UNIVERSITI TEKNOLOGI MARA

A SEMANTIC WEB ENABLED INTEGRATED SEARCH SERVICE FOR ELECTRONIC THESES AND DISSERTATIONS

HESAMEDIN HAKIMJAVADI

Thesis submitted in fulfilment of the requirements for the degree of

Master of Science

Faculty of Information Management

December 2011

ABSTRACT

In recent years, Electronic theses and dissertations (ETDs) are becoming an integral part of Institutional Repositories (IRs). However, multiplicity of data providers, as well as the variety of software solutions, standards, and protocols utilized for running these repositories, has led to some complexity in integration of ETD resources. Almost all of current data integration methods involve combining resources residing in different sources and providing users with a unified view of these data. Applying these methods in the domain of digital libraries led to development of a number of specialized integration methods (e.g. metadata harvesting, metadata aggregation, etc.) as well as some interoperability protocols (e.g. OAI-PMH, SWORD, z39.50, etc.). Nevertheless, this research unveiled that none of these methods and standards are capable of integrating ETD resources on the semantic level. In this study, 10 metadata integration methods and 8 interoperability protocols were evaluated from both theoretical and practical perspectives. For this purpose, we conducted two surveys (among 266 ETD archives and 136 ETD experts), and 2 comparative studies (among 15 IRs software solutions and 10 metadata integration methods). The results of the surveys indicated that the OAI-PMH is the most widely adopted interoperability protocols among ETD archives and IR software providers. On the other hand, the evaluation of metadata integration methods depicted that the metadata harvesting method is not capable of providing higher level of integration among ETD resources. Based on these results, a semantic web-based framework namely ETD Integrating System was designed. The framework consists of 5 steps, for each step a specific software tool was developed, so that together formed an information workflow system.

ACKNOWLEDGMENTS

I would like to express my full appreciation to everyone who assisted me with this work specially my supervisor Dr Mohamad Noorman Masrek. He has been extremely generous in the sharing of knowledge, and inspiring at every level. It had been not only a chance, but also a real privilege to work under his guidance. I am also deeply thankful to Dr Majid Sohrabi and Ahmad Tavakoli for their precious time and help with the assessment of reliability of proposed frameworks in this study.

Additional thanks to the participants who kindly dedicated their time and energy to this study.

Thank you all for making my first experience of studying abroad as a great adventure and memorable experience.

At last but not least, my whole-hearted thanks go to my wife for her care and strong support.

Table of Content

CHAPTER ONE: INTRODUCTION

Nowadays, there is a widespread agreement on the vital importance of openness and dissemination of scientific information resources on the web. Open Institutional Repositories are one of the most reliable types of these sources. Recently, among all types of IRs, the motion of generating Electronic Theses and Dissertations (ETDs), as a new genre of scholar documents, has achieved significant progresses. Universities provide free access to a huge number of ETD collections through their portals. However, the fast growth in the number of ETD repositories has caused new challenges for universities, which are heterogeneity of sources (Pyrounakis, Saidis, Nikolaidou, & Lourdi, 2004), lack of interoperability , and necessity of integration in order to provide a unified interface for access, search and browsing through different ETD repositories (Carlson, Ramsey, & Kotterman, 2010).

Semantic web as the next generation of web technology consists of standards, data expression language, and applications that could be used for integrating heterogeneous information sources (Shadbolt, Berners-Lee, & Hall, 2006). According to the definition of Sir Tim Berners-lee, who is the creator of the web, resources in the next generation of the web are structured in a way that are not just interpretable for the human end user, but also are processable for machines (Berners-Lee, 2002). In fact, being built upon the infrastructures such as Uniform Resource Locator (URL) for identification and Resource Description Format (RDF) for expression of resources on the web, one of the main goals of developing semantic web-based applications and standards is to provide a platform for integration of unstructured and semi-structured web resources.