Do Van Chau

Challenges of metadata migration in digital repository: a case study of the migration of DUO to Dspace at the University of Oslo Library

Supervisor: Dr. Michael Preminger

Oslo University College

Faculty of Journalism, Library and Information Science

DECLARATION

I certify that all material in this dissertation which is not my own work has been identified and that no material is included for which a degree has previously been coffered upon me.

……Do Van Chau……. (Signature of candidate)

Submitted electronically and unsigned

ACKNOWLEDGEMENTS

This work is finished with the supports from many persons in DILL program and at University of Oslo Library.

I am very grateful for valuable advice and enthusiasm from my supervisor, Dr. Michael Preminger. He has taken time and effort to read and comment on my work.

I would like to take this chance to thank all librarians and technical staffs at University of Oslo Library, librarians at Oslo University College and University of Cambridge Repository for sharing thoughts and comments in the questionnaires.

I also express deepest attitude to all professors in DILL program who have given interesting lessons for me. In particular, I would like to say thank you to Prof. Ragnar Nordlie for critical comments on my thesis during the seminars.

Finally, special love is given to my family and friends who are always beside me and give strong encouragement to me during the study.

ABSTRACT

This work is a study of challenges in the metadata migration, generally and with DUO as a case, thereby defining the appropriate strategy to convert metadata elements of DUO to Dspace in the migration project at UBO. The study is limited to DUO as a case study. DUO is currently using home-grown metadata elements while Dspace takes Dublin Core Metadata element set as a default metadata schema. Therefore, the challenges including risks and conflicts might be occurred in the metadata migration process from DUO database to Dspace. In order to minimize these risks and conflicts, the appropriate strategy for the DUO migration plays an important role.

To define the appropriate strategy and identify the challenges of metadata migration in DUO migration project, the structured interviews have been conducted to informants who play different roles in the DUO projects. Furthermore, the experiences of previous migration projects worldwide have also been consulted as well as the crosswalk of metadata elements in both DUO and Dspace were performed as well.

The results of this study indicate that creation of a custom schema for transferring metadata elements and their values from DUO database to Dspace is a suitable strategy among other strategies. Many kinds of risks and conflicts in the migration of metadata elements in DUO to Dspace were identified through this study such as *data loss, data distortion*, *data representation*, *synonyms*, *structure of elements set*, *null mapping* and *duplicate values.* From these issues, some recommendations have been made to control the challenges in the migration.

The findings in the thesis could be a useful reference for the DUO migration project and similar projects. The thesis might be used in the stage of decision-making for such future projects. Otherwise, the issues of the crosswalk from home-grown metadata elements to DCMES might provide evidences for other studies in this field.

Keywords: metadata migration, strategy and challenges, digital repository, DUO, Dspace.

TABLE OF CONTENT

LIST OF FIGURES AND TABLES

ABBREVIATIONS

AACR2 : Anglo-American Cataloguing Rules Second Revision

ANSI : American National Standard Institute

CCO : Cataloguing Cultural Objects

DC : Dublin Core

DCMES : Dublin Core Metadata Element Set

DCMI : Dublin Core Metadata Initiative

DOAR : Directory of Open Access Repositories

DUO : DigitaleutgivelservedUiO (Digital publication at University of Oslo)

EAD : Encoded Archival Description

ECCAM : Extended Common-Concept based Analysis Methodology

FGDC : Federal Geographic Data Committee metadata

IPL : Internet Public Library

IRs : Institutional repositories

LII : Librarian's Internet Index

MARC : MAchine-Readable Cataloging

MARC21 : MARC for 21st century

METS : Metadata Encoding and Transmission Standard

MODS : Metadata Object Description Schema

MR : Metadata repository

NISO : National Information Standards Organization

NSDL : National Science Digital Library

OAI : Open Archives Initiative

OAI-PMH : Open Archive Initiative – Protocol for Metadata Harvesting

OCLC : Online Computer Library Center, Inc.

PAP : The Picture Australia Project

RDF : Resource Description Framework

SQL : Structured Query Language

UiO : University of Oslo

UBO : University of Oslo Library

USIT : University Centre for Information Technology

XSLT : Extensible Stylesheet Language Transformations

XML : Extensible Markup Language

CHAPTER 1: INTRODUCTION

The chapter provides the background and statement of research problem as well as the aim of study and research questions. Afterwards, the scope of the study as well as the research methods is presented. Finally, an outline of the thesis is introduced.

1.1 Background

Metadata in digital institutional repositories (IRs) has been the subject of great concern from both research and practical communities. National Information Standards Organization (NISO), a non-profit association accredited by American National Standard Institute (ANSI) has provided a formal definition of metadata. According to the document titled Understanding metadata published by NISO in 2004, metadata is "*structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information*" (NISO, 2004, p.1). There are three main types of metadata introduced in this document: descriptive metadata, structural metadata and administrative metadata. Some functions of metadata are resource discovery, organizing electronic resources, interoperability, digital identification and archiving and preservation (NISO, 2004, p.1-2).

Park (2009) has conducted a study of the current state of research and practices on metadata quality in IRs. In her reviews, she did critical analysis of various issues related to metadata quality in IRs such as inconsistency, incompleteness and inaccuracy of metadata elements.

In addition to quality issues of metadata in IRs, Vullo, Innocenti and Ross (2010) have described multi-level challenges that digital repositories face towards policy and quality interoperability. These levels consist of organizational interoperability, semantic interoperability and technical interoperability. It was stated that "*there is not yet a solution or approach that is sufficient to serve the overall needs of digital library organizations and digital library systems*" (Vullo, Innocenti and Ross, 2010, p.3). By NISO (2004, p.2), *"interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality".* NISO (2004, p.2) also mentioned "*defined metadata schemes, shared transfer*

protocols, and crosswalks between schemes" as means to achieve the interoperability among different systems used in repositories. Two approaches for interoperability offered by NISO are cross-system search by Z.39.50 protocol and metadata harvesting via OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting) (NISO, 2004, p.2).

In a study of methodology for metadata interoperability and standardization, Chan and Zeng (2006) emphasized a proliferation of metadata schemas applied in IRs, "*each of which has been designed based on the requirements of particular user communities, intended users, types of materials, subject domains, project needs*"[1]. They proposed many kinds of methods to facilitate the migration and exchange of metadata among different metadata schemata and applications in IRs. These methods have been used to achieve or improve the interoperability among metadata schemata in IRs at three levels: repository level, schema level and record level. At repository level, efforts focus on mapping value strings associated with particular elements to enable cross-collection searching. At schema level, efforts focus on creating the communication among elements of metadata schemata. Methods used in this level include derivation, application profiles, crosswalks, switching-across, framework, and registry. At record level, efforts focus on integrating records through record migration and data reuse and integration. The results create new records based on combining values of existing records.

In practice, many important projects have been conducted to support the interoperability in different IRs worldwide such as the migration project at the Energy and Environmental Information Resources Centre (France), the Metadata Repository project at National Science Digital Library Metadata Repository, the migration project at University of Sydney Repository and the crosswalking project of Internet Public Library at Drexel University. These projects will be discussed in detail in chapter 2.

1.2 Problem statement

DUO (abbreviated from Norwegian name "DigitaleutgivelservedUiO") is a digital Institutional Repository at the University of Oslo (UiO), Norway. DUO was developed in 2000 in cooperation between University Centre for Information Technology (USIT) and the

[1] http://www.dlib.org/dlib/june06/chan/06chan.html

University of Oslo Library (UBO). Today, DUO includes electronic versions of theses, special assignments, doctoral dissertations and articles from UiO.

From 2010, UBO has decided to take Dspace (UBO, 2010) into use as a new platform for DUO migration because the old platform of DUO was obsolete. The project will establish new DUO as an open digital archive for the University of Oslo's total digital production. The project consists of three subprojects: Student Communication, Communication Research and Communication Media. DUO database was developed on home-grown metadata elements which had been chosen to meet specific needs of user communities at UBO, while Dspace is currently using standard Dublin Core Metadata Set. Hence, the definition of the migration of metadata elements in DUO to Dspace should be a requisite part of the migration project. Woodley (2008) has indicated that "*migration is accomplished by mapping the structural elements in the older system to those in the new system*" (p.7). She also found that "*there is often not the same granularity between all the fields in the two systems*" (p.7) because "*data fields in the legacy database may not have been well defined, or may contain a mix of types of information*" (p.7) .Thus, investigation of a suitable strategy of metadata mapping between DUO and Dspace is an important study before performing the real process of the migration of DUO to Dspace.

1.3 The aim of the study and the research questions

The study is an effort of identifying challenges in the metadata migration, generally and with DUO as a case, thereby defining the appropriate strategy to convert metadata elements of DUO to Dspace in the migration project at UBO. To achieve this aim, two following research questions are going to be regarded:

Research question 1: *What is the appropriate strategy to migrate metadata elements from DUO database to Dspace in light of current practices and the research available in this field?*

Research question 2: *In light of various issues experienced in previous metadata migration projects at different levels as well as issues particular to DUO, what are the challenges of metadata migration from DUO database to Dspace?*

1.4 Research methodology

In this study, DUO migration project at UBO is chosen as a case for investigation. Basing on this case, two techniques are going to be used to collect data: structured interview and crosswalk. The questionnaire contains both open-ended and closed-ended questions written in English language. The web based survey tool, SurveyMonkey is used to deliver the questionnaires to informants involved in the DUO project. Data collected from questionnaires are qualitative data because all questions were designed to get the opinions and experiences of informants about many kinds of research issues. Afterward, constant comparative analysis (Hewitt-Taylor, J., 2001, p.42) is used to analyze data gathered from questionnaires.

In addition to collecting data by questionnaire, previous studies and projects related to metadata migration in IRs are critically reviewed to gain the theoretical and practical background of the research issues. Then, the structure and semantics of metadata elements used in both DUO and Dspace are compared to develop a metadata crosswalk from DUO to Dspace. By this process, the conflicts of metadata elements in both systems are further defined.

1.5 Scope of the study

The strategies for metadata migration at schema level as well as the challenges of DUO migration project at UBO are the main foci of this study. The investigations of metadata migration from DUO to Dspace focuses on defining semantic mapping of metadata elements rather than matching of values associated with each element. Due to time and technical constraints, the study does not aim to conduct the experiments to examine the migration of metadata elements and their associated values at record level.

Otherwise, only informants involved in DUO migration project are consulted for this study.

1.6 Thesis outline

The content of thesis is presented in five chapters in addition to table of content, figures and tables, reference and appendices.

Chapter 1 presents the background and research problem statement as well as the aim of the study and research questions, brief introduction of research methodology and scope of study.

Chapter 2 gives a review of recent studies about various issues related to the topic of thesis such as metadata quality issues in IRs, metadata migration in theories and practices in IRs, semantic mapping of metadata schemata and conflicts in crosswalk.

Chapter 3 provides the justification of methods used in the research and the explanations of the ways these methods are going to be implemented to collect and analyze data.

Chapter 4 deals with the data collected by data analysis and discussions. Afterwards, findings of the research are summarized.

Chapter 5 presents the conclusions and recommendations for the research. It revisits the research questions set up in the beginning and lays out suggestions to solve the research issues and further studies related to topic.

CHAPTER 2: LITERATURE REVIEW

The chapter reviews recent studies in theory and practices related to institutional repositories (IRs) in academic libraries. Most of these studies are published recently in books, research papers, articles and reports from many sources such as Springer Link databases, D-Lib Magazine, Cataloging & Classification Quarterly, Emerald databases, etc. To find documents related to topic, some search engines were used including Google Scholar, and BYBSYS at Oslo University College Library and search functions integrated in Springer Link, Emerald databases. Afterward, ISI Web of Science was exploited to find more related documents based on citation retrieval. Several keywords have been used for searching documents. They are metadata migration, metadata migration, metadata translation, metadata issues, metadata quality, metadata crosswalk, metadata mapping, metadata integration, metadata challenge, metadata conflicts, and metadata semantics. Sometimes, scanning the reference list in one document can be a good way to reach to other interesting documents. Main focus of the reviews includes metadata quality issues in IRs, metadata migration in theories and practices in IRs, semantic mapping of metadata schemata crosswalk and challenges in metadata migration.

2.1 Metadata issues in institutional repository

2.1.1 Define institutional repository

Institutional repository becomes an essential infrastructure for scholar activities in universities on the world. This is evidenced by the development of thousand of IRs listed in DOAR (Directory of Open Access Repositories). Lynch (2003) defines IRs as: "*a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members*" (p.1).

Heery and Anderson (2005) developed a typology that provides a helpful framework for exploring IRs, as presented in Figure 1 below:

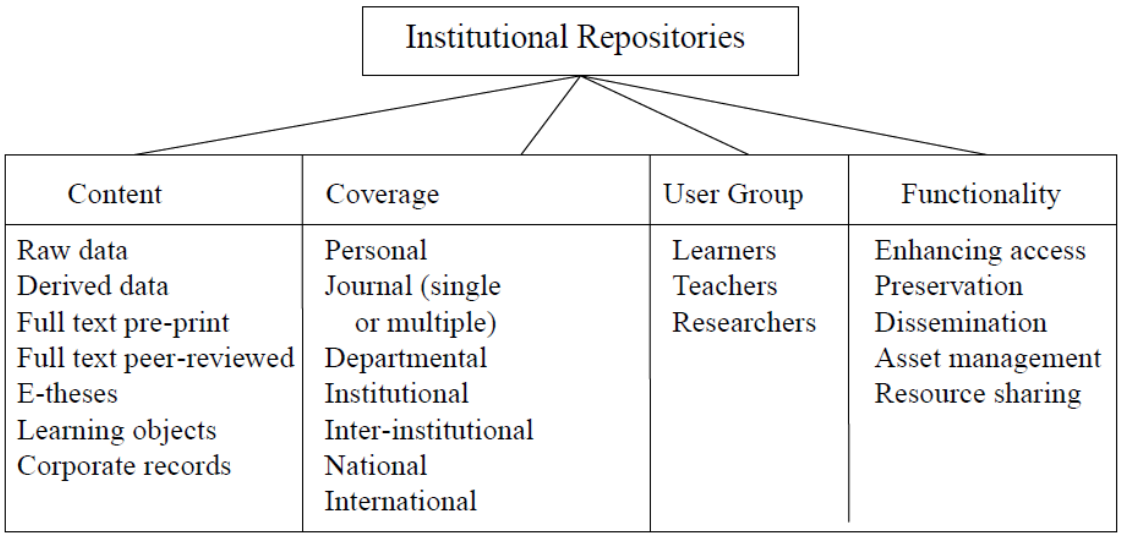

Figure 2.1: Typology of IRs (Heery and Anderson, 2005, p.17)

This framework presents four main focus of IRs including content, coverage, users and functionality.

2.1.2 Metadata quality issues in IRS

Almost concern about metadata quality in IRs is consistency. Bruce and Hillman (2004) stated a need to ensure elements are implemented in a way that is consistent with standard definitions and concepts in the subject or related domains. The authors also suggested that metadata elements should be presented to the user in consistent ways.

Park (2009) has defined the most common criteria for quality of metadata in institutional repository including completeness, accuracy and consistency.

The completeness of metadata elements can be evaluated by "*full access capacity to individual local objects and connection to the parent local collection(s). This reflects the functional purpose of metadata in resource discovery and use*" (Park, 2009, p.8). Furthermore, Zeng and Qin (2008, p.254) suggested that "*each project should set its own analysis criteria based on the functional requirements defined for its metadata system*" to evaluate the completeness of metadata functions in the system.

The accuracy (also known as correctness) of metadata elements "*concerns the accurate description and representation of data and resource content*" as well as accurate data input (Park, 2009, p.9). According to Zeng and Qin (2008, p.255-256), the accuracy of metadata elements could be measured in such various dimensions as:

- *"Correct content: metadata record represents resources correctly*
- *Correct format: correctness of element label and its values, data types, application of element syntax.*
- *Correct input: examines spelling, grammar, punctuation, word spacing, missing words or sections, foreign characters, etc.*
- *Correct mapping/integration: correct mapping of metadata elements in harvesting and crosswalks".*

Some tools such as content standards (Anglo-American Cataloguing Rules 2nd edition (AACR2), Cataloguing Cultural Objects (CCO), etc.), best practices guidelines provided by metadata standards and application profiles could be the best resources to check whether a metadata record correctly represents the content of resources.

The consistency of metadata elements can be measured by "*data value on the conceptual level and data format on the structural level*". The conceptual consistency *"entails the degree to which the same data values or elements are used for delivering similar concepts in the description of a resource"*. The structural consistency "*concerns the extent to which the same structure or format is used for presenting similar data attributes and elements of a resource*" (Park, 2009, p.10).

Zeng and Qin (2008, p.257) explained in detail many types of checking consistency in metadata migration such as consistent source links, consistent identification and identifier, consistent description of source, consistent metadata representation and consistent of data syntax.

Stvilia et al. (2004) divided metadata quality problem into six categories as following: lack of completeness, redundant metadata, lack of clarity, incorrect use of metadata schema or semantic inconsistency, structural inconsistency and inaccurate representation.

In another study of Electronic Theses and Dissertation metadata in digital repository at Drexel University which used Dspace, Janick and McLaughlin (2004) indicated the lack of specific metadata elements. These are date degree is awarded, type of degree, advisors and committee members, date of defense, and contact information for the author.

Other quality issues of metadata were also conveyed in many studies such as:

- Lack of contextual aspects of metadata: Metadata can be sparse or lack important contextual information particularly when that context is held at a collection level. Furthermore, there are no controlled vocabularies in subject headings and lack control of authority for author names (Chapman, Reynolds and Shreeves, 2009, p.3).

- Semantic overlap in several Dublin Core elements: type and format, source and relation, two qualifiers-part of and version of in element relation (Park, 2005).
- Inaccurate, incomplete and inconsistent usage of metadata elements in National Science Digital Library (NSDL). For example, the physical description field is either inaccurately used as format or description of Dublin Core elements; there is great confusion in employing the DC elements like type and format; the DC elements like source and relation are inconsistently used (Bui & Park, 2005, p.3).

Metadata quality is also specifically discussed in semantics by Park (2009) in *Metadata quality in digital repositories: a survey of the current state of the art.* The author specified various kinds of issues related to meaning of metadata in IRs as followings:

- The same meaning can be expressed by several different forms (e.g., synonyms) and the same forms may designate different concepts (e.g., homonyms) (p.5)
- The same concept can be expressed by different morpho-syntactic forms (e.g., noun, adjective, compound noun, phrase, and clause) (p.5)
- Different communities may use dissimilar word forms to deliver identical or similar concepts, or may use the same forms to convey different concepts (p.5)

Recently, in study of metadata best practices guidelines at Utah Academic Library Consortium, Toy-Smith (2010) emphasized that metadata consistency should be the primary consideration when developing digital collections.

2.1.3 Metadata interoperability in IRs

Park and Tosaka (2010) have conducted study of current state of metadata practices across digital repositories and collections by giving surveys for cataloging and metadata professionals in United States of America. They concluded that metadata interoperability still is a major challenge. The reason is *"a lack of exposure of locally created metadata and metadata guidelines beyond the local environments" (p.1).* Furthermore*, "homegrown locally added metadata elements may also hinder metadata interoperability across digital repositories and collections when there is a lack of sharable mechanisms for locally defined extensions and variants" (p.1)*

In this study, homegrown schemata and guidelines were defined as "*local application profiles that clarify existing content standards and specify how values for metadata elements are selected and represented to meet the requirements of a particular context" (p.6).* From this view, the authors investigated motivations for creating homegrown metadata

elements. The results showed that the desire to reflect the nature of local collection and the characteristics of target community of local collection are two main motivations beside constraints of local conditions and local systems.

In another study of metadata decisions for digital library projects, Zeng, Lee and Hayes (2009) reported that interoperability issues were highly concerned in most of libraries. "*Their concerns ranged from planning and mapping together various metadata templates to enable standards used by various communities interoperable within one discovery system*" (p.179)

2.2 Metadata migration in IRs from methodological point of view

Blanchi and Petrone (2001) defined metadata migration is "*a set of operations to translate the metadata contained in the digital object into another metadata schema*"[2].

In study of methodology for metadata interoperability and standardization, Chan and Zeng (2006) have defined three levels of metadata interoperability among IRs include schema level, record level and repository level. In the case of converting metadata from one schema to another, the authors suggested two methods including crosswalk at schema level and record migration at record level.

2.2.1 The crosswalk at schema level

A crosswalk is "*a mapping of the elements, semantics, and syntax from one metadata scheme to those of another*" (NISO, 2004, p.11). In similar view, Pierre and LaPlant (1998) stated *"crosswalk is a set of transformations applied to the content of elements in a source metadata standard that result in the storage of appropriately modified content in the analogous elements of a target metadata standard"*[3]. According to DCMI (Dublin Core Metadata Initiative) glossary, crosswalk is "*a table that maps the relationships and equivalencies between two or more metadata schemes. Crosswalks or metadata mapping support the ability of search engines to search effectively across heterogeneous databases*"[4]

[2] http://www.dlib.org/dlib/december01/blanchi/12blanchi.html
[3] http://www.niso.org/publications/white_papers/crosswalk/
[4] http://dublincore.org/documents/usageguide/glossary.shtml#C

Chan and Zeng (2006) asserted that crosswalks are by far the most commonly used method to enable interoperability between and among metadata schemata. In their views, crosswalks allow systems to effectively convert metadata elements from one schema to another.

The crosswalk commences with two independent metadata schemata. Then, equivalent or comparable metadata terms (elements and refinements) between those schemata are investigated. The predominant method used in crosswalk is direct mapping or establishing equivalency among elements in two schemata. The mapping refers to a formal identification of equivalent or nearly equivalent metadata elements or groups of metadata elements from two metadata schemata, carried out in order to facilitate semantic interoperability. The mechanism used in crosswalks is usually a chart or table that represents the semantic mapping of data elements in one metadata standard (referred as source) to those in another standard (referred as target) based on the similarity of function or meaning of the elements.

In general, two approaches have been used in crosswalk practice. The first is absolute crosswalk which requires exact mapping between the involved elements of a source schema and a target schema. Where there is no exact equivalence, there is no crosswalk. Absolute crosswalk ensures the equivalency (or closely-equivalent matches) of elements, but does not work well for data migration. The problem is that data values in non-mappable space will be left out, especially when a source schema has a richer structure than that of the target schema.

The other one, relative crosswalk is used to solve this problem. This way has been used to map all elements in a source schema to at least one element of a target schema, regardless of whether the two elements are semantically equivalent or not. The relative crosswalk approach appears to work better when mapping from complex to simpler schema (e.g., from MARC to DC, but not vice versa) (Chan and Zeng, 2006).

Pierre and LaPlant (1998) have indicated some problems in the crosswalk as well. According to their studies, crosswalk is a difficult and error-prone task requiring in-depth knowledge and specialized expertise in the associated metadata standards. Furthermore, obtaining the expertise to develop a crosswalk is particularly problematic because the metadata standards themselves are often developed independently, and specified differently using specialized terminology, methods and processes. Otherwise, maintaining

the crosswalk as the metadata standards change becomes even more problematic due to the need to sustain a historical perspective and ongoing expertise in the associated standards.

In the study, Chan and Zeng (2006) also mentioned some issues of the crosswalk between two independent metadata schema such as different degrees of equivalency including one-to-one, one-to-many, many-to-one, and one-to-none; no exact equivalents and perhaps overlap in meaning and scope of elements. Hence, data quality problem might occur in data migration based on crosswalk.

2.2.2 Record migration at record level

Chan and Zeng (2006) explained that the migration at record level was conducted when different projects had a need for integrating established metadata database. Recently, more projects have attempted to reuse existing metadata records and combine them (or their components) with other types of metadata records (or their components) to create new records. Two common methods for integrating or converting data values associated with specific elements/fields are migration and data integration.

Woodley (2008, p.7) also defined that "*data migration projects transfer the values in metadata fields or elements from one system (and often one schema) to another*". She mentioned a variety of reasons for data migration. For instance, when institutions want to upgrade to a new system, because the legacy system has become obsolete; or when they decided to provide public access to some or all of its content and therefore wishes to convert from a proprietary schema to a standard schema for publishing data.

Migration of metadata record

In this way, one metadata schema based a record including metadata elements and their data are converted to those in another schema. Some good projects of record migration are The Picture Australia Project (PAP) and National Science Digital Library (NSDL) (Chan and Zeng, 2006). In PAP, records from partner institutions are collected in a central location (the National Library of Australia) and then translated into a common record format which based on the Dublin Core metadata. Similarly, some records in NSDL were harvested from Alexandria Digital Library and later they were converted into DC records.

The major challenge in record migration is how to minimize loss or distortion of data. Zeng and Xiao (2001) found that mapping or converting became even more complicated when data values were involved. When the target record is more inclusive and has defined elements and sub-elements in greater detail than the source record, values in source record may need to be broken down into smaller units. For this reason, data values may be lost when converting from a rich structure to a simple structure. Zeng (2006) in recent study has provided strong evidence about the impact of the crosswalk based on real data migration on data quality when converting a large amount of data.

Metadata reuse and integration

Chan and Zeng (2006) presented that the components of a metadata record can be regarded as various pieces of a puzzle. They could be put together by combining pieces of metadata sources coming from different processes. They could also be used and reused piece by piece when new records need to be generated.

They also indicated the Metadata Encoding and Transmission Standard (METS) as the standard is used for packaging descriptive, administrative, and structural metadata into one XML document for interactions with digital repositories. Hence, it provides a framework for combining several internal metadata structures with external schemata.

Otherwise, The Resource Description Framework (RDF) of the World Wide Web Consortium was suggested as a data model to develop and share vocabularies with other communities.

In short, the selection of methods for metadata migration in IRs depends on status of metadata schemata being used and desired outcomes that the institutions want to reach.

2.3 Practices of metadata migration in IRs

A number of projects of metadata migration have been conducted in libraries worldwide so far.

Firstly, University of Sydney Repository had a project of migrating separated databases at faculties/units to Dspace. Those databases used various kinds of self-developed metadata elements stored on programs such as Filemaker, SQL or spreadsheet applications. Since metadata elements in those databases are quite different from the default Dublin Core Metadata Set in Dspace, four different choices of migration have been offered as following:

- Map original metadata elements to existing Dublin Core (DC) elements in Dspace

- Map original metadata elements to DC elements and create new qualifiers for DC elements
- Create a custom schema identical to the original metadata set
- Generate DC records as abstractions of the original metadata records and submit the original metadata records as digital object bit-streams

According to Brownlee (2009), each choice contains both advantages and disadvantages. The first choice has low submission and maintenance costs, OAI-PMH compliance and less effort on metadata schema customization but it might face with the loss of metadata granularity and data distortion. The second choice retains the granularity of original records and support harvesting via OAI-PMH but it has higher submission and maintenance costs and the challenge of DC registry management. The third choice avoids DC registry management issues whereas it requires much effort on customization of metadata schemata and OAI crosswalks as well as ongoing maintenance of Dspace index keys and project-specific schemata. The final choice keeps metadata records in their original format but it does not support the harvesting of original records. After the discussion, the University of Sydney Library has selected the fourth choice to apply for the project because it was thought to be coherent with primary preservation function of the repository. Furthermore, this choice might have least requirements for resources on ongoing maintenance of multiple schemata.

Secondly, the Internet Public Library (IPL), Drexel University (United States of America) had a project to convert local metadata elements stored in Hypatia (SQL database) to Dublin Core Metadata set. The IPL decided to develop a crosswalk between existing metadata elements and Dublin Core elements. To support this process, several activities were made for preparation including analysis the quantity and quality of the existing IPL metadata, creation of a new IPL metadata schema as an application profile of Dublin core, development of a new database structure and the development and testing of a new metadata creation and maintenance interface (Galloway, M. et al., 2009, p.1). In particular, the results of analytical comparison between IPL existing fields and Dublin Core Metadata Element set showed that there's no directly field to field mapping between two systems. The reasons for this issue were that fields in Hypatia database had different labels, definitions and the same data were represented in different ways. Otherwise, a number of fields were only used in Hypatia database and some of them had been no longer in use.

To prepare for the crosswalk, IPL has created a custom metadata schema by applying the concept of application profile. The custom schema contains existing IPL domain specific metadata elements and exploits Dublin Core Metadata Element set. It consists of four namespaces:

- *Dublin Core Metadata Element Set (version 1.1)*
- *Dublin Core Metadata Element Set Qualifier (2000)*
- *IPL-defined Metadata Element Set*
- *IPL-defined Metadata Element Set Qualifiers*

  (Galloway, M. et al., 2009, p.1)

IPL defined elements and qualifiers mostly focused on administrative and technical aspects of metadata. The custom schema at IPL specified element status and repeatability by taking the IPL context into account. Nevertheless, Galloway, M. et al., (2009, p.2) indicated that there were challenges in reaching consensus on metadata labels and element status within IPL Dublin Core compliance group. They also are working to develop further content designation rules and semantic aspects of the IPL custom metadata schema.

Regarding to IPL, Khoo and Hall (2010) have studied of metadata merger between IPL and the Librarian's Internet Index (LII) in which each library's metadata was mapped to Dublin Core to create new version of IPL (IPL2). From this process, they identified following challenges (p.2-4).

- Some metadata elements in the sources (IPL and LII) such as Former title, Sort title, Acronym, Alternate title and Alternate spelling were rarely used and unnecessary. There were many discussions about whether these elements should be used in IPL2. Finally, they were placed in custom administrative fields, "out of sight" of users.
- Many IPL collections had collection-level records but no item-level records for objects belong to those collections. This meant that there would no metadata for these objects mapped to DC.
- The collections are stored in both MySQL database and Filemaker Pro database so that they cannot be included in the same crosswalk process.
- Lack of controlled subjects headings in both IPL and LII.

Thirdly, the Energy and Environmental Information Resources Center has conducted a project of converting Federal Geographic Data Committee metadata (FGDC) into MARC21 and Dublin Core in OCLC's WorldCat. According to Chandler, Foley and Hafez (2000), the

migration included three steps. Firstly, a smaller number of elements referred to as "essential FGDC metadata" for a fully compliant FGDC record were selected. Criteria for selection including: required (mandatory) elements, search keys such as author, title, subject, date and commonly elements used by creators of FGDC metadata. Secondly, the crosswalk from FGDC to MARC21 and Dublin Core was developed. Finally, a converter program written in C was created to implement the migration.

Fourthly, Bountouri and Gergatsoulis (2009) proposed a crosswalk from Encoded Archival Description (EAD) to the Metadata Object Description Schema (MODS) comprising of three components. It includes creation of semantic mapping from EAD elements/attributes to MODS elements/attributes; mapping the hierarchical structure of EAD document to MODS and retaining in MODS the information inherited from the hierarchical structure of EAD document.

These following steps have been done to create a semantic mapping of the elements between EAD and MODS. Firstly, EAD and MODS's records were examined in elements and attributes, their semantics and scope notes. Secondly, semantic mapping among EAD fields and MODS fields were defined. Finally, some real-world examples were created to check the semantic correctness of mappings between EAD and MODS fields.

Two approaches were investigated to map the hierarchical structure of EAD documents to MODS. When there is a need to describe a single archival unit (e.g. a photograph) and provide some contextual information about its resources (e.g. collection of photographs), the standalone approach might be used. In this way, the record describing a photograph is related to the record representing the corresponding collection. On the other hand, if there is a need to provide users with a complete representation of the resources, records that include nested MODS records might be created (p.19).

For the case that the inherited information was not taken into account during the process of transforming an EAD document to MODS, considerable information may be lost. To cope with this issue, two different approaches were suggested by Bountouri and Gergatsoulis (2009, p.20-21). They are resulting MODS records embodying the inheritance property and constructing self-contained MODS records with respect to their information content.

Finally, National Science Digital Library had developed the Metadata Repository (MR) to convert metadata records harvested from various collections into Dublin Core records. By

Arms et al. (2003), MR "*holds collection-level metadata about every collection known to the NSDL and an item-level metadata record for each known individual item*" (p.228). Since it's difficult to establish a metadata standard that all collections in NSDL support, MR was designed to accept several preferred metadata that the collections will provide. Some of them are Dublin Core, Qualified Dublin Core, IEEE Learning Technology Standards Committee (IMS), MARC 21, Encoded Archival Description (EAD), etc. In addition to storing original metadata record harvested in MR, a Dublin Core record in a format called nsdl_dc is created for each object. Most of nsdl_dc records are created by crosswalks from original metadata records. Below is the mechanism to import metadata records into MR via OAI-PMH:

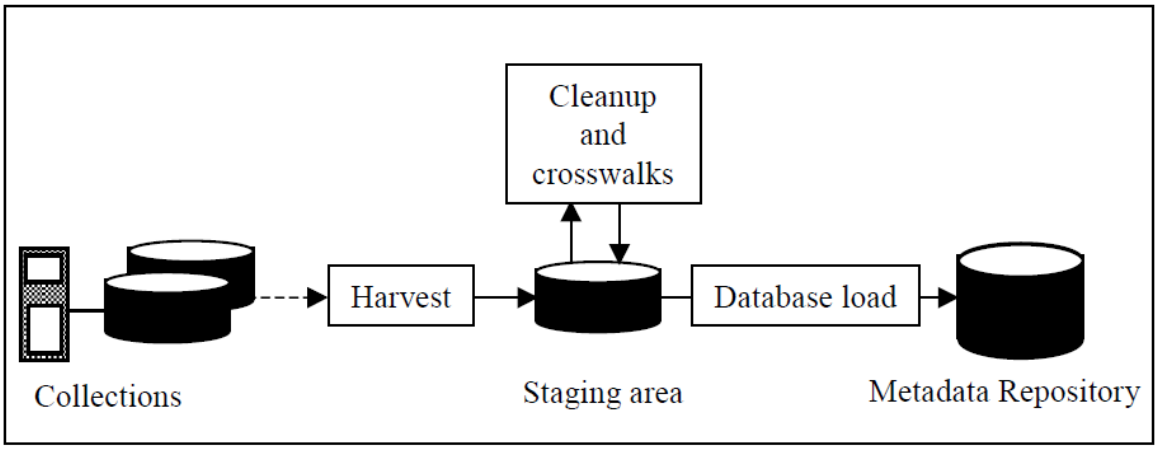

Figure 2.2: Import metadata record into MR via OAI-PMH

(Arms, et al., 2003, p.232)

MR at NSDL is designed as a relational database using the Oracle database software. The mechanism of importing metadata into MR begins by encoding in XML the original metadata records which are harvested from collections. When the records come to the staging area, they pass through three stages. Firstly, they are processed via cleanup step which includes "*combining ListRecords responses and possibly stripping off some of OAI-PMH wrapping*" (p.232). Secondly, a crosswalk is used to generate metadata record in nsdl_dc format. The crosswalks are implemented in XSLT (Extensible Stylesheet Language Transformations). They create XML files containing batches of records. Finally, the XML files are loaded into the database by Java programs. Thus, both the original metadata record and nsdl_dc record are stored together in MR.

2.4 Semantic mapping of metadata in crosswalk

2.4.1 Define semantic mapping

Semantic mapping is "*the process of analyzing the definitions of the elements or fields to determine whether they have the same or similar meanings*" (Woodley, 2008, p.3).

In technical view, Noy and Musen (2000) stated that mapping aims to establish correspondences among the source ontologies, and to determine the set of overlapping concepts, concepts that are similar in meaning but have different names or structure, and concepts that are unique to each of the sources.

2.4.2 Types of similarity/correspondences among schemata elements in semantic mappings

Masood and Eaglestone (2003) suggested Extended Common-Concept based Analysis Methodology (ECCAM). ECCAM define 2 types of semantic similarity among schema elements:

- Shallow similarity: two elements share common concepts among their intrinsic meanings.
- Deep similarity: two elements share common concepts among their intrinsic meanings in particular context.
  - ✓ The intrinsic semantics of a schema element is its meanings independent from the context within which it is used.
  - ✓ The in-context semantics of an element is its more specific semantics within the contexts in which the element is defined in schema. This in-context semantics are determined by intrinsic semantics and the contexts within which it is modeled.

In mapping assertion metamodel below, there are 4 types of relations: similar, narrower, broader and related to.

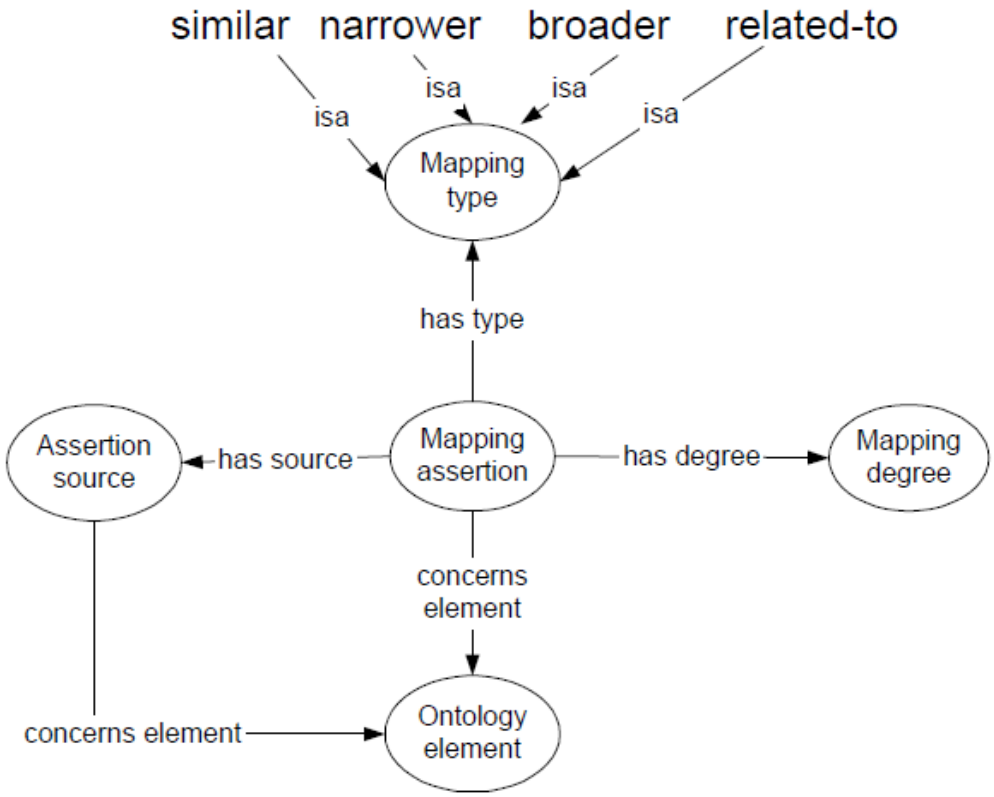

Figure 2.3: Mapping assertion metamodel (Hakkarainen, 1999)

There is one more type of relations, which is called dissimilar relation, added to the modified metamodel (Su, 2004, p.105).

Hakimpour and Geppert (2001) defined four levels of similarity relations as well:

- *Disjoint definitions*
- *Overlapping definitions*
- *Specialized definitions* (*sub concept* or *sub relation*)
- *Equal definitions*

From museum and archival practices, Lourdi, Papatheodorou and Nikolaidou (2006) identified specific "association" types correlating a couple of elements from the two different schemata:

- *equivalence:* for mapping elements that have the same meaning
- *refinement:* to express a relationship between an element and its qualifier following exactly the DC
- *Hierarchical*: to connect elements that can be considered as broader and narrower concepts.

2.4.3 Practice of semantic mapping in crosswalk

Lourdi, Papatheodorou and Nikolaidou (2006, p.16-17) demonstrated an effort to make the semantic mapping of metadata schemata in digital folklore collections. These collections belong to Greek Literature Department of the University of Athens. The researchers conducted the mapping by creating a table correlating the semantics of two different metadata schemata (vocabularies). For each metadata element of the source schema, they located a semantically related element of the target schema. In particular, they consider each metadata element as a topic and they define types of associations among metadata elements. An association correlates two metadata elements that belong to different schemata and each of the elements has a specific role in the association.

The mapping procedure follows these steps:

Firstly, they consider each metadata element as a "topic" with its own attributes, according to the metadata standard that comes from.

Then, they defined three topic types categorizing the elements of the two schemata: descriptive, administrative and structural metadata. Each metadata element is an instance of one of the above types.

Next, specific "association" types correlating a couple of elements from the two different schemata are formulated as following:

- *Equivalence:* mapping elements that have the same meaning
- *Refinement:* expressing a relationship between an element and its qualifier following exactly the DC
- *Hierarchy*: connecting elements that can be considered as broader and narrower concepts.

Finally, as each element in an association has a specific role, they have set the following couples of role types: *equivalent terms* for the "equivalence" association, *broader - narrower term* for the "hierarchical" and *element type – qualifier* for the "refinement" association.

Below is an example table of presenting roles and association types in mapping between the source (application profile for collection level) and the target (Dublin Core Collection Description Application Profile) (Lourdi, Papatheodorou and Nikolaidou, 2006, p.18).

(Note: DC CD AP: Dublin Core Collection Description Application Profile; ISAD: General International Standard Archival Description; ADL: the metadata model of Alexandria Digital Library; RSLP: Research Support Libraries Program; LOM: IEEE-Learning Object Metadata)

COLLECTION APPLICATION PROFILE		DC CD AP		ASSOCIATION TYPE
ELEMENT	ROLE	ELEMENT	ROLE	
(ISAD)_NOTE	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(ISAD)_LEGAL STATUS	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(ADL)_SCOPE/PURPOSE	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(DC)_SOURCE	NARROWER TERM	CUSTODIAL HISTORY	BROADER TERM	HIERARCHICAL
(RSLP)_LOCATION_PHYSICAL	EQUIVALENT TERM	IS LOCATED AT	EQUIVALENT TERM	EQUIVALENCE
(RSLP)_ACCRUAL STATUS	BROADER TERM	ACCRUAL_PERIODICITY ACCRUAL_POLICY ACCRUAL_METHOD	NARROWER TERMS	HIERARCHICAL
(LOM)_STRUCTURE	NARROWER TERM	CATALOGUE OR DESCRIPTION	BROADER TERM	HIERARCHICAL
(DCTERMS)_TABLE OF CONTENTS	QUALIFIER	CATALOGUE OR DESCRIPTION	ELEMENT TYPE	REFINEMENT
(DCTERMS)_RELATION	ELEMENT TYPE	ASSOCIATED PUBLICATION	QUALIFIER	REFINEMENT
(DC)_CONTRIBUTOR	NARROWER TERM	ABSTRACT	BROADER TERM	HIERARCHICAL
(RSLP)_LOCATION_DIGITAL	EQUIVALENT TERM	IS AVAILABLE VIA	EQUIVALENT TERM	EQUIVALENCE

Figure 2.4: Semantic mappings between collection application profile and Dublin Core Collection Description Application Profile

However, in this table, there is no clear explanation of the reason why element "ABSTRACT" in the target can be seen broader concept of element "(DC) _CONTRIBUTOR" from the source in mapping.

2.5 The challenges in metadata migration

Three types of conflicts in schema integration which belong to structural conflicts were studied by Batini and Lenzerini (1987, p.346) as following:

- *Type conflicts:* the same concept is represented by various forms/roles in different metadata schemata. This is the case when, for example, a class of objects is represented as an entity in one schema and as an attribute in another schema

- *Dependency conflicts*: the relations in group of concepts are expressed with different dependencies in more than one metadata schemata. For example, the relationship "marriage" between "man" and "woman" is expressed 1: 1 in one schema, but m: n in another schema.
- *Behavioral conflicts*: different insertion/deletion policies are assigned to the same class of objects between two schemata. For example, in one schema, class "department" may be allowed to exist without employees, whereas in another schema, deleting the last employee associated with class "department" leads to the deletion of the department itself. Note that these conflicts may arise only when the data model allows for the representation of behavioral properties of objects.

In similar point of view, Su (2004, p.85-86) in his study has categorized two types of conflicts in semantic mapping were terminology discrepancies and structural discrepancies.

The terminology discrepancies include:

- Synonym occurs when the same object or relationship is represented by different names/labels in component schemata.
- Homonym occurs when different objects or relationships are presented by the same name in the component schemata.

The structural discrepancies include:

- Type discrepancies arise when the same concept have been modeled using different data structure.
- Dependency discrepancies arise when a group of concepts are related among themselves with different dependencies in different schemata. For example, the relationship ProjectLeader between Project and Person is 1:1 in one schema, but m:n in another.

In study of metadata migration, Woodley (2008, p.7) has indicated some misalignments occurred during data migration include:

- There are no complete equivalent between metadata elements in source database and those in target database.

- It is difficult to distinguish between metadata elements that described original object and those that described object related information such as related image or digital surrogate.
- Data assigned in one metadata element in source schema may be mapped to more than one element in the target schema.
- Data is presented in separate fields in source schema may be placed in a single field in the target schema
- In a situation that there is no element in the target schema with an equivalent meaning with the source, unrelated information may be forced into a metadata element with unrelated or only loosely related content.
- When there is no consistency in entering data into records, it may not be possible to use same mapping mechanism for all records that are being converted.
- There may be differences in granularity and community specific information between the source and the target in migration.
- The source metadata schema may have hierarchical structure with complex relationships among elements while the target schema has flat structure or vice versa.

Furthermore, Chan and Zeng (2006) also found *"that data values may be lost when converting from a rich structure to a simpler structure".* In another study, Zeng and Qin (2008) addressed four most serious issues in metadata migration including "*(1) misrepresented data values, (2) valuable data values that are lost, (3) incorrectly mapped elements and data values, (4) missing elements*" (p.256).

In practice, Jackson, et al. (2008, p.11-14) have conducted some experiments to find out any changes in semantics and values in metadata migration from one metadata schema to another. They remapped original metadata records to Dublin Core at University of Illinois at Urbana Champaign to see which fields were most often incorrectly mapped. The results showed that publicly available crosswalks (e.g., Library of Congress' MARC to Dublin Core Crosswalk) do not always account for semantic values of elements, and may provide misleading mappings. Otherwise, among the fifteen simple Dublin Core elements, description, format, subject, and type fields show the most significant changes in numbers when remapped from the original harvested records. Multiple value strings in one element instance in the original records caused the increase in description and subject fields.

The authors also identified some kind of conflicts in metadata mapping to Dublin Core elements such as publication dates are mapped to the coverage field instead of the date field. Furthermore, information of different digital collections in the same IRs is placed in source instead of relation field. In another case, some records use the format field to describe the means of accessing the digital object, rather than the format of the object.
Finally, the authors conclude that original metadata records are rich in meaning in their own environment, but lose richness in the aggregated environment due to mapping errors and misunderstanding and misuse of Dublin Core fields. Also, mapping is often based on semantic meanings of metadata fields rather than value strings; and correct mapping could improve metadata quality significantly.

Park (2005) also conducted a pilot study to determine the accuracy of the mapping from cataloger-defined natural vocabulary field names (source) to Dublin Core metadata elements (target). Total of 659 metadata records from three digital image collections were chosen. Some evidences of incorrect and null mapping were identified. For example, "*physical field*" in source was either mapped to "*description*" and "*format*" in target; "*subject*" in target was mapped by various fields in source such as "*category*", "*topic*", "*keyword*", etc. Furthermore, some null mapping fields such as "*contact information", "note", "scan date", "full text*", etc. were identified as well.
From the results of this pilot study, the author strongly suggest "*the critical need for a mediation mechanism in the form of metadata mapping guidelines and a mediation model(e.g., concept maps)that catalogers can refer to during the process of mapping*" (p.8). The goal of this mechanism is increasing semantic mapping consistency and enhancing semantic interoperability across digital collections.

Conclusion

From reviews of studies of metadata migration and issues in IRs, some methods for converting metadata element and its values such as crosswalk, record migration and data reuse or integration are analyzed. Furthermore, the approaches for metadata migration based on experiences in practices are also discussed. Otherwise, many studies have found out critical issues in the crosswalk such as semantic conflicts and quality control of metadata in metadata migration from one metadata schema to other schemata. Those

theoretical background and experiences might be useful for defining appropriate strategy and make good preparation for DUO migration project at UBO.

CHAPTER 3: RESEARCH METHODOLOGY

The chapter addresses methodology and its deployment in this research. Sample population, data collection techniques and instruments are also explained. In particular, pilot study and afterward necessary adjustments as well as data analysis techniques are discussed as well.

3.1 Methodology

The research is based on qualitative methodology because it focuses on investigating the point of views from UBO librarians and outside experts as well as to analyzing the semantic of metadata elements being used in current DUO database. According to Strauss and Corbin (1990, p.19), "*qualitative methods can be used to uncover the nature of person's experiences with a phenomenon… and understand what lies behind any phenomenon about which little is yet known*". Since metadata migration from DUO to Dspace at UBO is a specific situation, the research methods used is case study. Pickard (2007, p.86) addressed that the purpose of a case study is to "*provide a holistic account of the case and in-depth knowledge of the specific through rich descriptions situated in context*". She further stated that "*using case studies is the most appropriate method when the purpose of the research requires holistic, in-depth investigation of a phenomenon or a situation from the perspective of all stakeholders involved"* (p.93).

The technique proposed to use to collect data is the structured interview. In addition to this primary technique, previous studies related to the topic and system documents about metadata used in DUO and Dspace are critically analyzed to gain fully and deep understandings of current research and practices available and the circumstances of the case study. Otherwise, the crosswalk of metadata elements in both DUO and Dspace is developed by using harmonization technique.

3.1.1 Structured interview

In discussion of Pickard (2007, p.175), Fontana and Frey (1994, p.363) defined "*structure interviewing refers to a situation in which an interviewer asks each respondent a series of preestablished questions with a limited set of response categories*".

Pickard (2007, p.175) introduced two forms of structured interview. The first is standardized, open-ended interview. In this interview, all respondents are asked the same, open-ended questions but they are allowed to respond in any way they feel comfortable with and with any kind of information they want to share with the researcher. In the second form, close and fixed-response interview, respondents receive the same questions and choose answers from a predetermined set of alternative choices. In practice, those forms of structured interview could be used together. In this study, two forms of structured interview are combined in use.

Also according to Pickard (2007, p.175), the major benefit of close and fixed-response interview is the visual and oral clues that researchers can pick up by listening and watching the respondent. She explained that researchers can learn a lot not only from what is said but also from how something is said. She stated that the interview is used to gain in-depth understanding of individual perceptions and when the nature of data is too complicated to be asked and answered easily (p.172). In case of metadata migration from DUO to Dspace, librarians and referred experts may have various attitudes/ideas about this process and expected outcomes. Therefore, it is important to explore those perspectives before ending up with a suitable strategy for this kind of migration.

In this study, the implementation of structured interview technique is proposed to be divided in two steps. Firstly, a well-structured questionnaire which consists of both closed questions and open-ended questions is composed and then distributed to the informants who involved in DUO migration project. Secondly, some informants will be picked up for interviews basing on their responses with the aims either to discover their experience regarding some important dimensions related to the case study or to clarify unclear information in their answers. Nevertheless, only one informant was interviewed by email to ask him exemplify his answers. Since some questions in the questionnaire prompted the informants to give their interpretation of the things that have not yet decided in the project, they refused to answer them. In this case, it's difficult to have more interviews with them.

3.1.2 The crosswalk

The crosswalk is "*a mapping of the elements, semantics, and syntax from one metadata scheme to those of another*" (NISO, 2004, p.13). In similar view, Pierre and LaPlant (1998) stated *"crosswalk is a set of transformations applied to the content of elements in a source*

metadata standard that result in the storage of appropriately modified content in the analogous elements of a target metadata standard"[5].

In studies of metadata interoperability, Chan and Zeng (2006) indicated that crosswalks are by far the most commonly used method to enable interoperability between/among metadata schemas. In particular, crosswalks allow systems to effectively convert data from one metadata standard to another. Therefore, crosswalk is carefully considered to apply in metadata migration from DUO to Dspace at UBO.

The crosswalk process including two steps is harmonization and semantic mapping.

Harmonization technique

In the definition by Pierre and LaPlant (1998), "*harmonization is the process of enabling consistency across metadata standards*"[6].

The purpose of harmonization is to develop successfully the crosswalks between metadata standards. Hence, it simplifies the development, implementation and deployment of related metadata standards through the use of common terminology, methods and processes.

The procedure of harmonization is as followings:

Firstly, common terminologies, properties and organization used in both source metadata schema and target metadata schema are defined. For terminology, formal definition of each term and share vocabularies prevent misinterpretation between two schemas are established.

Secondly, similarities and differences of properties used in both schemas are extracted. These properties of metadata element comprise of name, identifier, label, definition, data value (text/numeric/controlled vocabulary, etc.), obligation (mandatory/optional field), relationship (equivalent/hierarchy), and repeatable/unrepeatable field.

Finally, those data in the source and the target schemata should be presented in similar way in order for the mapping in crosswalk could be created easily.

[5,6] http://www.niso.org/publications/white_papers/crosswalk/

Developing the crosswalk by semantic mapping of metadata elements between the source and the target schemata

In the point of views of Pierre and La Plant (1998), this step involves specifying a mapping of each metadata element in source schema with a semantically equivalent metadata element in target schema. These mappings are often presented in tables or charts. There are some types of mapping as below:

One-to-one mapping: The element in source schema is corresponding to the element in target schema.

One-to-many mapping: The element in source schema may be made up of more than one value (for example, title element comprises of formal title, subtitle, title in second language, etc.) so that it can be mapped to more than one element in target schema. This situation often occurs in mapping from simple schema to complicated schema. In this case, the mapping requires specialized knowledge of the composition of the source element, and how it expands into multiple target elements.

Many-to-one mapping: This situation often occurs in mapping from complicated schema to simpler schema. For this case, the mapping should specify what to do with the extra elements. If all values of the source element are transferred to a single value in the target element, some rules are required to specify how the values will be appended together. Alternatively, if only one source element value is considered to map to element in the target, there is possibility of information loss. Hence, the resolution should indicate the criteria for selection of element values, for instance, important value or common value.

Null mapping: The element in the source cannot find corresponding element in target schemas. In this situation, qualifiers may be created in target schema.

There are some exceptional cases which require special specifications for the crosswalk. For instance, an element that is both hierarchical and repeatable in the source is mapped to an element that is not both same hierarchy and repeatable.

Pierre and LaPlant (1998) analyzed that a complete or fully specified crosswalk consists of both a semantic mapping and a metadata migration specification. The metadata migration

specification contains the transformations required to convert the metadata record content from the source into a record content in the target.

The crosswalk between the source and the target are presented in the composite table for easy comparison. In this way, the element from the source will have the correspondent element in the target. Type of mapping of elements between the source and the target schemas is indicated as well.

3.2 Sampling technique

Snowball sampling is used to choose respondents for structured interview because it helps to indentify key informants for this research. Furthermore, it is hard to explore all suitable informants for interview at the first time. In this study, snowball sampling technique is applied as following. Firstly, the introduction letter presenting purpose and objectives of the study is sent to people who are involved in DUO migration project at UiO. Those people include director and vice director of the library, director of information technology unit, director of research department, chief engineer, consultants and Dspace administrators at Oslo University College and Cambridge University Repository. Afterwards, these people will recommend other persons who can contribute information to the research. This searching strategy continues until all suitable people for study are covered.

3.3 Data collection instrument

The instrument selected to collect data is online questionnaire.

The questionnaire is designed to collect ideas, attitudes or comments about research issues from respondents at UBO and outside. It has both closed questions and open-ended questions.

The structure of questionnaire consists of the introduction, three sections and respondent profile described below:

The introduction gives guidelines for respondent about how to make an answer for the questions.

Section 1: Strategy for metadata migration

This section includes positioning questions about motivations, the approach, influence factors and the strategy for metadata migration from DUO database to Dspace.

Section 2: Metadata migration from DUO to Dspace

The respondents are asked specific questions about the reuse of metadata elements in DUO database, the usage of Dublin Core elements and the configuration of metadata registry in Dspace.

Section 3: Conflicts/risks in metadata migration from DUO to Dspace

This part investigates the respondent's perceptions and interpretation based on their experiences about the possible type of conflicts/risks in metadata migration as well as how should the library prepare to control these conflicts/risks.

The final part in the questionnaire asks for respondent's profile such as name, position/role and email address. The information from respondent is declared to be kept secret and it is only used for further discussion about the study.

For distribution, the questionnaire is designed on computer and delivered to informants at UBO and outside by Survey Monkey, an online survey tool. Online survey tool is selected because of its convenience to recipients. It increases the capability of reaching to potential respondents, especially in using snowball sampling technique. Also, it saves time, cost and efforts for both researcher and participants. Nevertheless, there are some threats in online survey as well such as technical problems or low response rate because of the incompatibility of end-user computer and the lack of physical interactions with informants.

In short, the process of developing the questionnaire includes the following steps:

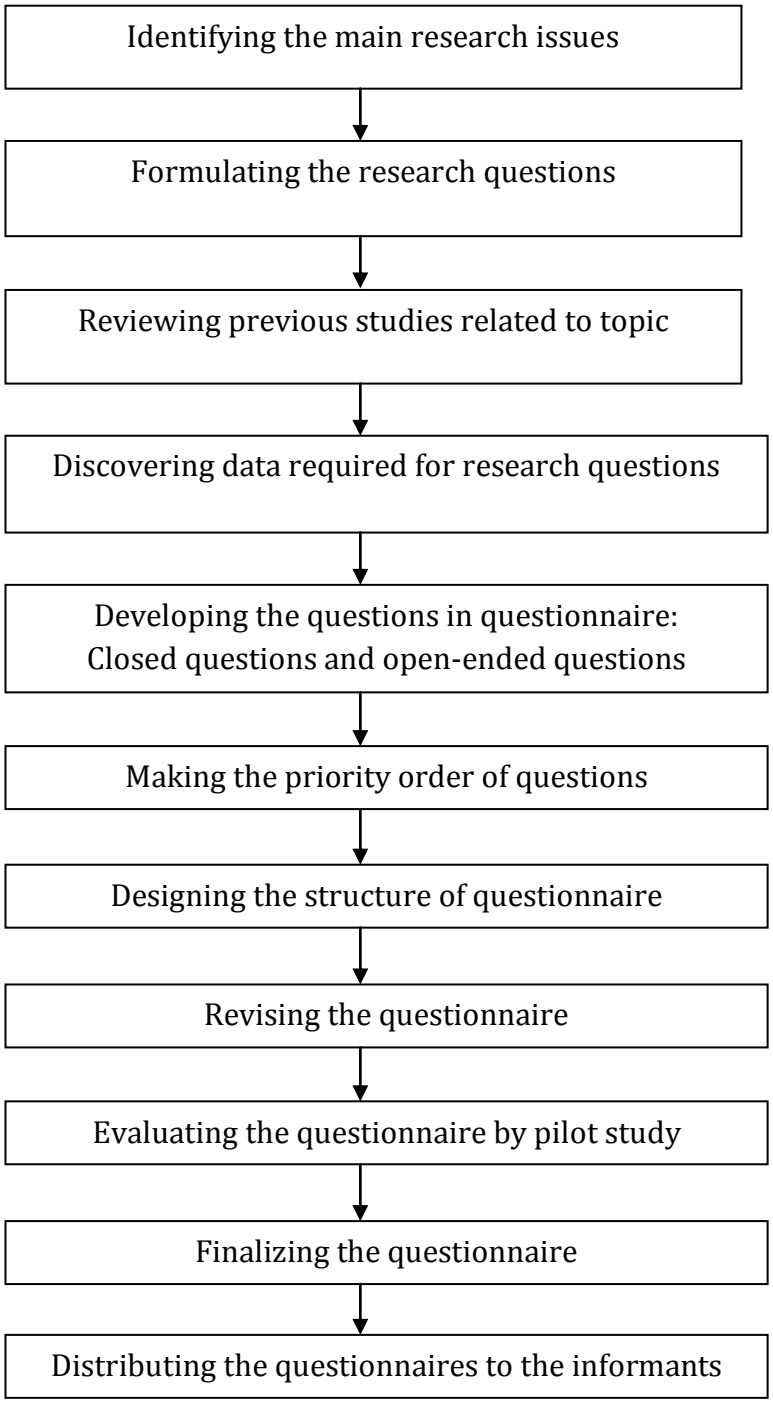

Figure 3.1: Steps to developing the questionnaire

3.4 Pilot testing

Pilot testing is a very important task in study. It helps the researcher make necessary adjustments before official data gathering. In this study, the online questionnaire was tested by a digital services librarian at Oslo University College. She has expertises in digital

repository and Dspace. She is also invited to become a member of consultant committee for DUO migration project at UBO.

The comments and recommendations from pilot respondent focus on rephrasing the questions to reduce ambiguity, changing the three scale names from "very important, less important, no need" to "definitely use, maybe use, won't use" and merging two questions in one.

3.5 Data analysis methods

The data gathered from structured interviews are mainly qualitative data because all the questions focus on finding out the perception and interpretation of respondents. A method called constant comparative analysis is used for coding and categorising data. Constant comparative analysis is "*one method that can be used to identify broad themes and patterns, or categories that emerge from qualitative research studies*" (Hewitt-Taylor, 2001, p.42). This method comprises of three steps including coding, categorizing and clustering.

For coding, each question in the questionnaire is attributed a code which represents a theme that data is associated with. The code is identified by name, definition and abbreviation. Afterwards, data is placed under that code and some notes such as question number, name of respondent are also taken.

For categorizing, when the coding process is finished, the codes that contain common opinions are merged together to form categories. Simultaneously, data placed under each code is also joined together.

Finally, these categories are clustered around each research question to identify which categories could be answer for research issues. Some categories may be related to more than one research question. If categories do not fit to any research issues in the study, it might be used for further research recommendation.

The results of harmonization process are organized in the tables which has many columns reflecting the semantics and content of metadata element such as: element label, qualifiers (for DC), definitions and refinements. Then, the crosswalk between the source and the target are presented in the composite table for easy comparison. In this way, the element

from the source will have the corresponding element in the target. Types of mapping of elements between the source and the target schemas are indicated as well.

3.6 Limitations of the research

Some limitations of the research are addressed below:

Firstly, the answers from informants may be not adequate for clarifying research issues because the DUO migration project at UBO is in an early stage. Therefore, it is hard for informants to interpret many things which have not yet happened in reality.

Secondly, documents describing metadata elements in DUO database are written in Norwegian language. Hence, it is translated to English by some tools such as Google translate, dictionary. Nevertheless, understanding clearly and thoroughly content of these documents is difficult because information sometimes is translated incorrectly.

Thirdly, all the questions and answers are written in English so that the informant may feel uncomfortable to express the ideas. Furthermore, some technical terms can be difficult for informant to understand. Otherwise, English language is also a barrier for researcher to conduct the interview with participants.

Finally, some informants are so busy with the work that they might not take enough time to answer the questions or they will refuse to participate in the study.

3.7 Ethical consideration

The anonymity of the informants was stated clearly in the questionnaires. The names of informants were also coded in the presentation of data analysis and findings. The data collected from the questionnaires were only used for the study purpose. The questionnaires containing answers from the informants are not available in appendices to ensure the confidentiality.

CHAPTER 4: DATA ANALYSIS AND FINDINGS

The chapter presents the data analysis and findings in four sections. The first section is the analysis of data collected from the online questionnaires with librarians at UBO and experts outside. The second section is the harmonization process of metadata elements within both DUO and Dspace based on the analysis of documents describing structures and meanings of metadata elements in these systems. Then, the results of two first sections are combined to develop the crosswalk of metadata element set in DUO and Dspace in the third section. The final section will summarize the findings of research.

4.1 The analysis of data collected by online questionnaires

The online questionnaires have been delivered to 20 informants who are involved in different roles of DUO project including Project, Steering, Line managers, Reference DUO student/academic and Reference DUO media. Totally, there are six informants who have expertise in DUO project, replied the answers to the questionnaires via SurveyMonkey, an online survey tool. The remaining informants refused to give responses to questionnaires with the reason that they do not have specialized knowledge to this project.

The table below gives brief description of replied informants' profile. Their names are coded because the confidentiality was assured to them. All their original answers are put in quote. More additional explanations to clarify their words are placed in square brackets.

Informants	Role	Institution
#H	Vice director	University of Oslo Library (UBO)
#K	Head engineer of new DUO project	University of Oslo Library (UBO)
#To	Consultant	University Center for Information Technology (USIT), University of Oslo
#E	Manager, DUO reference group	Dspace Cambridge Repository
#T	Digital services librarian, DUO reference group	Oslo University College
#M	Software engineer	USIT

Table 4.1: The profile of informants

The responses of all informants stored in SurveyMonkey were exported to PDF file to keep the structure of questionnaire. Each question in the questionnaire is assigned with a theme that is representative for various answers associated in it from informants. Then similar themes are clustered in category which focuses on finding answers for research issues. There are three following major categories generalized from data collection.

- Strategy of converting DUO metadata elements to Dspace
- Customization of metadata elements in Dspace
- Challenges of metadata migration from DUO to Dspace

4.1.1 Strategy of converting DUO metadata elements to Dspace at UBO

The informants were asked to reflect their opinions on important aspects related to strategy of converting metadata elements in DUO to Dspace. The answers from informants are divided into four themes including motivations of migration, migration approaches, factors influencing to metadata migration and choices of migration.

4.1.1.1 Motivations of migrating DUO to Dspace

There are two motivations for which, the decision of migrating DUO to Dspace was made. The first comes from the fact that technical platform of DUO currently cannot meet the requirements of maintenance and development in future at UBO. It is said that "*the existing DUO technical platform is being deprecated*" (#K). Furthermore, "*DUO was developed in programming environment that all web-application in UiO (University of Oslo) shall leave*" (#M). Therefore, technical limitations of DUO might be one of the important reason that it was not received the support to use anymore.

The second motivation is common use, easy customization and interoperable capability of Dspace for which it has been chosen to replace for the position of DUO. This statement is generalized from the answers of most of the informants, for example:

- "*Almost every institution in Norway use the DSpace software for their institutional repository: easier to share code, no longer necessary to develop own software*" (#T)
- "*All other universities in Norway except NTNU [Norwegian University of Science and Technology] use Dspace*." (#M)
- "[Dspace is] *common software platform for nearly all repositories in Norway. It is also used extensively worldwide and open source software*." (#H)

- "*Interoperability, cooperation with other institutions (DSpace is very common in Norwegian universities and colleges), highly customizable open source software, free, durability*" (#To)
- "*DSpace functionality is well suited to our needs, it is open source and can be customized to interoperate with other systems at the University of Oslo, the total cost of ownership is minimal, and we can cooperate with other DSpace institutions both nationally and internationally*." (#K)

From above responses of the informants, technical limitations of current DUO and interoperability of DUO with other institutional repositories in Norway in future are two important motivations that lead to the project of migrating DUO to Dspace.

4.1.1.2 Migration approaches

It's interesting that informants have proposed two different approaches for converting metadata elements in DUO to Dspace.

The first approach is completely change metadata elements in DUO to fit with default Dublin Core Metadata Element Set (DCMES) in Dspace. Two informants (#T, #E) thought this is a suitable proposal because "*there is no reason to mix metadata schemes*" (#T) in both DUO and Dspace and "*ideally one should follow relevant standards such as DC (Dublin Core)*" (#E).

It's obviously true that following this approach, Dspace based DUO database at UBO can achieve the interoperability with other institutional repositories in Norway and on the world as well. Nevertheless, DUO has many local elements because it was developed internally to meet specific needs of local users at UBO. It is possible that some important local metadata elements might not find corresponding elements in DCMES. In this case, these elements and their values may be lost or mapped incorrectly during the migration. This is the risk that should be considered carefully in selection of this approach for converting DUO to Dspace.

Therefore, the informants who suggested this kind of approach have given the reminder about this issue. That is "*the library should of course make sure they keep all the metadata values in the migration*" (#T) and "*a workaround may be useful if valuable information is held in the original formats*" (#E).

In proper care of above issue, two other informants have suggested another approach for migration. The first idea is "*keep the original metadata elements intact*" (#To). The reason given was that "*this would be the ideal solution to avoid losing existing information but what kind of metadata we want to keep is not yet decided*" (#To). In similar point of view, the second idea is "*keep only important local elements*" and "*local administrative data are not interesting to keep*" (#M). From these words, it is understood that metadata elements that help to identify and access digital objects in DUO should be kept. Other elements related to administration of tables in relational database of DUO can be removed.
This approach strongly supports the preservation of metadata elements and their values in current DUO so that information loss can be prevented during the migration. There are, however, questions of how many original metadata elements should be kept, the maintenance cost and the interoperability of DUO with other systems after migration.
The rest informants (#H, #K) provided no ideas about this question.

In short, two approaches for converting metadata elements in DUO to Dspace are completely change metadata elements in DUO to DCMES in Dspace and keeping original metadata elements or important local elements in DUO in the migration. The first approach mostly focuses on achieving the interoperability of DUO with other similar systems in Norway and in the world while the second approach pays more attention to preservation of local elements and their values in DUO. Each of them has both positive and negative aspects that should be examined carefully during the selection of an applicable method of migration.

4.1.1.3 Factors influential to the selection of strategy for converting DUO to Dspace

Informants were asked to rate a set of predetermined factors influential to the selection of strategy for migrating DUO to Dspace. Otherwise, there was also a space for them to add additional factors. Three scales for evaluation are most important, important and least important.

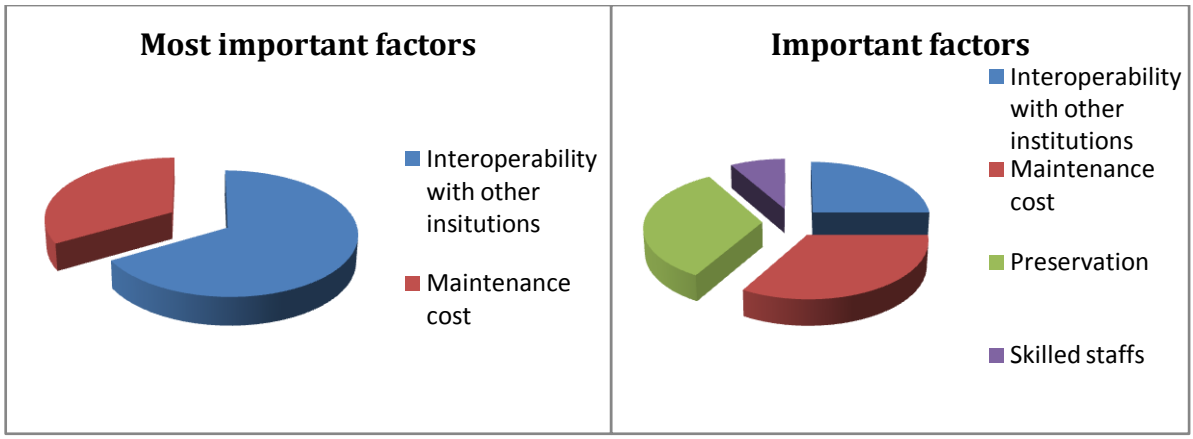

Figure 4.1: Factors influential to strategy of migration

From the charts above, all informants (excluding #T) considered preservation, maintenance cost, interoperability with other institutions and skilled staffs as important factors that influence strategy of migrating DUO to Dspace. In particular, two informants evaluate that interoperability with other institution is the most important factor while one informant has the same evaluation with maintenance cost as well. It is interesting that preservation is thought of as the least important factor by one informant (#M) who suggested keeping important local elements in current DUO.

Today, the interoperability of institutional repositories is increasingly concerned because it allows harvesting and easily sharing data among different repository. Tennant (2001) said that interoperable repositories provide the ability *"to discover through one search what digital objects are freely available from a variety of collections, rather than having to search each collection individually*". To obtain this goal, the repositories should be developed on popular standards such as DCMES and OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Dspace is open source software which supports these standards. In addition to interoperability, preservation of local metadata elements and their values is considered carefully in migration because these elements meet specific requirements of users' community at institution. Maintenance cost and skilled staffs are also noteworthy because they impact successful and sustainable development of DUO after migration.

It is significant to see from the ratings that the informants desire to achieve both interoperability and preservation goals which seem not to be at same direction in the strategy of migrating DUO to Dspace in addition to maintenance cost. This desire is quite understandable because the strategy of migration should be able to allow the new database

of DUO in Dspace at UBO to communicate with similar systems of other institutions in Norway. Furthermore, this strategy should avoid losing important values in the existing database of DUO as well.

In short, factors influence to the selection of strategy for migrating DUO to Dspace should be count on the decision making process of the strategy and procedures of migrating DUO to Dspace.

4.1.1.4 Migration choices

From the experiences of migration of self-developed databases to Dspace which were developed in different programs at University of Sydney Repository, four choices for migrating metadata elements and their values in those databases were generalized (Brownlee, 2009, p.4-6). These choices are:

- Map metadata elements in the original database to existing DCMES in Dspace
- Map metadata elements in the original database to existing DCMES in Dspace and create new custom qualifiers for Dublin Core elements.
- Create a custom schema identical to the metadata elements set in original database
- Generate Dublin Core based records as abstractions of the original metadata records and submit the original metadata records as digital object bit-streams.

The above four choices were mentioned as reference when asking the informants about their opinions or suggestions as to the good possible method for converting metadata elements in DUO to DCMES in Dspace. Two of them were pointed out as the methods to apply in migration project at UBO.

- Map DUO data elements to qualified Dublin Core elements in Dspace and create new qualifiers for default Dublin Core elements in Dspace. (#E, #T, #K)
- Create a custom schema in Dspace identical to DUO metadata elements (#To, #M)

The remaining informant (#H) did not provide any idea about this question.

From the analysis by Brownlee (2009, p.4-6), each of above choices has both positive and negative aspects that need to be checked carefully.

For mapping DUO data elements to qualified Dublin Core elements in Dspace and create new qualifiers for default Dublin Core elements, DUO data elements are transferred to Dspace as default DC elements and remaining elements are mapped to new DC qualifiers. This way has some advantages. Firstly, the granularity of original records and contextual meanings of data are retained so that the recreation of the original records in the future may be supported. Secondly, it does not require too much effort for configuration or maintenance of the DSpace index keys, customized metadata schemata or OAI crosswalks. Finally, records would be fully searchable via default Dublin Core indexing and harvestable via default OAI-PMH.
In addition to these advantages, the library might face some challenges when this method is chosen in the migration project. The first is submission and maintenance costs as well as requiring additional and ongoing recordkeeping and maintenance procedures. The second, when qualifiers of Dublin Core elements proliferate, management of the central registry may be a difficult task.

Another choice for migration suggested by the informants (#To, #M) at UBO is creating a custom schema in Dspace identical to DUO metadata elements. In this way, a custom schema distinct from default Dublin Core is created in Dspace and DUO data elements are transferred to Dspace in their original forms. This choice was preferred by some informants because they thought "*this [way] will ensure that we get all metadata we want*" (#To). Furthermore, "[the library] *need to create more elements in Dspace to handle journals in a proper way because the default metadata set in Dspace doesn't handle journals. These elements may be in bib_work* [a table in current DUO]: *magtitle, magyear, magvolume, magpart, magfirstpage, maglastpage*" (#M).

Also, this choice of migration has both strong points and weaknesses. On the good side, the original forms of important local metadata elements in DUO can be kept in the migration to Dspace. Moreover, this way avoids the challenge of management of central registry present in first choice, by enabling partitioning and separate maintenance of each custom schema.
Nevertheless, it requires efforts in configuration and ongoing maintenance of DSpace index keys, customized metadata schemata and OAI crosswalks. Otherwise, a proliferation of project-specific schemata may require accompanying recordkeeping and maintenance. Therefore, if this choice of migration is used, higher cost and human resource should be

paid than the first choice. That's why the informant who suggested the choice worried "*it is a question of resources and costs and we don't know yet if this strategy is possible*" (#To).

In short, two choices of migration have been suggested by informants. One focuses more on changing metadata elements in current DUO to a standard metadata schema such as qualified Dublin Core in Dspace while the granularity of records in DUO is still kept. Another choice tries to keep the original forms of metadata elements of records in the original database by creating the custom schema in Dspace identical to data elements in DUO. Both choices bring with their advantages and disadvantages.

4.1.2 The usage of metadata elements in Dspace

4.1.2.1 The ways of customization of metadata elements in Dspace

Two major ways of customizing metadata elements in Dspace are suggested by informants. The first way is creating new qualifiers for default Dublin Core metadata set (#T, #E). The second way is using additional metadata schemata and then developing a custom schema in Dspace (#To, #K, #M).

In the first way, "*it would probably extend the existing DC schema in order to maintain similar metadata support to what DUO could do today"* (#E). It is also emphasized that "*if the requirements were different though, (because different types of content would be deposited), defining separate namespaces would be worth exploring. Because DC is hardcoded in some areas of the code base though, this has to be carefully managed*". (#E)

For the second way, "*probably, custom schemas will impose added customization to a number of DSpace components (indexing, OAI-PHM harvesting/crosswalks among others) and it is yet unclear [that] whether library should go down that road or not*" (#K). And "*use additional metadata schema in order not to throw valuable data in DUO*" (#M). Moreover, "*we want to offer (as in DUO now) that you can export bibliographicinformation to reference manager, endnote etc. The Dspace solution is to put everything in a single field and that is not a very good solution. You can't export and differentiate fields use different ways of citation*" (#M).

Thus, by the first way, only default Dublin Core metadata set in Dspace is used as standard metadata schema for the migration and new qualifiers might be added to enable Dublin Core element to fit with data elements transferred from DUO. In second way, metadata

elements from different schemata might be combined to create a custom metadata schema in Dspace which can be mapped as closely as possible from data elements of original records in DUO.

4.1.2.2 Usage of qualified Dublin Core in Dspace

According to the latest documents approved by Dublin Core Usage Board in 2005, qualified Dublin Core Metadata Set has 15 original elements plus 6 additional refining elements and many qualifiers for each element. Some elements of them are used regularly and they can be considered as mandatory elements while the other elements are rarely used as optional elements.

Therefore, the informants are asked to give their opinions about the usage of elements of qualified Dublin Core in Dspace at three levels: definitely use, maybe use and won't use.

The results are presented in the following chart:

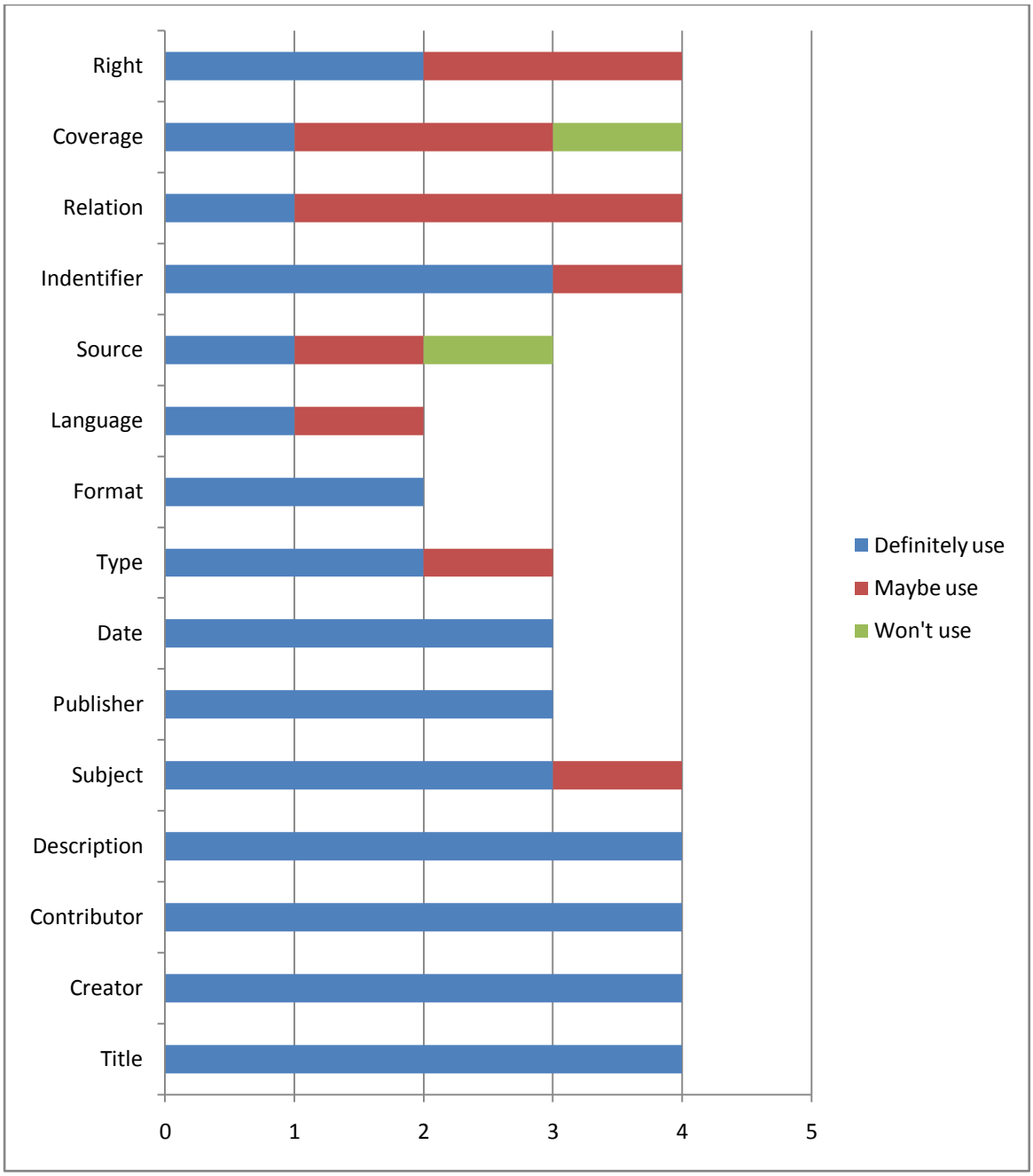

Figure 4.2: Usage of qualified Dublin Core in Dspace

From figure 4.2, most of the elements in qualified Dublin Core receive high "*definitely use*" support for the mapping with data elements transferred from DUO. However, some elements of Dublin Core such as *source* and *coverage* are much less supported.

The results might be used to serve for selection of metadata elements of Dublin Core in the mapping process such as developing the crosswalk between fields in DUO and Dublin Core elements in Dspace or creating a custom schema in Dspace in order to support the migration of DUO to Dspace.

4.1.2.3 Reuse of metadata elements in the existing DUO database during the migration

Matching relational tables describing fields and their values in the current DUO database with qualified Dublin Core metadata elements set in Dspace, it can be seen that many elements in DUO might not find corresponding elements in Dspace. These elements may be local elements which were developed to meet specific needs of user community at UBO. This situation raises a question whether these elements in DUO should be migrated to Dspace or not. In the case that it's necessary, there is also a next question that which elements of them should be reused or extended in the migration. Due to those concerns, the informants were prompted for opinions or suggestions of the reuse of local elements in DUO in the form of three levels of usage: definitely use, maybe use and won't use. The results are shown in following chart:

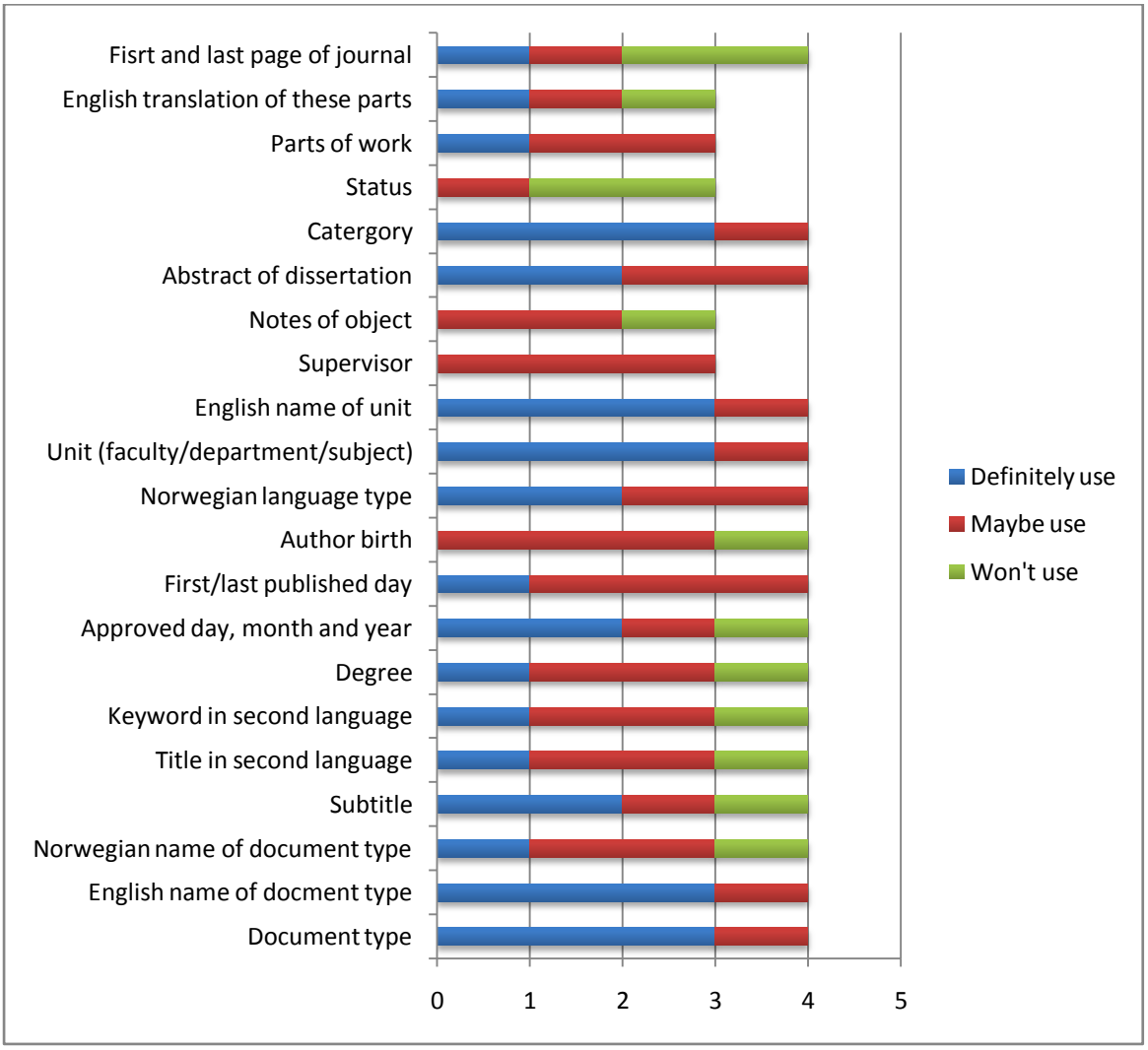

Figure 4.3: Reuse of metadata elements in DUO

Analysis of the above results, some elements such as *document type, English name of document type, unit, English name of unit, category, subtitle, approved day, month and year, Norwegian language type and abstract of dissertation* are strongly recommended by at least half of the informants to reuse in the migration to Dspace. They might be understood as mandatory elements which should be used in the migration.

In opposite side, a few elements such as *first and last page of journal, status* are also suggested for "*won't use*" by half of the informants. Therefore, they might be not necessary to be included in new form of DUO in Dspace.

Finally, remaining elements received only few "*definitely use*" votes or more "*maybe use*" votes. They might be considered as optional elements for usage in selected cases of the migration.

In short, this section has presented two different ways of metadata customization in Dspace to map with metadata elements transferred from DUO. One way is creating new qualifiers for default Dublin Core metadata set while the other is using additional metadata schemata and then developing a custom schema in Dspace. The results of the survey from the informants have also suggested which metadata elements of qualified Dublin Core in Dspace should be used in custom metadata as well as which local metadata elements in DUO should be reused in the migration.

4.1.3 Challenges in metadata migration from DUO to Dspace

4.1.3.1 Risks and conflicts in metadata migration from DUO to Dspace

In previous studies, many kinds of risk and conflicts of converting metadata elements and their values from one system to another in repositories were indicated. At early stage, Batini and Lenzerini (1987, p.346) have found three types of conflicts in schema integration were type conflict, dependency conflict and behavioral conflict.

Su (2004, p.85-86) categorized two types of conflicts in semantic mapping: terminology discrepancies and structural discrepancies.

Chan and Zeng (2006) mentioned the risk of data loss or data distortion in the migration from a complicated metadata schema to simple schema. They also warned against some conflicts in the mapping process among various metadata schemata such as different

degrees of equivalency including one-to-one, one-to-many, many-to-one, and one-to-none mappings; no exact equivalents and perhaps overlap in meaning and scope of elements.

In study of metadata migration, Woodley (2008, p.7) has indicated that some misalignments occur during data migration from including no data match, partly data match, overlap mapping, incorrect data presentation, etc.

The results of those studies are used as hypotheses to examine the possible risk and conflicts in metadata migration from DUO to Dspace at UBO. The informants were asked for their opinions and interpretation of risk and conflicts that may occur in process of mapping data elements in DUO to Dspace. The following risk and conflicts have been interpreted from the questionnaires:

- Data loss: metadata values can be lost in migration (#T, #To)
- Data distortion: contextual meaning of data lost in migration (#T)
- More trouble with differences in DUO metadata accumulated over the years, done differently by each cataloguer and so on (#T)
- Different representation: Data in separated fields in DUO may be in a single element of DC in the Dspace. (For example: moth approved, year approved, first published, last published, creation date (DUO) = date (DC)) (#E)
- The complicated structure of elements set in DUO database and flat structure of Dublin Core in Dspace (#E)
- Synonym: different terminologies for the same value. (For example: Date (Dublin Core) =CREATION DATE (DUO), Description (DC) = Abstract (DUO), Subject (DC) = Keyword (DUO)) (#To)
- The duplicated value because some values are automatically created by Dspace. For example: file format, submission date, etc. (#M)

Two informants did not provide any ideas about this problem (#H, #K).

As DUO has a complicated structure with many local data elements while qualified Dublin Core in Dspace has a flat structure with fewer metadata elements, the risk of data loss and data distortion possibly happen in the migration from DUO to Dspace. Furthermore, DUO database was developed internally by USIT in cooperation with UBO so that labels of fields, values and rules in DUO do not follow standards like Dublin Core in Dspace. Therefore, one

may expect conflicts in terminologies, values and data presentation between the two systems.

Although it's hard to interpret exactly what kind of risk and conflicts will occur in the migration of DUO to Dspace, the above predictions are important to prepare thorough plan for successful migration.

4.1.3.2 Control of risks and conflicts in metadata migration from DUO to Dspace

Both identification and control of risks and conflicts in metadata migration from DUO to Dspace represent challenges to the library.

One informant suggested that "*many of the above risks can be avoided by careful preparations, ascertaining that the metadata in both systems are well understood and mapped as best possible. Also, by not working on the original data but copies will allow a test transfer to take place and problems and errors can be discovered and dealt with before the full migration and transfer is done*" (#E).

In more detail process, (s) he suggested that "*sample single records can be mapped manually using Excel [program] to discover initial problems. Test migrations on larger samples and later in the process on the entire collection will allow a controlled process in terms of handling problems/mapping errors. Test careful at every stage, by manually comparing selected single records - if available automated processes for checking should also be implemented*" (#E).

By this idea, a careful plan before the migration is the most important thing. Then, a pilot migration should be run firstly with sample data to check occurred problems and errors in this process. If errors were discovered, they would be fixed. All the problems in the pilot process will be studied as lessons learnt before the full migration. If these above procedures are properly implemented, the process of migration will be controlled carefully and the expected outcomes would be achieved.

For planning, a suggestion is that "*the library needs to plan everything in advance, have competent staff, do a thorough cleaning and quality control of metadata, know enough about the Dspace software*" (#T). In the same point of view, another informant suggested the

library should "*make a lightweight implementation plan outlining activities including who is responsible for what and when each activity should take place (table form is good for this)*" (#E). Otherwise, it's necessary to have close cooperation among librarians at UBO and other staffs at UiO as one informant said "*the migration of data will be a collaboration process between the project group, the database technicians at USIT and the DUO technical staff*" (#K).

Additional suggestion is creating *date.publishedfirst, date.publishedlast, date.created* in Dspace if library wants to keep the separated fields from DUO intact (#T). This configuration helps to overcome the conflict of different data representation in both systems. For instance, data in separated fields in DUO may be in a single element of DC in Dspace. As an example, some fields in DUO such as *month approved, year approved, first published, last published and creation date* might be mapped to just one element like *date* in Dublin Core.

In conclusion, the section 4.1 has presented the analysis of data in the questionnaires for informants. The data convey important suggestions on the strategy for metadata migration from DUO to Dspace at UBO, the customization of metadata schemata in Dspace and the challenges in the migration from DUO to Dspace at UBO. The results comprise good contribution to defining and controlling the migration of metadata elements from DUO to Dspace at UBO in a proper way.

4.2 Harmonization of metadata elements in DUO and Dspace

Pierre and LaPlant (1998) have defined "*harmonization is the process of enabling consistency across metadata standards"*. The purpose of harmonization is to successfully develop the crosswalks between metadata schemata.

The results of the harmonization process are organized in the table which has columns reflecting the semantics and content of metadata element such as: element labels, qualifiers (for DC elements), definitions and refinements.
Before the harmonization of metadata elements in both DUO and Dspace is developed, it's crucial to understand the structures of both the DUO database and the default Dublin Core schema in Dspace.

Documents describing the structure of DUO database were issued in Norwegian in 2007 by USIT, the organization creating DUO. According to these documents, DUO is a relational database with tables/fields for describing and accessing objects and tables/fields for the administration of the database.

Totally, there are 16 tables: BIB_WORK, BIB_LANGDESCR, BIB_ORGUNIT, BIB_XMLMETADATA, BIB_INSTANCE, BIB_CLASSIFICATION, ASSOCIATION TYPE, Works Association, BIB_CLASSES, BIB_ACTUAL USERS, BIB_EDITOR, BIB_LANGUAGE, BIB_LOGTEXTTABLE, BIB_LOGTABLE, DOCUMENT TYPE and SCIENCE. Detail description of fields in those tables see appendix 1 in appendices.

The complicated relation among these tables is depicted in the following diagram:

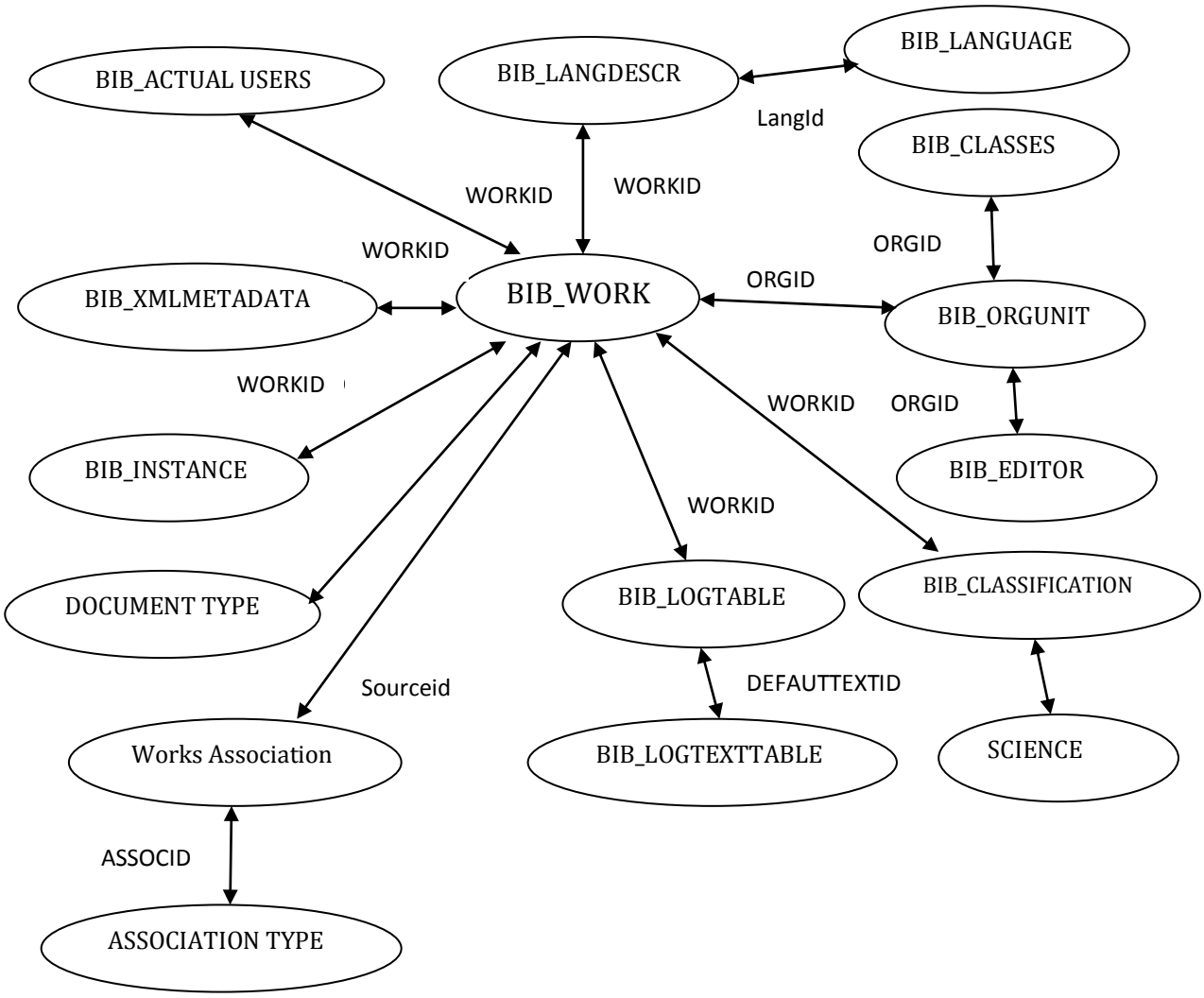

Figure 4.4: Relations among tables in DUO database

For qualified Dublin Core, the document which described Dublin Core qualifiers, two categories of qualifiers, and lists instances of qualifiers was approved by the Dublin Core Usage Board in 2005. The qualifiers listed in this document were generally identified in working groups of Dublin Core Metadata Initiative. It is said that the implementers can develop additional qualifiers for use within local applications or specific domains.
There are two classes of qualifiers. Firstly, element refinement makes the meaning of elements in Dublin Core more specific or narrower. Secondly, encoding scheme identify schemes that aid in the interpretation of an element values. These schemes include controlled vocabularies and formal notations or parsing rules.
List of Dublin Core elements and their qualifiers see appendix 3 in the appendices.

Nevertheless, default Dublin Core Metadata Set in Dspace has been adapted. It is not compliant with original qualified Dublin Core and some deviations have been made in a few elements. For instance, the qualifier "*author*" of element "*contributor*" is used to indicate a person or an organization that is responsible for the content of the resource instead of the element "*creator*". The element "*creator*" is only used for harvested metadata. The list of Dublin Core elements and their qualifiers in Dspace metadata registry is presented in appendix 2 in appendices.

Below is the presentation of the harmonization between metadata elements in DUO and qualified Dublin Core Metadata Set (DCMES) in Dspace

DUO fields	**Definitions**	**DCMES in Dspace**	**Definitions**
TITLE	Title of document	Title	The name given to the resource
SUBTITLE	Under title of document	*Alternative*	Any form of the title used as a substitute or alternative to the formal title
ALTTITLE	Title in second language		
AUTHORLIST	List of authors, separated by #	Creator	An entity primarily responsible for making the content of the resource *Note: used only for harvested metadata*
		Contributor *Advisor* *Author* *Editor* *Illustrator*	Entity responsible for making contributions to the content of the resource Use primarily for thesis advisor

ABSTRACT	Summarize the content of the resource Papers related to the content of the resource	Description *Table of Contents* *Abstract* *Sponsorship* *Uri*	An account of content of the resource A list of subunits of the content of the resource A summary of the content of the resource Information about sponsoring agency Uniform Resource Identifier pointing to description of the object
KEYWORDS	Free keywords	Subject	The topic of the content of the resource. Typically, a subject will be expressed as keywords/key phrases/classification codes that describe the topic of the resource
ALTKEYWORDS	Free keywords in second language		
CLTYPE	Specify classification schema		
		ddc *lcc* *lcsh* *mesh*	Dewey Decimal Classification number Library of Congress classification number Library of Congress Subject Headings Medical Subject Headings
		Publisher	The entity responsible for making the resource available
CREATION DATE FIRSTPUBLISHED LASTPUBLISHED MONTHAPPROVED YEARAPPROVED	Date in which the document was created First time the document was published Last time the document was published Month in which the document is approved Year in which the document is published	Date *Created* *Available*[1] *Issued*[1] *Submitted* *Accessioned*[1] *Copyrighted*	Date will be associated with the creation or availability of the resource. Date of creation of intellectual content if different from date.issued Date that the resource will become available to the public Date of publication or distribution Date of submission of the resource Recommend for theses/dissertations. Date Dspace takes possession of object Date of a statement of copyright
DOCMENT TYPE OAI ENGNAME NORNAME	Category of objects (article, report, book chapter, conference paper, dissertation…) Type name is defined to map OAI harvesting English name for document type Norwegian name for document type	Type (image, sound, text…)	Type includes terms describing general categories, functions, genres, or aggregation levels for content
XML TEXT INSTFORMAT	Xml stream with metadata PDF or HTML	Format *Extent* *Medium*	The physical or digital manifestation of the resource. Typically, Format may include the media-type or dimensions of the resource The size or duration of the resource The material or physical carrier of the resource
LangId ENGNAME NORNAME	ISO 6392 code for language English name of language Norwegian name of language	Language *ISO 639-2RFC 3066*	A language of the intellectual content of the resource
		Source	A reference to a resource from which the present resource is derived *Note: Only use for harvested metadata*

FilePath	URL for the full text document	Identifier	An unambiguous reference to the resource within a given context
MAGTITLE	The title of journal	*Citation*	A bibliographic reference for the resource. Recommended practice is to include sufficient bibliographic detail to identify the resource as unambiguously as possible, whether or not the citation is in a standard form
MAGYEAR	The published year of journal		
MAGVOLUME	The periodical volume		
MAGPART	The journal number		
MAGFIRSTPAGE	The home page of journal		
		Govdoc	A government document number
		Isbn	International Standard Book Number
		Issn	International Standard Serial Number
		Sici	Serial Item and Contribution Identifier
		Ismn	International Standard Music Number
MAGLASTPAGE	Last page of journal		
INSTDESCR	Attach a brief description of the file, which comes up on title page (such as it is a corrected version	Relation	A reference to a related resource
		Isversionof	References to earlier version of object
		Hasversion	References to later version of object
		Ispartof	The described resource is a physical or logical part of the referenced resource.
TEXTFROM	Part of series	*ispartofseries*	Series name and number within that series.
TEXTFROMENGLISH	English translation		
TEXTTO	The series holding/contains	*Haspart*	The described resource includes the referenced resource either physically or logically.
TEXTTOENGLISH	English translation		
Referee	Specify if the document is refereed	*Isreferencedby*	Pointed to by referenced resource
		Uri	References Uniform Resource Identifier for related item
		Coverage	The extent or scope of the content of the resource
		Spatial	Spatial characteristics of the intellectual content of the resource.
		Temporal	Temporal characteristics of the intellectual content of the resource
		Right	Information about rights held in and over the resource
		Access Rights	Information about who can access the resource or an indication of its security status.
		License	A legal document giving official permission to do something with the resource
YEAROFBIRTH	The birth year of author		
TUTOR	Supervisor		
ORGNAME	Name of unit		
ORGTYPE	Specify the type of unit (faculty, institute…)		
NORWEGIAN DISPLAY	Norwegian name that appears in the interface		
ENGLISH DISPLAY	English name that appears in the interface		
UNIT CODE	Unit code		
SCIENCE	The discipline of unit		
CONTENT	For series of booklets		

[1]: default use by system

Table 4.2: Harmonization between fields in DUO and default Dublin Core in Dspace

The results show that labels of fields in DUO and metadata elements in Dublin Core are quite different. Labels of fields in DUO are not assigned in consistent rules. The numbers of elements in DUO is greater than the ones in DCMES as well so that some elements in DUO might be mapped to one element in DCMES and many elements in DUO will not find the correspondent elements from DCMES.

4.3 The crosswalk of metadata elements in DUO and default Dublin Core in Dspace

The crosswalk is developed on the approach in section 4.1.1.2 that important local elements of DUO should be kept in the migration to Dspace as well as the converting method is mapping DUO data elements to existing Dublin Core elements in Dspace and remaining elements are mapped to new DC qualifiers.

The crosswalk of metadata elements in both systems are presented in the composite table for easy comparison. In this way, the element from the source (DUO) will have the correspondent element in the target (Dspace). Type of mapping of elements between the source and the target schemas is indicated as well.

DUO fields	Semantic mapping	Qualified DC elements	Types of mapping
TITLE	→	Title	Many to one
SUBTITLE	→	*Alternative*	
ALTTITLE	→		
AUTHORLIST	→	Contributor *author*	Many to one
TUTOR	→	*advisor*	
ABSTRACT	→	Description *Table of Contents* *Abstract*	One to one
KEYWORDS	→	Subject *lcsh, mesh, ddc, lcc.*	Many to one
ALTKEYWORDS	→		
CLTYPE	→		
CREATION DATE	→	Date *Created* *Issued* *Submitted*	Many to one
FIRSTPUBLISHED	→		
LASTPUBLISHED	→		
MONTHAPPROVED	→		
YEARAPPROVED	→		

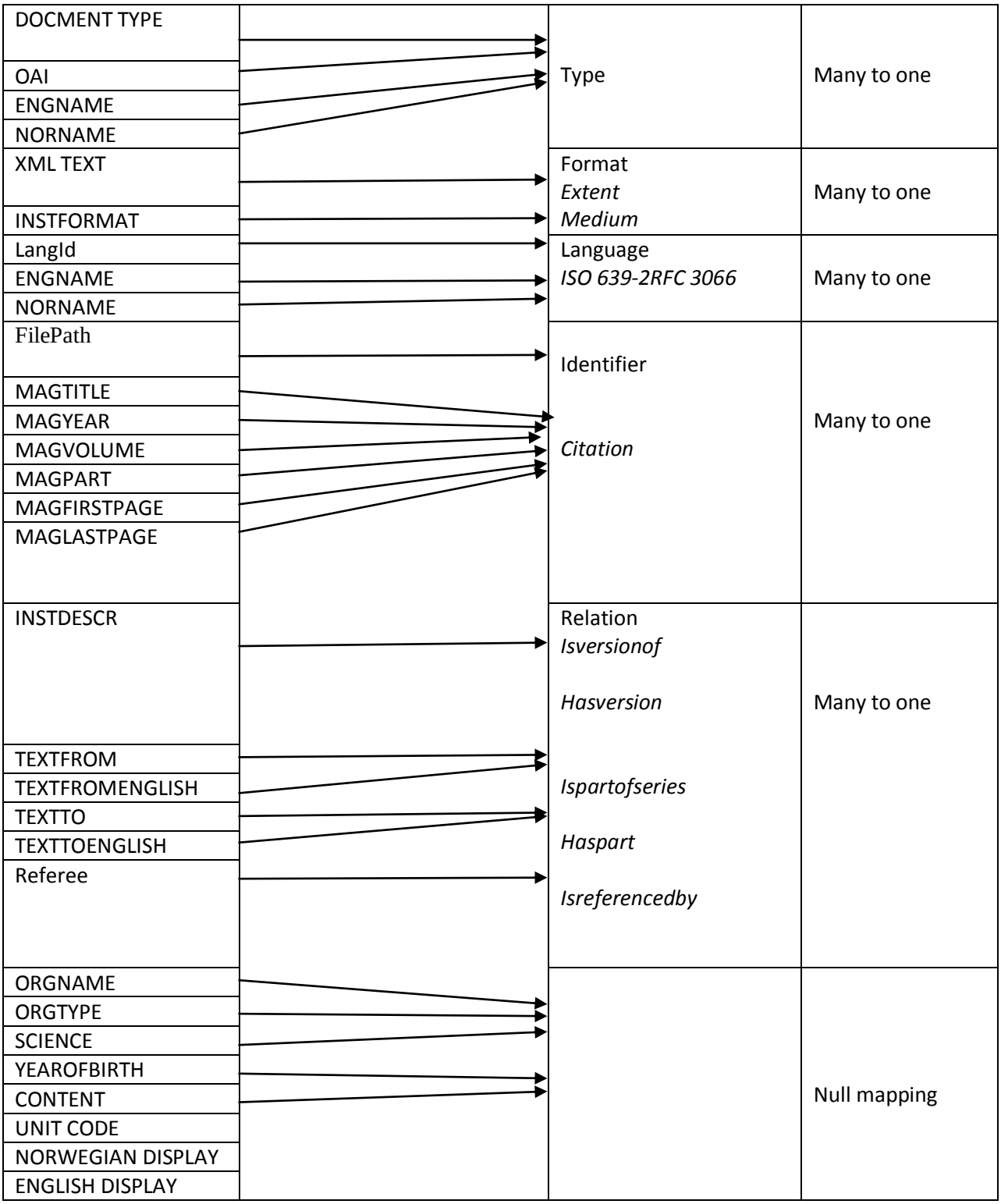

Table 4.3: The crosswalk of metadata elements in DUO and Dspace

Analysis of table 4.3, some elements in DUO presented at the bottom of table cannot map to any elements and qualifiers in Dublin Core. From the results of in section 4.1.2.3, three elements out of them: ORGNAME, ORGTYPE, and SCIENCE were strongly suggested for

reuse in Dspace. Basing on the semantic meanings of Dublin Core elements in the harmonization, those three elements might be suitable to map to element *publisher* of DCMES. The remaining elements such as YEAROFBIRTH, TUTOR, and CONTENT are recommended as "*maybe use*" only. The other elements including NORWEGIAN DISPLAY, ENGLISH DISPLAY, and UNIT CODE were used as administrative elements in DUO so that they can be removed.

Another issue identified in the table of crosswalk is conflicts of mapping metadata elements. Some kinds of conflicts are discussed below:

- Terminology conflict: different labels used for the same concept in descriptions of fields/elements in DUO and Dspace. For example: SUBTITLE, ALTTITLE (in DUO) = Title.Alternative (in DC), AUTHORLIST (in DUO) = Contributor.author (in DC), KEYWORD (in DUO) = Subject (in DC), etc.
- Null mapping: some elements in DUO cannot find the correspondent elements in Dublin Core. List of those elements is presented in bottom in table 4.3.
- Many to one mapping: many elements in DUO are forced to be mapped to one element in DC. Therefore, data in separated fields in different tables of DUO are placed in one element and its qualifiers in DC. This causes issues of data representation and danger of data loss and data distortion.

  For example:

  KEYWORDS, ALTKEYWORDS, CLTYPE (in DUO) = Subject (in DC); CREATION DATE, FIRSTPUBLISHED, LASTPUBLISHED, MONTHAPPROVED, YEARAPPROVED (in DUO) = Date.created, Date.issued, Date.modified, Date.accepted (in DC).
- The conflict between the rich structure among metadata elements of relational database in DUO and the simple structure of Dublin Core elements metadata set in Dspace. Therefore, many elements in DUO become qualifiers of one element of DC in Dspace.

This section has discussed a mechanism of metadata mapping in converting DUO to Dspace. The crosswalk table 4.3 is developed from the combination of results collected from section 4.1.2 and analysis of metadata meanings by harmonization in section 4.3. The situation of metadata mapping from DUO to Dspace is evaluated to identify possible risks and conflicts during the migration. It might give better understanding of metadata issues in order to have a careful plan before the real migration.

4.4 Findings of the study

The investigation of data in the questionnaires and the harmonization process has revealed some important findings for the study. Two major findings are strategy for converting metadata elements in DUO to Dspace and the challenges occurred in this process.

4.4.1 Strategy for converting metadata elements in DUO to Dspace

The important components of this strategy include motivations, approaches, influence factors and methods of the migration.

Firstly, the motivation for migrating DUO database to Dspace is the technical limitations of current DUO platform and the prominent capacities of Dspace such as common use, easy customization and interoperability with other systems of repositories in Norway.

Secondly, two approaches for the migration have been proposed. They are presented as following:

- Completely change the metadata elements in DUO to fit with default Dublin Core Metadata Element set in Dspace.
- Keeping elements of original records in DUO during the migration.

There is a remarkable emphasis that only important local elements should be kept in the migration.

Thirdly, some major factors that influence to the strategy of migration were listed. They are interoperability with other institutions in Norway, maintenance cost, preservation and skilled staffs. In particular, the two first factors were evaluated as the most important factors.

Finally, two choices of migration in the case of DUO are suggested. The first is mapping DUO data elements to qualified Dublin Core elements in Dspace and create new qualifiers for default Dublin Core elements in Dspace. The second is creating a custom schema in Dspace identical to DUO metadata elements. The analysis of the two choices in 4.1.1.4 showed that each of them entails both advantages and disadvantages in the application.

To define which local elements in DUO and metadata elements of DCMES in Dspace should be used in the mapping process, the questionnaires has been given to get the opinions of informants. The results indicated that most of the metadata elements of Dublin Core should be used. Such elements in DUO as *document type, English name of document type, unit, English name of unit, category, subtitle, approved day, month and year, Norwegian language type and abstract of dissertation* were strongly recommended for the reuse. A few elements such as *first and last page of journal, status* are suggested for not being used. The remaining elements are advised to be used where appropriate.

For better understanding of the mapping process of metadata elements at schema level in the migration from DUO to Dspace, the crosswalk table 4.3 has been developed basing on the above recommendations and the semantics of metadata elements analyzed in the harmonization.

4.4.2 Challenges of metadata migration from DUO to Dspace

Some kinds of risks and conflicts in the metadata migration from DUO to Dspace have been judged by informants in questionnaires. Two risks for data of records of DUO mentioned in the migration are data loss and data distortion. In addition to risks, various forms of conflicts in metadata mapping between data elements in DUO and metadata elements of default Dublin Core in Dspace are data representation, synonym, structure of elements set and duplicated value.

The crosswalk table 4.3 also provides evidence of the above conflicts such as terminology conflicts (e.g. synonymy) and structural conflicts. Furthermore, the table discloses additional conflicts of metadata mapping in both systems such as null mapping and many-to-one mapping.

To control risks and conflicts of metadata mapping process in the migration, some recommendations were given in the questionnaires.

Firstly, a thorough planning before the migration is the most important thing. The plan includes competent staffs, cleaning and quality control of metadata, expertise of Dspace, procedures in the migration.

Secondly, a pilot migration should run firstly with sample data to check occurred problems and errors in this process. If errors were discovered, they would be fixed. All the problems in the pilot process should be studied as lessons learnt before the full migration. Test careful at every stage, by manually comparing selected single record or automated processes.

Finally, a more extensive customization for the metadata registry in Dspace should be made. For example, from the results in crosswalk table, more qualifiers should be created to existing elements of Dublin Core in Dspace.

The above findings are going to be used for finding the answers for research questions in next chapter.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

This chapter presents the usage of findings of the study to answer for research questions. Then some recommendations as well as suggestions for further research are provided.

5.1 Treatment of the research questions

The major aim of this study has been discussion of the appropriate choices for converting metadata elements in DUO to Dspace at UBO and prediction of challenges that UBO might face in this process. To achieve this purpose, two research questions have been formulated at the beginning of the study:

Research question 1: *What is the appropriate strategy to migrate metadata elements from DUO database to Dspace in light of current practices and the research available in this field?*

Research question 2: *In light of various issues experienced in previous metadata migration projects at different levels as well as issues particular to DUO, what are the challenges of metadata migration from DUO database to Dspace?*

5.1.1 What is the appropriate strategy to migrate metadata elements from DUO database to Dspace in light of current practices and the research available in this field?

In the methodology of metadata interoperability between two schemata, Chan and Zeng (2006) have proposed two methods including the crosswalk at schema level and record migration at record level. They stated that the crosswalk is a common method used in converting metadata elements between two schemata. As there might be various situations which require different degrees of mapping of schemata, two approaches have been suggested in the crosswalk. An absolute crosswalk requires the exact mapping of elements between two schemata whereas a relative crosswalk allows mapping of many elements in a source schema to at least one element of a target schema, regardless of whether the two elements are semantically equivalent or not. Hence, in the migration of metadata from richer structure schema to simpler schema, the relative crosswalk would be the suitable choice. In record migration, one schema based a record including metadata elements and their values are converted to those in another schema. This method was conducted when different projects had a need for integrating established metadata databases.

In practice, a number of projects of metadata migration have been conducted in libraries worldwide. The first is a project of migrating separate databases stored in faculties/units to Dspace at the University of Sydney IRs. The second is a crosswalking project of local metadata elements stored in Hypatia (SQL database) to Dublin Core Metadata set at the Internet Public Library (IPL), Drexel University (United States of America). The third is the migration of Federal Geographic Data Committee metadata (FGDC) into MARC21 and Dublin Core in OCLC's WorldCat at the Energy and Environmental Information Resources Centre (France). The final is a metadata repository project at National Science Digital Library (NSDL) in order to convert metadata records harvested from various collections into Dublin Core records. Those projects have been discussed in detail in section 2.3 of chapter 2.

For the migration project at UBO, from the answers of the informants in the questionnaires, the proper way to convert metadata elements from DUO to Dspace is understood as one that "*should follow relevant standards such as DC [Dublin Core*]" (#E) as well as "*keep all the metadata values in the migration*" (#T) and "*not to throw valuable data in DUO*" (#M). Basing on those recommendations from the informants, the approach for the migration should include translating metadata elements in DUO to DCMES in Dspace and keeping important elements of DUO in the migration.

To translate metadata elements in DUO to new database in Dspace, two strategies of migration were proposed in the questionnaires. The first is mapping DUO data elements to default Dublin Core elements into Dspace and creating new qualifiers for Dublin Core elements in Dspace. The second is developing a custom schema in Dspace identical to DUO metadata elements.

By the first strategy, DUO metadata elements are transferred to Dspace as default Dublin Core elements and remaining elements are mapped to new Dublin Core qualifiers. This strategy allows metadata elements in DUO to be converted to standard metadata as Dublin Core Metadata Set so that the interoperability of new form of DUO with other institutional repositories in Norway as well as OAI-PMH services among them are supported. The granularity of the original records in DUO is also retained in the migration by this strategy (see more in section 2.3, chapter 2). Nevertheless, there is concern about data representation in Dublin Core records because the original records in DUO have richer

structure than Dublin Core based records in Dspace. As one informant said "*The Dspace solution is to put everything in a single field and that is not a very good solution. You can't export and differentiate fields use different ways of citation* [e.g. Endnote, ProCite, and Reference Manager]" (#M). Indeed, the results in table 4.3 showed that many fields in DUO are forced to be mapped to one Dublin Core element and its qualifiers in Dspace. Theoretically, data of these fields will be represented in repeatable fields of Dublin Core metadata set. However, it's not quite sure that all data come into Dspace in exact representation because sometimes, data might be accumulated in one field in a Dublin Core record or data are filled in wrong fields. Otherwise, some elements of original records in DUO cannot find corresponding elements of Dublin Core in Dspace and they might be missed in the migration.

Regarding the second strategy of migration, a custom schema distinct from default Dublin Core is created in Dspace and DUO data elements are transferred to Dspace in their original forms. This strategy ensures that the original forms and values of important metadata elements in DUO can be kept in the migration to Dspace. On the other hand, if this strategy is chosen, much effort and human resource goes on configuration and ongoing maintenance of the DSpace index keys, customized metadata schemata and OAI crosswalks (see more in section 2.3, chapter 2). There is also a concern of the number of metadata elements in original records of DUO should be kept as well as the interoperability of custom schema of new DUO database with other schemata used in repositories in Norway.

From the above discussion, none of these strategies of migration perfectly meet the requirements of a good way for converting metadata elements from DUO to Dspace. The current status of DUO database and the migration project at UBO might be similar to the case at IPL project, Drexel University. The results of analytical comparison between IPL existing fields and Dublin Core Metadata Element set (Galloway, M. et al., 2009, p.1) or the harmonization table 4.2 and crosswalk table 4.3 between metadata elements in DUO and DCMES in Dspace shows that there's no direct one-to-one mapping between the two systems. Thus, from the interpretation of the expectations of the informants in the questionnaires, and the circumstance of metadata elements in DUO and Dspace as well as experiences at IPL project, the strategy of creating a custom schema might be the suitable one for metadata migration from DUO to Dspace at UBO. This way, a custom schema contains both Dublin Core elements and specific elements from the existing DUO database

is created in Dspace. Then DUO fields are crosswalked to this custom schema during the migration. The results presented in figures 4.2 and 4.3 could provide good references to decide the reuse of DUO metadata elements and the usage of Dublin Core elements in the customized schema for the migration project at UBO. The projects discussed in section 2.3 of chapter 2 might give good experiences of creating the custom schema and developing the crosswalk from one schema to another in the metadata migration. As Brownlee (2009) addressed, the strategy of creating a custom schema provides the possibility of keeping original record values in the migration and avoids Dublin Core registry management issues because of DC qualifiers proliferation. However, it requires that the skilled staffs at UBO pay much effort developing a customized schema, OAI crosswalks and costs of ongoing maintenance of Dspace index keys. In particular, the interoperability or standardization of a custom schema should be assured.

In case UBO wants to have DUO follow standard DCMES in Dspace and keep only important and selected original elements in DUO, the strategy of mapping DUO data elements to existing Dublin Core elements in Dspace and creating additional qualifiers for Dublin Core elements might be considered. As the informant #E suggested, "*it would probably extend the existing DC schema in order to maintain similar metadata support to what DUO could do today".* This way, Dublin Core elements and qualifiers should be customized in order for the important data of original records in DUO to be transferred to Dublin Core records after the migration.

5.1.2 In light of various issues experienced in previous metadata migration projects at different levels as well as issues particular to DUO, what are the challenges of metadata migration from DUO database to Dspace?

By experiences of the challenges occurring in the implementation of the strategies mentioned in section 5.1.1, the results in the questionnaires and the crosswalk table 4.3, the challenges of the migration of metadata elements from DUO to Dspace are implied in forms of risks and conflicts of metadata elements and their values.

Two risks in the migration are data loss and data distortion because the structure of DUO records is more complicated than the Dublin Core based records in Dspace. In case of mapping DUO fields to existing Dublin Core elements in Dspace, some unmapped data elements and their values might not be transferred to Dspace during the migration process

at UBO. Otherwise, values of DUO fields could be filled in the wrong place in Dspace. These issues can cause the loss of data values and data meaning in original records of DUO after they are converted to Dublin Core based records in Dspace.

In previous studies, many types of metadata conflicts in mapping among schemata were investigated. Batini and Lenzerini (1987) indicated type conflict, dependency conflict and behavioral conflict in metadata schema integration. Su (2004) categorized two types of conflicts in semantic mapping of metadata elements including terminology discrepancies and structural discrepancies. Woodley (2008) has found some misalignments occurring during data migration such as not equivalent between metadata elements, one-to-many mapping, many-to-one mapping, null mapping, inconsistency in data representation, hierarchical structure versus flat structure, etc. (see also section 2.5, chapter 2).

From the results of the questionnaires and table 4.3, various forms of conflicts in metadata mapping between fields in DUO and metadata elements of DCMES in Dspace are interpreted as *data representation*, *synonyms*, *structure of elements set*, *null mapping* and *duplicate values*, respectively.

For *data representation*, data in separate fields in DUO have to be mapped to one element of Dublin Core and its qualifiers in Dspace. The results in the crosswalk table 4.3 had provided evidence that many fields in a DUO record were mapped to one element of a Dublin Core record by semantic mapping of metadata between both systems.

For *synonyms*, different labels are used for the same concept in the descriptions of fields/elements in DUO and Dspace so that they can lead to misunderstandings of terminology in the system during the mapping process. For example: SUBTITLE, ALTTITLE (in DUO) = Title.Alternative (in DC), AUTHORLIST (in DUO) = Contributor. author (in DC) etc.

For *structure*, there is the conflict between the rich structure of fields of relational database in DUO and the simple structure of Dublin Core elements metadata set in Dspace. Therefore, many fields in DUO become qualifiers of one element of DC in Dspace.

For *null mapping*, some fields in DUO cannot find the corresponding elements in Dublin Core. A list of those fields is presented in the bottom of table 4.3.

Otherwise, some similar values existing in DUO such as file format, submission date, etc. are automatically created by Dspace. This can cause duplicate values in new database of DUO.

5.2 Recommendations

From the discussion about different choices of migration by Brownlee (2009) and the results in the questionnaires, it seems there's still no perfect strategy of migration which does not incur challenges to the library. Therefore, whatever strategy is applied in the migration of metadata elements from DUO to Dspace, control of risks and conflicts must be implemented. Some recommendations for preparation of the migration at UBO are presented below based on suggestions of the informants in the questionnaires and the previous studies in this field.

Firstly, a thorough planning before the migration is the most important thing. The plan should include competent staffs, cleaning of unused/redundant fields in existing DUO database, gaining expertise of Dspace and procedures in the migration process.

To have competent staffs included in the migration project, it's significant to establish collaboration among librarians and technical staffs at UBO with experts at USIT and consultants from other institutions in Norway and worldwide. For instance, the experienced experts from the migration project at the University of Sydney IRs (Australia) and the crosswalking project at the Internet Public Library (IPL), Drexel University (United States of America) could provide good advices to project members at UBO. The collaboration process will exploit best ideas from the experts to solve different issues occurring in the DUO migration. Therefore, UBO can minimize many kinds of errors and risks in the real migration. In fact, the preparations for DUO migration at UBO are being operated in this direction. From the answers of the informants and information about the project published on UBO's webpage, many groups consisting of project managers, metadata librarians, technicians group and reference group have been established to handle different packages in the project.

Since DUO database has various kinds of fields and some fields are no longer used, those unused fields, administrative fields and other unnecessary fields should be listed in the plan to remove before the migration. In addition to DUO database preparation, the usage of Dspace requires expertise to have more customization on it in order to meet specific needs from the library because Dspace is open source software.

Lastly, the procedures to operate the migration should be carefully discussed in the plan to control risks and conflicts in the real migration. By the suggestion in section 4.1.3.2, chapter 4 of the informants, those procedures for DUO migration could include selecting DUO records sample, testing the migration on small sample, discovering occurring problems, fixing those problems, testing and fixing on the larger sample and finally performing the migration on real collection.

Secondly, a control mechanism of metadata quality during the migration to new database in Dspace should be established. "*Quality control involves testing, checking and sampling of records to ensure adherence to quality objectives and to determine where, when and how quality failures occur*" (Zeng and Qin 2008, p.263). As discussed in section 2.1.2 of chapter 2, the most common criteria for quality of metadata in institutional repositories are completeness, accuracy and consistency. Hence, the control mechanism could use those criteria for metadata quality evaluation and define procedures to check the quality of metadata elements and their values in DUO records during the migration.

Thirdly, a pilot migration should be run firstly with sample data to check problems and errors occurring in this process. If errors were discovered, they would be fixed before the full migration. The check should be based on comparison of a pair of records in the source and record in the target. Checking list could include number of elements transferred, correct mapping of elements and their values, exact data representation in fields, number of missed elements, types of errors and the reasons of errors, etc. Although there are tools for automatic checking process, those tools might not cover all problems in this process. Therefore, the testing at this pilot stage still needs to be controlled carefully by the experts.

Finally, more customization for the metadata registry in Dspace should be made in order to create the correspondence between metadata elements in both DUO and Dspace in the migration. For instance, from the results in crosswalk table 4.3, more qualifiers should be added to existing elements of Dublin Core in Dspace if the strategy of mapping DUO data elements to default Dublin Core in Dspace is in use. If the strategy of creating a custom schema in Dspace is implemented, separate namespaces would be defined and those namespace are customized in application profile.

5.3 Further research

This study aims to find a suitable strategy to migrate metadata elements in DUO to Dspace at schema level. Hence, studies on different aspects of the migration from DUO to Dspace would be welcome.

Firstly, the preparation at UBO for DUO migration to Dspace should be further studied because the thesis has been conducted in an early stage of the project. The results of the study are supposed to show important aspects of the preparation such as the process of decision-making for DUO migration strategy; skilled staffs, experts involved and their roles in the project; the plan of migration with specific procedures; the control mechanism during the converting process, etc. By reviewing previous studies in this field, it seems that there is almost no paper about preparations for migration projects. Hence, it's necessary to conduct further studies about the preparation/planning in other migration projects. Those investigations will provide best practices for similar projects in future.

Secondly, it is significant to evaluate the quality of metadata elements of new DUO in Dspace. Such a study should focus on measuring completeness, accuracy and consistency of metadata elements and their values in new DUO records to check whether previous DUO records are transferred correctly and completely during the migration. Otherwise, the performance of new DUO in Dspace after the migration is the important thing in need of exploration. This study might consider whether a new version of DUO works well by testing functionality and services offered by the previous version of DUO such as publishing, searching, downloading, exporting records in various citation formats and extended services in Dspace like sharing and harvesting records via OAI-PMH with other repositories in Norway.

Thirdly, the implementation of crosswalk, metadata schema customization and other methods in the real migration process, and the outcomes should be further investigated to provide best practices for other projects in future because such studies are too few in the research available in this field. The errors occurring in other migration projects and the solutions to deal with them should be explored more systematically as well.

Finally, findings in the thesis could be a useful reference for DUO migration project and similar projects in which libraries/institutions plan to convert home-grown metadata based on local databases to Dspace or metadata standard based on other systems.

Discussing different choices for metadata migration and identifying various issues related to risks and conflicts of metadata elements in the migration, the thesis might be used in the stage of decision-making for such future projects. Otherwise, the issues of the crosswalk from home-grown metadata elements to DCMES might provide evidence for other studies in this field.

REFERENCES

Arms, et al. (2002). A case study in metadata harvesting: NSDL, *Library Hi Tech*, 21 (2), pp.228 – 237.

Batini, C. &Lenzerini, M.(1984). A methodology for data schema integration in the entity relationship model, *IEEE Transaction Software in English*, 10(8), pp. 330-350.

Blanchi, C. &Petrone, J. (2001). *Distributed interoperable metadata registry, D-Lib Magazine,* 7(12). Retrieved on March 14th, 2011 from http://www.dlib.org/dlib/december01/blanchi/12blanchi.html

Bountouri, L. &Gergatsoulis, M.(2009). Interoperability between archival and bibliographic metadata: an EAD to MODS crosswalk, *Journal of Library Metadata*, 9(1), 98-133.

Brownlee, Rowan (2009). Research data and repository metadata: Policy and technical Issues at the University of Sydney Library, *Cataloging & Classification Quarterly*, 47(3), 370 — 379.

Bruce, T.R. and Hillmann, D.I. (2004). The continuum of metadata quality: defining, expressing, exploiting, Cited in Weagley, J., Gelches, E. & Park, J. (2010). *Interoperability and metadata quality in digital video repositories: A study of Dublin Core, Journal of Library Metadata*, 10(1), 37-57.

Chan, L.M. & Zeng, M.L.(2006). *Metadata interoperability and standardization - a study of methodology part I: Achieving interoperability at the schema level,* D-LIB Magazine, 12(6). Retrieved on March 14th, 2011 from http://www.dlib.org/dlib/june06/chan/06chan.html

Chan, L.M. & Zeng, M.L.(2006). *Metadata interoperability and standardization - a study of methodology part II: Achieving interoperability at the schema level*, D-LIB Magazine,

12(6). Retrieved on March 14th, 2011 from http://www.dlib.org/dlib/june06/zeng/06zeng.html

Chandler, A., Foley, D. & Hafez, A.M.(2000). *Mapping and Converting Essential Federal Geographic Data Committee (FGDC) Metadata into MARC21 and Dublin Core*, D-LIB Magazine, 6(1). Retrieved on March 22th, 2011 from http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january00/chandler/01chandler.html

Chapman,J.W., Reynolds,D. & Shreeves,L.R. (2009). Repository metadata: approaches and challenges, *Cataloguing and Classification Quarterly*, 47(3), 309-325.

Dublin Core Metadata Initiative (2007). *Using Dublin Core – Dublin Core qualifiers*. Retrieved on January 14th, 2011 from

http://dublincore.org/documents/usageguide/qualifiers.shtml

Galloway, M. et al.(2009). *Crosswalking IPL metadata to Dublin Core*. Retrieved on March 14th, 2011 from https://www.ideals.illinois.edu/bitstream/handle/2142/15289/khoo.pdf?sequence=2

Hakimpour, F. & Geppert, A.(2001). Resolving semantic heterogeneity in schema integration: an ontology based approach, *FOIS '01 Proceedings of the international conference on Formal Ontology in Information Systems*, 297-308.

Hewitt-Taylor, J. (2001). Use of constant comparative analysis in qualitative research, *Nursing standard*, 15 (42), 39-42.

Jackson, et al. (2008). Dublin Core Metadata Harvested Through OAI-PMH, *Journal of Library Metadata*, Volume 8 (1), p. 5-21

Khoo, M. & Hall, C.(2010). Merging metadata: a socio-technical study of crosswalking and interoperability, *JCDL*, 361-364.

Lourdi, I., Papatheodorou, C., & Nikolaidou, M.(2007). A multi-layer metadata schema for digital folklore collections, *Journal of Information Science*, 33(2), 197-213.

Lubas, R. L. (2009). Defining best practices in electronic thesis and dissertation metadata, *Journal of Library Metadata*, 9 (3), 252-263.

Lynch, Clifford A. (2009). *The institutional repository in 2010 ... and beyond Institutional, Paper submitted to the Online Information Conference 2009*. Retrieved on March 14th, 2011 from http://atmire.com/presentations/OnlineInformation2009/Repositories-in-2010.pdf

Masood, N. & Eaglestone, B.(2003). Component and federation concept models in a federated database system**,** *Malaysian Journal of Computer Science,* 16(2), 47-57

NISO (2004). *Understanding metadata*. Retrieved on March 14th, 2011 from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

Park, J.R. & Tosaka, Y.(2010). Metadata creation practices in digital repositories and collections: schemata, selection criteria, and interoperability, *Information Technology and Libraries*, 104-116.

Park, J.R. (2009). Metadata quality in digital repositories: a survey of the current state of the Art, *Cataloguing & Classification Quarterly,* 47(3), 213-228.

Park, J.R. (2005). *Semantic interoperability across digital image collections: a pilot study on metadata mapping*, Paper presented at the Canadian Association for Information Science (CAIS) 2005 Annual conference. Retrieved on April 6th, 2011 from http://idea.library.drexel.edu/handle/1860/928

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

Pierre, M. &LaPlant, W.P. (1998). *Issues in crosswalking content metadata standards*. Retrieved on March 14th, 2011 from http://www.niso.org/publications/white_papers/crosswalk/

Su, Xiomeng (2004). *Semantic enrichment for ontology matching: Doctoral thesis, Department of Computer and Information Science*, Norwegian University of Science and Technology, 219p.

Shreeves, S. L. et al. (2005). *Is "quality" metadata "shareable" metadata? The implications of local metadata practices for federated collections*. Retrieved on March 14th, 2011 from http://mailer.fsu.edu/~bstvilia/papers/ACRL.pdf

Strauss, A. & Corbin, J. (1990). *Basics of qualitative research: ground theory, procedures and techniques.* California: Sage Publications, Inc.

Table descriptions of DUO (originally published in Norwegian language) (2007), University Center for Information Technology.

Tennant, R. (2001). Different paths to interoperability, *Library Journal*, 126(3), 118-119.

Toy-Smith, V. (2010). UALC best practices metadata guidelines: a consortia Approach, *Journal of Library Metadata*, 10 (1), 1-12.

Vullo, Innocenti & Ross (2010). Interoperability for digital repositories: towards a policy and quality framework, *OR2010 – 5th International Conference on Open Repositories*, 1-12.

Woodley, M. S. (2008). Crosswalks, metadata harvesting, federated searching, metasearching: using metadata to connect users and information. In Baca, M.(ed.), *Introduction to metadata*. Getty Research Institute, Los Angeles, online 3.0 ed.

Yen, Bui and Jung-ran Park. (2006). An Assessment of Metadata Quality: A Case Study of the National Science Digital Library Metadata Repository. In CAIS/ACSI 2006 *Information Science Revisited: Approaches to Innovation*, HaidarMoukdad (ed.). Proceedings of the 2006 annual conference of the Canadian Association for Information Science. Retrieved on March 14th, 2011 from http://www.caisacsi.ca/proceedings/2006/bui_2006.pdf.

Zeng, M.L., Lee, J. & Hayes, A.F.(2009). Metadata decisions for digital libraries: a survey report, *Journal of Library Metadata*, 9 (3), 173-193.

Zeng, M.L. & Qin, J. (2008). *Metadata*. London: Facet Publishing.

Zeng, M.L. & Xiao, L. (2001). Mapping metadata elements of different format. *Proceedings of the E-libraries* (pp. 91-99). New York: Information Today, Inc.

APPENDICES

APPENDIX 1: TABLES DESCRIPTIONS OF DUO (University of Oslo Library)

(Originally published in Norwegian in 2007 by University Center for Information Technology)

BIB_WORK table

Column name	Data type	Commentary
AUTHORLIST	VARCHAR2	List of authors, separated by #
BIBSYSID	VARCHAR2	Link to Bibsys's object id
DOCUMENT TYPE	VARCHAR2	Value retrieved from Document type table
ISEDITED	NUMBER	If value is 1 means for editing, otherwise null
ORGID	NUMBER	Link to study units which may be department or faculty
URN	VARCHAR2	Taken from national library
WORKID	NUMBER	Station identifier which ties all the surrounding tables, coupled to a sequence
CLAUSE	NUMBER	Restricted or not
YEARAPPROVED	NUMBER	Year in which the document is published
MONTHAPPROVED	NUMBER	Month in which the document is approved
CREATION DATE	DATE	Date in which the document was created
FIRSTPUBLISHED	DATE	First time the document was published
LASTPUBLISHED	DATE	Last time the document was published
LAST EDITOR	VARCHAR2	Last person who has changed the submission
TUTOR	VARCHAR2	Supervisor
BIBSYSID	VARCHAR2	Link to Bibsys object ID
FILESTOCOPY	NUMBER	Internal flag indicates whether to copy files from server to archive
HTMLINCLUDE	VARCHAR2	URL HTML embed in the page
ISHTMLINCLUDE	NUMBER	Flag to say about HTML to integrate
SORTAUTHOR	VARCHAR2	The field is used to write the author name in a way that it can be sorted properly.
VIDEOURL	VARCHAR2	URL of the video
CLASS CONNECTED	NUMBER	Pointer for ID in BIB_CLASSES
APPEND	CLOB	Used for different markup
YEAROFBIRTH	NUMBER	The birth year of author
INBIBSYS	NUMBER	Specify whether the object is registered in Bibsys or not
TITLEPAGEAUTHOR	VARCHAR2	
LOAD COUNT	NUMBER	Number of times that document was downloaded
MAGTITLE	VARCHAR2	The title of journal
MAGYEAR	NUMBER	The published year of journal
MAGVOLUME	VARCHAR2	The periodical volume
MAGPART	VARCHAR2	The journal number
MAGFIRSTPAGE	NUMBER	The home page of journal
MAGLASTPAGE	NUMBER	Last page of journal
FRIDAID	NUMBER	Frida ID
FROM_04_TO_07	NUMBER	Download period

Referee	NUMBER	Specify if the document is refereed

BIB_LANGDESCR table

Column name	Data type	Commentary
ID	NUMBER	Coupled to a sequence
KEYWORDS	VARCHAR2	Free keywords
LangId	VARCHAR2	ISO 6392 code for language
SUBTITLE	VARCHAR2	Under title of document
TITLE	VARCHAR2	Title of document
WORKID	NUMBER	Link to BIB_WORK
ISBN	VARCHAR2	
BLACHTITLE	VARCHAR2	Option to sort title in different way
ABSTRACT	CLOB	Summary
ALTTITLE	VARCHAR2	Title in second language
ALTSUBTITLE	VANCHAR2	Subtitle in second language
ALTKEYWORDS	VARCHAR2	Free keywords in second language

BIB_ORGUNIT table

Column name	Data type	Commentary
ORGID	NUMBER	ID unit
ORGNAME	VARCHAR2	Name of unit
EMAIL	VARCHAR2	Unit email address
URLPATH	VARCHAR2	Specify the path to file
ISUSED	NUMBER	No longer used
CLASSIFICATION PAGE	VARCHAR2	No longer used
PARENT	NUMBER	Parent ID
ORGTYPE	VARCHAR2	Specify the type of unit (faculty, institute,…)
MULTI LANGUAGE	NUMBER	Specify whether the submission can put the proposed title, etc in more than one language
PUBLISHCOUNT	NUMBER	Specify how many documents are published on …
NORWEGIAN DISPLAY	VARCHAR2	Norwegian name that appears in the interface
ENGLISH DISPLAY	VARCHAR2	English name that appears in the interface
UNIT CODE	VARCHAR2	Unit code
SCIENCE	VARCHAR2	The science discipline

BIB_XMLMETADATA table

Column name	Data type	Commentary
ID	NUMBER	Linked to the sequence
WORKID	NUMBER	Linked to BIB_WORK
YEAR	NUMBER	Year
Faculty	VARCHAR2	Name of faculty
INSTITUTE	VARCHAR2	Name of any institute
SUBJECT	VARCHAR2	Name of any profession

XML TEXT	LONG	Xml stream with metadata

BIB_INSTANCE table

Column name	Data type	Commentary
FilePath	VARCHAR2	URL for the full text document
INSTDESCR	VARCHAR2	Attach a brief description of the file, which comes up on title page (such as it is a corrected version)
INSTFORMAT	VARCHAR2	PDF or HTML
InstID	NUMBER	Sequence controlled counter
LangId	VARCHAR2	Language code – not applicable
WORKID	NUMBER	Link to BIB_WORK
REPROPRINT	NUMBER	Flag indicates that the document is printed on repro
CHECKSUM	VARCHAR2	MD5 checksum is generated when link is established and the document is copied to the archive

BIB_CLASSIFICATION table

Column name	Data type	Commentary
CLID	NUMBER	Sequence-driven ID
CLTYPE	NUMBER	Specify classification schema
CLVALUE	VARCHAR2	Classification code
WORKID	NUMBER	Linked to BIB_WORK

ASSOCIATION TYPE table

Column name	Data type	Commentary
ASSOCID	NUMBER	Identifier
TEXTFROM	VARCHAR2	Part of series
TEXTTO	VARCHAR2	The series holding/contains
EXPLANATION	VARCHAR2	Description of the association in the case of
TEXTFROMENGLISH	VARCHAR2	English translation
TEXTTOENGLISH	VARCHAR2	English translation

Works Association table

Column name	Data type	Commentary
CONTENT	CLOB	For series of booklets
ASSOCID	NUMBER	Link to association type, describe the type of relationship they are
ID	NUMBER	Sequence controlled id
SINKID	NUMBER	Workid objective
BLACK CODE	VARCHAR2	Sort code is used to sort series of booklets by series title
Sourceid	NUMBER	Workid for source

BIB_CLASSES table

Column name	Data type	Commentary
VARIETY NAME	VARCHAR2	Used to manage order coal
CLASS NAME	VARCHAR2	Name of coal
ORGID	NUMBER	Link to studies unit
ID	NUMBER	Identifier

BIB_ACTUAL USERS table

The data model

Column name	Data type	Commentary
WORKID	NUMBER	Linked to BIB_WORK
LOGIN NAME	VARCHAR2	Userid to the student
ID	NUMBER	Sequence controlled id

BIB_EDITOR table

Column name	Data type	Commentary
USERNAME	VARCHAR2	Userid to user
ID	NUMBER	Identifier here is no sequence
ORGID	NUMBER	Studies unit linked to BIB_ORGUNIT
UNIT	VARCHAR2	User role

BIB_LANGUAGE table

Column name	Data type	Commentary
FREQUENTLY USED	NUMBER	Help user choose between all sorts of language
ID	NUMBER	Identifier
ENGNAME	VARCHAR2	English name of language
LONG CODE	VARCHAR2	ISO 6392 letters code
NORNAME	VARCHAR2	Norwegian name of language
OPTIONAL	VARCHAR2	Not used, identical to the long code
TWOLETTER	VARCHAR2	Two letter code of ISO 6392

BIB_LOGTEXTTABLE table

Column name	Data type	Commentary
DEFAULTTEXT	VARCHAR2	The text is inserted into the log for a specific here
ID	NUMBER	Id link for

BIB_LOGTABLE table

Column name	Data type	Commentary
DEFAULTTEXTID	NUMBER	Link for id in BIB_LOGTEXTTABLE
EDITOR	VARCHAR2	Name of administrator who made the incident
LOGDATE	DATE	Time
LOGID	NUMBER	Sequence controlled id
LOGTEXT	VARCHAR2	Opportunity to comment on here
UserID	VARCHAR2	Userid to the administrator
WORKID	NUMBER	Link to work

DOCUMENT TYPE table

Column name	Data type	Commentary
OAI	VARCHAR2	Type name is defined to map OAI harvesting
ENGNAME	VARCHAR2	English name for document type
ENGTITLE	VARCHAR2	Title in English with document type
NORNAME	VARCHAR2	Norwegian name for document type
NORTITLE	VARCHAR2	Title in Norwegian with document type

SCIENCE table

Column name	Data type	Commentary
CODE	VARCHAR2	The code to use in the classification
NAME_NORWEGIAN	VARCHAR2	Norwegian name
NAME_ENGLISH	VARCHAR2	English name
CODE_LEVEL	NUMBER	Come from Frida
OWNER	VARCHAR2	Parent node – the top level

APPENDIX 2: DEFAULT DUBLIN CORE METADATA REGISTRY IN DSPACE (ver.1.5.2)

Retrieved on April 25th, 2011 from:
http://www.dspace.org/1_5_2Documentation/ch15.html#docbook-appendix.html-dublincoreregistry

Element	Qualifier	Scope Note
contributor		A person, organization, or service responsible for the content of the resource. Catch-all for unspecified contributors.
contributor	Advisor	Use primarily for thesis advisor.
contributor[1]	Author	
contributor	Editor	
contributor	illustrator	
contributor	Other	
coverage	Spatial	Spatial characteristics of content.
coverage	temporal	Temporal characteristics of content.
creator		Do not use; only for harvested metadata.
date		Use qualified form if possible.
date[1]	accessioned	Date DSpace takes possession of item.
date[1]	available	Date or date range item became available to the public.
date	copyright	Date of copyright.
date	Created	Date of creation or manufacture of intellectual content if different from date.issued.
date[1]	Issued	Date of publication or distribution.
date	submitted	Recommend for theses/dissertations.
identifier		Catch-all for unambiguous identifiers not defined by qualified form; use identifier.other for a known identifier common to a local collection instead of unqualified form.
identifier[1]	Citation	Human-readable, standard bibliographic citation of non-DSpace format of this item
identifier[1]	Govdoc	A government document number
identifier[1]	Isbn	International Standard Book Number
identifier[1]	Issn	International Standard Serial Number
identifier	Sici	Serial Item and Contribution Identifier
identifier[1]	Ismn	International Standard Music Number
identifier[1]	Other	A known identifier type common to a local collection.
identifier[1]	Uri	Uniform Resource Identifier
description[1]		Catch-all for any description not defined by qualifiers.

description[1]	Abstract	Abstract or summary.
description[1]	provenance	The history of custody of the item since its creation, including any changes successive custodians made to it.
description[1]	sponsorship	Information about sponsoring agencies, individuals, or contractual arrangements for the item.
description	statementofresponsibility	To preserve statement of responsibility from MARC records.
description	tableofcontents	A table of contents for a given item.
description	Uri	Uniform Resource Identifier pointing to description of this item.
format[1]		Catch-all for any format information not defined by qualifiers.
format[1]	Extent	Size or duration.
format	Medium	Physical medium.
format[1]	mimetype	Registered MIME type identifiers.
language		Catch-all for non-ISO forms of the language of the item, accommodating harvested values.
language[1]	Iso	Current ISO standard for language of intellectual content, including country codes (e.g. "en_US").
publisher[1]		Entity responsible for publication, distribution, or imprint.
relation		Catch-all for references to other related items.
relation	isformatof	References additional physical form.
relation	Ispartof	References physically or logically containing item.
relation[1]	ispartofseries	Series name and number within that series, if available.
relation	Haspart	References physically or logically contained item.
relation	isversionof	References earlier version.
relation	hasversion	References later version.
relation	isbasedon	References source.
relation	isreferencedby	Pointed to by referenced resource.
relation	requires	Referenced resource is required to support function, delivery, or coherence of item.
relation	replaces	References preceding item.
relation	isreplacedby	References succeeding item.
relation	Uri	References Uniform Resource Identifier for related item.
rights		Terms governing use and reproduction.
rights	Uri	References terms governing use and reproduction.
source		Do not use; only for harvested metadata.
source	Uri	Do not use; only for harvested metadata.

subject[1]		Uncontrolled index term.
subject	classification	Catch-all for value from local classification system. Global classification systems will receive specific qualifier
subject	Ddc	Dewey Decimal Classification Number
subject	Lcc	Library of Congress Classification Number
subject	Lcsh	Library of Congress Subject Headings
subject	Mesh	MEdical Subject Headings
subject	Other	Local controlled vocabulary; global vocabularies will receive specific qualifier.
title[1]		Title statement/title proper.
title[1]	alternative	Varying (or substitute) form of title proper appearing in item, e.g. abbreviation or translation
type[1]		Nature or genre of content.

[1]Used by system: do not remove

APPENDIX 3: DUBLIN CORE METADATA INITIATIVE - DUBLIN CORE QUALIFIERS

(Approved in 2007 by the Dublin Core Usage Board)

Retrieved on May 6th, 2011 from: http://dublincore.org/documents/usageguide/qualifiers.shtml

DCMES Element	Element Refinement(s)	Element Encoding Scheme(s)
Title	Alternative	-
Creator	-	-
Subject	-	LCSH MeSH DDC LCC UDC
Description	Table Of Contents Abstract	-
Publisher	-	-
Contributor	-	-
Date	Created Valid Available Issued Modified Date Accepted Date Copyrighted Date Submitted	DCMI Period W3C-DTF
Type	-	DCMI Type Vocabulary
Format	-	IMT
	Extent	-
	Medium	-
Identifier	-	URI
	Bibliographic Citation	-
Source	-	URI
Language	-	ISO 639-2RFC 3066
Relation	Is Version Of Has Version Is Replaced By Replaces	URI

	Is Required By Requires Is Part Of Has Part Is Referenced By References Is Format Of Has Format Conforms To	
Coverage	Spatial	DCMI Point ISO 3166 DCMI Box TGN
	Temporal	DCMI Period W3C-DTF
Rights	Access Rights	-
	License	URI
Audience	Mediator Education Level	-
Provenance	-	-
Rights Holder	-	-
Instructional Method	-	-
Accrual Method	-	-
Accrual Periodicity	-	-
Accrual Policy	-	-

APPENDIX 4: THE INTRODUCTION LETTER

Dear Sir/Madam,

My name is Van Chau Do, Vietnamese student. I am studying Master program in Digital library learning (DILL) at Oslo University College. I have had an internship at University of Oslo Library (UBO) since November, 2010. During that time, I am interested in the project of migration DUO database to Dspace. I found that current DUO database is using structure of data elements which are quite different to Qualified Dublin Core Metadata integrated in Dspace. Therefore, I decide to write the thesis titled "Define metadata migration at schema level from DUO database to Dspace at University of Oslo Library".

My thesis aims to identify a strategy to map data elements in DUO database to Dublin Core standard in Dspace prior to the migration. Conflicts of metadata elements in the migration will also be discussed to find the possible ways to control them. To achieve these aims, I would like to kindly survey by questionnaire the ideas from UBO librarians who are involved in DUO project.

I will send the online questionnaire to you in next few days. I would greatly appreciate if you can spend few minutes to provide the answers for the questions. I hope that my thesis will contribute to the migration project at your institution.

Best regards,

Van Chau Do

APPENDIX 5: THE ONLINE QUESTIONNAIRE

Dear Sir/Madam,

I would greatly appreciate if you can spend time to provide the answers for following questions. Your responses are used only in this master's thesis. Please stick (for making a choice) or fill information (for blank box). Note: It is fine not to answer all questions.

I hope that my thesis will contribute to the migration project at your institution.

Thank you very much for your help!

STRATEGY FOR METADATA MIGRATION

1. To what extent should metadata elements of DUO records be kept in migration?

- ○ Keep the original metadata elements intact
- ○ Only important local elements
- ○ Completely change to Dublin Core elements

Please specify other ideas and explain more for your choice

2. By your opinion: What are the most important reasons/motivations for migrating DUO database to Dspace?

3. Why was Dspace chosen for DUO migration?

4. Which factors influence the selection of strategy for migrating DUO to Dspace?

	Most important	Important	Least important	Not important
Interoperability with other institutions	○	○	○	○
Preservation	○	○	○	○
Maintenance cost	○	○	○	○

Please specify other factors and explain more your choice

5. By your opinion, what is the best possible strategy for migrating data elements from DUO database to default Dublin Core in Dspace?

- ◉ Map DUO data elements to qualified Dublin Core (DC) elements in Dspace. (Explain: DUO data elements are transferred to Dspace as qualified DC elements)

- Map DUO data elements to unqualified DC elements in Dspace
- Create new qualifiers for default DC elements in Dspace. (Explain: DUO data elements are transferred to Dspace as default DC elements and remained elements is mapped to new DC qualifiers)
- Create a custom schema in Dspace identical to DUO metadata elements. (Explain: DUO data elements are transferred to Dspace in their original forms)

Please specify another choice and explain reasons for your choice

6. What do you think of using additional metadata schema in Dspace (in addition to default Dublin Core) to map with DUO data elements in migration?

METADATA MIGRATION FROM DUO TO DSPACE

7. Which metadata elements of qualified Dublin Core in Dspace will the library use?

	Definitely use	Maybe use	Won't use
Title			
Creator/Author			
Contributor/Co-author			
Description/Abstract			
Subjects/Keywords			
Publisher			
Date			
Type (image, sound, text...)			
Format (physical/digital form of object)			
Language			
Source (where content is derived)			
Identifier (URL, ISBN, DOI,...)			
Relation (part/version of)			
Coverage (spatial/temporal topic in object)			
Rights (license)			

Please specify other ideas or explain more your choice

8. By your opinion: what is the best way to configure metadata in Dspace to fit with data elements in DUO?

- Create new qualifiers for default Dublin Core metadata set
- Using additional metadata schemes and create application profile

Please specify other ways or explain more your answer

9. Which elements in the current DUO database should be reused or extended in Dspace (in addition to default Dublin Core elements)?

	Definitely use	Maybe use	Won't use
Year of birth (author)			
Document type			
English name for document type			
Norwegian name for document type			
Subtitle			
Title in second language			
Keyword in second language			
Degree			
Approved day, month and year			
First/last published day			
Norwegian language type			
Unit (faculty/department/subject)			
Norwegian/English name of unit			
Supervisor/mentor/tutor			
Notes of object			
Abstract of dissertation			
Category (of research paper)			
Status			
Parts in periodical series/research work			
English translation of these parts			
First and last page of journal			

Please specify other elements or explain more your choice

CONFLICTS/RISKS IN METADATA MIGRATION FROM DUO TO DSPACE

10. In your ideas, what are possible risks/conflicts in metadata migration from DUO database to DSpace?

- ☐ Data loss: metadata values can be lost in migration
- ☐ Data distortion: Contextual meaning of data is lost
- ☐ No correspondence of metadata elements between two systems. (For example: year approved, month approved, advisor, degree, etc. in DUO)
- ☐ Synonym: different terminologies for the same value. (For example: Date (Dublin Core) =CREATION DATE (DUO), Description (DC) = Abstract (DUO), Subject (DC) = Keyword (DUO))
- ☐ Homonym: same terminology but different meanings. (For example: document type, subject, etc. in DUO)
- ☐ Homonym: same terminology but different meanings. (For example: document type, subject, etc. in DUO)
- ☐ Different representation: Data in separated fields in DUO may be in a single element of DC in the Dspace. (For example: moth approved, year approved, first published, last published, creation date (DUO) = date (DC))
- ☐ Language barrier because default language in Dspace is English.
- ☐ The complicated structure of elements set in DUO database and flat structure of Dublin Core in Dspace
- ☐ The duplicated value because some values are automatically created by Dspace. For example: file format, submission date, etc.

Please specify other risks/conflicts and explain more your choice

11. How do you think should these risks/conflicts be controlled?

12. How should the library prepare (planning; staff; metadata cleaning and preparation; metadata quality control mechanism; technology, etc.) for migrating DUO database to Dspace?

13. If you have more ideas/comments about my topic, please feel free to write here.

14. Please provide your contact information (The information is only used for further discussion. Your identification is kept secret).

Name:	
Position:	
Email Address:	