Web Archiving in NDL

National Diet Library

1 Introduction

As you know, in the last CDNLAO meeting we agreed to cooperate on four common issues and the National Diet Library (hereinafter NDL) is assuming the leadership for Web archiving.

First, I would like to report on the NDL Web archiving project, and then I will introduce what we are planning to do in CDNLAO as the leader country on this issue.

2 Web Archiving in NDL

In Japan, the number of the Internet users increased 7.6 times and the Internet diffusion rate has reached nearly 70% in the last decade. Nowadays, many people are searching for necessary information first on a computer. It seems that looking for information with cell phones is also increasing rapidly. However, such digital information does not exist permanently and immutably on the web: on the contrary, it will mostly disappear or change. Moreover, unlike print books, we will never be able to find it again once it disappears.

NDL has implemented a Web archiving project called "WARP" since 2002 and collected 2,448 titles of websites. It includes sites of central and local government, corporations, organizations, and universities as well as websites of events such as the 2002 FIFA World Cup Korea/Japan and festivals. Adding 1,827 titles of online journals, 4,275 titles in total have been stored and made available on our website (as of February 2009).

The project is undertaking selective Web harvesting based on contracts with right holders. There are various approaches to collecting websites, but they fall roughly into two categories: selecting sites individually or collecting sites comprehensively under a specific condition.

NDL has also considered a comprehensive approach because a selective approach requires much more labor for selection and contracts, while the scope of collection is quite limited.

We consulted outside experts and got a recommendation in 2004 that NDL should establish a system with legal force including restrictions of copyrights to enable the collecting of a wide range of Web information regardless of the content.

Following the recommendation, NDL set up a draft and invited comments from the public in 2005. We received various comments: whether NDL should collect all Japanese domains including private sites or not, is it right that whole domain harvesting carries the risk of collecting illegal Internet information? Then, we revised the draft and limited the coverage to the sites of central and local government, academic and educational bodies, associations and organizations.

Nevertheless, we failed to win the consensus of related parties. Therefore, in 2008, we decided to narrow down the coverage again to the websites of central and local government and some academic institutions such as national and public universities. We are planning to legislate it within this year.

System environment is also important to implementing Web archiving. NDL has been developing a system for acquisition and long-term preservation, which will be in operation in early 2010.

3 Tasks on Web Archiving in NDL

Now, let me talk of what kind of problems NDL has with Web archiving. The first is the limited scope of harvesting.

As I mentioned earlier, if the Diet of Japan passes the bill, Web harvesting will be limited to a small scale. This means that some of Japan's digital cultural heritage will disappear forever. (We hope to preserve all Japanese websites some day. Although the Internet Archive collects websites from all around the world comprehensively, we should not fully depend on it.)

Concerning selective harvesting, it is hard to choose valuable sites or works from a vast amount of information. Another problem is that Japanese information is not always sent from a server in Japan. This is one of the international collaboration issues.

Next, I would like to talk about technical issues. For preservation of bit sequences, it is

easy to refresh them when a recording medium becomes obsolete. However, that is not enough to make them visible. The problem is that digital information always depends on some sort of format and the format also depends on a particular environment. (For example, file formats of MS-WORD rely on its relevant OS and PCs.) What should we do to open the file in twenty years' time? Indeed, the rapid change of information technology and dynamic expansion of digital information makes us dizzy!

There are many other issues I have not mentioned here. But whatever the problem is, regional and international cooperation would be helpful to find out the best solution.

4. International Efforts for Web Archiving

Efforts for web archiving have already begun at a global level.

We can see national libraries, archives, universities and research institutions take the initiative. But what information to acquire varies from country to country. Some countries acquire their country's websites inclusively; others acquire websites selectively under their own themes and standards. Preservation and capture methods also vary; to preserve the frameworks of websites by saving hyperlinks or to preserve per work. As for capture methods, automated accumulation through hyperlinks can be done by developing a web crawler. Or you can ask authors or webmasters to send the information.

In order to implement Web archiving worldwide, the International Internet Preservation Consortium (IIPC) was started up in 2003 by Internet Archive and 11 national libraries. It now has 38 participants. From CDNLAO members, Australia, China, Korea, New Zealand, Singapore and Japan join in it. NDL joined in April 2008. IIPC's objectives are to promote the development of interoperable tools and technology for Web archiving and also to promote standardization in order to encourage international use of resources.

NDL will contribute not only globally through the IIPC but also internally by introducing the interoperable tools and technology in Japan.

NDL also tries to promote cooperation on digital archiving with national libraries in China and Korea. We will particularly work together on 1) normalization of metadata standards, 2) integrative information providing service, 3) cooperation for long-term preservation of digital contents.

5. Suggestion at CDNLAO 2009

We recognize that the issues on Web archiving should be basically shared with all the CDNLAO members although we are in different stages and situations. Therefore we would like to conduct a questionnaire survey and find out our future direction through its results as FY 2009 challenge. The procedure is as follows:

Step 1) Invite participation in the survey and comments on its question items
Step 2) Fix questionnaire items
Step 3) Conduct the survey
Step 4) Share the results
Step 5) Sort out collaboration themes from the results

NDL has already asked for cooperation on Step 1. The questionnaire items we suggested are shown on this slide:

1) How widely is the Internet spread in your country? What is the Internet diffusion rate? Do you have an approach for spreading Internet use?
2) Do you have a legal deposit or other legal systems for Web archiving? If not, please let us know the reason and future prospect.
3) Do you or will you have a Web archiving project? If you have, what is the scope of acquisition? Do you have a Web archiving system?
4) What are the challenges or obstacles to implementing the Web archiving project? Do you have legal or technical problems or others?
5) Does your country have a framework to share roles in acquisition, preservation and provision of Web information? In that case, what is your role? What institutions are you working together with?

6) What is the scope of acquisition of Web information in your country? What are the criteria to identify the scope? For example do you employ bulk archiving or selective archiving of specified fields? What is the frequency of archiving?
7) How do you think we can cooperate in the Web archiving? What do you expect for other CDNLAO members in cooperating in Web archiving?
8) How do you provide or utilize accumulated Web contents? Please let us know how you use, how you provide them on the Internet, etc. Please give us examples.

We would like to proceed in this way and seek future directions. Please give us your frank comments. We are looking forward to an active exchange of views. Thank you very much for your attention.