

Developing National Taiwan University Web Archiving System

Kuang-hua Chen
Department of Library and
Information Science,
National Taiwan University

National Taiwan University Library

No. 1, Sec. 4, Roosevelt Rd., Da-An Dist., Taipei 10617, TAIWAN

+886-2-3366-2963

khchen@ntu.edu.tw

Yen-liang Chen
National Taiwan University Library

+886-2-3366-2287

vicchen@ntu.edu.tw

Peng-fung Ting
National Taiwan University Library

+886-2-3366-4585

jerryting@ntu.edu.tw

ABSTRACT

Web archiving is an emerging concept whose main purpose is to preserve websites with cultural or historical significance. This paper discusses the development and implementation of the National Taiwan University Web Archiving System (NTUWAS), which was developed by the National Taiwan University Library. In order to help readers unfamiliar with web archiving, we give a brief survey of current activities of web archiving in the world. We then describe the implementation and functionalities of NTUWAS, as well as the challenges and difficulties that we encountered.

Categories and Subject Descriptors

H3.7 [Information Storage and Retrieval]: Digital Libraries – collection, system issues.

General Terms

Design, Management.

Keywords

Collection Development, Digital Archive, Web Archiving

1. INTRODUCTION

The announcement of Mosaic set up a milestone of the Internet development in 1993. Internet reached to users of all kinds of fields after the debut of the Mosaic. Most peoples could easily use Mosaic, and the followers like Netscape Navigator, Internet Explorer, Opera, Firefox, Safari, and so on. The content creators had more routes to publish their works. As a result, hundreds of millions of websites containing miscellaneous information appeared. Various institutions constructed their own websites which demonstrated their official information and related contents. These websites and web pages, thus, possess the prompt, learned, skilled, cultural, and historical meanings. As a result, the Internet mirrors the epitome of the current political, social, economic, and cultural significances in some senses. Studying and analyzing contents of websites in terms of temporal aspect make information

This work is licence under an Attribution- NonCommercial -NoDerivs 2.0 France Creative Commons Licence.

further full of progressive significances. We could say that the valuable materials of Internet in every time slice will be the treasure for the future uses and studies.

However, the average lifetime of information or resources on the Internet is actually a very short period. According to a survey, the average lifetime of a website is only 44 days [1]. To preserve websites or digitalized documents containing valuable and researchable components has been the indispensable responsibility of a library. National Taiwan University Library (hereafter, NTUL) has concerned web archiving for a long time. From 2006, we started to build the web archiving system and regulated the scope of the collections. We made efforts on selecting, collecting, and preserving the digital data about National Taiwan University, the Taiwanese governments, the Taiwanese political parties, the well-known persons, the research issues, and the significant events. Moreover, we established guidelines to manage the collections, and we also framed a classification scheme for organizing websites and featured various functionalities for the users. We expect to initiate the trend to preserve the Internet assets in Taiwan.

The main purpose of this paper is to introduce National Taiwan University Web Archiving System (NTUWAS) and the ways we think about to develop and operate this system. The rest of this paper is organized as follows. In Section 2, the current activities of web archiving will be briefly described. Section 3 will explain the implementation and functionalities of NTUWAS. The challenges and difficulties we encountered will be discussed in Section 4. Section 5 is the conclusions.

2. ACTIVITIES of WEB ARCHIVING

The concept of web archiving situated at an initial stage in Taiwan. In contrast, Internet Archive, an unprofitable organization established at San Francisco, has been devoted to collecting all kinds of digital materials for the potential applications or researches since 1996 [2]. All kinds of websites are the targets of Internet Archive. National Library of Australia also built PANDORA and joined the web archiving in 1996 [3]. The United States of America, the United Kingdom, and Japan executed the related projects consequently. Several famous web archiving projects/systems are briefly introduced in the following.

2.1 Internet Archive

Internet Archive manages to collect all types of websites extensively in order to maintain the powerful mirrors of Internet in the world. The tool which Internet Archive used to fulfill this purpose is Wayback Machine. Internet Archive archives websites by the topics, such as Asian Tsunami, Hurricanes Katrina and Rita, the 911 event, and so on. Besides, Internet Archive organized a forum and built an FAQ to strengthen the interaction with the users and to response the common questions of users. However, Wayback Machine only features a search function between two dates without supporting the keyword search. The visitors could only search the website by entering its URL. Wayback Machine has collected overall 85 billion pages which is about 200 TB of data [2], but many contents are incomplete.

2.2 LCWA

LCWA (Library of Congress Web Achieves) is the project of Library of Congress for preserving the American websites. MINERVA (Mapping the Internet Electronic Resources Virtual Archive) is the early name of this project. It classifies the digital resources clearly after collecting the related websites according to certain concerned topics. For instance, the election topic is divided into candidates, governments, parties, public opinions, and so on. The topics, like the 911 event and the crisis in Darfur are continually collected. As for the technical background, the acquisition tool which LCWA applies is the Heritrix web crawler [4]. The HTML pages, images, PDF, Flash, voices, and films are collected completely. LCWA has stored up data in excess of 70 TB and 36,000 websites until May, 2007 [5]. Please refer to [6] for details.

2.3 PANDORA

PANDORA was built initially by National Library of Australia in 1996. Now, it is built in collaboration with nine other Australian libraries and cultural collecting organizations. PANDORA Digital Archiving System (PANDAS) focuses on collecting the websites in Australia and categorizes the websites to eighteen subjects, including arts, education, politics, sciences, history, and so on. The users could easily browse archived websites by the alphabetical order. The web crawler adopted by PANDORA is HTTrack. Please refer to [7] for detailed information.

2.4 UKWAC

UKWAC (UK Web Archiving Consortium) originated in June, 2006 was invested and operated by six institutions including British Library (Lead partner), The National Archives, National Library of Wales, National Library of Scotland, JISC, and Wellcome Trust. UKWAC had been starting to collect websites related to UK from 2005. The participating institutions could have their own policies for archiving. For example, Wellcome Library concentrates on the medical websites while British Library focuses on the critical cultural, historical or political issues. Because UKWAC adopted PANDAS, its web archiving system has a similar structure. In addition, UKWAC carried out web archiving after reaching the agreement with the owners of websites. Please refer to [8] for more detailed information concerning UKWAC.

2.5 WARP

WARP (Web Archiving Project) was proposed by the National Diet Library of Japan. WARP was a pilot project starting in 2002

and matured in 2006. WARP has collected more than 3,000 websites. In the beginning, WARP collected the websites of governments, universities, and special topics. Then it explored the scope to unofficial institutions. Users could recommend websites and owners of websites can also contact project's staff for archiving their websites. All archived websites have been proved by the mechanism of verification and authorization. WARP harvests each website at least once per year and features keyword search function. Interestingly, the library classification scheme is used in WARP project. Therefore, the archived websites are well organized in a pre-defined framework. [9]

2.6 Web Infomall

Web Infomall was set up by the Internet laboratory of Peking University. The growth of archived websites is about forty-five millions of web pages per month and the total number of web pages is three billions until April, 2008. It also features special collections by topics like LCWA. However, the searching function is not convenient, since users have to key in the complete URL to search related web pages. [10]

To sum up, LCWA features intensive contents concerning America-related information with various topics. In contrast, Internet Archive features extensive contents. WARP is sophisticated in terms of usage of library classification scheme but with the clear on-line helps. Both PANDORA and UKWAC operated in a joint-force mode. Each participating institution has its definite responsibility and different collections. Web Infomall is an increasingly growing system with high potential. Table 1 summarizes aforementioned web archiving activities.

3. DEVELOPMENT of NTUWAS

NTUWAS (<http://webarchive.lib.ntu.edu.tw>) has been operated from 2006 and demonstrates the current structure and functionalities through a succession of improvements. In general, we assign metadata for each archived website to provide most useful information to users based on the policies of collection development. As the matter of fact, these metadata are the core of functionalities. This section introduces the users' environment, collections and classification scheme, search features, and recommendation features of NTUWAS which are fully supported by the metadata. In addition, we describe the features of management system of NTUWAS in this section as well.

Figure 1 shows the homepage of NTUWAS. NTUWAS demonstrates a friendly environment for helping visitors obtain the information easily and rapidly. A brief overview of our system components is described as follows. The part (1) and part (2) of Figure 1, "Features" and "Time Capsule", promote some commendatory collections by the running pictures. To let the visitors acquire the information systematically, the part (3), part (5), and part (6) of Figure 1, "Categories", "Research Issues", and "Significant Events", show three different perspectives to the collections. If our visitors have no ideas with the categories, they can use the search functions, the part (7) of Figure 1, to find out all related information directly. The part (4), "New Downloads" and "Top Hits", lists the latest archived websites and the popular ones. In addition, NTUWAS accepts the recommendations of visitors to complement our archives. In short, after crawling the archived websites, NTUWAS categorizes them based on our own classification scheme. We consider this kind of design much more appropriate for users.

Table 1. Summary of Web Archiving Activities

Project/System	Country	Web Crawler	Characteristics
Internet Archive	USA	Heritrix	Extensive and global
LCWA	USA	Heritrix	Intensive and topical
PANDORA	Australia	HTTrack	Joint force with search feature by categories
UKWAC	UK	HTTrack	Joint force with different collections and policies
WARP	Japan	Wget	Feature on-line helps and Lib. classification scheme
Web Infomall	China	N/A	Focus on Chinese web pages with quick growth

3.1 Users' Environment

The part (1) of Figure 1 shows typical pictures randomly when users visit NTUWAS. Each picture represents one of the archived websites. This part demonstrates archived websites and manages to give our visitors visual impression. The visitors will have a quick understanding of the contents of NTUWAS. If the visitors click on the picture, NTUWAS will direct them to the collected website as Figure 2 shows. The visitor could also easily turn to other versions of this website archived in different dates by choosing a date in the pull-down menu as shown in the part (1) of Figure 2. However, sometimes the archived websites are not complete. Therefore, we design a feature to collect the feedbacks from our visitors. The visitors who would like to comment could click on the button as shown in the part (2) of Figure 2. The part (3) of Figure 2 shows three buttons. The first button help visitors go back to the previous page and the third button help the visitor go to the next page. If the visitor clicks on the second button, the system will take him/her to the information page which demonstrates the detailed information about this archived website as Figure 3 shows. Some metadata are automatically added by system but valuable information such as classification tags and descriptions are added by our staff as shown in Figure 3. In addition, all versions of the archived website with the archived time are clearly listed based on the temporal order.

The part (2) of Figure 1 shows time capsule of websites which is a distinguished feature of NTUWAS. Time capsule demonstrates a specific website in a running manner which shows the different versions of this specific website archived in different dates. The feature exhibits the changing nature of a website along with the time. Users will perceive this website in a historical feeling and understand the dynamical changing contents in the social, political, and economic contexts. Users could pause or continue the running pages of this website. When users click on the picture, the system will direct them to the specific collection according to the chosen date.

The part (3) of Figure 1 shows the classification scheme of NTUWAS. We will discuss it in the next subsection.

In order to feature further better experience for our potential users, NTUWAS provides the latest ten archived websites and the top ten popular archived websites as shown in the part (4) of Figure 1. Users could click on any one website on the fly while browsing the top latest or the top popular lists.

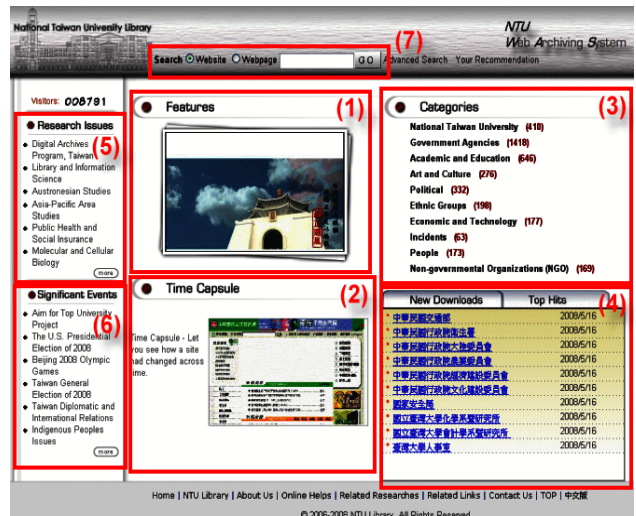


Figure 1. The Homepage of NTUWAS



Figure 2. One of the Archived Website

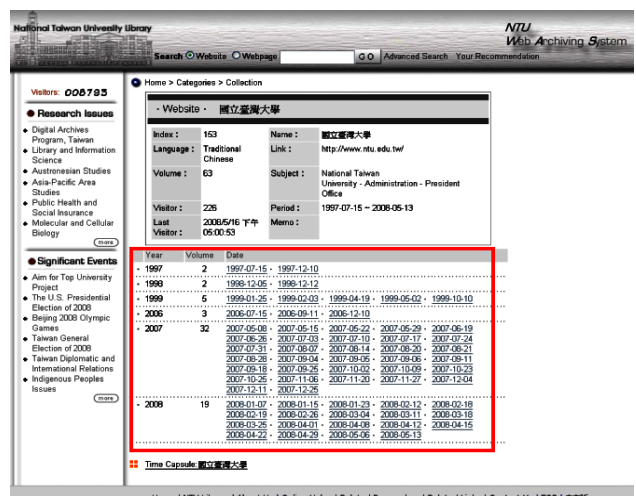


Figure 3. The Detailed Information of Archived website

3.2 Collections and Classification Scheme

3.2.1 Collections

In order to establish meaningful web collections, we decide to archive websites based on the from-inner-to-outer perspectives, which means we have to archive the websites of NTU first and then those of community, Taiwan, China, and other countries. Being the most important university library in Taiwan, NTUL is responsible for preserving valuable digital assets. To begin with, all digital resources produced by NTU should be archived completely. The next step to which NTUWAS goes is to archive Taiwan-related websites. NTUWAS collects digital resources of governments, educational organizations, politicians and political parties, the cultural organizations or events, important persons, aboriginal tribes, and cutting-edge technologies. The last not the least, NTUWAS expands the targets to websites of China and other countries. As a result, ten categories shown in Table 2 are figured out to fulfill the specialty and consideration. The ten categories also serve as the top-level categories of the classification scheme which will be discussed in the next subsection.

Several rules are set up for crawling websites. Firstly, the content of targeted website should be valuable, especially from the historical, cultural, social, educational, and academic viewpoints, for the applications and researches in the future. Secondly, in order to reduce the unnecessary or repeated websites, we focus on the original websites rather than derivative ones. In addition, the websites with many commercial and entertainment contents are not considered.

3.2.2 Classification Scheme

NTUWAS has harvested around three thousand and nine hundred websites from 2006. In order to organize these collected websites, a classification scheme with cataloguing rules is designed for archived websites. For the general public, they may like to access some particular websites and the related websites without sophisticated classification scheme. Therefore, the websites registered in NTUWAS are catalogued using a classification scheme with only three levels. Users could find the related websites hierarchically or broadly. Each website is assigned one class tag at present. As shown in the part (3) of Figure 1 and Table 2, there are ten categories in the first level. The volume of websites contained in each top-level category is shown in the bracket following each category tag. When users click on the category tag, all websites belonging to the designated category will be shown as Figure 4. In addition, a tree-like classification scheme including all-level classes is located at the right-hand side as shown in Figure 4. The visitor could use this tree structure to find websites directly.

Although each website is categorized to one class in terms of its characteristics, several websites of different classes may be related. For example, the official website of a baseball player may be assigned to the class “the famous people” and the official website of the Olympic Movement belongs to the class “the international organization”. If the player joins the 2008 Beijing Olympic Game, the two websites will be correlated. Therefore, an issue-based list and an event-based list as the snapshot shown in the part (5) and (6) of Figure 1 are designed to serve this purpose. The part (5) exhibits well-known issues such as the digital archive projects in Taiwan, the researches of Asian culture, and the applications of the nanomaterials. The part (6) provides the well-concerned

events such as the president election of Taiwan, the 2008 Beijing Olympic Game, and the 921 earthquake in Taiwan.

Table 2. Collections of Ten Main Categories

Category	Collections
National Taiwan University	Websites of programs, projects, faculty, and students in NTU.
Government Agencies	Websites of central and local governments.
Academic and Education	Websites of universities and academic institutions; websites containing teaching materials.
Art and Culture	Websites with contents about literature, art, ancient objects, and customs, and the related institutions.
Political	Websites of the political parties and organizations.
Ethnic Groups	Websites of ethnic groups in Taiwan; websites for women and ethnic minorities.
Economic and Technology	Websites of various issues and institutions of economy and technology.
Incidents	Websites of critical events or hot issues in Taiwan, such as judicial reform.
People	Websites of people with important contributions and influences.
Non-governmental Organization	Websites such as labor unions, unprofitable organizations, etc.

3.3 Search Features

As many web archiving systems do, NTUWAS features the search service as well. But unlike other systems, NTUWAS is equipped with complete full-text searching functionality. As shown in the part (7) of Figure 1, two kinds of “basic searches” are provided to users. One is the “website search” which allows users to search metadata; the other is “webpage search” which allows users to execute full-text searching. Visitors could search titles, URLs, and metadata of the registered websites using metadata search feature. In contrast, full-text search helps visitors search the contents of web pages of every website. Sometimes, the reported results using basic searches are not discriminative. NTUWAS features “advanced search” which allows users to carry out Boolean search for various fields or contents. Figure 5 shows the snapshot of advanced search.

The visitor could submit six keywords at most in the advanced search. These keywords will be searched based on the designated fields as the part (2) of Figure 5 shows. These keywords could be connected using Boolean operators (AND, OR, and NOT) as shown in the part (1) of Figure 5. In addition, the advanced search of NTUWAS could be restricted to a specified level of classification scheme, designated “research issues”, or particular “significant events”. The part (3) of Figure 5 demonstrates these functionalities. While users select the top-level classes, the second level classes will be shown. The third level classes work in the same way. All selected conditions are combined using AND operators implicitly in part (3).

3.4 Recommendation Features

In addition to the websites selected by staff based on policies of the collection development, NTUWAS accepts the users' recommendation as well. While users visit NTUWAS, they could recommend valuable websites by clicking on the "Your Recommendation" button. Figure 6 demonstrates the web page of users' recommendation. The whole recommendation process is very simple. All the needed information for the recommended websites is their URLs. Furthermore, users can also recommend appropriate class tags and review the recommended websites. The operating flow for users' recommendation will be explained in the next subsection.

3.5 Management System

A good management mechanism will absolutely extend the lifetime of an information system and increases its quality. The backend of NTUWAS features management functionalities for our staff. Administrative staff could perform their jobs effectively through the helps of various management functions. Four main modules are discussed here.

3.5.1 Maintain Websites and Classes

The "website maintenance" is the most basic and the most primary function. Website maintenance allows staff to add, modify, and delete the metadata for websites. The information for the NTUWAS users to review has to be edited in website maintenance module.

Similarly, the "class maintenance" allows staff to maintain classification scheme and the class tags. After adding or modifying classification scheme, registered websites could be allocated to the designated classes. Through modules of website and class maintenance, our staff could contribute their professional knowledge as much as possible and greatly help our users.

3.5.2 Schedule Harvestings

As an automatic web archiving system, NTUWAS has to harvest websites regularly or on specific schedule. The whole harvesting process of NTUWAS is shown in Figure 7. The "AutoTrack" in Figure 7 will generate a harvesting list daily based on the schedule registered in metadata database. This list will be submitted to HTTrack, an open-source web crawler, for harvesting the target websites. The administrative staff has to decide when to harvest target websites. There are three options for setting up schedule of harvesting websites. The first option is to harvest target websites regularly. The second option is to harvest target websites immediately. The third option is to harvest target websites between two pre-setup dates. The following explains each of the harvesting scheduling.

● Regular Harvesting

The contents of some websites change frequently. For example, "Department of Drama and Theatre, National Taiwan University" updates their website frequently and features many contents about exhibitions and performances. Administrative staff of NTUWAS decides to harvest this website monthly. Of course, different websites should have their own appropriate periods to match their different characteristics. In addition, multiple regular schedules could be set up easily in NTUWAS.

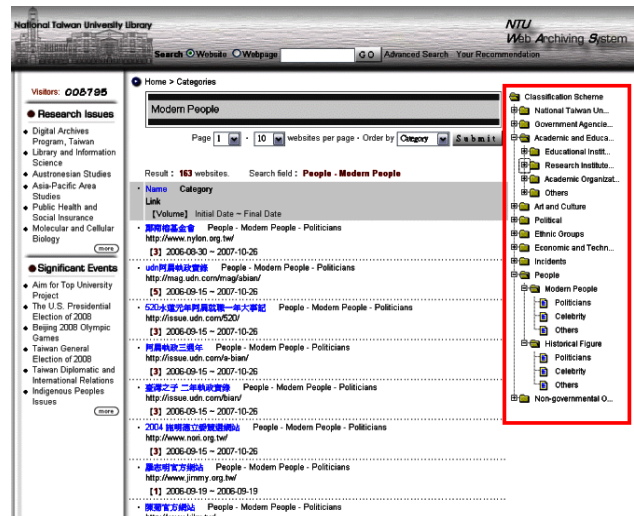


Figure 4. The Classification of Websites

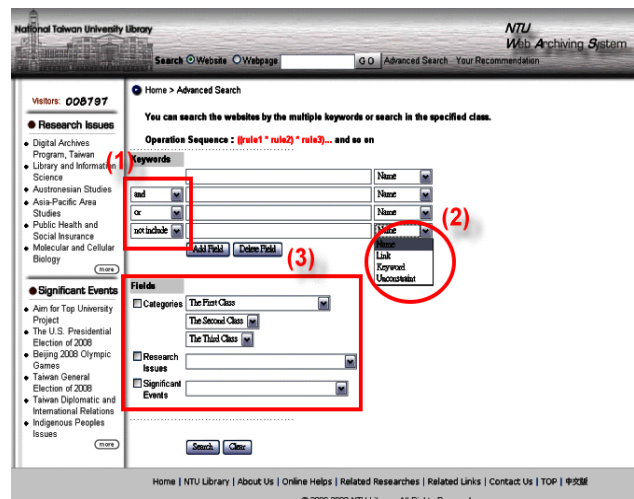


Figure 5. Advanced Search of NTUWAS

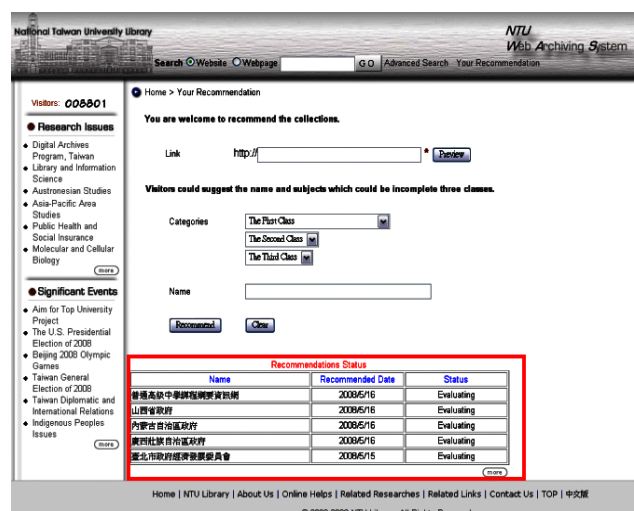


Figure 6. Recommendation of NTUWAS

- Immediate Harvesting

As the matter of fact, the updating frequency of websites could not be predictable. Harvesting websites based on high frequency will waste system resources. In addition, the contents of the websites are changed quickly due to some particular events. For example, some organizations in Taiwan change their names due to political or financial reasons. If we do not harvest this kind of websites immediately, the older version of websites will be lost forever. Therefore, NTUWAS features “Immediate Harvesting” functions which allows administrative staff to click on “harvest right now” button to issue harvest task.

- Pre-setup Harvesting

“Pre-setup Harvesting” is designed for known events which will take place in the near future. For example, the elections or the Olympic Games are known events which will last for a known interval. Administrative staff could set up a series of specified dates in advance to harvest these websites.

3.5.3 Process Recommendations

This part explains how the staff and NTUWAS system deal with users’ recommendations. This process consists of three stages. The first is examination, the second is harvesting, and the third is verification. Figure 8 shows this process. In the step 1, our administrative staff has to examine the contents of recommended website. NTUWAS then will check if this website has been harvested ever before. The staff will catalogue this website by adding the mandatory metadata, such as the class tag, and discuss the appropriate schedule for harvesting it. After setting the schedule, NTUWAS will be ready for harvesting the website in the step 2. While finishing harvesting task, NTUWAS will initiate verification task in the step 3. NTUWAS checks whether the harvested website is complete with helps of staff. If it is not complete, NTUWAS will go back to step 2 and re-harvest this target website.

3.5.4 Register Statistics

Gathering statistics of usage will contribute greatly to the improvements of system functionalities. NTUWAS, of course, records the activities occurring in the duration of users’ visiting, including number of visitors, maximum of visitors, duration time of each visiting, and something the like. These statistics will be used to make decisions, improve functionalities, and select significant websites.

The functionalities fulfilled by management system help staff to manage and operate NTUWAS. Most importantly, we reduce numerous labors and costs and maintain high-performance operation at the same time.

4. Challenges and Difficulties

While carrying out NTUWAS project, we are confronted by many challenges and difficulties. Here we discuss four major ones and would like to share our experiences in dealing with them.

4.1 Quality of Archived Websites

In order to preserve valuable websites, NTUWAS comprises an open-source web crawler, HTTrack, to crawl target websites. Whether the website is composed by ASP files, JSP files, or the other types, HTTrack can transform all files into HTML files and restructure the website accordingly. All images, audio files

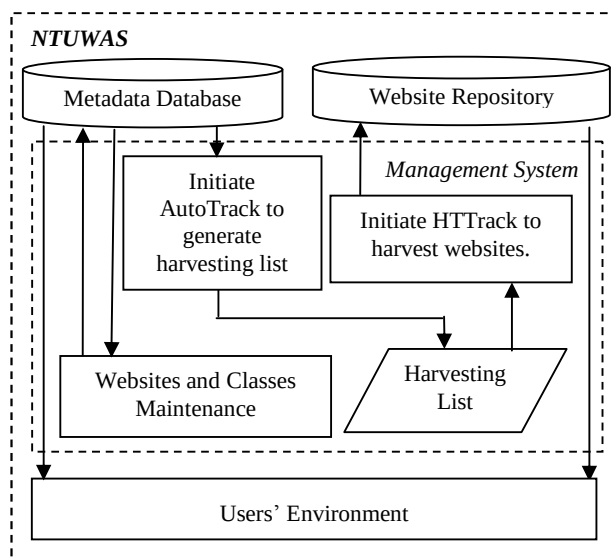


Figure 7. Harvesting Procedure in NTUWAS

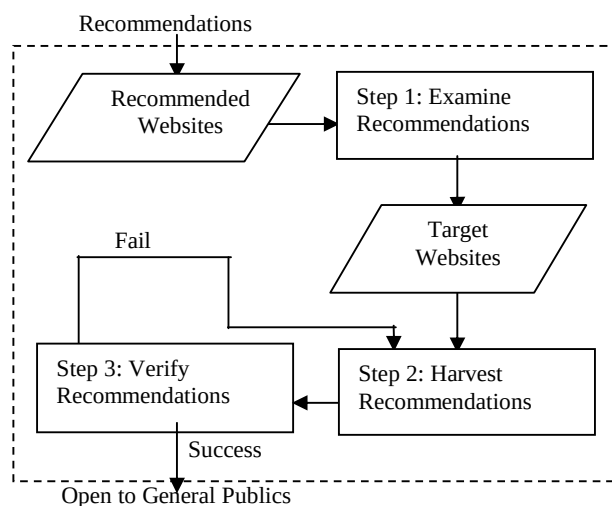


Figure 8. Processing Users' Recommendations

embedded in the web pages will be harvested at the same time. In general, it is very convenient that AutoTrack just generate a list of website URLs, and then HTTrack could harvest those target websites.

Unfortunately, HTTrack still exhibits some shortcomings. Firstly, it is difficult for HTTrack to analyze FLASH files. Although HTTrack could harvest the FLASH files successfully, HTTrack cannot read and rewrite the hypertext references embedded in the FLASH files. Thus, when the visitor clicks the FLASH URL, the browser will be directed to an empty file. In fact, HTTrack may lose many web pages because of the unknown FLASH URLs. To overcome the problem, we have to add those missed files manually and redirect the visitors to the correct pages by the added files. Actually, this is a common problem of several web archiving systems.

The second problem we met is the scope of harvesting. In the current setting, HTTrack just crawls the internal files of the website and leaves the external files alone. This kind of setting prevents HTTrack from crawling files endlessly and then avoids unnecessary harvesting. However, the external files are subject to change. If these external files are modified or removed, some registered websites will be influenced directly. Take a real case as an example. Administrative staff found some archived websites in NTUWAS became plain texts without style format suddenly. The staff checked to see what was going on and found that the external CSS file to which these websites referred was removed.

Unknown dynamic URLs are the third shortcoming. Many URLs of the web pages are produced dynamically by JAVASCRIPT code, like the AJAX technique. Because there are no standard rules to analyze the JAVASCRIPT code, the dynamic URLs are often missed.

The fourth problem for preserving websites with high quality is the change of the website URLs. Many websites may change their URLs when they were greatly revised. In most cases, NTUWAS cannot receive these related messages immediately and then will harvest erroneous or empty websites. When our staff finds out this problem, much information has been missed. To remedy this critical shortcoming, the verification feature of NTUWAS discussed in Subsection 3.5.3 has to be revised in the near future. That is to say, we will verify websites at each harvesting rather than at the first harvesting as the current design. In fact, adopting some semi-automatic process can greatly improve the quality of archived websites.

4.2 Policies and Strategies

Frankly speaking, how to decide the target websites and how to catalogue archived websites are subjective and sometimes controversial. It is not possible to archive all the contents in the Internet for an academic library like NTU Library. We have to setup the manifest goal, clarify the mission, draw up policies, and make actionable strategies. Obviously, it is difficult to make everybody happy. The course of negotiation and compromise takes times but it is worthy of being done.

The practice of catalogue is another tough challenge. To our knowledge, there are no guidelines available for cataloguing websites at present. In contrast, AACR2 has exhibited practical rules for cataloguing various materials from periodicals to monographs within the framework of MARC. We have to design metadata format and cataloging guidelines by ourselves. In addition, web archives as a part of libraries' collections, we have to identify its role within the context of collection development.

4.3 Allocation of System Resources

Actually, the main system and the web crawler will disturb each other, if they are executed in the same computer system. Web crawler consumes lots of computing resources, because it has to crawl the large volume of target websites. It would be better of having an independent server to execute main system and leave web crawler running on another system.

A common issue which all web archiving system encountered is the usage of disc volume for archived websites. The size of a website is unpredictable before hand and a website may be repeatedly harvested for its different versions. Therefore, a web archiving system needs enormous disc space. For instance, the

collections of "Internet Archive" have consumed more than 200 TB. A project of web archiving should have a manifest goal and analyze the needed disc space. Furthermore, a smart mechanism to optimize the harvesting frequency for each website is important. After all, harvesting websites with unnecessary high frequency does waste disc space. Of course, it is best to devise a practical approach to monitor the changing of websites and to harvest them automatically. This exhibits another challenge which is necessary to be overcome in the future.

4.4 Cooperation of Different Professionals

Web archiving is a new and important mission of the library. The library could not disregard this issue. A successful web archiving system needs the well cooperation of technicians and librarians. The tasks of verification of recommendations, classification of websites, and so on truly rely on the librarians. In contrast, system operation and collection maintenance are the job of technicians. The well cooperation between librarians and technicians could result in a successful web archiving system, but there often exists a contradiction among two professions. How to coordinate people from different professions is a critical task.

5. CONCLUSIONS

The newspapers in ten years ago could be found in the library, but the e-papers might not. The preservation of information on the Internet has exhibited a critical issue in the information exploding environment. Much influential information like those documents published by the scholars and scientific resources may disappear along with the time. Furthermore, the websites or weblogs are popularly applied in the political elections. Many candidates announce their policies, opinions, schedules, and achievements on their websites or weblogs. The websites or weblogs will be almost closed at the end of the elections and will provide no contents henceforth. If those significant websites are not archived today, some remembrance will be blank in the future. Not only will the scholars lose important contents for researches, but also the public will lose the passing moments.

In order to take the responsibility to preserve cultural assets for the changing information in digital form, it is necessary for libraries to consider websites as parts of contents of digital repository. In other words, collecting the tangible Internet documents and building the characteristic collections have become one of principal goals of a university library. However, the scope of Internet is unlimited. It is not possible to preserve websites without appropriate blueprints. Libraries have to establish guidelines, strategies, and policies for web archiving to meet their own needs. University libraries could also draw up subjects in terms of characteristics of universities. For instance, a funded institute, "Taiwan's Indigenous People Resource Center", is established in NTU Library. Therefore, NTUWAS contains many aboriginal tribes' related websites to emphasize our feature of collections.

As the matter of fact, historical, cultural, financial, and technical issues should be considered in building and developing a web archiving system. Professionals of different disciplines have to work together to achieve the goal. NTUWAS as the first web archiving system in Taiwan, it demonstrates the possibility to preserve the academic, cultural and historical materials using technical approach. The quick development and sharp changing of Taiwan from the end of 20 century are very special and important

both in spatial and temporal aspects. If the libraries in Taiwan could cooperate together to collect websites and share the digital assets, it will be a meaningful progress.

6. ACKNOWLEDGMENTS

Authors thank National Taiwan University Librarian, Professor Jieh Hsiang, for his support to NTUWAS project. Authors are also grateful to the members of NTUWAS developing team at National Taiwan University Library. They are Su-Chuan Chang, Mei-lin Kuo, Shao-Wei Juan, Ting-fai Ho, Ming Yu, and Wei-chen Liao. The research presented in this paper is partly supported by Excellent Research Projects of National Taiwan University.

7. REFERENCES

- [1] Library of Congress. 2003. Web Capture & Archiving. Retrieved April 22, 2008, from <http://www.loc.gov/acq/devpol/webarchive.html>
- [2] Internet Archive. N.D. Frequently Asked Questions: The Wayback Machine. Retrieved April 22, 2008, from http://www.archive.org/about/faqs.php#The_Wayback_Machine
- [3] National Library of Australia. 2008. About PANDORA. Retrieved April 22, 2008, from <http://pandora.nla.gov.au/about.html>
- [4] Library of Congress. N.D. Technical Background. Retrieved April 22, 2008, from <http://www.loc.gov/webcapture/technical.html>
- [5] Library of Congress. 2007. Web Capture: FAQs. Retrieved April 22, 2008, from <http://www.loc.gov/webcapture/faq.html>
- [6] Arms, W. Y., Adkins, R., Ammen, C., and Hayes, A. 2001. Collecting and Preserving the Web: The Minerva Prototype. RLG DigiNews, 5(2). Retrieved June 20, 2007, from <http://www.rlg.org/preserv/diginews/diginews5-2.html>
- [7] McPhillips, S. 2004. PANDORA: Technical Details. Retrieved June 26, 2007, from <http://pandora.nla.gov.au/pandoratech.html>
- [8] Bailey, S. and Thompson, D. 2006. UKWAC: Building the UK's First Public Web Archive. D-Lib Magazine, 12(1). Retrieved June 26, 2007, from <http://www.dlib.org/dlib/january06/thompson/01thompson.html>
- [9] WARP. 2006. Introduction to WARP Project. Retrieved April 22, 2008, from http://warp.ndl.go.jp/WARP_Intro.html (in Japanese)
- [10] Infomall. 2002. Introduction to Web Infomall. Retrieved April 22, 2008, from <http://www.infomall.cn/> (in Chinese)