Linked Data for Electronic Theses and Dissertations of the Thai Academic Digital Collection

Yuttana Jareonruen
Doctor of Philosophy program in Information Studies,
Khon Kaen University, Thailand
tinybronko@gmail.com

Supervisor: Kulthida Tuamsuk
Information and Communication Department,
Faculty of Humanities and Social Sciences,
Khon Kaen University, Thailand
kultua@kku.ac.th

ABSTRACT

This paper presents the research concept on the development of linked data for electronic theses and dissertations (ETD) of the Thai Academic Digital Collection (TDC). The research focusing on the ETD collections and will use the features of Linked Data to design the data sets for TDC and develop an interoperating system that enhance the sharing of digital information resources from different university libraries in the network and enable the end-users to more effectively retrieve and access information resources in TDC’s ETD collections.

Keywords: Linked Data, ETD, Digital Collection, Thai Academic Digital Collections

INTRODUCTION

The Thai Academic Digital Collection (TDC) project was founded by the Commission of Higher Education (CHE) of Thailand in 2000. The main objective was to support the academic libraries in creating the digital collections of the institutional publications such as theses and dissertations, research reports, journal articles, rare books, teaching materials, and institutional archives in order to promote information sharing and accessing among the members of the ThaiLIS (Thai Library Integrated System) network. The TDC project currently holds the digital collections of more than 404,000 items and the numbers of items increase every year. There are more than 2,100,000 users from 174 academic institutions and the public. The statistic of items downloading has been more than 20 million times per year, of which the highest statistic was 32,863,268 times in 2015. The project has two phases, and the first phase focuses on the digital collection of Electronic Theses and Dissertations or ETD.

The report of TDC project (Commission of Higher Education, 2015) revealed that though the digital collections have been the important resources for accessing the Thai universities’ publications, the operation problems have been occurred in five aspects:
acquisitions, digitization, metadata, right management and information retrieval. The main problem is information retrieval which directly effect on the end users. This is because the lack of tool for information discovery, lack of tool for citation and metric management, and lack of accurate link between digital objects.

In addition to directly operations problems that mention above, DC project also has a few problems such as data standards does not meet the needs of users and technology (Google scholar, semantic search and Linked Data), no mechanism or engine to handle the references and citation of digital theses and dissertations (Walailak University, 2014). Moreover, the operating of ETD project in each university in Thailand is also different, for example some universities operated the project as individual institutional repository (IR) and the others used TDC software as a core system and upload the digital version of ETD to the central database of TDC. The central database plays role as a digital repository which collects the digital outputs of members of TDC and supports the archiving and long-term preservation of the institution's intellectual outputs.

Most of Thai Academic Digital Collections collect the digital files in PDF format which are converted from Microsoft Word files, therefore the stored items look much like the traditional printed version of original theses and dissertations (Ammarukleart, 2007). These ETD items can be viewed via with the free Adobe Acrobat reader and can be searched by using with different version of PDF reader. At present, the TDC collections can be uploaded into the central database in other formats such as HTML, images, multimedia, source code and interactive applications so that the information is better accessible and distributed.

TDC has also become available online, but, for the most part, it is not yet an integral part of the Web. Suitability of the traditional and established standards for the communication of this information in the Web environment is disputable.

In order to increase the contributions of digital theses and dissertation collection effectively, the CHE (Commission of Higher Education, 2015) has pointed out that the project needs to develop or revise the following five elements: the framework for system operation, the appropriated system architecture, the engine for digital objects management, the standards for system interoperability and the good solution for right management.

Linked Data is a method of publishing structured data so that it can be interlinked and become more useful through semantic queries. It builds upon standard web technologies, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried (Soylu, Mödritscher and De Causmaecker, 2012). Recently, Linked Data has gained great attentions and become continuously growing in the uses for interlinking data, especially ETD, among the academic library networks as a tool for information sharing and accessing. Linked Data is also a tool for expanding the traditional digital collection services by creating the new data set with special features that enable the interoperation of ETD and linkage of metadata resources among different institutions.

LITERATURE REVIEW

An overview of Digital Collection

Digital Collection is an organized assembly of digital information objects in different formats like text documents, still or moving images, data, voice, etc., or a combination of all these and be systematically described in a structured format (metadata) using standard tools such as thesauri, controlled vocabularies (subject headings) and taxonomies (Tapaswi, 2010).
The evaluation parameters for the digital collections could very well be set based on the NISO's framework namely “The Framework of Guidance for Building Good Digital Collections” (NISO, 2007). This NISO's framework composes of four criteria for good practices in the following four core types of entities:

- **Collections** (organized groups of objects)
- **Objects** (digital materials)
- **Metadata** (information about objects and collections)
- **Initiatives** (programs or projects to create and manage collections)

Digital collection is the combination of content and services, and plays role as a tool at the center of intellectual activity having no logical, conceptual, physical, temporal, or personal borders or barriers on information. To build a good digital collection, the institutions need to create for an appropriate reference model. A reference model is an abstract framework for understanding significant relationships among the entities of some environment, and for the development of consistent standards or specifications supporting that environment (Tapaswi, 2010).

The OASIS Reference Model for SOA (Service Oriented Architecture) provides a common language for understanding the important features of SOA but does not address the issues involved in constructing, using or owning a SOA-based system (Specification, O. A. S. I. S. Committee, 2012). The intended audiences of OASIS model for digital collection and digital library are: enterprise architects, standards architects and analysts, decision makers and stakeholders/developers.

**An overview of ETD life-cycle management**

ETD (Electronic theses and dissertations) is a subset of digital collection, which is also a subset of IR (Institutional Repositories). ETD is defined as a master's thesis or doctoral dissertation in electronic form, rather than paper format. ETD also refers to a master’s thesis or doctoral dissertation that is archived and circulated electronically, not in the printed and bound formats (Cayabyab, 2015). ETD have gained huge attention of institutions and organizations worldwide as a technological tool for the graduate research outputs to be communicated globally and play a major role as tool to empower universities to unlock their information resources.

In 2011 researchers from the worldwide Networked Digital Library of Theses and Dissertations (NDLTD), began studying the production, dissemination, and preservation of ETD (Schultz & Krabbenhoeft, 2014). The scope of their works increased to encompass the roles and responsibilities of core stakeholders in the ETD lifecycle including: students, faculty members, administrators, technologists, commercial vendors, and librarians. The resulting for Lifecycle Management of ETDs has addressed the areas of interests for ETD program planners, managers, and curators. The activities in ETD life-cycle are:

- ETD program planning
- ETD creation, submission and ingest
- ETD access and retrieve
- ETD archiving and preservation
- ETD evaluation and assessment

In summary, the guideline has identified a number of access restrictions and embargoes that can be applied to ETD. The benefits of the digital collection of ETD should be considered prior to applying worldwide distribution to an ETD’s availability.
An overview of Linked Data

Linked Data is a data publishing technique that uses common web technologies to connect related data and make them accessible on the web. It relies mainly on identifying resources with (HTTP) Uniform Resource Identifiers (URI), and, using standards such as the Resource Description Framework (RDF), in providing data about these resources and connecting them to other resources on the web (Berners-Lee, Handler, & Lassila, 2006).

In order to enable Linked Data applications to discover datasets as well as to ease the integration of data from multiple sources, Linked Data publishers should comply with a set of best practices which comprised of three areas (Bizer, Heath, & Berners-Lee, 2009).

Linking: by setting SKOS (Simple Knowledge Organization System) and RDF (Resource Description Framework) links, data providers (university or library) connect their datasets into a single global data graph which can be navigated by applications and enables the discovery of additional data by following RDF links.

Vocabulary Usage: the best practices advise publishers to use terms from widely-used vocabularies in order to ease the interpretation of their data. If data providers use their own vocabularies, the terms of such proprietary vocabularies should be dereference-able into their RDF schema or OWL (Web Ontology Language) definitions.

Metadata Provision: Linked Data should be as self-descriptive as possible, and thus include metadata. An important form of metadata is provenance metadata describing the origin of datasets and enabling applications to assess their quality. The best practices also advise to provide licensing metadata and dataset level metadata.

METHODOLOGY

This research aims at development of Linked Data for TDC focusing on the ETD collections. The research will use the features of Linked Data to design a data set for TDC and develop an interoperating system that enhance the sharing of digital information resources from different university libraries in the network and enable the end-users to more effectively retrieve and access information resources in TDC’s ETD collections.

Research Objectives

1. To identify the current situations of the life cycle management of digital collections in Thai universities.
2. To design the dataset structure for the electronic theses and dissertations (ETD) collections of TDC.
3. To develop Linked Data for the ETD collections of TDC.

Research Conceptual Framework

The researchers will comply the following knowledge and theories for the development of the research conceptual framework: the concept of digital collection (NISO, 2007), the ETD life cycle management, the Digital Library Reference Models, the 5-stars of bibliographic linked data model (Janowicz, Hitzler, Adams, & Kolas, 2014) and SKOS. The detail of research conceptual framework is shown in Figure 1

Research Methodology

The study uses research and development approach which comprises of the following steps:
1. A documentation analysis and a survey research will be conducted in order to identify the current situations of the life cycle management of digital collections in Thai universities. The study will base on the digital collection development of NISO (NISO, 2007) and the ETD life-cycle management of The Network Digital Library of Theses and Dissertation (NDLTD) (Schultz & Krabbenhoeft, 2014) which comprises of 5 aspects: project initiative, process of creation, submission and ingestion of digital collections, access and retrieval issues, long term preservation of the collection and the evaluation and assessments.

2. A survey research to identify the users’ requirements and then design the dataset structure for the electronic theses and dissertations (ETD) collections of TDC.

3. Development of a Linked Data system for digital collection management by using NISO digital collection frameworks, digital libraries reference models and linked data technologies. The system will have 5 modules including RDF and ingestion engine, search and browse, content management, user interface for manage linked data and inter-operational tools for external systems. The system will be implemented and tested by the end users.

The research conceptual framework is shown as follow

**Expected Outcomes**

The results of this research will provide 3 outputs: (1) the frameworks of academic digital collections management for Thai universities (2) the system prototype that include of datasets and data structures for the TDC Linked Data (3) Linked Data for Thai Academic Digital Collections with 5-stars bibliographic linked data rank that can contribute systemically and effectively. In addition it is the tool for gathering, accessing, using and preserving of the knowledge in Thai Academic Digital Collection.
REFERENCES


Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, and Dave Kolas. .(2014) Five stars of linked data vocabulary use. *Semantic Web 0*. .4-1


Matt Schultz, and Nick; Skinner Krabbenhoeft. .(2014) Guidance Documents for Lifecycle Management of ETDs.) Matt Schultz, editor (*Educopia V1.0*). .21-5


