Developing Integrated Theses and Dissertations System and Improving University Information Infrastructure: The Korean Experience

ETD2001
The Fourth International Symposium on Electronic Theses and Dissertations

Yong-Hyo, Lee[+], Yong-Soon, Kim[*], Dae-Joon, Hwang [+]

[+]School of Electrical Engineering and Computer Engineering, Sungkyunkwan University
300 Chunchun-dong, Jangan-Ku, Suwon, 440-746, Korea
[*]Researcher of Korea Education & Research Information Service, Arirang Tower, 1467-80, Seocho-Dong, Seoch-gu, Seoul, 137-070, Korea

Abstract

In Korea, more than 45,000 master theses and doctoral dissertations (TDs) are produced every year, but it was difficult to share and reuse them among the researchers and graduate students due to lack of integrated TDs search system and accessible digitalized database. This paper provides an overview of the project developing national Digital Library of master Thesis and doctoral Dissertation (DLTD) system. This is a substantial project to enhance the value of Research Information Service System (RISS), which was designed as national digital library aimed for portal research information service system and developed in 1998 in order to support not only each university library but also all the researchers across in Korea. The DLTD system was designed and implemented as a nationally accessible web based Digital Library service system with full-text of theses and dissertations produced with member universities in Korea. The major goals in this project is to enhance universities' information infrastructure through collaborative work with universities and to increase the scholarship by letting students make use of digital library and share information through developing integrated search system with full-text of thesis and dissertations. During this first year 2000, DLTD project has grown rapidly, opening its service system with over 20 members. This paper deals with various issues associated with DLTD project and shows our experiences of developing system with more than 90 member universities; Korea's unique situations and backgrounds, status of this project, intellectual copyright issues, and related studies of DRM(Digital Rights Management).

Key words: TD (Thesis and Dissertation), Copyright, Integrated Retrieval System, Metadata, DL (Digital Library), DLTD (Digital Library with Theses and Dissertation), DRM(Digital Rights Management)

1. Introduction

The DLTD project is a national collaborative work with member universities in Korea aimed at improving the research environment through digital library with recently published research outcomes of master theses and doctoral dissertations. In this project we focused on the increasing the number of access to full text of research outcomes that each member university produce.

KERIS (Korea Education & Research Information Service), the government organization under the Ministry of Education, started an initiative with DLTD system to integrate and expand RISS (Research Information Service System) in April 1999. KERIS announced the development of DLTD in order to enhance the RISS service system including full text of student's research outcomes, or TDs. Over 90 universities out of 104 universities that produce their theses and dissertations in Korea participated in this project. A subcommittee for the project was made and all the policies and strategies for the project were set up. System development started from December 1999 and service was launched on late May 2000. It was the first collection of digital file of TD which can be shared without copyright issues. Presently service is being operated in Web page http://www.riss4u.net.

1.1 Backgrounds

Although approximately 40,000 master and 3,500 doctoral degrees are awarded every year in Korea, graduate students and scholars suffer from re-using these valuable research outcomes due to lack of integrated search system and unready to digitalized. Researchers have difficult time in searching one more sites before finding the one they want. Most are unaware of any related works recently completed.

As of April 1999, KERIS announced starting this DLTD project. Before that we surveyed the Korean situation two times in December 1998 and August 1999 to design an efficient DLTD integrated service model based on Korean situation. After announcing the DLTD project, more than 90 universities wanted to attend this project as member universities.

Among 104 universities that produce TDs in Korea half of universities(49%) were collecting digital file of TD from their students. Only 22% of universities were clearing the intellectual copyright that is critical for resource sharing. At the same time shortfalls and shortage of budget was the biggest problem in university libraries. Expenditure ratio for digitizing TDs and managing service system was so constrained. The portion of digitized TD was 28%. Less than 3% TD can be accessed by the internet only within the campus without copyright issue. When we design a DLTD system, more than 64% of member universities wanted to keep their full text TD database only in their local university system and have a right to control in the access of database from out of campus.

A number of universities, in Korea, developed their own TD service systems for their students and faculties. The electronic TD format for internet service were so diverse; TIFF(Tag Image File Format), PDF(Adobe Acrobat Portable Document Format), author edited word processor files such as DOC and HWP(Hanguel Word Processor), HTML(HyperText Markup Language), Latex, XLX, SGML(Standard Generalized Markup Language), etc. While PDF has been adopted as representative format for worldwide digital libraries of TDs[1][2], CJK fonts were not supported technically till Adobe Acrobat 4.0, 1999. That has delayed its usage and uniformity of electronic file format for internet documents exchange. This leads to diversity of document file exchange format, systems configuration and protocols and accompanies hindrance of compatibility among heterogeneous TDs service systems[3].

1.2 Goals and Aims

We put the our goals during the first year of this project changing the submission form from paper to electronic format, clearing the copyright issues, making integrated retrieval system and full-text available on web-based internet service.

The rationale behind this project is to provide national standardization regarding digitizing TD and to lead universities to form a solid foundation of information infrastructure.

The primary issue for the success of this project was to persuade universities to clear the intellectual copyright in order to share that TDs through internet. In 1998 December, 49 universities collect ETDs as word processor file format from their students but it is impossible to share with the ordinary researchers out of campus. For this project, KERIS made the copyright statement terms and conditions to solve the copyright issues.

We put the rationale behind this project as follows:

1. To build a integrated DLTD search system
2. To enhance the university information infrastructure through collecting the TD form from paper to digital form and clear the copyright issue
3. To help researcher’s effort to find adequate full text for their research activity effectively and efficiently
4. To define standardizations related to building digital library system, such as defining metadata sets.

This paper will give an overview of the DLTD system focused on establishing integrated master and doctoral dissertation service system so far and describes some of issues and processes we experienced.

2. DLTD Project

2.1 The role of each partner

There are three interested parties in this project; KERIS, universities, graduate students. The relationship among the three partners and the life cycle of TDs is shown in Fig 2.

KERIS is the center of this project. The role of KERIS consists of three fields; building a DLTD service system, operating the subcommittee in order to decide the regulation for this project, and defining standards.

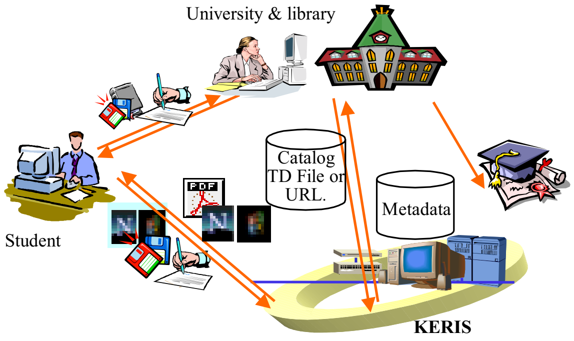

Fig 1. Life cycle of TD

Each university is responsible for collecting digital forms of TDs from their students, handling the copyright clearance, and sending TDs related information to KERIS in order to help to build DLTD service system. The copyright holder is the author. If the author does not sign the copyright statement when they submit the paper, then TD is to be shared only abstract.

Author, graduate students write their paper for graduation as requirements for getting a degree. After defense, they will submit their publication to graduate university with paper based TD and digital file of TD. The university will review and examine the digital file if there is no error or missing contents such as table, picture, or content body. If there is no problem in their submitting file, students can finally obtain a degree. After collecting TDs in the university, each library makes a catalog of TD for their library service, and proceed to build a database of TDs. KERIS collects each catalog information of TD and either digital file or address of TDs such as URL. The DLTD service is provided through web-page service and authors can access and reuse the other research outcome easily. In DLTD Web service page, the online submitting function is provided.

2.2 Strategic Issues and Standards

The success of the DLTD project depends on adopting a sensible strategy that involves regulations and standards for sharing and ensuring interoperability of participants systems.

During the software testing and modification phase, project team proposed a set of draft standards. The core regulation was defining internet document exchange format such as PDF, TD Metadata element set based on Dublin Core Metadata set, and converting table between TD Metadata and KORMARC.

Defining intellectual copyright terms and conditions

Defining languages for representing intellectual property usage terms and conditions were considered the first thing to be done for storing the digitalized TDs, providing TDs through network transmission. Since the meaning of the “copy” and “copyright law” in interpreting , copyright law deals with only a physical or material object not electronic object in Korea. It was needed to mention conversion into PDF for electronic service through internet, archiving and transforming TDs to ensure conversion of original file into another digital form.

Table 1 describes the handling information from the author, licensor, not to harm author’s intellectual copyright and protect their digital rights against abuse of research outcomes.

Categories	Information
Licensor Profile	Name, Address, Phone Number, FAX Number, Contact Email address
Agreement license Information	Copyright effective date, Copyright expiration date Title of Work, Category of Work, Publish year, University name, the number of Pages

Table 1. Copyright Statements Language

PDF file format

Most universities does not want to collect the TDs in PDF form in order to keep the flexibility when they built the TD Database system and archiving. Considerations behind this decision was that there was no guarantee that Adobe Acrobat Reader will continue to offer free and in addition, it was not yet defined as an international standard. A number of universities hesitate in choosing their internet document exchange file format as PDF.

Image format of TIFF and PDF was popular in digitalizing printed old version of TDs. For the traditional type of TD, printed hardcopies of TD was digitalized through scanning each page of TD. While in the case of digital file of TDs, universities chose various types of format because PDF did not fully support the two bytes of language such as CJK (Chinese, Japanese, Korean). This leads to practice diverse format; XML, SGML, HTML, TIFF, word processor file itself that student authored and edited when the university build their TD service system. When we defined the internet document exchange format, most considerations were user’s accessibility, international standardization, and information management. For users generality, simplest, device independence, and scalability were considered. For international standardization. Information management aspect, information format’s persistency and managing cost were also considered.

Korean metadata set for TDs

Table 2 summarized the Korean TD Metadata Element set based on Dublin Core Metadata and mapping information between these qualifiers and KORMARC. DC metadata element set consists of 15 elements. These definitions are officially known as Version 1.1. Korean TD metadata is defined as 12 elements. The definitions utilize a formal standard for the description of DC metadata elements. This formalization helps to improve consistency with other metadata communities and enhances the clarity, scope. Three elements, SOURCE, RELATION, COVERAGE, were not defined as Korean TD metadata element set.

Dublin Core Metadata Set	Definition of Korea TD Metadata Set with Qualifier	Notes	KORMARC	Usage
TITLE	DC.TITLE	Title (Korean)	245 $a	M
	DC.TITLE.PARALLEL	Title (English)	245 $x	O
	DC.TITLE.SUBTITLE	Subtitle	245 $b, 245 $c	O
CREATOR	DC.CREATOR.PERSONALNAME	Author name	100 $a, 700 $a	M
	DC.CREATOR.PERSONALNAME.ALTERNATIVE	Alternative author name	700 $a	O
	DC.CREATOR.PERSONALNAME,AFFILIATION	Organization which author belongs	502 $b	O
	DC.CREATOR.PERSONALNAME.EMAIL	Author's email address		O
	DC.CREATOR.PERSONALNAME.HOMEPAGE	Author's homepage		O
	DC.CREATOR.PERSONALNAME,MAJOR	Major or Department	502 $c	O
SUBJECT	DC.SUBJECT	Subject of TD	653 $a, 650 $a, 056 $a, 082 $a, 080 $a	O
DESCRIPTION	DC.DESCRIPTION	Keywords or phrases describing the subject or TD	510 $a	O
	DC.DESCRIPTION.ABSTRACT	Abstract	520 $b	O
PUBLISHER	DC.PUBLISHER	Entity responsible for making this work available in its present form	260 $b	M
	DC.PUBLISHER.PLACE	The place of publisher	0081503	O
	DC.PUBLISHER.ALTERNATIVE	Alternative name of publisher		O
	DC.PUBLISHER.EMAIL	Publisher's Homepage		O
	DC.PUBLISHER.HOMEPAGE	Country name		O
CONTRIBUTOR	DC.CONTRIBUTOR.PERSONALNAME	Contributor's name		O
	DC.CONTRIBUTOR.PERSONALNAME.ALTERNATIVE	Alternative name of contributor		O
	DC.CONTRIBUTOR.PERSONALNAME.EMAIL	Contributor's email address		O
	DC.CONTRIBUTOR.PERSONALNAME.AFFILIATION	Organization which contributor belongs		O
	DC.CONTRIBUTOR.PERSONALNAME.ROLE	Role of Contributor		O
	DC.CONTRIBUTOR.PERSONALNAME.HOMEPAGE	Contributor's homepage address		O
	DC.CONTRIBUTOR.CORPORATENAME	Group contributor's name		O
	DC.CONTRIBUTOR.CORPORATENAME.ALTERNATIVE	Alternative name of group contributor		O
	DC.CONTRIBUTOR.CORPORATENAME.EMAIL	Group contributor's email address		O
	DC.CONTRIBUTOR.CORPORATENAME.AFFILIATION	Organization which Group contributor belongs		O
	DC.CONTRIBUTOR.CORPORATENAME.ROLE	Role of Group contributor		O
	DC.CONTRIBUTOR.CORPORATENAME.HOMEPAGE	Group contributor's homepage address		O
DATE	DC.DATE.CREATED	Date committee approve the paper in final form	0080704	M
	DC.DATE.METADATACREATED	Date of input metadata first version		M
	DC.DATE.METADATAMODIFIED	Date of update metadata latest version		O
TYPE	DC.TYPE	Category of TD(Master or Doctoral)	5020 $a 5021 $a	M
FORMAT	DC.FORMAT	TD's data format		O
	DC.FORMAT.PAGE	TD's total page number based on printed copy	300 $a	O
IDENTIFIER	DC.IDENTIFIER	Unique ID of the TD	856 (URL)	O
SOURCE	-	-		
LANGUAGE	DC.LANGUAGE	Description language of the TD	0083503, 041 $a	O
RELATION	-	-		
COVERAGE	-	-		
RIGHT	DC.RIGHTS	Accessibility		O
	DC.RIGHTS.STARTS	Start date of accessibility		O
	DC.RIGHTS.ENDS	End date of accessibility		O

- : Not defined, M: mandatory, O: Optional

Table 2. TD Korean Metadata Set with Qualifiers and KORMARC Tag Translation

2.3 System Architecture

Fig. 2 illustrates system architecture of DLTD system. The main module of this system consists of two parts; Converter between MARC and DC and ARPA module for protect digital resources, TDs. University keeps the catalog information of each TD as MARC format. KERIS adopt Korean TD metadata set based on DC. Converter module converts KORMAC records from each university into Korean TD DC metadata. If university holds Full text information of TD, KERIS collects and stores catalog and URL of TD. If university allows KERIS to keep the full text of TD in KERIS's server, KERIS collects the catalog and URL information of TD.

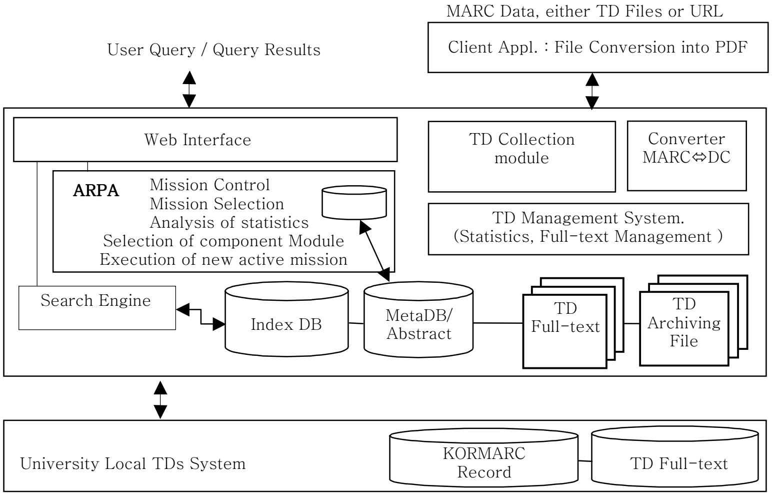
MARC Data, either TD Files or URL
Client Appl. : File Conversion into PDF
User Query / Query Results
Web Interface
TD Collection module
Converter MARC⇔DC
ARPA
Mission Control
Mission Selection
Analysis of statistics
Selection of component Module
Execution of new active mission
TD Management System. (Statistics, Full-text Management)
Search Engine
Index DB
MetaDB/ Abstract
TD Full-text
TD Archiving File
University Local TDs System
KORMARC Record
TD Full-text

Fig 2. System Architecture

Related Study: Protection scheme of digital contents based on mobile agent(ARPA)

The mechanism for intellectual property protection is different from that of the intrusion detection technology, and it has to make it possible that traces any illegal use of digital resources whether on-line or off-line and real time preventive protection. We are in the middle of implementing and examining ARPA system as protect scheme for digital contents.

In addition, it must keep and manage accumulated logging data(such as copying attempts, place, contents, time) in order to prove illegal access to digital resource in question. To protect digital resources adaptively, ARPA model is designed to meet the following requirements

- A system available under both on-line and off-line environments.
- Maintain independence from representative of any media types
- Support dynamic and adaptive protection for environment
- Real-time statistic results analysis and maintenance that are based on the Internet and the Intranet.
- Real-time management.
- Proactive protection over digital resources
- Support consistent application on active and passive resources

To achieve the above goals, the main features of the ARPA functions are as follows.

- Mission Control: Controlling of protected resources and conditions for the protection of files, directories, memories, ports, processes, threads, and etc when the system is either on-line or off-line.
- Mission Selection: Support individual selection of types of protected resources and conditions of protection
- Analysis of statistics: Collection and analysis of results transferred from agents
- Decision of types of executed job: Decision of types of job which is to be done at the system to protect resources
- Selection of component module: Selection of component module needed to job that is to be done at the system
- Execution of new active mission: Execution of new active mission based on results

3. Conclusion

3.1 Difficulties and Success

Several problems and obstacles in this project were found: a lack of standards, digital file collection and verification, unstable service cause from the low performance of each local university. The lack of standard of cataloging and full-text processing in digital library was the most urgent problem to work out. The variety of internet document format requires a middleware for handling different file format. The contents of submitted TD file are fully depends on students. Table and picture were missed in some TDs. Two-thirds member universities in this project deposit and archive TD only in their own full-text server. Therefore service performance cannot be controlled. Responsive service is up to local server.

Evaluations for this project could be measured through various ways and sophisticated manners. In this paper we provide the data of what percentage of universities are attending this project, how many TDs were gathered, usage statistics and what percentage university they follow the new regulations and standards since opening of our first service system. According to the survey in the middle of this project in November 1999, 45 universities replied that they hope to participate in this project. In copyright terms and conditions, 15 universities adopt the copyright terms and conditions proposed in this project, while 17 universities adopt some forms of amendments to the copyright terms and conditions proposed in this project. The rest of 10 universities will continue to use their own style of copyright terms and conditions. This means 32 universities could open and share their resources without any copyright issues. Overall 40.3%, or 42 universities will treat the copyright issues. Compared to the beginning of this project, 18.2%, 19 universities are increased. 34 out of 45 universities answered that standards should be adopted such as TOC DTD. 24 universities have a plan to investigate their student's TD file when they submit the file but the 21 universities couldn't not afford to do this due to lack of manpower in the meantime.

As of July 2000, 19.2% universities who produce the TD, 20 universities out of 104 universities provided 17,939 TDs for this project and these can be accessed through web without copyright issue. With the help of 20 member universities in Korea, DLTD system could provide around 18,000 TDs through the internet. Every month, more than 20,000 access of search page were recorded.

There are many issues those are not canvassed in this paper. More work and research will need to be done on issues such as improving service performance and the long term archiving of electronic theses. Progress has been rapid and reinforced by a number of member university libraries' efforts.

References

[1] Edward A. Fox, John L. Eaton, Gail McMillan, National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources,' D-Lib Magazine, September 1996.

[2] Tony Cargnelutti, Fred Piper, Karen Kealy. The Australian Digital Theses (ADT) Pilot Project: the trials, tribulations and (some) successes, EDUCAUSE in Sydney, April 1999.

[3] James Powell, Edward A. Fox, Multilingual Federated Searching Across Heterogeneous Collections, D-Lib Magazine, September 1998.