Web Archiving Effort in National Library of China

Qu Yunpeng
National Library of China
No.33 Zhongguancun Nandajie, Haidian District
Beijing, China
quyp@nlc.gov.cn

ABSTRACT
In this paper we introduce the effort in Nataional Library of China in recent years, including resources accumulation, software development and works in Promotion Project in China.we have developed a platform for Chinese web archiving. And we are building some sites to propagate our works to the nation. At last we figure out some questions about the web archiving in china.

Categories and Subject Descriptors
http://www.acm.org/class/1998/

General Terms

Keywords
National Library of China, Web archiving

1. INTRODUCTION
Nowadays the web resources in Chinese language are growing in a very fast speed. According to the <29th China Internet Development Statistics Report> from CNNIC, up to Dec. 2011, the total number of sites in China reached 2.3 million. The number of pages reached 88.6 billion, with the Annual growth rate of over 40%. [1]

However, the web resources are also disappearing rapidly. The Chinese web resources are the civilization achievement of Chinese people and the important part of the digital heritage of Chinese culture. They need to be preserved and protected. In 2003, WICP (Web Information Collection and Preservation) Project was found and some experiments were done. After a few years of researches and tests, we made some progress in Web Archiving.

2. Related Research
In the 1990s, web information archiving was focused by some institutions and organizations. The libraries, archives, research institutes started to do research and experiments on Web Archiving. The national library all over the world took Web Archiving as their duty-bound mission. In the Europe and American, some national libraries found their Web Archiving projects, for example, the Mineval [2] project of Library of Congress, the Pandora of National Library of Australia [3], the Kulturarw project of National Library of Sweden [4], and so on. They accumulated much valuable experience for us.

Up to now, the preservation of Chinese web resources are still in the stage of Theoretical research and testing phase. In many colleges and research institutes, digital preservation is carried as an issue. There are two main testing projects for Web Archiving in China. One is the Web Infomall of Peking University. The other is the WICP (Web Information Collection and Preservation) project of National Library of China. The Web Infomall is carried by the Computer Networks and Distributed Systems Laboratory of Peking University, with the support of national '973' and 985 projects. It is a Chinese web history storage and access system. It collected the Chinese web resources from 2002 with the amount of 3 billion pages, and now going with the speed of 45 million pages per day. In the site of Web Infomall, users can view the past pages, and view the selected events pages[5]. In 2003, WICP in National Library of China was started to preserve the web resources.

3. The WICP Project
3.1 The Introduction
In the early 2003, WICP project was found and a team was established. The team consisted of the reference librarians, cataloguers and network managers. 4 problems needed to be solved by the team

i. To find the problems in the collection, integration, catalogue, preservation and access of the web resources, and find the answers to them
ii. Experimentally, to harvest those web information that can reflect the development progress of the politics, culture, finance and society of our country, and provide access to those long-term preserved.
iii. To find the objects, the policy and the measures of the Web Archiving in NLC (National Library of China), so the technical routes and policies can be made accordingly.
iv. To find the scheme of the business and to promote the integration of the web archiving business.

In the early stage, the main jobs are the researches and the software testing. In 2005, with the cooperation with Peking University, we preserved 19968 governmental sites which were registered in China and the domain name ended with 'gov.cn'. From 2006 we start to collect by ourselves. Up to now, we have a collection of web resources about 20TB, including 70 events from 2542 sites and 80000 government sites*harvest.

In 2009, we put the site 'China Events' on web and it can be accessed through internet. 'China Events' are based on the events crawling and preservation mentioned previously mentioned. 'China Events' are organized by the important historical events, selecting the news from archived resources and form multi-events contents. Users can search the metadata and browse the archived web sites. The events in 2008 consist of 10 events, such as the southern snow damage, the National People's Congress and Chinese People's Political Consultative Conference, the 5.12 Wenchuan Earthquake, the Olympics in Beijing, the Paralympics in Beijing, the launching of Shenzhou VII spaceship and so on.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. iPRESS2012, Oct 1-5, 2011, Toronto, ON, Canada. Copyright 2012, Digital Curation Institute, iSchool, University of Toronto.

The events in 2007 consist of 8 events including the 85th anniversary for Communist Party, the 10th anniversary for the return of Hong Kong., the Special Olympics in Shanghai, the 17th CPC National Congress, and Chang'e-1 lunar probe and so on. The events in 2006 include 7 events. They are Construction of new Rural, the Eleventh Five Year Plan, the 2006 International Culture Industries Fair, the 70th anniversary of Long-March, the Opening of the Qinghai-Tibet Railway and so on.

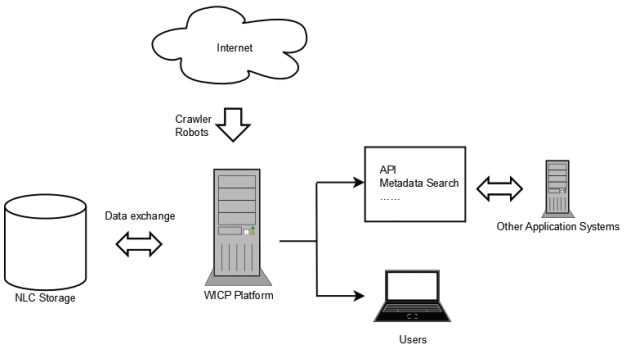

Figure 1. the framework of WICP

Figure 1 shows the framework of WICP. We collect resources from the internet using crawlers and robots, and put the preserved data to the storage. Other systems can use the data and exchange with WICP by AIP and other methods.

3.2 The Harvesting Policies

Comparing to the traditional information resources, web resources are tremendous, wide-spread and quickly increased. Also they are free to publish, and come from all kinds of sources. So they have the characteristics of complication and unevenness. So we consider that we need not to preserve them all. So according to the function of NLC being the repository of the nation's publications and a comprehensive research library, after a adequate research, we decide to use 'full harvest' policy for government sites mirroring and 'selecting harvest' for news preserving. In the ordinance of literature selecting of NLC, it wrote that the selecting of web resources should be done by event; the great or important events about the politics, finance, culture, sports, science and education should be focused.

4. The Distributed Harvesting Platform

4.1 The Motivation

In 2006, the focus of the project was turned to the technical problems in web archiving. After the comparing between the well-known sites and testing on the open source software, we decide to use the software that IIPC provide, including Heritrix, Nutchwax and Wayback.

After a few years of using, we find it inconvenient. There are some points:

i. The guiding documents and the language are written in English, it is not easy for Chinese people to understand the exact meaning.

ii. The open source softwares have their own functions, Heritrix for crawling, Nutchwax for full-text indexing and Wayback for url indexing and accessing. So if we want to finish the whole task, we should switch between softwares from time to time. Especially when we have several servers, it is a annoying job to handle all the stuff in these servers.

iii. The analyzer in Lucene in Nutchwax does not perform good for Chinese language.

iv. Some jobs have to be done outside the software, such as the cataloguing, the authorization for crawling, the statistics, and so on.

So, to solve these questions, we start to design a framework for a integrated platform. It covers all the function that heritrix, nutchwax and wayback have, and make them a smooth workflow including cataloguing and statistics. It can manage the servers in a distributed environment so that we do not have to change from one server to another from time to time. Its UI are in Chinese, supporting multi-language switch. The framework is show in figure2.

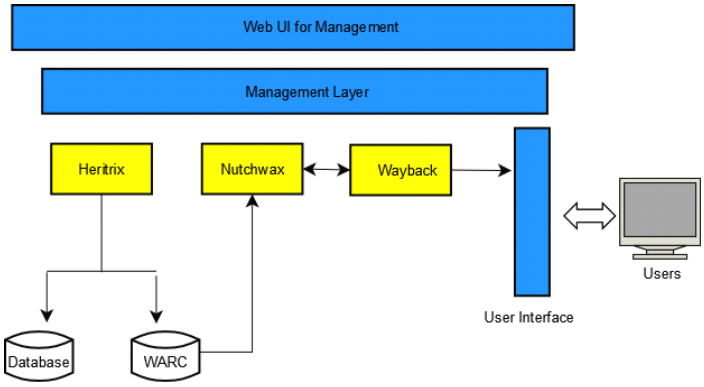

Figure 2. The framework of the Platform

4.2 The framework

After talking to some experts in computer and networks, we decided to put the platform on a distributed storage, for the performance of the platform and easy extension of the space. The Figure 3 is the primary design of the platform. The central controller is the kernel of the platform. Several crawlers are connected to the controller, saving the WARC to the Distributed storage through the controller. The specific information about tasks and crawlers are saved to a database. After indexing, users can access to the pages by the controller.

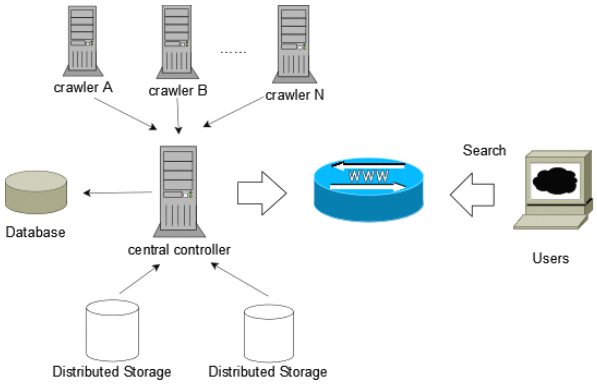

Figure 3. design of the platform

We do research on the open source softwares, and finally decided to program our platform based on Web Curator Tool[6]. It is an open-source workflow management application for selective web archiving. It is designed for use in libraries and other collecting organizations, and supports collection by non-technical users while still allowing complete control of the web harvesting process. It is integrated with the Heritrix web crawler and

supports key processes such as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata. WCT was developed in 2006 as a collaborative effort by the National Library of New Zealand and the British Library, initiated by the International Internet Preservation Consortium. From version 1.3 WCT software is maintained by Oakleigh Consulting Ltd, under contract to the British Library. WCT is available under the terms of the Apache Public License.

However, Web curator tool did not support management of multiple crawlers, and did not run on a distributed storage. So we need make some changes.

First, we must connect the crawlers to the controllers, so that the controller can check the status of the crawlers and assign tasks to them. In the implementation we use socket to connect the controller and the crawlers.

Second, we need to build a distributed storage and put the controller on them. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-avaiability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-availabile service on top of a cluster of computers, each of which may be prone to failures [7]. So we decide to use Hadoop as the distributed storage environment. Figure 4 is the implementation of the platform.

Third, we need change the analyzer to support a good performance on Chinese.

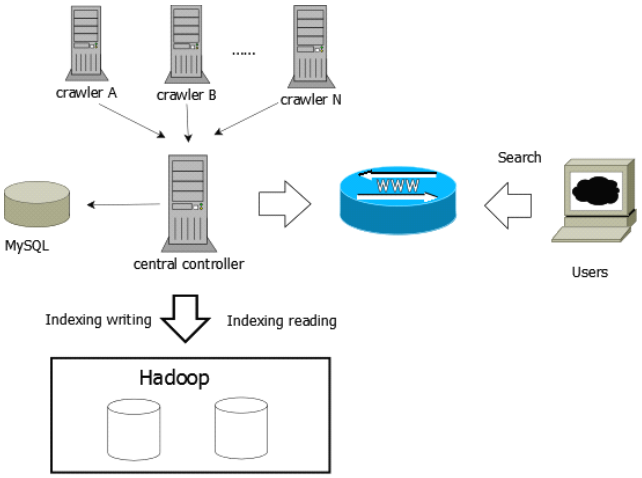

Figure 4. Implementation of the platform

4.3 Functions

Now the platform is under development, but the functions are ok.

It has all the functions WCT have, such as permission, harvesting, configuration and so on. Figure 5 is the main page of the platform. There is a new module in the UI, the cataloguing, which page is in the figure 6.

Figure 5. Management UI

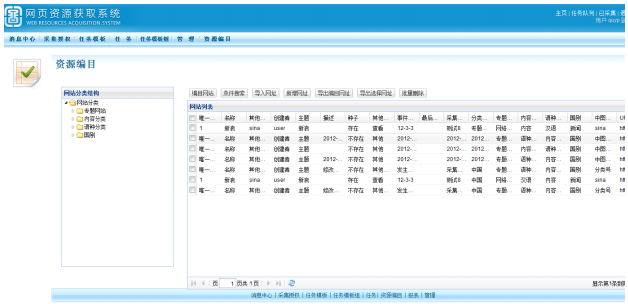

Figure 6. The Cataloguing module

NLC has developed a draft for internet items cataloguing, but it is still trying by experts. In this platform we only adopt several core elements in DC, because the key point of the platform is not on cataloguing. The platform provide a template of excel file. Librarians using the platform can fill in the excel file the metadata and submit to the platform once.

In the Configuration Module we can see the status of the multiple crawlers. When making the target, librarian can choose which crawler to run the task. If not, after 5minutes the system will assign a crawler to run automatically.

5. The promotion Project

The Promotion Project of Digital Library was established by the Ministry of Culture, Ministry of Finance.NLC is main force of the project. The project will apply VPN, digital resources, servers and application systems to the public libraries all over the nation, provincial, municipal and county level.

The platform was included in the application systems of the Promotion Project, as an optional one. This is a good opportunity for us to propagate the web archiving idea and our progress in this field. We have done some effort for the project to propagate itself nationwide. We are building a site for the knowledge of the web archiving and rebuilding the site of 'China Events'. In a few months we will give a training session in the conference of the Promotion Project, for the platform to the librarians all over China.

Web archiving is not an easy job for the public libraries in china. It needs great investment. It needs high- power servers, large capacity storage, enough bandwidth, and many human forces. The Promotion Project solves some of the financial problems, so this is a good opportunity for us to make web archiving understood, accepted and tried. The platform promotion is the first step of it. Next we will organize the libraries and archives together to do web archiving.

6. Conclusion and Future Works

In this article we present the progress of the web archiving effort in National Library of China. We have started the project for about 10 years, and have accumulate 20TB archived resources and much experience. In recent years we made several step, such as the software platform and works in Promotion Projects. But there is still long way to go, both for NLC and for Web archiving in China.

6.1 Legal Problems

The legal problem is the first obstacle that libraries meet when they are harvesting, preserving and using web resources. There are many conflicts against the existing copyright law. In order to solve this problem, many countries permit the deposit of web resources by legislation, such as Denmark, Sweden [8], France, Finland, etc. But in China, the copyrights and existing deposit regulations do not cover the web resources. And there is a blank for deposit of internet publications. That means there is no special laws and regulations for deposit of web resources. In order to form the national deposit system, the web resources should included in the deposit range of the 'Chinese Library Law'. The web resources could be preserved completely by the law.

6.2 Cooperation

Web archiving action involves multiple factors, including policies, laws, finance, technique and management. It is large-scaled, heavy-invested, complex and persistent. Single institute could not take the heavy response and take the hard job. So we need to coordinate all the society resources by the means of cooperation. The starting stage of the web resources preservation lacks the unified planning and no institute is specified to take the responsible to preserve the web resources. So the situation is, some resources are collected by different institutes for several times and human power and money are wasted. Meanwhile, large amount of web resources are left unprotected. In many countries, national libraries are the main force to preserve and archive the web resources. They bring us a lot of references and Inspirations. NLC now are trying to form a cooperation system for the preservation and archiving of web resources, and those are mentioned above.

7. References

[1] 29th China Internet Development Statistics Report [EB /OL], [2012-06-11]. http://www.cnnic.net.cn/dtygg/dtgg/201201/W020120116337628870651.pdf

[2] Library of Congress Web Archives Minerva [EB /OL].[2012-06-11]. http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html

[3] PANDOR [EB /OL]. [2012-06-11]. http://pandora.nla.gov.au/

[4] The Kulturarw3 Project — the Swedish Royal Web Archive [EB /OL].[2009-03-01]. http://www.emeraldinsight.com/Insight/ViewContentServlet?contentType=Article&Filename=/published/emeraldfulltextarticle/pdf/2630160205.pdf

[5] Web InfoMall [EB /OL].[2009-03 -02].http://www.infomall.cn/

[6] Web Curator Tool[EB /OL].[2009-03 -02], http://webcurator.sourceforge.net/

[7] Apache hadoop[EB /OL].[2009-03 -02],,http://hadoop.apache.org/

[8] Zhong Changqing，Yang Daoling. Legal Issue in Web Resource Preservation [J].Information studies: theory and application，2006（3）：281 – 284

[9] Wang Ting, Wu Zhenxin, Gao Fan. Analysis of the International Collaboration Mechanism of the Preservation of Network Information Resources [J].Library Development，2009（3）：6 – 13