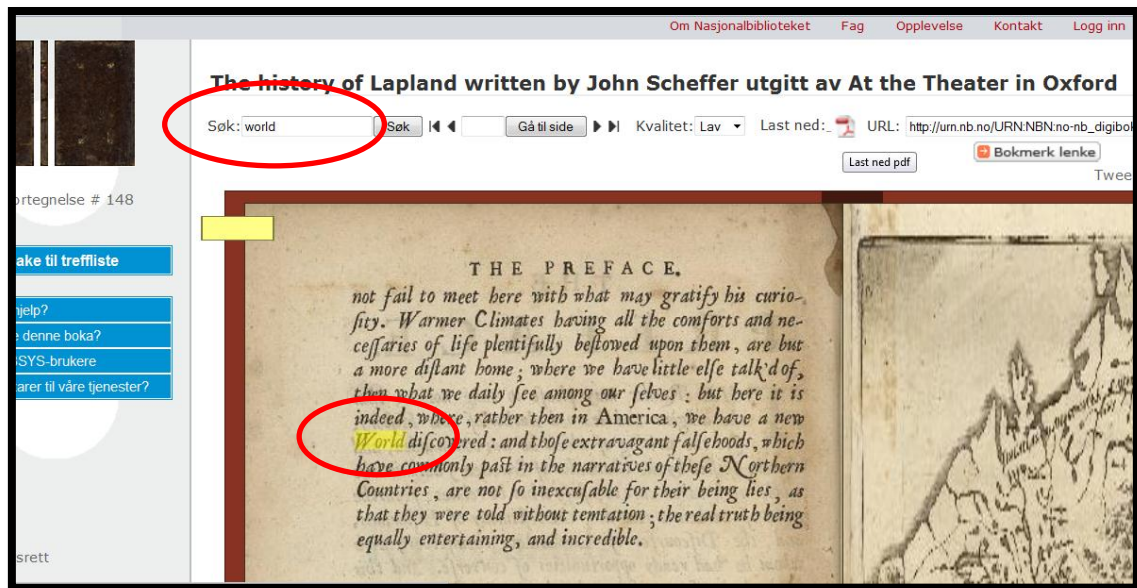


FIGURE 6: example of how after applying an OCR it is possible to make searches by content.

This “packages” are stored in a long term preservation system. Three copies are made; one on disk and two on tapes.

Another challenge was to try to decide to use a format for digital files that could cover all the NB needs of storage, display, etc. That is why NB staff chose JPEG2000 format for photographs and XML based schemes for descriptive and structural metadata. Metadata files are stored in the same folder as the digital object. High resolution image files are stored as JPEG2000 because it is a “lighter” format than TIFF and allows them to create from it copies in other formats, so the TIFF files generated during scanning process are deleted. Then this storage copies are stored in a long term storage repository. They also generate lower resolution JPEG2000 files which are stored in their image system for the purpose of being displayed in the internet for the final users.

Concerning interoperability the NB staff stated that they have not experienced any interoperability problem. They use OAI-PMH tool (Open Archives Initiative – Protocol for Metadata Harvesting). In addition to that they have implemented services which will make any other institution able to implement their own application when they search in NB database/search engine and results are returned in XML files.

In the NB they do not have explicit guidelines about how to carry out all these processes but they have some rules about how to manipulate materials, how to manage the OCR application, etc.

They are not planning to change their work policies and routines for the moment. What NB is doing right now is trying to establish new working lines for new types of materials. I.e. at the moment they are trying to implement a production line for film scanning.

See appendix 2 for an example of a MARC21 record of an analog item and a digitized item in NB.

4.3.1.2 Digital objects' production line:

In the next pages the digital object production line will be described and analyzed. As stated before, the analysis will focus on books and text documents (not in audiovisual content objects).

Books

All books belonging to the NB's collection had already been catalogued in Bibsys database as analog objects and they have a barcode associated to an ID. At the same time this ID is associated to a set of metadata stored in Bibsys database. In Bibsys descriptive metadata are stored following MARC21 metadata scheme.

When the NB staff, are about to scan a book, in the first stage they scan the barcode of the analog book and they get back its ID, which allows them to recover the already existing descriptive metadata of the physical object (book in this case). This search and recover of existing metadata is an automatic process carried out by an application implemented on their system. The user (NB staff) receives the set of metadata and this application displays them for the user. Then they only have to check that the recovered data join the book and accept it so they can go on with the process. At the same time, the recovered metadata are stored in the local system of the scanning workstation. By doing that they get the ID of the physical object and its set of descriptive metadata that will be added to the metadata set of the new digital object that is about to be generated.

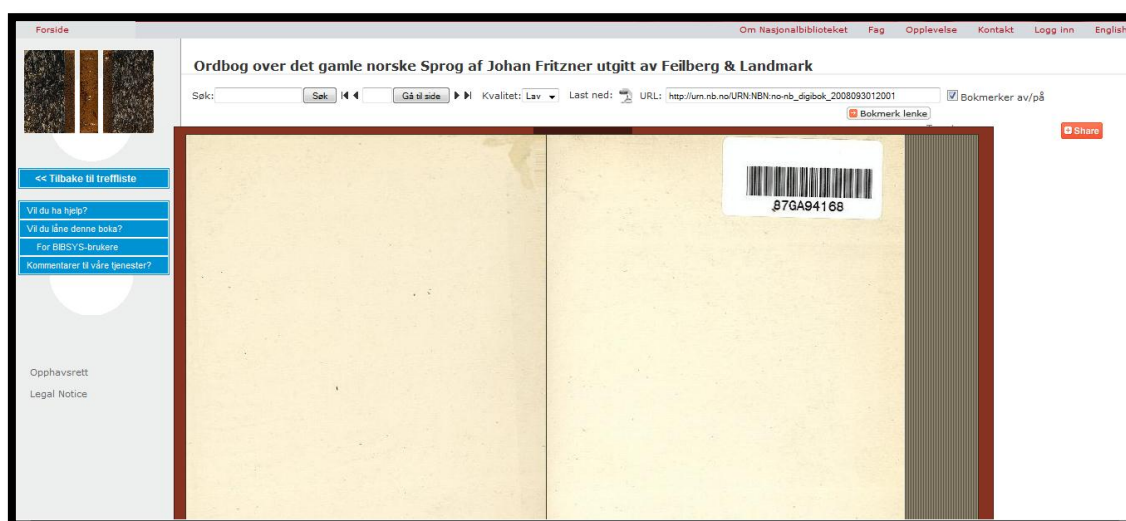


FIGURE 7: detail of how the final user sees the digitized book. In this caption the barcode with the physical books id can be seen.

It is in that stage when the ID for the new digital object is generated and the catalog files are prepared, where the digital images generated during the scanning process will be stored. Then they create an XML file where they record both the physical object's ID and new digital object's ID. This file is very important, that is the link between both objects (analog and digital). When this file is generated the scanning process begins, starting from books' covers. A barcode is printed out with the new digital ID and they stick it in the cover. After that (the object already has two IDs) the book is moved to the next scanner where the content is scanned. The result of this process will be TIFF format images which are always migrated to JPEG2000 high resolution images, those will be the storage copies. An OCR process is also applied in which they analyze the content and its structure and generate xml files following METS-ALTO scheme in order to allow further searches by content.

Now they have both ID's (analog and digital objects'), structural metadata, descriptive metadata, METS-ALTO (xml) files with the content description, master copies (JPEG2000) and display copies (also JPEG2000 but not so high quality as master copies). See appendix 3 for a METS-ALTO file of NB.

After that, the last step of the production line is to tell the catalog that "a certain physical book with a certain ID has been scanned and must be linked with a certain digital object with a certain ID". This means adding structural metadata. They use a tool which sends all the digital object's data to the catalog and the catalog automatically creates a new record for the digital object and links both ID's (using

the URN of the digital object). By doing that they get in their “traditional” catalog the references to their physical books and, in the case of books that have been digitized, the ID (URN – name) of the digital object, in this case the digital version of the physical book.



FIGURE 8: inside the red circle the persistent link based on the URN of the current page.

Books, newspapers and periodicals (text documents) are scanned and TIFF images are obtained, one image per page. All the image files are stored in the same folder (called i.e. “digibok_2007022801018”), named with their correspondent page number (i.e. “digibok_2007022801018_0001.tiff”). Then they apply the OCR, which extracts the text content and will generate an XML file (following METS-ALTO scheme) with the same name as the TIFF files’ name but with .xml extension (i.e. “digibok_2007022801018_0001.xml”). Then, they convert all TIFF files to JPEG2000 format, which is the storage format. They chose JPEG2000 for its flexibility, as it lets them create high quality images (for storage) and lower quality images (for display).

When the scanning process is finished and all the stages of the production line are complete, the catalog will hold the information for both the physical and digital objects. Their retrieval and information presentation system is based on a search engine that harvests all the data from the catalog (including new digital objects) and stores data in the search engine, which makes the final user able to search even for content in a book (thanks to METS-ALTO metadata files). So the presentation system, in the end, harvests everything from the catalog, the catalog is where everything is stored. This search engine uses OAI protocol for metadata harvesting

Books and periodicals are cataloged in BIBSYS catalog system.

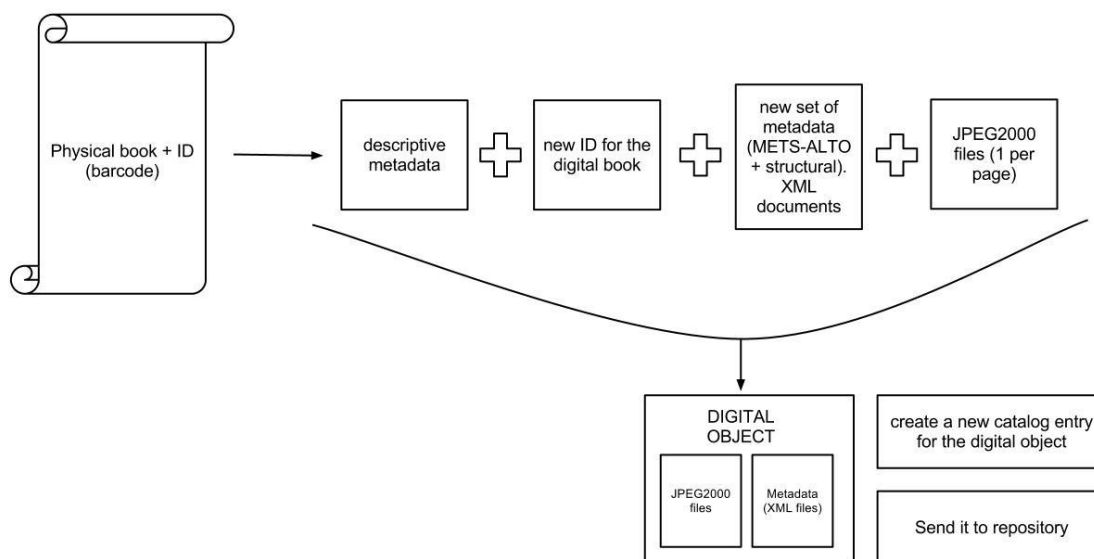


FIGURE 9: NB's digital object production line.

Photographs

For photographs a different catalog (MAVIS) is used. In the case of photographs the staff is not as strict as with books on how to register physical photographs in their catalog. They create a catalog record with a set or a collection of photographs but not a record for each one of the pictures, so each photograph does not have its own metadata. In that case the challenge is to represent each one of these photos. They can extract from their MAVIS catalog metadata in XML format and then move these files to a separate system (PhotoStation software) which is implemented with the purpose of maintaining and managing metadata (adding tags, etc.). In that way the NB staff can take a set of digital photographs that have been digitized and add more metadata elements, i.e. tags... When metadata have been added, the metadata are exported from PhotoStation as an XML file and are sent back to the catalogue.

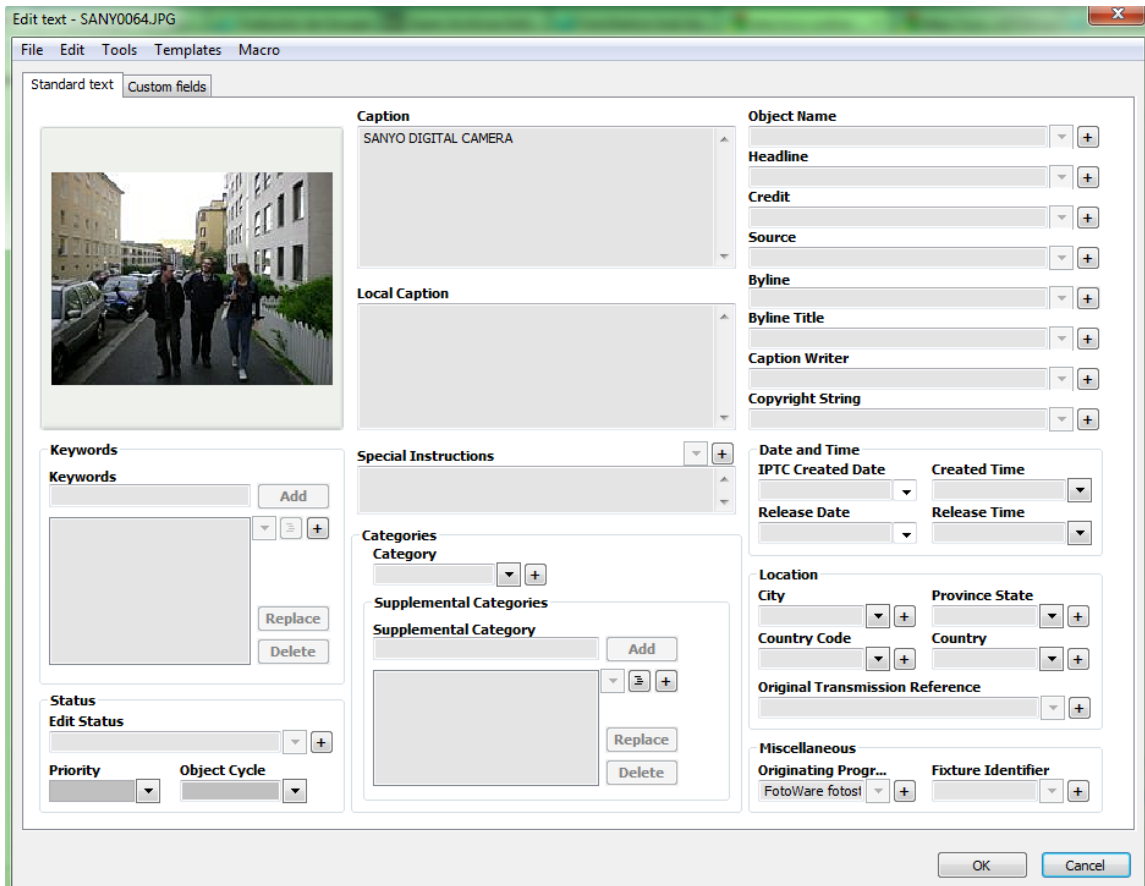


FIGURE 10: Photostation's default tool for editing and adding metadata. These metadata fields can be customized.

PhotoStation software has several metadata editors which NB staff specialized in cataloging use to manually add descriptive metadata. The underlying scheme in this software is XMP (Adobe Systems Incorporated, 2005), which embeds metadata to the digital photograph using XML language.

One of the reasons they chose MAVIS cataloging system is because it is quite flexible when representing relations between objects (photographs), so they can create collections (by photographer, subject, place, year...) and add descriptive metadata as a collection, at the same time as adding metadata to each individual digital photograph. When the digitization process of photographs is complete they will have collections (with their own ID, the same as in the physical collection and metadata) and inside each collection an asset of digital photographs (also with their new born ID which is the URN of the digital photograph and its own metadata).

Newspapers

Three different working lines are used for newspapers. One for paper (analog) newspapers, which are digitized in the NB. Another one for their newspaper microfilms collection and a third line for their “born digital” (delivered to the NB as pdf’s) newspaper collection. All of these 3 different materials are digitized (in case of paper and microfilm) and OCR-analyzed. They create JPEG2000 files for storage and in the case of the born digital newspapers they also store the original pdf’s. As with books, they generate one JPEG2000 file and one METS-ALTO (xml file with content) file for each page, and one METS file with descriptive, administrative and structural metadata (production info., etc.). In the case of digital born newspapers, NB has an agreement with some of the newspaper houses to deliver them these pdf’s with an xml file with a good and complete set of metadata useful for NB. They agreed with this houses a rule for file naming where the different parts of the name would represent metadata: name of newspaper, data, number, section, etc. that make up the complete set of descriptive metadata that the NB needs. That is how the solved the metadata issue for newspapers. Once they extract these metadata they store them in the MAVIS catalog.

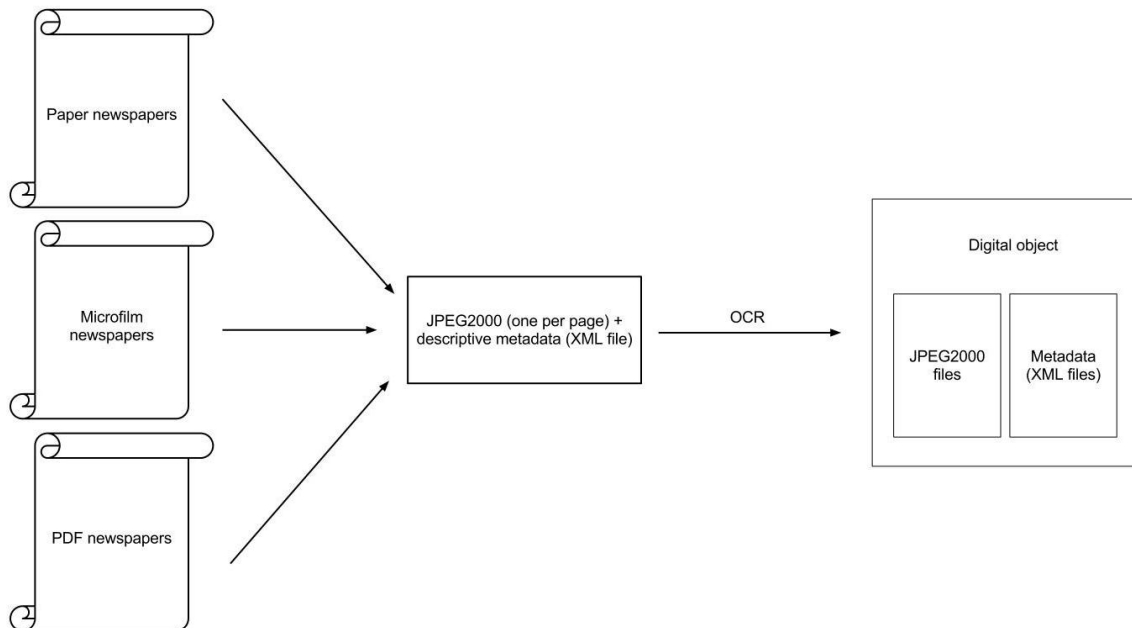


FIGURE 11: digital newspapers production line in NB.

4.3.1.3 Problems/Challenges

One of the first problems the National Library staff faced was to decide what metadata scheme to use; they defined that task as quite hard. Other quite hard problems they faced during this process were that they found several confusing metadata concepts, difficulties determining which metadata elements were really useful for users and staff, the lack of data on the digitized materials, the need of high qualification and skills on the part of the staff and not having enough available documentations. On the other hand, in National Library's case, budget was not a problem.

Norway's National Library has quite a lot of guidelines about metadata creation and implementation. This task is done by staff and is highly automatized. The main sources for adding metadata are digitized materials themselves and their packaging. In few cases fieldwork is also a source for adding metadata.

The descriptive metadata elements they consider more important are ID of the physical object, ID (URN) of the digital object and other descriptive metadata of both objects. They are used for making searches as these descriptive metadata contain information of the object (title, author, year...) and of the content (applying an OCR they get METS-ALTO .xml files which locate content and have content information that allows searches by words or phrases which are part of the book's content). For NB staff structural metadata elements are also very important, as the ones that represent the relation between the digital object (ID and URN) and the physical object (ID) and relation with other entities. In addition to metadata stored in the catalogue, the complete set of metadata is saved in a METS xml file and stored in the preservation storage

The informant stated that they did not have important problems. But they did find some challenges and he emphasized as an important challenge to link the physical object to the digital object in their catalog. (In the case of books it was "quite easy" because all their collection was registered in the catalog so every physical book had its own and unique ID previously, not like in photos, sometimes they did not have an ID for each photograph, but one ID for a whole collection/set of photographs).

The NB staff decided to create a document with the object ID, the title, author and year of book. They have one of these documents for each analog edition of the book (in case they have more than one copy). When they started digitizing books they decided to represent the digital object at a lower level. So right now if they wished to digitize different editions of the same item (book) and tell the system that this specific edition of the physical book corresponds to that specific digital copy, it is not possible because the digital record is not in the same level. For part of the staff this represents a problem. That is something they learnt when they started carrying out all the process and now they are discussing to improve this aspect.

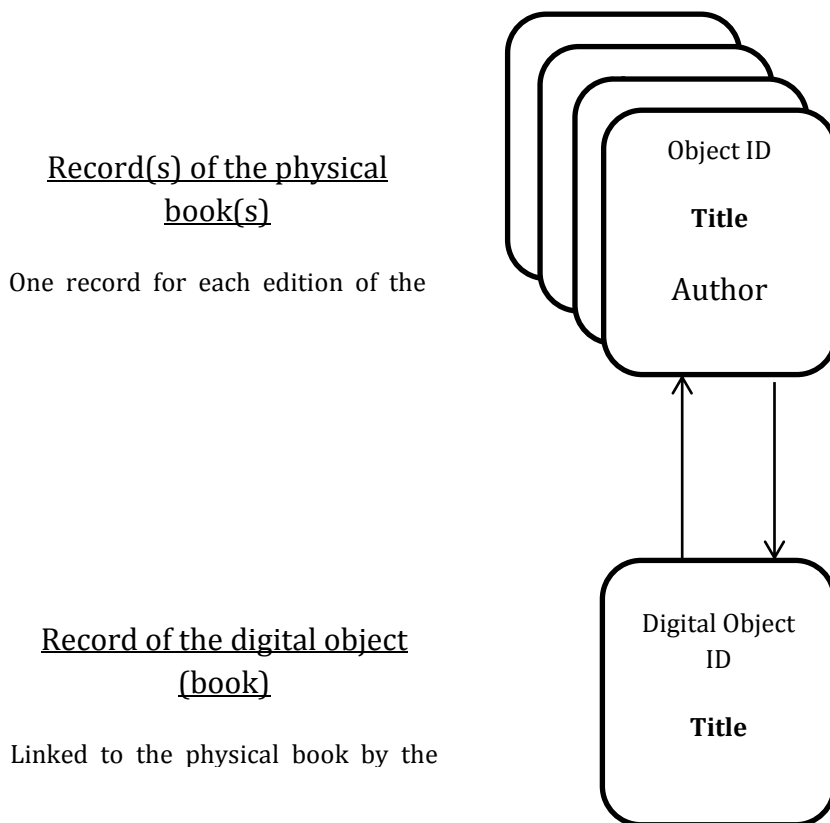


FIGURE 12: relationship between the analog and the digital object in NB.

With photographs the challenge is what has been stated before. Physical objects (negative or positive images) have no their own individual ID, so when digitizing these photos they need to generate an ID and upgrade this new data in the catalog at the same time as they link it to its URN.

The last years they have been improving the production line of digital photographs. In the beginning they digitized a big amount of photographs without having a system to implement descriptive metadata tags to the new digital objects. Right now they are about to implement a solution that extracts all metadata of a collection as an XML and import it to PhotoStation and tag an asset of photos at the same time (not one by one) in a semi-automatic way.

But for the rest of issues, they feel happy and satisfied with their digitizing policies and practices. General opinion and feeling is that they are doing things pretty well comparing with other national libraries.

4.3.2 NORWEGIAN PETROLEUM MUSEUM

The Norwegian Petroleum Museum is an institution which has a great old photography and brochure collection. They have been digitizing this collection for the last 13 years and their digital collections can be consulted online.

The main reasons that led them to digitize their collection are to preserve originals, improve accessibility, support educational and research activities and increase information sharing.

The Petroleum Museum cooperates with Nasjonalbiblioteket. The Oil Museum is digitizing photographs and objects (they take pictures of objects and catalog them), while the book collection is being digitized by the Nasjonalbibliotek.

Three main reasons led this institution to develop digitization projects in collaboration with Nasjonalbiblioteket: first of all they wanted to create a complete photography collection which documents all the recent history on oil exploitation in Norway. The second reason to collaborate was that the project was part of ABM-utvikling. NMU and the directorate for librarians joined forces sometimes in the 90s to form ABM-utvikling. This cooperation was later dissolved (2010), and the former NMU was taken over by the Culture Council. The third reason was the OAI (Open Archives Initiative). This is a Web protocol that makes it possible to harvest the data from different databases and make them searchable. It was a challenge for involved institutions to expand their knowledge on metadata management (storing, searching) and it was a great opportunity to put this new techniques into practice.

This institution has very few guidelines about how to create and implement metadata. The process of creating and implementing them is carried out by the staff (cataloger, archivist, curator, IT staff...) and as the informant stated, it is not an automatic process at all (not even in combination with staff), everything is manually made by personnel.

The most important sources for adding metadata are researchers and field work in combination with the consult of books and documentation, and not the digitized material itself.

4.3.2.1 Some technical issues

The database systems used for their digitized collections are Primus and Asta and the cataloging system used for their digital collection is ABM-Utviklings feltkatalog fra 2002 (ABM, NMU, 2002).

Primus (based on Oracle database) database is used as the main tool for managing their catalog. Primus is a museum collection managing tool, especially for historical and art collections but it can also be used with other type of materials. It provides






solutions for cataloging artifacts, buildings, photographs, sound, etc. (ABM-Utvikling, 2006). So it fits perfectly with Petroleum Museum's needs.

That system was chosen mainly because it is widely used in Norway and it is supported by a leading organization, NMU (Norsk Museum Utvikling, now a part of the Norwegian Culture Council, Norsk kulturråd). In their decision a main feature taken in account was that it is an extensible system, as well as easy to use, it supports information sharing, and they had a previous experience with it. In this point it is very important to mention that the information needed for the research regarding very technical aspects concerning how the digital object is managed and stored could not be answered by the Petroleum Museum informant. As the database and repository system, Primus, is externalized and managed in another institution and another city (Kultur IT in Lillehammer, Norway).

The informant was also asked about what factors he considers that will influence in their scanning practices. Between the factors considered very important there are the participation in joint programs, users' needs and staff commitment. In a lower lever other important factors influencing their digitization projects would be the technological development, development of policies and other technical matters as metadata standards development.

Staff of Petroleum Museum had some basic training and they discussed with other museums and institutions in order to design their digitization project. Leading institutions gave them support, in this case the Nasjonabibliotek and also they are part of a network of science museums (Technical Museum in Oslo....). So in case of some help or guidance is needed or have any registration problems during digitization tasks they can discuss them and get recommendations from other institutions.

Eksempler Høgskolen i Oslo

	NOMF-00128.00	Foto	Statfjord_A
	NOMF-00431.Q	Foto	Statfjord A. Besøk av samlingskonsulent Jan R. Johansen og konservator Jone Johnsen fra Norsk Oljemuseum.
	NOMF-02751.700	Foto	Kranlekteren Saipem 7000 i arbeid med å løfte bort moduler fra DP 2-plattformen på Friggfeltet. Her er det modul M31 som heises fra DP 2 og over til lekterens dekk.
	NOMF-02774.340	Foto	I sjuende etasje på Statfjord C står det en del bøker på gangen. Disse kan fritt lånes blant de ansatte.
	NOMF-02776.140	Foto	Verksted i modul W23 på Statfjord C. Her fikses og lages det aller meste innen blikk, diverse sveising og metallarbeid - på nye ting og gjenstander som trenger vedlikehold. Gunleiv Hadland fra Norsk Oljemuseum er på omvisning.

[Tilbake](#)

Museumsnummer:	NOMF-00431.Q
Motivbeskrivelse:	Statfjord A. Besøk av samlingskonsulent Jan R. Johansen og konservator Jone Johnsen fra Norsk Oljemuseum.
Avbildet person :	Statfjord A
Avbildet person :	Johansen, Jone
Avbildet person :	Johansen, Jan. R
Emneord :	Norsk Oljemuseum
Emneord :	#Statfjord
Historikk :	Bildene i denne serien er duplikatkopier og bildene kan derfor ha variert kvalitet.
Datering :	1981 - 1981
Datering :	1981 - 1981
Produsent :	Johansen, Jone
Fotograf :	Ukjent
Rettingssinnehaver/eier av original :	Norsk Oljemuseum
Produksjonssted / Eksponeringssted (F) :	9942, Statfjord feltet, Statfjord A
Andre opplysninger :	Besøk av samlingskonsulent Jan R. Johansen og konservator Jone Johnsen xxx Emnekode:xxx
Registrert	24.04.2003, Import
Eksemplar	NOMF-00431AB.Q.1
	Diaspositiver plast
	18x28cm



FIGURE 13: PRIMUS XML export from Petroleum Museum. Each photograph's descriptive metadata are displayed in these captures.

4.3.2.2 Digital objects' production line

Upon acquiring new objects (photographs for their collection), they are registered and cataloged in their Primus database. Afterwards, when digitizing one of these objects the previous metadata stored in Primus can be reused. Materials are registered directly in Primus database, both physical photographs and digital ones. When digitizing an object, the staff of the museum link the new digital object to the original photograph's record, so once this is done the previously existing metadata of the original photograph are shared with the new digital photograph's record.

If possible, the museum staff get in contact with the photographer, donor or somebody related to the photographs to extract as much information about the content as possible (place, people, date...) (descriptive metadata) and add this information to Primus database, where all the images are stored (high resolution images, negatives, etc.).

The interface is not difficult to use (there is not specific technical training needed for staff), the only problem they stated they might find is that there are many fields to fill so it can be confusing for the staff. In order to solve this problem they specified some guidelines for the staff in order to standardize the metadata fill in process.

They have a tool to export metadata stored in their Primus database to the digital museum automatically.

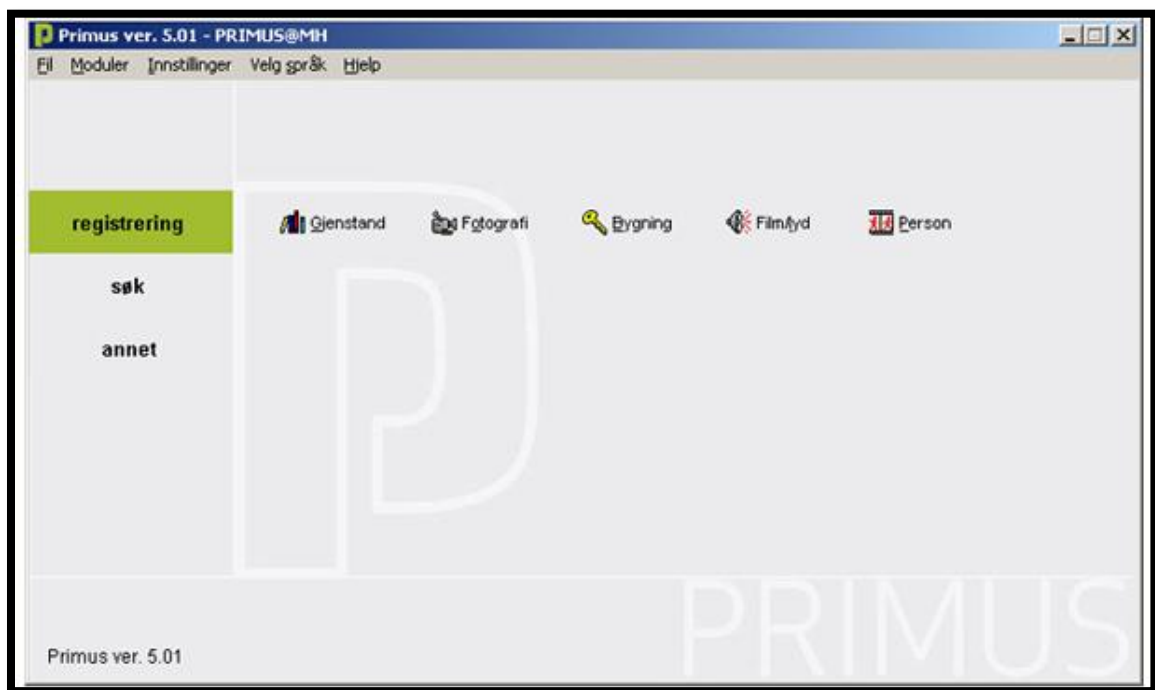


FIGURE 14: that is how PRIMUS interface looks like.

The informant stated that it is very important to create a good digital catalog structure in order to have a good starting before digitizing their collections. They have few specific guidelines (for filling in fields in their database system) but for the rest of activities they just follow national official standards and advices for photo registration: *Digitalisering av fotosamlinger* (ABM-Utvikling, 2009) and *Standard for fotokatalogisering* (ABM-Utvikling, 2008).

4.3.2.3 Problems/Challenges

The informant was asked about the problems they had faced during the metadata implementation process. He stated that the choice of what metadata schema to use was not a big deal as they do not create metadata packs themselves, they catalog their collection with Primus (and that's all), also confusing metadata concepts or the determination of which metadata elements are useful for users and staff were not a problem for the same reason. Among the main problems they faced during this stage there were the lack of data on existing materials, not enough availability of documentations and an insufficient budget.

There are some projects which try to exchange metadata with the Nasjonalbibliotek. It was defined by the informant as a technical challenge to get metadata from the digital museum to the Nasjonalbibliotek. An exchange format which allows them to collect data had to be designed. The problem was that in the beginning it was possible only to export the whole collection hosted in the digital museum, and they wanted to export only certain digital objects. The challenge was to be able to choose a part of the collection (i.e. a topic) and only export metadata of these records. Right now they are working to solve that problem. Several metadata sets may be created for each project in order to be able to select and export only determinate groups of digital objects and not all. They are working on creating a specific "topic" related metadata in order to allow the staff make the system export only certain metadata sets to this or that project.

In the case of Petroleum Museum budget was a "problem" in a way because it is limited for them, so when deciding what software and hardware use for their digitization practices they had to try to make the best choice taking in account their budget limitations.

4.3.3 ATENEU BARCELONÈS

L'Ateneu Barcelonès (AB from now on) is one of the main institutions in Catalonia (Spain) whose main purpose is to protect and spread local culture 50 years ago.

This institution has an important library and archive and they have been digitizing and publishing online part of their collection since 2006. The types of materials

they digitize are books, manuscripts, *incunables*, photographs and tapes. They collaborate in digitization projects with Biblioteca de Catalunya (National Library of Catalonia).

The main reasons to start digitizing their collection were to preserve the originals, to improve accessibility, to support educational and research activities and increase information sharing.

In their opinion in the near future their digitizing practices will be strongly influenced by new developments on scanning collections and the management of digital collections. Also the existing increasing number of digital objects will be a decisive factor that will influence the digitizing policies they follow. Users' needs and participating in joint programs with other institutions will also be very important factors. In a not so strong way, other factors will also influence them, as metadata standards and practices, new technological achievements and ongoing knowledge and skills, staff's commitment... On the other hand, they state that administrative changes in their institution will not influence their scanning policies in an important way.

It is important to emphasize that for this institution budget is a very important obstacle so they invest part of their efforts trying to establish a priorities policy of what to digitize. Anyways, they hope that in the future they will be able to digitize the rest of their collection, but it make take years to get the sufficient budget for everything.

When the informant was asked about their general feeling about digitization practices he stated to feel completely comfortable and to see it just as any other technical process.

4.3.3.1 Some technical issues

The cataloging database in which they store their collection is OPAC (Online Public Access Catalog). And the used cataloging system is Digibib (software created by the Spanish company Digibís). This software is based on standard metadata management of bibliographic and electronic resources as MARC21, MARC-XML, Dublin Core, METS and PREMIS. They chose this software because of the use of standards and because it is highly orientated to facilitate interoperability (one of its characteristics is that it fits with *Europeana's* specifications).

The metadata scheme used by AB is MARC21, Dublin Core and MODS. The main reasons that led AB to choose these schemes were its flexibility, extensibility, it supports information sharing, it is easy to use, it is widely used by many other institutions worldwide and because it is supported by leading organizations. Since the beginning they chose these schemes, although they did not have any other

experience with it before. They have a metadata preservation policy consisting on the generation of quality and standard metadata and making at least two copies of the digital object and its related metadata which are stored in disks, one of them stored in an external institution, the National Library of Catalonia.

The informant from the AB stated that they never have had interoperability problems concerning metadata. They participate in joint projects such as a cooperative project with the National Library of Catalonia and Google Books, an old press digitization project (Biblioteca Virtual de Prensa Histórica) in cooperation with other public libraries of Spain and other projects with university libraries. That shows up that in this institution work hard thinking on making available interoperability, that is why they use standards as long as it is possible for every process.

4.3.3.2 Digital objects' production line

In that institution they stated to have few guidelines about how to carry out the digitization process of collections. The process of creating and implementing metadata is carried out by the institution's staff itself. When adding metadata the main sources they use are materials themselves, researchers and fieldwork, all of them are important sources.

For generating a digital object the production line does not differ substantially from the processes described before for the other two institutions studied in the current research.

Their first step is to select the works for digitization. As said before, they have designed a priority policy so not all their collection is being scanned in the current project. When they have selected the object they scan it and afterwards they check all the TIFF files generated in order to confirm that the quality is correct. They upload this new digital object in MDC (*Memòria Digital de Catalunya*, an open access cooperative repository of Catalonia) and at this same stage they extract metadata from their LIS, Digibib, and reuse them to attach them to the digital object (together with the new descriptive, structural and administrative metadata). In the Digibib record of the object they add the link (structural metadata) to the repository with the digital object.

Finally, for preservation purposes they store two copies in two different hard disks which are in the National Library of Catalonia (the deposit is called COFRE).

4.3.3.3 Problems / Challenges

Among the problems found during the digitization process the informant did not underline any relevant issue concerning metadata or other technical factors, expect storage for preservation purposes. They admitted that the only technical problem they found was having space for preservation storage. That problem was solved thanks to a collaboration agreement with the Library of Catalonia, a leading institution which has more budget and technical resources, and there two of the master copies of each digital object are stored in their deposit (COFRE).

Another challenge identified during the study of this institution's digitization practices is the lack of descriptive metadata of the content of the documents. As explained before in the case of the National Library of Norway they use METS-ALTO metadata to describe the textual content of their digitized documents. That enables the staff and the final users to make searches by content. In the documents digitized by AB, they do no generate this kind of metadata.

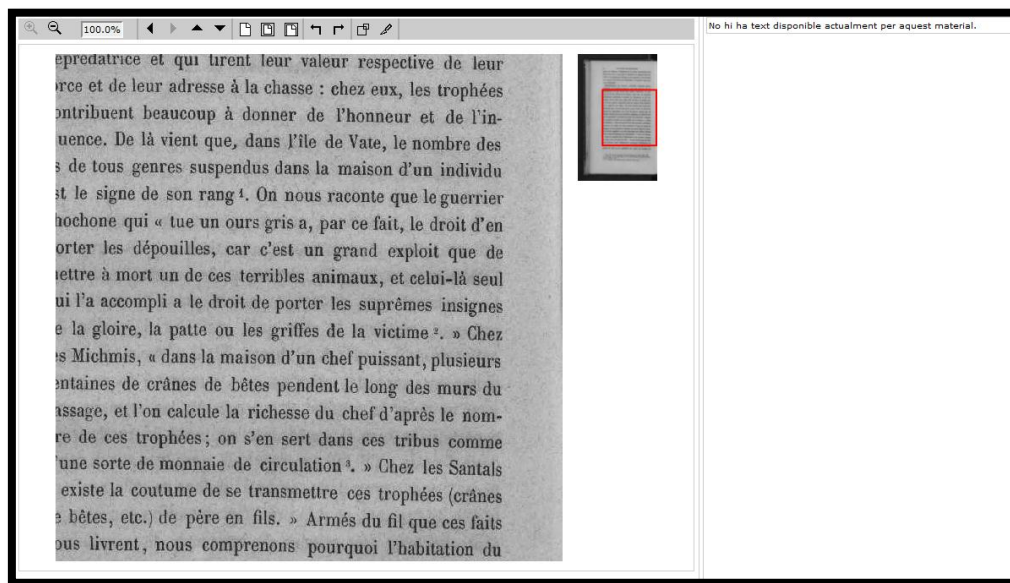


FIGURE 15: display view of a digitized document of Ateneu Barcelonés.

Above an example of how the system displays a digitized document to the final user. An image (JPEG) file can be seen but the text can not be selected. The White space in the right is reserved for text transcription (if in further projects they decide to add content metadata using an OCR process).

CHAPTER 5: CONCLUSIONS AND DISCUSSION

5.1 INTRODUCTION

In this chapter findings are set out, research questions are discussed, challenges in digitization are identified and suggestions for further researches are made.

5.2 FINDINGS

After analyzing different institutions' answers to the questionnaire and the interview and collected data, the main general idea we get is that for the studied institutions, digitization does not mean a big challenge or effort hard to cope with. All of them state that they work with international standards designed by leading institutions and as far as they follow these standards none of the institutions has reported any significant problem due to malpractice.

The only issue in which institutions differ is in budget. This issue appears to have a huge importance in digitization projects' design. The available budget for the project limits the far-reach of the results. As seen in chapter 4, data analysis, the National Library of Norway is the only institution that has no budget problems. That becomes evident if we focus on the results of their digitization projects: they digitize absolutely everything so they do not need to invest time and efforts designing a priorities policy; they create very rich metadata, including METS-ALTO, so they improve significantly search yield for final users and enable more services (possibility to select part of text, etc.)...

All of them follow more or less the same digital object production line, it does not matter if the institution is big or small or the kind of objects they digitize. Their working line from a general point of view is as follows:

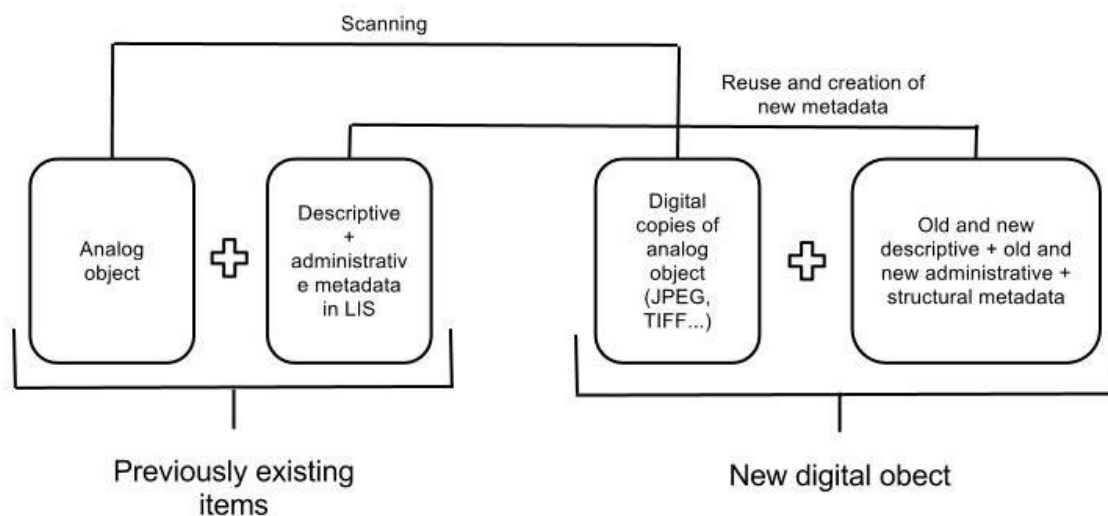


FIGURE 16: digital object production line in general terms.

5.3 RESEARCH QUESTIONS

5.3.1 How do the studied institutions carry out the digital object production process?

This section answers the current state of the digitization policies and practices in the studied institutions, such as if they have their own guidelines or follow leading institutions recommendations, if they participate in joint projects with other institutions, if they have a digitization prioritization needs and policies or not, the importance of interoperability for them, preservation policies, etc.

As stated in chapter 4 in Nasjonalbiblioteket (chapter 4.3.1) they do not have explicit guidelines about how to carry out the digitization process. The same happens with the Petroleum Museum (chapter 4.3.2) and Ateneu Barcelonès (chapter 4.3.3). When saying that they do not have explicit guidelines it means that these institutions do not have a document which can be consulted when staff needs to look up for advice. The only institution that stated having some explicit guidelines was the Petroleum Museum, which has guidelines concerning the metadata implementation process. They give staff advice about how to fill in the metadata fields in Primus interface.

Having explicit guidelines or advices may help to avoid situations in which a staff member does not know how to do a process and is compelled to take his/her own decisions. As a result it may take away the results of this process from the standards and best practices.

In line with this question of guidelines, there is the need of doing things as standardize as possible in all the institutions studied as all of them participate in joint programs with other memory institutions.

Concerning database and repository management there are also differences on who manages them. In NB they manage their own databases and repositories as well as master copies. On the other hand, the Petroleum Museum has outsourced this service to Kultur IT (Primus database creators and managers) so they really do not have direct control of the digital objects they create. Same happens in Ateneu Barcelonès with their preservation copies. They store them in an external institution, the National Library of Catalonia, so they really do not have direct control of them. Anyway, as in these two cases they have tight collaboration projects with the other institutions, it is supposed that the possibility of problems would be low.

It is important to emphasize the importance of budget for each institution. While NB has not problems in that sense, and so it has no needs for designing a priorities policy, the other two studied institutions had this problem.

Budget is a very important issue when talking about priority policies. The institutions which had budget limitations had to design a priority policy which influences the following aspects of a digitization project: acquiring a cheaper software for scanning and managing the digital collection (see chapter 4.3.2 Norwegian Petroleum Museum), deciding what objects will be digitized and which ones will not (see chapter 4.3.3 Ateneu Barcelonès), deciding time and staff invested in digitization (as time and staff may had also to manage other works carried out in the institution), maintenance of a digital store or repository and preservation policies (see also chapter 4.3.3 Ateneu Barcelonès), etc.

5.3.2 What is the life-cycle of metadata during a digitization process? By who, how and when are generated?

About the information and knowledge they need for cataloging their digital materials and adding metadata, all of them stated that the source material itself is the main information source together with some field work that is done by staff (documentation, etc). Most metadata are created in a semi-automatic way (exporting and reusing analog object's metadata) and manually by staff (descriptive metadata of content, i.e. digitized photographs by Petroleum

Museum). The different time and effort investment they made in this process is evident when looking to the metadata associated to each digital object from different institutions. As stated in 5.3.1, this difference in effort investment seems to be linked to budget availability.

NB is clearly the institution which most effort invested in creating quality metadata. They “reuse” previous existing descriptive metadata (analog objects’ metadata) but they make a big effort extending metadata and creating new metadata files such as METS-ALTO (see chapter 4.3.1). Thanks to this initial effort and resource and time investment, they generate quality metadata wrapped in the same digital package with the digital object (JPEG2000 files) and they guarantee the further search, access and management (both for staff and final users via the Internet) using URNs. The two smaller institutions (at the same time the ones with more budget limitations), the Norwegian Petroleum Museum and Ateneu Barcelonès, did allegedly not invest such an effort in the creation of so extended descriptive metadata and also do not guarantee long term access with the use of URNs. The exception in Norwegian Petroleum Museum are the text documents, they are digitized by the National Library of Norway so the generated digital object has the same characteristics as the rest of text documents digitized by NB.

The three studied institutions stated to have preservation policies consisting of creating two or three master copies and storing them. Again the difference is between the National Library and the other two smaller institutions concerning who takes care of these storage copies. The NB as it has more economical resources takes care of its own storage systems. The other two smaller institutions delegate their copies and repository management to other companies or institutions (see chapter 4.3.2 and 4.3.3).

But from a general and technical point of view, all institutions follow the same practices concerning metadata life-cycle: they create them during the digitization process, link them to the digital object and store them all together with the digital copy of the object. The technical differences between one and other institutions is not so centered in how they create and manage metadata (except the storage), but in what kind of metadata they create.

5.3.3 What are the main differences between institutions?

The three studied institutions carry on the digital object production line in a quite similar way. In that aspect differences between institutions are basically the kind of metadata added to the digital objects and the preservation policies. These differences are basically because of the budget available, and not so much because of having a different opinion or perspective about how metadata and a digitization process should be.

As stated in 5.3.2 the effort invested in creating metadata is higher in NB. It is due to the budget available. The other two institutions showed to be concerned about creating quality metadata and the most complete possible but means at the moment do not allow them, that is the case of Ateneu Barcelonès. Preservation policy is also linked to budget. In the case of Ateneu Barcelonès that service had to be externalized to a trusting and leading institution in Catalonia, the National Library.

5.3.4 What is the best way to generate and implement metadata during a digitization process?

This question corresponds to one of the main objectives of this thesis: try to establish a good practice guide or recommendations list.

All the informants stated that they feel satisfied with the way they have been working and creating and managing metadata of their digital collections. They perceive it as one more technical process among other processes they carry out (see as an example chapter 4.3.3). That is because they did not find any problems or obstacles managing their collections after digitizing and creating metadata. They all use standards and no institution uses its own metadata scheme, so they should avoid major obstacles such as interoperability problems, or at least, the problems they may find will be shared with the rest of institutions that use standards, and fixing further problems will be a responsibility of the leading organizations that create those standards.

With the collected data it is difficult to establish a definitive recommendation list or guidelines. But some ideas arouse which could be hypothesis for further research. Following the suggestions for further research extracted from this study are explained.

5.4 FURTHER RESEARCH

The same study approach made in this research could and should be done again but trying to avoid the main problems and obstacles found this time. It would be recommendable to have many informants for each institution and working in diverse digitization stages, instead of having only one. If time is not a limitation the questionnaire should be sent to more staff of the studied institutions and if possible interview them. Having more than one opinion about the same process can enrich in a substantial way the research results. The findings would more

accurately show the current state of digitization and metadata creation and management.

Further studies should focus on improving metadata practices and establishing more accurate needs for each type of collection and institution. A possible research could center on the objective of designing a guidelines book for a specific institution which has some special need, for example. A starting hypothesis for a further research could be the following: having explicit guidelines (a document which any staff member can consult whenever needed) may be helpful for staff to guide them on how to carry out the processes. This document could be printed or digital, but always accessible for staff. The main purpose of these guidelines should be to standardize the processes inside the workplace and make them explicit for all the staff involved in digitization. Some leading institutions have published general guidelines. Then a particular institution involved in a digitization project, before starting may write their own guidelines based on those general and in other institutions' work. This may be the perfect situation in which a "global convergence" of schemes and procedures exists in the digital cultural heritage management. Starting from a national perspective towards an international one. As ABM-Utvilking (2010) recommends it would be desirable to create a guidance and a framework for digitization projects that is loose enough to reflect the different needs within the ALM sector, but also strict enough to lead to some minimum guarantees. Standardization, the use of generic infrastructure and cooperation between institutions provide opportunities for better balanced costs, greater independence and a better utilization of IT infrastructure.

In the same study line it would also be interesting to make research on metadata practices not from the point of view of institutions and managers, but from the point of view of final users. Final users can be very different and so may need different kind of information about the digital object they are consulting. They can be advance users or not, users with IT knowledge or not, they may need a lot of information (professionals, historians, photographers, artists...), etc. Depending on users profile metadata needs may be different, it would be interesting to make a research and ask different users if they feel satisfied with the current state of digitization and metadata and ask them if when they make an Internet search in a digital library or museum they find what they want and expect.

Another suggestion for further researches is to make a deeper study of to which extent budget can influence the quality of projects. Of course it influences on the quantity of projects carried out by an institution, but it would be interesting to analyze if the institutions with less budget carry out projects in a different way and if it has influence on the final result taking in account more aspects apart from metadata (already done in this research). The starting hypothesis could be that it may be positive to try to invest as much as possible budget in the scanning and metadata generating and storing processes (instead of investing it in other

activities). This is of vital importance to ensure long lasting digital objects. The main purpose in this case would be to avoid the need of re-scanning or re-cataloguing items after little time. In cultural institutions budget seems to be one of the main obstacles: through the development of a good IT architecture an organization reduces costs and increases efficiency. At the same time investment should be made to modernize and maintain standards in the system, so that value creation can remain large.

Furthermore, open access is a major challenge. Users should be guaranteed access to the service regardless of the technology and vendor they choose (both institutions and users). Carrying out a research in which the access to digitized collections is analyzed could be very useful and enriching. It is evident that memory institutions are investing big efforts on digitization, but how is the access to these new digital collections? Is it as expected? A research covering these issues could be very interesting.

REFERENCES

ABM, NMU (2002). Feltkatalog for kunst og kulturhistoriske museer.
<http://www.abm-utvikling.no/publisert/tidligere-utgivelser/nmu3-2002.pdf>

ABM-Utvikling (2006). Cultural heritage for all.
http://www.abm-utvikling.no/publications/abm-publications/32_eng_web.pdf

ABM-Utvikling (2008). Standard for fotokatalogisering.
<http://www.abm-utvikling.no/publisert/abm-skrift/abm-skrift-44-fotokatalogisering>

ABM-Utvikling (2009). Digitalisering av fotosamlinger.
<http://www.abm-utvikling.no/publisert/abm-skrift/abm-skrift-55>

ABM-Utvikling (2010). Åpen og samordnet tilgang til kulturarven. 2. utgave.
<http://www.abm-utvikling.no/publisert/abm-skrift/abm-skrift-66-1>

Adobe Systems Incorporated (2005). XMP Specification.
<http://partners.adobe.com/public/developer/en/xmp/sdk/XMPspecification.pdf>

Biblioteca de l'Ateneu Barcelonès.
<http://biblioteca.ateneubcn.cat/web/continguts/ca/index.html>

Avery, J.M. (2010). The Democratization of Metadata: Collective Tagging, Folksonomies and Web 2.0. Library Student Journal 5,
<http://www.librarystudentjournal.org/index.php/ljsj/article/view/135>

Baca, M. (2008). Introduction to metadata : pathway to digital information (Los Angeles, Calif.; Hove: Getty Research Institute ; Roundhouse [distributor]).

Biblioteca Nacional de España (2010). Biblioteca Digital Hispánica.
<http://www.bne.es/es/Catalogos/BibliotecaDigital/>

Caplan, P. (2003). Metadata fundamentals for all librarians (Chicago: American Library Association).

Cornwell Library (2012). Cornell Library Windows on the Past.
<http://cdl.library.cornell.edu/>

Day, M., Huber, E., and Gaviria, J.N.. (2010). IMPACT Best Practice Guide: Metadata and the large-scale digitisation of text.
<http://www.impact-project.eu/uploads/media/IMPACT-metadata-bpg-pilot-1.pdf>

Digitalt Museum.

<http://www.digitaltmuseum.no/>

Dorner, D. (2000). Cataloging in the 21st century—part 2: digitization and information standards. *Library Collections, Acquisitions, and Technical Services* 24, 73–87.

Europeana.

<http://www.europeana.eu/portal/>

Garnes, K., Landøy, A., Repanovici, A., Bagge, B.-A., Åsmul, A.B., Kongshavn, H., Sivertssen, S., Tonning, A.S.V., Torras, M.-C., Skagen, T., et al. (2006). Aspects of the Digital Library (Alvheim & Eide).

<https://bora.uib.no/bitstream/1956/1821/1/Aspects%20of%20the%20digital%20library.pdf>

Gorman, G.E., Clayton, P., Shep, S.J., and Clayton, A. (2005). *Qualitative research for the information professional : a practical handbook* (London: Facet).

Hesse-Biber, S.N., and Leavy, P. (2011). *The practice of qualitative research* (Los Angeles, Calif.: SAGE Publications).

Hodge, G.M. (2004). *Understanding metadata* (Bethesda, Md.: NISO Press).

Hughes, L.M. (2004). *Digitizing collections : strategic issues for the information manager* (London: Facet Publishing).

I. H. Witten, D.B. (2009). *How to build a digital library* (Morgan Kaufmann Publishers Inc).

ISO/IEC 11179, Information Technology, M.R. Home Page for ISO/IEC 11179 Information Technology -- Metadata registries.

<http://metadata-standards.org/11179/>

Kelly, B. (2006). *Choosing a Metadata Standard For Resource Discovery*.

<http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-63/html/>

Kvale, S. (1996). *Interviews : an introduction to qualitative research interviewing*. (London: Sage).

Lincoln, Y.S., and Guba, E.G. (1985). *Naturalistic inquiry* (Beverly Hills, Calif.: Sage Publications).

Marty, P.F., Rayward, W.B., and Twidale, M.B. (2003). Museum informatics. *Annual Review of Information Science and Technology* 37, 259–294.

Miles, M.B., and Huberman, A.M. (1994). *Qualitative Data Analysis: An expanded sourcebook* (California: London & Thousand Oaks).

Minerva Project (2003). Digitisation Guidelines.
<http://www.minervaeurope.org/guidelines.htm>

Minerva Project. List of good practices in digitisation.
<http://www.minervaeurope.org/listgoodpract.htm>

NasjonalBibliotek, B. Bokhylla.
<http://www.nb.no/bokhylla>

Nasjonalbibliotek - Digitalisation Policies.
<http://nb.no/content/download/5928/38289/version/1/file/Digitalisation+Policies.pdf>

National Information Standards Organization (2007). A framework of guidance for building good digital collections a NISO recommended practice (Baltimore, MD: National Information Standards Organization (NISO)).
<http://www.niso.org/publications/rp/framework3.pdf>

Norwegian Petroleum Museum.
<http://www.norskolje.museum.no/index.asp>

Open Archives Initiative.
<http://www.openarchives.org/>

Open Archives Initiative – Protocol for Metadata Harvesting.
<http://www.openarchives.org/pmh/>

Pickard, A.J. (2007). Research methods in information. (London: Facet publishing).

University of North Carolina (2012). Documenting the American South homepage.
<http://docsouth.unc.edu/index.html>

Yin, R.K. (2009). Case study research: design and methods (Los Angeles, Calif.: Sage Publications).

APPENDICES

APPENDIX 1

PREMIS scheme used by the National Library of Spain:

M (Mandatory) / O (Optional) / R (Repeatable) / NR (No Repeatable)
1.1 objectIdentifier (M, R)
1.1.1 objectIdentifierType (M, NR)
1.1.2 objectIdentifierValue (M, NR)
1.2 objectCategory (M, NR)
1.3 preservationLevel (O, R) [representation, file]
1.3.1 preservationLevelValue (M, NR) [representation, file]
1.4 significantProperties (O, R)
1.5 objectCharacteristics (M, R) [file, bitstream]
1.5.1 compositionLevel (M, NR) [file, bitstream]
1.5.2 fixity (O, R) [file, bitstream]
1.5.2.1 messageDigestAlgorithm (M, NR) [file, bitstream]
1.5.2.2 messageDigest (M, NR) [file, bitstream]
1.5.3 size (O, NR) [file, bitstream]
1.5.4 format (M, R) [file, bitstream]
1.5.4.1 formatDesignation (O, NR) [file, bitstream]
1.5.4.1.1 formatName (M, NR) [file, bitstream]
1.5.4.1.2 formatVersion (O, NR) [file, bitstream]
1.5.4.2 formatRegistry (O, NR) [file, bitstream]
1.5.4.2.1 formatRegistryName (M, NR) [file, bitstream]
1.5.4.2.2 formatRegistryKey (M, NR) [file, bitstream]
1.5.4.2.3 formatRegistryRole (O, NR) [file, bitstream]
1.5.5 creatingApplication (O, R) [file, bitstream]

1.5.5.1	creatingApplicationName (O, NR) [file, bitstream]
1.5.5.2	creatingApplicationVersion (O, NR) [file, bitstream]
1.5.5.3	dateCreatedByApplication (O, NR) [file, bitstream]
1.5.6	inhibitors (O, R) [file, bitstream]
1.5.6.1	inhibitorType (M, NR) [file, bitstream]
1.5.6.2	inhibitorTarget (O, R) [file, bitstream]
1.5.6.3	inhibitorKey (O, NR) [file, bitstream]
1.6	originalName (O, NR) [representation, file]
1.7	storage (M, R) [file, bitstream]
1.7.1	contentLocation (O, NR) [file, bitstream]
1.7.1.1	contentLocationType (M, NR) [file, bitstream]
1.7.1.2	contentLocationValue (M, NR) [file, bitstream]
1.7.2	storageMedium (O, NR) [file, bitstream]
1.8	environment (O, R)
1.8.1	environmentCharacteristic (O, NR)
1.8.2	environmentPurpose (O, R)
1.8.3	environmentNote (O, R)
1.8.4	dependency (O, R)
1.8.4.1	dependencyName (O, R)
1.8.4.2	dependencyIdentifier (O, R)
1.8.4.2.1	dependencyIdentifierType (M, NR)
1.8.4.2.2	dependencyIdentifierValue (M, NR)
1.8.5	software (O, R)
1.8.5.1	swName (M, NR)
1.8.5.2	wVersion (O, NR)
1.8.5.3	swType (M, NR)
1.8.5.4	swOtherInformation (O, R)
1.8.5.5	swDependency (O, R)
1.8.6	hardware (O, R)

1.8.6.1 hwName (M, NR)

1.8.6.2 hwType (M, NR)

1.8.6.3 hwOtherInformation (O, R)

1.9 signatureInformation (O, R) [file, bitstream]

1.9.1 signature (O, R)

1.9.1.1 signatureEncoding (M, NR) [file, bitstream]

1.9.1.2 signer (O, NR) [file, bitstream]

1.9.1.3 signatureMethod (M, NR) [file, bitstream]

1.9.1.4 signatureValue (M, NR) [file, bitstream]

1.9.1.5 signatureValidationRules (M, NR) [file, bitstream]

1.9.1.6 signatureProperties (O, R) [file, bitstream]

1.9.1.7 keyInformation (O, NR) [file, bitstream]

1.10 relationship (O, R)

1.10.1 relationshipType (M, NR)

1.10.2 relationshipSubType (M, NR)

1.10.3 relatedObjectIdentification (M, R)

1.10.3.1 relatedObjectIdentifierType (M, NR)

1.10.3.2 relatedObjectIdentifierValue (M, NR)

1.10.3.3 relatedObjectSequence (O, NR)

1.10.4 relatedEventIdentification (O, R)

1.10.4.1 relatedEventIdentifierType (M, NR)

1.10.4.2 relatedEventIdentifierValue (M, NR)

1.10.4.3 relatedEventSequence (O, NR)

1.11 linkingEventIdentifier (O, R)

APPENDIX 2:

MARC21 records of NB

Paper Book	Electronic book
<pre> <srw:record> <srw:recordSchema> marcxchange </srw:recordSchema> <srw:recordPacking> xml </srw:recordPacking> <srw:recordIdentifier> 000740195 </srw:recordIdentifier> <srw:recordPosition> 20 </srw:recordPosition> <srw:recordData> <marc:record format="MARC21" type="Bibliographic"> <marc:leader> 99999 am a2299999 c 4500 </marc:leader> <marc:controlfield tag="001"> 000740195 </marc:controlfield> <marc:controlfield tag="003"> NO-TrBIB </marc:controlfield> <marc:controlfield tag="005"> 20070418105947.0 </marc:controlfield> </pre>	<pre> <srw:record> <srw:recordSchema> marcxchange </srw:recordSchema> <srw:recordPacking> xml </srw:recordPacking> <srw:recordIdentifier> 070555729 </srw:recordIdentifier> <srw:recordPosition> 16 </srw:recordPosition> <srw:recordData> <marc:record format="MARC21" type="Bibliographic"> <marc:leader> 99999 am a2299999 c 4500 </marc:leader> <marc:controlfield tag="001"> 070555729 </marc:controlfield> <marc:controlfield tag="003"> NO-TrBIB </marc:controlfield> <marc:controlfield tag="005"> 20101115105947.0 </marc:controlfield> <marc:controlfield tag="007"> </pre>

<pre> <marc:controlfield tag="007"> ta </marc:controlfield> <marc:controlfield tag="008"> s2000 no nob </marc:controlfield> <marc:datafield tag="015" ind1=" " ind2=" "> <marc:subfield code="a"> 0010939 </marc:subfield> <marc:subfield code="2"> nbf </marc:subfield> </marc:datafield> <marc:datafield tag="020" ind1=" " ind2=" "> <marc:subfield code="a"> 8202196027 </marc:subfield> </marc:datafield> <marc:datafield tag="082" ind1=" " ind2="#"> <marc:subfield code="a"> 387.52 </marc:subfield> </marc:datafield> <marc:datafield tag="082" ind1=" " ind2="#"> <marc:subfield code="a"> 387.5 </marc:subfield> </marc:datafield> <marc:datafield tag="082" ind1=" " ind2="#"> <marc:subfield code="a"> 387.524 </pre>	<pre> cr </marc:controlfield> <marc:controlfield tag="008"> s2000 no nob </marc:controlfield> <marc:datafield tag="020" ind1=" " ind2=" "> <marc:subfield code="a"> 8202196027 </marc:subfield> </marc:datafield> <marc:datafield tag="082" ind1=" " ind2="#"> <marc:subfield code="a"> 387.52 </marc:subfield> </marc:datafield> </pre>
--	--

<pre> </marc:subfield> </marc:datafield> <marc:datafield tag="082" ind1=" " ind2="#"> <marc:subfield code="a"> 387.524 </marc:subfield> </marc:datafield> <marc:datafield tag="100" ind1="0" ind2="#"> <marc:subfield code="a"> Johnson, P å l Espolin </marc:subfield> </marc:datafield> <marc:datafield tag="245" ind1="1" ind2="0"> <marc:subfield code="a"> Med hurtigruta nordover </marc:subfield> <marc:subfield code="c"> P å l Espolin Johnson </marc:subfield> </marc:datafield> <marc:datafield tag="250" ind1=" " ind2=" "> </pre>	<pre> <marc:datafield tag="100" ind1="0" ind2="#"> <marc:subfield code="a"> Johnson, P å l Espolin </marc:subfield> </marc:datafield> <marc:datafield tag="245" ind1="1" ind2="0"> <marc:subfield code="a"> Med hurtigruta nordover </marc:subfield> <marc:subfield code="c"> P å l Espolin Johnson </marc:subfield> <marc:subfield code="h"> elektronisk ressurs </marc:subfield> </marc:datafield> <marc:datafield tag="250" ind1=" " ind2=" "> <marc:subfield code="a"> </pre>
--	--

<pre> <marc:subfield code="a"> 6. oppl. [i.e. ny utg.] </marc:subfield> </marc:datafield> <marc:datafield tag="260" ind1=" " ind2=" "> <marc:subfield code="a"> Oslo </marc:subfield> <marc:subfield code="b"> Cappelen </marc:subfield> <marc:subfield code="c"> 2000 </marc:subfield> </marc:datafield> <marc:datafield tag="300" ind1=" " ind2=" "> <marc:subfield code="a"> 112 s. </marc:subfield> <marc:subfield code="b"> ill. </marc:subfield> <marc:subfield code="c"> 27 cm </marc:subfield> </marc:datafield> <marc:datafield tag="500" ind1="#" ind2="#"> <marc:subfield code="a"> 1. utg. Oslo : Boksenteret, 1994 </marc:subfield> </marc:datafield> </pre>	<pre> 6. oppl. [i.e. ny utg.] </marc:subfield> </marc:datafield> <marc:datafield tag="260" ind1=" " ind2=" "> <marc:subfield code="a"> Oslo </marc:subfield> <marc:subfield code="b"> Cappelen </marc:subfield> <marc:subfield code="c"> 2000 </marc:subfield> </marc:datafield> <marc:datafield tag="300" ind1=" " ind2=" "> <marc:subfield code="a"> 112 s. </marc:subfield> <marc:subfield code="b"> ill. </marc:subfield> <marc:subfield code="c"> 27 cm </marc:subfield> </marc:datafield> <marc:datafield tag="500" ind1="#" ind2="#"> <marc:subfield code="a"> 1. utg. Oslo : Boksenteret, 1994 </marc:subfield> </marc:datafield> <marc:datafield tag="500" ind1="#" ind2="#"> <marc:subfield code="a"> </pre>
---	--

<pre> <marc:datafield tag="610" ind1="2" ind2="4"> <marc:subfield code="a"> Hurtigruten </marc:subfield> </marc:datafield> <marc:datafield tag="650" ind1="#" ind2="4"> <marc:subfield code="a"> Hurtigruta </marc:subfield> </marc:datafield> <marc:datafield tag="651" ind1="#" ind2="7"> <marc:subfield code="a"> Vestlandet </marc:subfield> </marc:datafield> <marc:datafield tag="651" ind1="#" ind2="7"> <marc:subfield code="a"> Nord-Norge </marc:subfield> </marc:datafield> <marc:datafield tag="653" ind1=" " ind2=" "> <marc:subfield code="a"> sj ø reiser vestlandet nord norge norskekysten hurtigruta kystfrakt </marc:subfield> </marc:datafield> <marc:datafield tag="776" ind1="0" ind2="#"> </pre>	<pre> Elektronisk reproduksjon </marc:subfield> </marc:datafield> <marc:datafield tag="610" ind1="2" ind2="4"> <marc:subfield code="a"> Hurtigruten </marc:subfield> </marc:datafield> <marc:datafield tag="651" ind1="#" ind2="7"> <marc:subfield code="a"> Vestlandet </marc:subfield> </marc:datafield> <marc:datafield tag="651" ind1="#" ind2="7"> <marc:subfield code="a"> Nord-Norge </marc:subfield> </marc:datafield> <marc:datafield tag="653" ind1=" " ind2=" "> <marc:subfield code="a"> sj ø reiser vestlandet nord norge norskekysten hurtigruta kystfrakt </marc:subfield> </marc:datafield> <marc:datafield tag="776" ind1="0" ind2="#"> <marc:subfield code="w"> </pre>
---	--

<pre> <marc:subfield code="w"> (NO-TrBIB)070555729 </marc:subfield> </marc:datafield> <marc:datafield tag="852" ind1=" " ind2=" "> <marc:subfield code="p"> 00sa08447 </marc:subfield> <marc:subfield code="a"> NB/BEV </marc:subfield> <marc:subfield code="h"> b </marc:subfield> </marc:datafield> <marc:datafield tag="852" ind1=" " ind2=" "> <marc:subfield code="p"> 00a020093 </marc:subfield> <marc:subfield code="a"> GUNNERUS </marc:subfield> <marc:subfield code="b"> qB </marc:subfield> <marc:subfield code="h"> 36744 </marc:subfield> </marc:datafield> </pre>	<pre> (NO-TrBIB)000740195 </marc:subfield> <marc:subfield code="z"> 82-02-19602-7 </marc:subfield> </marc:datafield> <marc:datafield tag="852" ind1=" " ind2=" "> <marc:subfield code="p"> 07sg01413 </marc:subfield> <marc:subfield code="a"> NB/DIG </marc:subfield> <marc:subfield code="b"> nbdigi </marc:subfield> </marc:datafield> <marc:datafield tag="856" ind1=" " ind2="0"> <marc:subfield code="u"> http://urn.nb.no/URN:NBN:no-nb_digibok_2007022801018 </marc:subfield> </marc:datafield> <marc:datafield tag="886" ind1=" " ind2=" "> <marc:subfield code="2"> bibsysm </marc:subfield> <marc:subfield code="a"> 009 </marc:subfield> <marc:subfield code="b"> \$aNordomr ã </pre>
--	---

<pre> <marc:datafield tag="852" ind1=" " ind2=" "> <marc:subfield code="p"> 00ud05017 </marc:subfield> <marc:subfield code="a"> UBIN </marc:subfield> <marc:subfield code="h"> 387.524 Joh </marc:subfield> </marc:datafield> <marc:datafield tag="852" ind1=" " ind2=" "> <marc:subfield code="p"> 00ga15370 </marc:subfield> <marc:subfield code="z"> (ib.) (Til bruk p å NB Oslos lesesal) </marc:subfield> <marc:subfield code="a"> NB/BRU </marc:subfield> <marc:subfield code="h"> NA/A 2000:4849 </marc:subfield> </marc:datafield> <marc:datafield tag="852" ind1=" " ind2=" "> <marc:subfield code="p"> 00sd25745 </marc:subfield> <marc:subfield code="a"> </pre>	<pre> dene </marc:subfield> </marc:datafield> </marc:record> </srw:recordData> </srw:record> </pre>
---	---

<p>NB/DEP</p> <p></marc:subfield></p> <p></marc:datafield></p> <p><marc:datafield tag="852" ind1=" " ind2=" "></p> <p><marc:subfield code="p"></p> <p>00sd26781</p> <p></marc:subfield></p> <p><marc:subfield code="a"></p> <p>NB/DEP</p> <p></marc:subfield></p> <p></marc:datafield></p> <p><marc:datafield tag="852" ind1=" " ind2=" "></p> <p><marc:subfield code="p"></p> <p>00kj35870</p> <p></marc:subfield></p> <p><marc:subfield code="a"></p> <p>UHS</p> <p></marc:subfield></p> <p><marc:subfield code="h"></p> <p>387.5 Joh</p> <p></marc:subfield></p> <p></marc:datafield></p> <p><marc:datafield tag="852" ind1=" " ind2=" "></p> <p><marc:subfield code="p"></p> <p>06xd02262</p> <p></marc:subfield></p> <p><marc:subfield code="a"></p> <p>HIVE</p> <p></marc:subfield></p> <p><marc:subfield code="h"></p> <p>387.524 Joh</p> <p></marc:subfield></p>	
--	--

<pre></marc:datafield> <marc:datafield tag="886" ind1=" " ind2=" "> <marc:subfield code="2"> bibsysm </marc:subfield> <marc:subfield code="a"> 009 </marc:subfield> <marc:subfield code="b"> \$aNordomr å dene </marc:subfield> </marc:datafield> </marc:record> </srw:recordData> </srw:record></pre>	
--	--

APPENDIX 3:

METS-ALTO (XML) document belonging to a digitized book by NB. Highlighted in red some of the words (content) of one of the book's pages and their "location" in the page:

```
<?xml version="1.0" encoding="UTF-8"?>
<alto xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:noNamespaceSchemaLocation="//Produksjon5/docworks/docWORKSshare/schema/alto-1-2.xsd"
xmlns:xlink="http://www.w3.org/TR/xlink">
  <Description>
    <MeasurementUnit>mm10</MeasurementUnit>
    <sourceImageInformation>
      <fileName>//Produksjon5/ocr-
output/Nordomradene/digibok_2007022801018/images/digibok_2007022801018_0010.tiff</fileName>
    </sourceImageInformation>
    <OCRProcessing ID="OCRPROCESSING_1">
      <preProcessingStep>
        <processingSoftware>
          <softwareCreator>CCS Content Conversion Specialists GmbH,
Germany</softwareCreator>
          <softwareName>CCS docWORKS</softwareName>
          <softwareVersion>6.0-8.16</softwareVersion>
        </processingSoftware>
      </preProcessingStep>
      <ocrProcessingStep>
        <processingSoftware>
          <softwareCreator>ABBYY (BIT Software), Russia</softwareCreator>
          <softwareName>Finereader</softwareName>
        </processingSoftware>
      </ocrProcessingStep>
    </OCRProcessing>
  </Description>
  <Styles>
    <TextStyle ID="TXT_0" FONTSIZE="9" FONTFAMILY="Times New Roman" FONTSTYLE="italics"/>
  </Styles>
</alto>
```

```

<TextStyle ID="TXT_1" FONTSIZE="12" FONTFAMILY="Arial" FONTSTYLE="bold italics"/>

<TextStyle ID="TXT_2" FONTSIZE="24" FONTFAMILY="Arial"/>

<TextStyle ID="TXT_3" FONTSIZE="12" FONTFAMILY="Times New Roman"/>

<TextStyle ID="TXT_4" FONTSIZE="11" FONTFAMILY="Fraktur"/>

<TextStyle ID="TXT_5" FONTSIZE="7" FONTFAMILY="Times New Roman"/>

<ParagraphStyle ID="PAR_CENTER" ALIGN="Center"/>

<ParagraphStyle ID="PAR_LEFT" ALIGN="Left"/>

</Styles>

<Layout>

<Page ID="P10" PHYSICAL_IMG_NR="10" HEIGHT="2595" WIDTH="1859" PRINTED_IMG_NR="8">

    <TopMargin ID="P10_TM00001" HPOS="0" VPOS="0" WIDTH="1859" HEIGHT="130">

        <TextBlock ID="P10_TB00001" HPOS="672" VPOS="107" WIDTH="505"
HEIGHT="20" STYLEREFS="TXT_0 PAR_CENTER">

            <TextLine ID="P10_TL00001" HPOS="673" VPOS="108" WIDTH="504"
HEIGHT="19">

                <String ID="P10_ST00001" HPOS="673" VPOS="108"
WIDTH="66" HEIGHT="19" CONTENT="MED" WC="0.99" CC="100"/>

                <SP ID="P10_SP00001" HPOS="739" VPOS="127"
WIDTH="25"/>

                <String ID="P10_ST00002" HPOS="764" VPOS="108"
WIDTH="210" HEIGHT="19" CONTENT="HURTIGRUTA" WC="0.97" CC="0101011110"/>

                <SP ID="P10_SP00002" HPOS="974" VPOS="127"
WIDTH="26"/>

                <String ID="P10_ST00003" HPOS="1000" VPOS="108"
WIDTH="177" HEIGHT="19" CONTENT="NORDOVER" WC="0.99" CC="10100001"/>

            </TextLine>

        </TextBlock>

    </TopMargin>

    <LeftMargin ID="P10_LM00001" HPOS="0" VPOS="130" WIDTH="106" HEIGHT="2326"/>

    <RightMargin ID="P10_RM00001" HPOS="1720" VPOS="130" WIDTH="139"
HEIGHT="2326"/>

    <BottomMargin ID="P10_BM00001" HPOS="0" VPOS="2456" WIDTH="1859"
HEIGHT="139">

        <TextBlock ID="P10_TB00002" HPOS="915" VPOS="2459" WIDTH="22"
HEIGHT="30" STYLEREFS="TXT_1 PAR_CENTER">

            <TextLine ID="P10_TL00002" HPOS="916" VPOS="2459" WIDTH="21"
HEIGHT="30">

```

```

                                <String ID="P10_ST00004" HPOS="916" VPOS="2459"
WIDTH="21" HEIGHT="30" CONTENT="8" WC="1.00" CC="0"/>
                                </TextLine>
                                </TextBlock>
                                </BottomMargin>
                                <PrintSpace ID="P10_PS00001" HPOS="106" VPOS="130" WIDTH="1614" HEIGHT="2326">
                                <TextBlock ID="P10_TB00003" HPOS="374" VPOS="177" WIDTH="1109"
HEIGHT="80" STYLEREF="TXT_2 PAR_CENTER">
                                <TextLine ID="P10_TL00003" HPOS="374" VPOS="177" WIDTH="1109"
HEIGHT="80">
                                <String ID="P10_ST00005" HPOS="374" VPOS="195"
WIDTH="418" HEIGHT="62" CONTENT="HUNDRE" WC="1.00" CC="000000"/>
                                <SP ID="P10_SP00003" HPOS="792" VPOS="257"
WIDTH="78"/>
                                <String ID="P10_ST00006" HPOS="870" VPOS="177"
WIDTH="187" HEIGHT="78" CONTENT="ARS" WC="1.00" CC="000"/>
                                <SP ID="P10_SP00004" HPOS="1057" VPOS="257"
WIDTH="81"/>
                                <String ID="P10_ST00007" HPOS="1138" VPOS="192"
WIDTH="345" HEIGHT="63" CONTENT="SEILAS" WC="1.00" CC="000000"/>
                                </TextLine>
                                </TextBlock>

```

APPENDIX 4

Questionnaire:

Some introductory questions

What kind of memory institution do you work for?

- Library
- Museum
- Archive
- Other:

For how long has your institution digitized its collection? (Months, years...)

Does your institution participate in a joint digital collection development project with other institutions?

- No
- Yes
- Other:

What are the source materials of your digitized collection? (Books, manuscripts, photographs...)

What cataloging database system do you use for your digitized collection?

What was (were) the main reason(s) for digitizing the collection?

- to preserve the original
- to improve accessibility
- to support educational/research activities
- to increase information sharing
- Other:

Is (Are) your digital collection(s) published online?

- Yes
- No

- Not now, but we plan to do it

The followings describe problems faced during the implementation of metadata in digitization projects. Please identify the rating scale of each problem

Hard to decide which metadata schemes to use (descriptive cataloging and subject cataloging)

1 2 3 4

It never was a problem It was a great problem

Several confusing metadata concepts (metadata types, mapping, crosswalk...)

1 2 3 4

It never was a problem It was a great problem

Difficult to determine which metadata elements are useful for users and staff

1 2 3 4

It never was a problem It was a great problem

Not enough existing data on the materials

1 2 3 4

It never was a problem It was a great problem

Need of high knowledge and skills on the part of staff

1 2 3 4

It never was a problem It was a great problem

Not enough available documentations

1 2 3 4

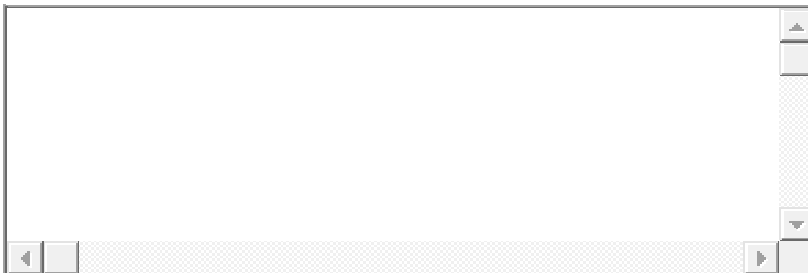
It never was a problem It was a great problem

Insufficient budget

1 2 3 4

It never was a problem It was a great problem

If you had/have other problems not mentioned above, please state them



Does your institution have guidelines about how to carry out the creation and implementation of metadata during digitization processes of collections?

- No
- Few
- Quite a lot

- Many

Who creates and implements metadata during digitization process?

- Staff (cataloger, archivist, curator, IT staff...)
- It is an automatic process
- Both (staff and automatic)
-

What is the importance of each information source for adding metadata to the digital items? (chose each element's importance from 1-not important- to 4 -very important-)

	1	2	3	4
Material itself or the packaging	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Researchers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fieldwork	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Do you use any other information source?

Which metadata scheme is used for your digital collections?

Why the above-mentioned scheme was chosen? (chose each element's importance from 1-not important- to 4 -very important-)

	1	2	3	4
It is flexible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It is extensible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It is easy to use	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

	1	2	3	4
It supports information sharing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It is widely used	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It is supported by leading organizations	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
You had a previous experience with it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Is there any other reason for choosing that scheme?

Please, list the metadata elements you consider more important:

Descriptive metadata (to facilitate discovery, identification, selection, access...)

Structural metadata (to define how to display the digital resource and its relationship with other resources)

Administrative metadata (to manage the collection)

In your opinion, which factors will influence your collection scanning practices in the near future? (chose each element's importance from 1-not important- to 4 -very important-)

	1	2	3	4
Administrative infrastructure changes in your institution	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Development of policy on scanning and digital collection management	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Metadata standards	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The increasing number of digital objects	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Ongoing knowledge and skills	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Technology	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Users' needs	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Participating in a joining program	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Staff commitment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Name other factors that you think may influence in your institution's collection scanning practices

Does your institution have any metadata preservation policy?

- Yes
- No
- No, but planning to design one

If you have a preservation policy, what does it consist in?

A large, empty rectangular text input field with a thin black border. On the right side, there is a vertical scroll bar with a small upward-pointing arrow at the top and a downward-pointing arrow at the bottom. The scroll bar is currently at the top position, indicating the text area is empty.

APPENDIX 5

Semi-structured interview guide:

- First of all: could you please describe in a general way your digital object production line?
- What specific metadata scheme do you use for each kind of digital object (photography, books, manuscripts, journals, magazines, etc..)?
- Your metadata schemes are _____. You emphasized its positive aspects: flexibility, interoperability, widely used, supported by leading organizations... But did you find any problem or difficulties while creating, implementing or managing metadata of your digitized collections? (mention technical problems, difficulties to understand-teach to staff, missing elements, or any aspect that you consider it could be improved).
- Do you consider these schemes simple and easy to use for the staff? Did your institution offer any special training for staff? If so, what kind of training?
- How is your data entry interface? Is it usable?
- What is the metadata scheme used for your non digital objects? How do you export these files to the new digital object's metadata files? Do you have an import tool?
- What difficulties did you find exporting these records to the new digital object's metadata files?
- Participation in joint digitizing programs: in what programs do you participate? How do you distribute the work? What of the institution takes care of metadata creation, management and preservation?
- Interoperability. Have you ever had any interoperability problems with other institutions you collaborate or provide information to? What problems? For example in joint digitization projects? Or other situations...
- Metadata elements. Your chosen scheme fulfills all your needs or you miss any metadata element? What do you miss? Is it extensible? (allows you extension for particular needs)

- Technology: did you have any technological problems during digitization? What problems? (concerning software and hardware)
- How do you feel about cataloguing digital objects? Does it require the same effort from staff as cataloguing analog materials? more, less...?
- Do you have digitization guidelines? What do they consist in?
- You stated that staff is metadata creators... could you describe a little bit more specifically this process and the kind of staff involved in each phase? Problems faced during metadata creation?
- Do you get direct support from any leading institution? What? Would you like to get it?
- Would you suggest any comment, feeling or recommendation for digital object cataloging? (concerning metadata standards, database, interoperability, or any other aspect you would like to mention. Feel free).