

UTEM ACADEMIC PUBLICATION DASHBOARD



UNIVERSITI TEKNIKAL MALAYSIA MELAKA

BORANG PENGESAHAN STATUS TESIS*

JUDUL: UTEM Academic Publication Dashboard

SESI PENGAJIAN: 2015/2016

Saya LIM BOON HEE

(HURUF BESAR)

mengaku membenarkan tesis (PSM/Sarjana/Doktor Falsafah) ini disimpan di perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dengan syarat-syarat kegunaan seperti berikut:

1. Tesis dan projek adalah hakmilik Universiti Teknikal Malaysia Melaka.
2. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat Salinan untuk tujuan pengajian sahaja.
3. Perpustakaan Fakulti Teknologi Maklumat dan Komunikasi dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.
4. ** Sila tandakan (/) SULIT

(Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA RASMI 1972)

TERHAD

(Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan)

TIDAK TERHAD

(TANDATANGAN PENULIS)

Alamat tetap: 90, Jalan Sri Temiang 4,

Taman Sri Temiang

84000 Muar, Johor

Tarikh: 25/8/2016

(TANDATANGAN PENYELIA)

ZURINA SAAYA

Nama Penyelia

Tarikh: 25/8/2016

Catatan: * Tesis dimaksudkan sebagai Laporan Akhir Projek Sarjana Muda (PSM)

** Jika Tesis ini SULIT atau TERHAD, sila lampirkan surat daripada pihak berkuasa.

UTEM ACADEMIC PUBLICATION DASHBOARD



This report is submitted in partial fulfillment of the requirements for the
Bachelor of Computer Science (Computer Networking)

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY
UNIVERSITI TEKNIKAL MALAYSIA MELAKA
2016

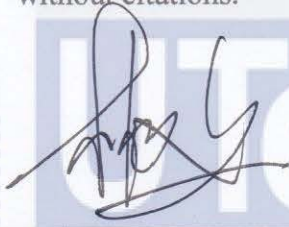
DECLARATION

I hereby declare that this project report entitled

UTEM ACADEMIC PUBLICATION DASHBOARD

is written by me and is my own effort and that no part has been plagiarized

without citations.

STUDENT :  Date: 25/8/2016
(LIM BOON HEE)

اونيورسيتي تیکنیکل ملیسيا ملاک

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

I hereby declare that I have read this project report and found this project report is sufficient in term of the scope and quality for the award of Bachelor of Computer Science (Computer Networking) With Honours.

SUPERVISOR :  Date: 25/8/2016
(DR. ZURINA BINTI SA'AYA)

DEDICATION

Everybody in this world should learn to program a computer, because it teaches you how to think



ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor (Dr. Zurina Saaya) for guiding me on my researches throughout the whole project, as well as all of my course mates for sharing all the knowledges and helping me throughout the whole project.



ABSTRACT

Academic electronic information archival has evolved far beyond the simple days of creating contents and storing them into local respiratory systems. Nowadays, academic information are already gathered and available in one of many online sites. For instance, an author's academic publications can be retrieved from Google Scholar website. Google Scholar is more preferable because Google Scholar keeps a record on the citation index for each publication and the citation index are kept on track all the time. The citation index can later be used to give the h-index score of the author. Such information is essential to measure the research competency at the institution level.

Currently, UTeM acquires publication records through internal system called URIS system which is a platform that keeps record of all research details of publication information. However, this system requires academicians to submit their publication details into the system manually which could lead to issues such as inefficient data collection, missing data, delayed submission, etc. In this project, a web crawler will be built to retrieve the academic publication information of UTeM staff. With the availability of an effective web crawler for publication data, UTeM can monitor scholarly information in a better way and plan towards increasing the publication index among academic.

ABSTRAK

Sistem arkib maklumat akademik telah banyak berkembang. Kini, dengan hanya menyimpan rekod maklumat penerbitan dalam sistem repositori adalah tidak mencukupi. Informasi akademik telah dikumpulkan daripada laman web yang umum contohnya di laman web Google Scholar. Informasi seseorang penulis dapat diakses dengan mudah di laman web Google Scholar. Ia menjadi pilihan utama kerana indeks petikan yang terdapat di Google Scholar sering dikemaskini. Indeks petikan yang terdapat di Google Scholar kemudiannya akan digunakan untuk mendapatkan indeks-h seseorang penulis itu. Informasi sebegini amat penting untuk mengukur kompetensi penyelidikan institusi. Malaysia Research Assessment Instrument (MyRA), telah digunakan untuk menilai kapasiti penyelidikan sesebuah Institusi Pengajian Tinggi (IPT). Oleh itu, indeks petikan penerbitan UTeM harus dikumpul dan dikemaskini dari semasa ke semasa untuk menilai kompetensi UTeM di antara IPT di Malaysia.

Pada masa ini, UTeM mendapatkan rekod penerbitan melalui sistem dalaman yang dikenali sebagai URIS yang merupakan sebuah platform untuk mengemas kini rekod penerbitan. Namun begitu, sistem tersebut hanya berfungsi dengan memasukkan butiran secara manual dan akan menyebabkan masalah seperti penangguhan, kehilangan data, dan sebagainya. Dalam projek ini, sebuah “web crawler” akan dibina untuk mendapatkan maklumat penerbitan staf UTeM secara automatik. Dengan adanya “web crawler” ini, UTeM akan dapat memantau informasi penerbitan dengan lebih baik dan membuat perancangan untuk meningkatkan indeks penerbitan dan kedudukannya di MyRA.

TABLE OF CONTENTS

CHAPTER	SUBJECT	PAGE
	DECLARATION	I
	DEDICATION	II
	ACKNOWLEDGEMENTS	III
	ABSTRACT	IV
	ABSTRAK	V
	TABLE OF CONTENTS	VI
	LIST OF TABLES	X
	LIST OF FIGURES	XII
	LIST OF ABBREVIATIONS	XIV
CHAPTER I	INTRODUCTION	
	1.1 Introduction	1
	1.2 Problem Statement (PS)	2
	1.3 Project Question (PQ)	2
	1.4 Project Objective (PO)	3
	1.5 Project Scope (PS)	3
	1.6 Project Contribution (PC)	4
	1.7 Thesis Organization	4
	1.8 Conclusion	6
CHAPTER II	LITERATURE REVIEW	
	2.1 Introduction	7
	2.2 Fact and Findings	8
	2.2.1 Domain	8
	2.2.2 Related Work	16
	2.3 Critical review	18
	2.3.1 Comparison with Existing System	18
	2.3.2 Project Requirement	20
	2.3.2.1 Software Requirement	20

2.4	Proposed solution	26
2.5	Conclusion	27
CHAPTER III	PROJECT METHODOLOGY	
3.1	Introduction	28
3.2	Methodology	29
3.3	Project Milestone	31
3.4	Conclusion	35
CHAPTER IV	ANALYSIS AND DESIGN	
4.1	Introduction	36
4.2	Problem Analysis	36
4.3	Requirement Analysis	37
4.3.1	Data Requirement	37
4.3.2	Functional Requirement	43
4.3.3	Non-functional Requirement	46
4.3.4	Other Requirement	47
4.3.4.1	Software Requirement	47
4.3.4.2	Hardware Requirement	48
4.4	High-Level Design	49
4.4.1	System Architecture	49
4.4.2	User Interface Design	50
4.4.2.1	Navigation Design	50
4.4.2.2	Input Design	51
4.4.2.3	Output Design	52
4.4.3	Database Design	57
4.4.3.1	Conceptual and Logical Database Design	57
4.5	Detailed Design	58
4.5.1	Software Design	58
4.5.2	Physical Database Design	61
4.5.2.1	Database Connection	61

	4.5.2.2 Create Google Scholar Author Table	62
	4.5.2.3 Create Google Scholar Article Table	62
	4.5.2.4 Create Citation by Year Table	63
	4.5.2.5 Create Citation index Table	63
	4.5.2.6 Create Google Scholar Author Article Table	64
	4.5.2.7 Create UTeM Staff List Table	65
4.6	Conclusion	65
CHAPTER V	IMPLEMENTATION	
5.1	Introduction	66
5.2	Software Development Environment Setup	67
5.3	Software Configuration Management	68
	5.3.1 Configuration Environment Setup	68
5.4	Version Control Procedure	72
5.5	Implementation Status	74
5.6	Conclusion	74
CHAPTER VI	TESTING	
6.1	Introduction	75
6.2	Test Plan	76
	6.2.1 Test Organization	76
	6.2.2 Test Environment	77
	6.2.3 Test Schedule	78
6.3	Test Strategy	78
6.4	Test Design	79
	6.4.1 Test Description	79
	6.4.2 Test Data	81
6.5	Test Result and Analysis	81
6.6	Conclusion	89
CHAPTER VII	PROJECT CONCLUSION	
7.1	Introduction	90

7.2	Project Summarization	90
7.3	Project Contribution	92
7.4	Future Works	92
7.5	Conclusion	93



LIST OF TABLES

TABLE	TITLE	PAGE
1.1	Summary of Problem Statement	2
1.2	Summary of Project Questions	2
1.3	Summary of Project Objectives	3
1.4	Summary of Project Contributions	4
2.1	Comparison of UTeMAIR and Scholar.py	19
2.2	Python HTML parser comparison	21
2.3	Python robots.txt parser comparison	22
2.4	Javascript chart library comparison	24
2.5	Local server comparison	24
2.6	Hardware requirement of server	25
3.1	Project milestone	31
3.2	Project Gantt Chart	34
3.3	System Development Gantt Chart	34
4.1	Data dictionary of google_scholar_author table	38
4.2	Data dictionary of google_scholar_citation_by_year table	39
4.3	Data dictionary of google_scholar_citation_index table	40
4.4	Data dictionary of google_scholar_article table	41
4.5	Data dictionary of google_scholar_author_article table	42
4.6	Data dictionary of utem_staff_list table	42
4.7	Software requirement	47
4.8	Hardware requirement	48
4.8	Table of input design	51
5.1	Version control of UTeMAIR crawler	72
5.2	Version control of UTeMAIR A-PD	73
5.3	Implementation status of UTeMAIR crawler	74
6.1	Test Organization	76

6.2	Item description of test structure	77
6.3	Test Schedule	78
6.4	Description of test case of UTeMAIR crawler	79
6.5	Description of test case of UTeM A-PD dashboard	80
6.6	Test result for manual_scraping	81
6.7	Test result for manual_update	82
6.8	Test result for automated crawling process	87



LIST OF FIGURES

DIAGRAM	TITLE	PAGE
2.1	Web Crawling Architecture	9
2.2	Properties of a crawler(Shetty et al. 2012)	11
2.3	Examples of robots.txt file from w3schools	13
2.4	Google reCAPTCHA survey	14
2.5	Example to retrieve an article with a keyword and the author name	16
2.6	Search based on publication cluster id	17
2.7	Code fragment of Scholar.py to generate url for data retrieval	19
2.8	result of robotparser and reppy with the same url	23
3.1	Project Methodology	29
4.1	Context Level Diagram of UTeMAIR Crawler	43
4.2	UTeMAIR crawler DFD level 0	44
4.3	Data Flow Diagram Level 1 for process 4.0 of UTeMAIR crawler	45
4.4	System Architecture	49
4.5	Navigation design of UTeM Academic Publication Dashboard	50
4.6	Running UTeMAIR crawler	52
4.7	UTeMAIR crawler updater	53
4.8	Retrieving information of an author manually	53
4.9	Updating information of an author manually	54
4.10	Scheduling UTeMAIR crawler	54
4.11	Example chart of UTeMauthors total citations per year	55
4.12	Comparison of data in UTeM Academic Publication Dashboard	55
4.13	An overview of publication information of all author	56

4.14	Overview of all author citation by year	56
4.15	ERD diagram of UTeMAIR repository system	57
4.16	System flowchart of UTeMAIR crawler	58
4.17	Database connection by UTeMAIR crawler using Python	61
4.18	Database connection by UTeM A-PD using PHP	61
5.1	httpd-xampp.conf file	68
5.2	config.inc.php file	69
5.3	Add user account	69
5.4	Port forwarding rules	70
5.5	Dynamic DNS	70
5.6	php script to test MySQL database connection	71
5.7	Output of the php script mentioned on the image above	71
6.1	Structure of testing environment	77

LIST OF ABBREVIATIONS

API	-	Application Programming Interface
CAPTCHA	-	Completely Automated Public Turning test to tell Computers and Humans Apart
DNS	-	Domain Name Server
HTML	-	Hypertext Markup Language
HTTP	-	Hypertext Transfer Protocol
IDLE	-	Integrated Development Environment
IP	-	Internet Protocol
IPT	-	Institusi Pendidikan Tinggi
IT	-	Information Technology
JSON	-	JavaScript Object Notation
MyRA	-	Malaysia Research Assessment
OCR	-	Optical Character Recognition
PHP	-	Hypertext Preprocessor
REP	-	Robot exclusion protocol
SQL	-	SEQUEL (Structured English Query Language)
UI	-	User interface
URIS	-	UTeM Research Information System
URL	-	Universal Resource Locator
WAN	-	Wide Area Network
WWW	-	World Wide Web
UTeM A-PD	-	UTeM Academic Publication Dashboard
UTeMAIR	-	UTeM Academic Information Retrieval
XML	-	Extensible Markup Language

CHAPTER I

INTRODUCTION

1.1 Introduction

With the evolution of academic electronic information archival system, creating contents and storing them into local respiratory systems is no longer useful. Such important information is essential to an institute. By just storing the information in the electronic information archival system without making full use of it, it is just a waste in computer resources. Nowadays, with the evolution of information technologies, processing information becomes handier and these information could be processed to provide more detailed comparison of information to measure the research competency at the institution level so that plans and measures could be taken to improve the institution.

Currently, UTeM acquires publication records through URIS system which is a platform that keeps record of all research details of publication information. However, this system requires academicians to submit their publication details into the system manually which could lead to issues such as inefficient data collection, missing data, delayed submission, etc. In this project, a web crawler will be built according to the robots exclusion standard or also known as robot exclusion protocol to retrieve the academic publication information of UTeM staff from Google Scholar and Scopus to make sure the web crawler operates in an ethical and good manner to prevent any denial of service attack on the intended server. With the availability of an effective web crawler for publication

data, UTeM can monitor scholarly information in a better way and plan towards increasing the publication index among academics and ultimately improve its ranking.

1.2 Problem Statement (PS)

Table 1. 1 Summary of Problem Statement

PS	Problem Statement
PS ₁	The academic publication by UTeM authors needs to be gathered and updated on timely basis in order to keep track of its research progress.

1.3 Project Question (PQ)

Table 1. 2 Summary of Project Questions

PS	PQ	Project Question
PS ₁	PQ ₁	How can a tool can be used to crawl the required services in order to extract publication data of UTeM staff?
	PQ ₂	How to ethically crawl the web sites in order to extract publication data of UTeM staff?
	PQ ₃	How to integrate, summarize and visualize the publication data that are gathered by crawling tool?

Project Objective (PO)

Table 1. 3 Summary of Project Objectives

PS	PQ	PO	Project Objective
PS ₁	PQ ₁	PO ₁	To develop a web crawler to retrieve publication information of UTeM staff from academic online resources such as Google Scholar and Scopus.
		PO ₂	To extract the information from the retrieved data and create a local repository to store the information.
	PQ ₂	PO ₃	To make sure the web crawler operates in an ethical manner.
	PQ ₃	PO ₄	To develop a platform to summarize and visualize in the publication data.

1.4 Project Scope (PS)

I. Web Crawler

The web crawler will be developed using Python by following the robots exclusion standard.

II. Dashboard

The dashboard will be developed to structure the unstructured data crawled by the web crawler to visualize it.

III. Source of publication information

The publication information will be retrieved from the web site of Google Scholar and Scopus.

1.5 Project Contribution (PC)

Table 1. 4 Summary of Project Contributions

PS	PQ	PO	PC	Project Contribution
PS ₁	PQ ₁	PO ₁	PC ₁	Proposed a web crawler to crawl publication information in a timely manner to keep the information up to date.
		PO ₂	PC ₂	Proposed to store and update publication information into local repository.
	PQ ₂	PO ₃	PC ₃	Proposed an ethical method to crawl data from web sites containing publication information.
	PQ ₃	PO ₄	PC ₄	Proposed a dashboard to visualize the publication information from the local repository.

1.6 Thesis Organization

Chapter 1: Introduction

This chapter introduces the project and the project background briefly including the project problem statement, project question, project objectives, project scope, and project contributions.

Chapter 2: Literature Review

This chapter gives a preview to the literature review to the project. Citations from articles of previous work will be provided as well. This chapter includes previous work examples, critical review of current problem and justification to the problem, and also proposed solution.

Chapter 3: Project Methodology

This chapter provides the methodology of the development process of the project. The project methodology will be divided into stages to be carried out and the project milestones will explain the actions and plans of each stage prior to the end of the project.

Chapter 4: Analysis and Design

This chapter will provide the preliminary design and the detailed design of the project. A few analysis will be done in this chapter including problem analysis which will be visualized to describe the flow of current system or business based on problem statements mentioned earlier in Chapter 1. The other analysis is the requirement analysis which will be broken down into 4 categories (data requirement, functional requirement, non-functional requirement and others requirement). For the design of the system, it consists of 2 parts which are high-level design and detailed design. High-level design will be consisting of system architecture, UI design and database design. Detailed design will include software design and physical database design.

Chapter 5: Implementation

This chapter describes the activity involved in the implementation phase of the system development consisting of software development environment setup, software configuration management and implementation status.

Chapter 6: Testing

This chapter describes the activities involved in the testing phase of the system briefly including the testing strategies adopted. This chapter consists of test plan, test strategy, test design and test results and analysis.

Chapter 7: Project Conclusion

This chapter summarize the project by describing the objectives achieved and conclude the results gained from this project. The weakness and strength of the project are also stated here. This chapter includes project summarization, project contribution, project limitation and future works.

1.7 Conclusion

The key motivation for designing web crawler is to retrieve web pages and add their representations to a local repository. Crawlers are computer programs that roam the web with the goal of automating specific tasks related to the web. However, web crawlers often causes a denial of service attack to the destined server. So, it is important that the standards in web crawler are met. The aim of this project is to retrieve data related to academic publications by UTeM staffs. The web crawler will primarily crawl data from Google Scholar and Scopus.

CHAPTER II

LITERATURE REVIEW

2.1 Introduction

Academic publication is a main part in academic research and scholarship. Mostly the publication is published as academic journal article, book or thesis in offline and online form. Google Scholar and Scopus are some example of focused search engine specifically for retrieving online academic publication. These search engine is freely accessible and it indexes the full text or metadata of scholarly publication across different types of publishing formats and disciplines. Web crawling is one of the many techniques that can be used to extract information from these search engines.

Web crawler has been around since the largest information space, the World Wide Web (www) gains its popularity. The web crawler is often described as an internet bot was developed since then to retrieve all sorts of information from the web.

Since then, a wide variety of theories and methodologies of web crawlers are available to explain the operation and functionalities of a web crawler. However, only a few themes will be covered in this review which are web crawler, ethical web crawling, CAPTCHA, and web APIs.

A variety of context of related topics are presented in the journals and article, but this chapter will primarily focus on a focused crawler.