Submitted on: 10/06/2015

Lessons learned from twelve years' operation of the Web ARchiving Project (WARP)

Kosuke Murakami
Digital Library Division, Kansai-kan of the National Diet Library, Kyoto, Japan
E-mail address: mur-k@ndl.go.jp

Abstract:

The National Diet Library (NDL) has been operating the Web ARchiving Project (WARP) since 2002, to collect and keep available for future access websites published in Japan. This paper describes the purpose of, history behind, and system used for this project, and introduces actual case studies to demonstrate the challenges faced in fulfilling the potential of this project.

WARP has been attempting to create a comprehensive archive of websites published by public agencies in Japan, as prescribed in the 2010 revision of the NDL Law. It also archives, with permission of the publishers, the websites of private universities, websites promoting cultural or international events held in Japan, and websites related to the Great East Japan Earthquake. As of March 2015, the archived content reached 85,764 items, comprising 533 TB of data and 3.1 billion files. WARP was created using Open Source Software (OSS), such as Heritrix, Wayback and Solr, with some original software and user interfaces.

Publications significant for public use, which are included in the collected websites, are cataloged individually, and made accessible together with other digitized materials. WARP metadata can also be searchable via other integrated search services. Some public agencies even guide their users to WARP in order to ensure access to older information that is no longer available on their own websites.

Since it does not seem practicable for individual public libraries in Japan to conduct web archiving on their own, the NDL will take a step further in promoting WARP within the framework of digital resource sharing programs. We consider this an important part of the NDL's mission as a national library responsible for disseminating cultural heritage through configuration of platforms and networks for digital resource sharing.

Keywords: web archiving, Open Source Software, digital resource sharing, national library, Japan

1. Importance of web archiving for the NDL

As of December 2013, more than 100 million Japanese use the Internet, which is roughly 83% of the total population. The use of commercial e-books and e-journals continues to increase gradually, and much information by public agencies is available on websites under Japan's "open government" policy. The Internet is already a significant part of Japan's social infrastructure.

One salient feature of information on the Internet is how frequently it is updated. Unlike paper-based materials, most of the information on the Internet will be lost forever if deleted. Some websites disappear when the publisher goes out of business. Yet the information may be quite valuable as a cultural heritage, academic resource, and national recording or publication.

The NDL is responsible for collecting and preserving books and other materials, which are fruit of intellectual activities by Japanese citizens, in order to assist members of the National Diet in the performance of their duties as well as to provide library services for the executive and judicial branches of the government and for the Japanese public. Handing down these materials to future generations is another essential aspect of the NDL's mission. Thus, the NDL has a clear responsibility to collect and preserve as well as assure future accessibility for not just traditional library materials but any and all information on the Internet, as well.

Since most libraries in Japan seem to have neither the budget nor the human resources to archive web content for themselves, the NDL, as the sole owner of these related resources, will endeavor to perform this important role.

2. Brief history of web archiving at the NDL

The NDL began studying ways to collect information from the Internet in the late 1990s, and WARP (http://warp.da.ndl.go.jp/) was launched in 2002 to collect websites published in Japan and keep them accessible. At the time of its launch, WARP was a selective, permission-based collection of websites, such as public agencies, universities, cultural or international events such as the 2002 FIFA World Cup, traditional festivals, and open access e-magazines. After several years of trial operation, regular operation began in 2006.

In 2009, the NDL Law and the Copyright Law were amended to enable institutionalized acquisition (i.e., without requesting permission) of websites published by national or municipal governments and related institutions. These laws came into force on April 1, 2010. Since then, the NDL has been using WARP to collect websites of national government agencies on a monthly basis and those of other public agencies on a quarterly basis. Permission is still required, however, to make the archived websites available to users via the Internet. 70% of public agencies have granted permission to allow access to archived versions of their websites via the Internet, while those of the other 30% are available only on the premises of the NDL.

Initially, the NDL considered archiving websites comprehensively, including those of private organizations and individuals. After careful consideration of the impact this might have on freedom of speech and the risk of collecting illegal and harmful information, the scope of

acquisition targeted only public sector websites. The NDL thereafter worked to create another approach to institutionalized acquisition while continuing to collect websites with permission of private universities, cultural or international events held in Japan, and websites related to the Great East Japan Earthquake.

The NDL Law and the Copyright Law were once again amended in 2012, and since July, 2013, private publishers have been obliged either to send to the NDL or allow the NDL to collect online publications that are carried via the Internet. Online publications are defined as corresponding to books or serials, in specific, which are in specific file formats (PDF, EPUB or DAISY) or having bibliographic identifiers (ISBN, ISSN or DOI). Under this form of institutionalized acquisition, the target is limited to specific publications rather than entire websites. But the institutionalized acquisition has not yet come into full force, and at this time, the NDL collects only online publications available for free and without DRM. The NDL continues to study technological issues related to institutionalized acquisition of priced or DRM-protected materials and is negotiating with stakeholders to achieve an effective means of acquisition.

The NDL has been also collecting electronic versions of doctoral dissertations granted after April 2013.

3. Improvement of the WARP system

WARP has archived 85,764 items, comprising 533TB and 3.1 billion files as of March 2015. Figure 1 shows growth of the collection. This chapter describes the system with which WARP collects and preserves these websites.

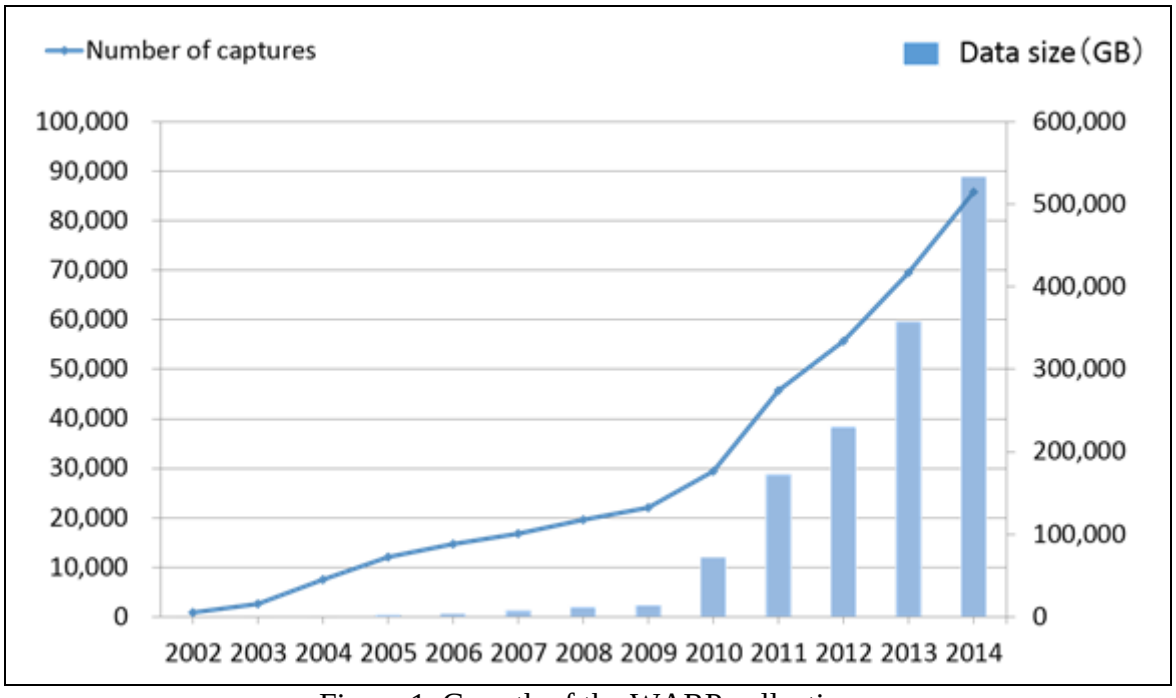

Figure 1. Growth of the WARP collection

The NDL launched the third generation of the WARP system in January 2013. It was built using Open Source Software (OSS), including the Heritrix web crawler, the Wayback

browser, and the Solr search engine. Figure 2 gives an overview of the main components used in WARP.

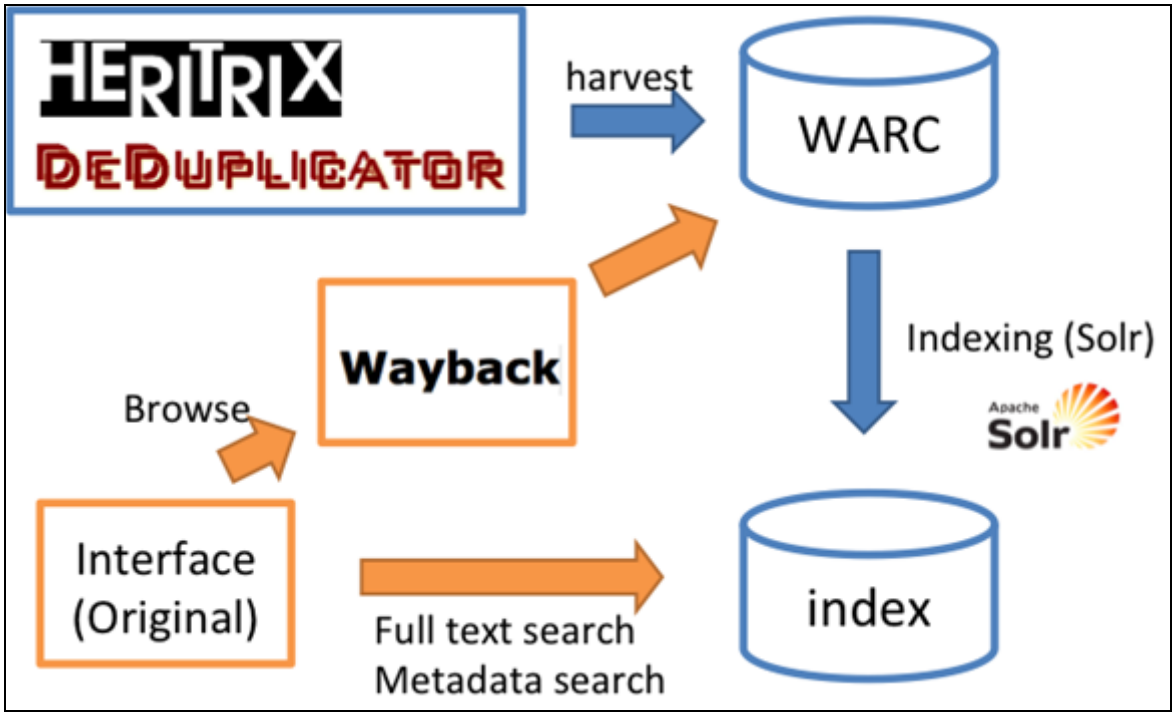

Figure 2. Overview of the WARP mechanism

The first component is Heritrix, a web crawler for collecting data from websites. DeDuplicator is an add-on module for Heritrix, which reduces the amount of duplicate data collected in a series of snapshot crawls. During a three-month trial operation, the NDL found that DeDuplicator reduced the volume of collected data by 70%.

WARP includes NDL-developed software for controlling up to 120 scheduled jobs by Heritrix at a time. Although an OSS called Web Curator Tool was used in the former system, it was unable to sustain performance when a large number of jobs were performed simultaneously, so we stopped using it.

Collected files are converted to and maintained in the Web ARChive file format (WARC), an international standard (ISO 28500:2009). Formerly, it was necessary to reconvert the archived files from the WARC file in order to browse them. But Wayback is able to browse WARC files directly from the WARP system, which obviates the need to have two different formats for one page. This reduces the necessary storage by half, and in combination with DeDuplicator has reduced the total storage by 85%.

The last component, Solr, is a search engine for making indexes and enabling full text search.

The WARP user interface was developed by the NDL. Users can find archived content via full text search as well as browse a map or directory trees of the organizations. The NDL also promotes its web archives by highlighting selected content as *This month's feature* (in Japanese). The NDL has also prepared a guide to its web archiving system and a diagram that visualizes how the websites of municipal governments in Japan are interrelated.

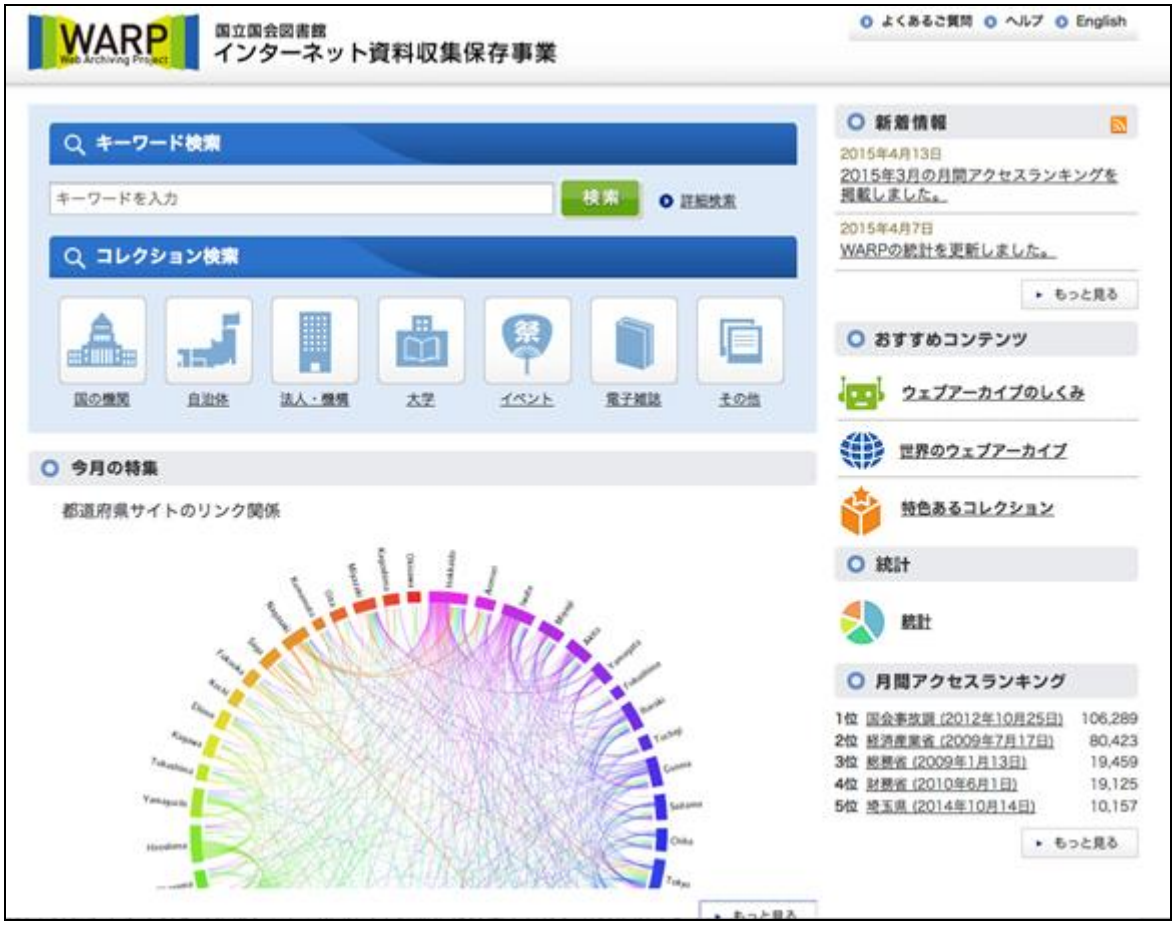

Figure 3. WARP website (http://warp.da.ndl.go.jp/)

4. Some case studies showing how WARP is used

The NDL has been working to make its archived websites more useful than merely serving as an historical record.

Publications significant for public use, such as individual documents published by government agencies and e-magazines on the websites collected by WARP are extracted and cataloged individually to facilitate search and browsing, so that they can be used seamlessly with their precedent materials and other digitized materials. As of April 2015, the NDL had cataloged roughly 280 thousand such publications, which are included in the national bibliography and can also be accessed via the NDL Digital Collection, which contains 2.8 million digitized materials.

Metadata for the websites in WARP is also searchable via NDL Search, an integrated search service for a variety of digital data held by the NDL and other Japanese institutions.

Also, metadata for websites related to the Great East Japan Earthquake is searchable via the NDL Great East Japan Earthquake Archive (HINAGIKU). HINAGIKU is a portal site which enables integrated search of records and reports on the earthquake held by public institutions, private organizations, and the media as well of research published by universities, academic societies, and research institutes. HINAGIKU was built to hand down all of the records and lessons related to the Great East Japan Earthquake to future generations and to utilize them

for the restoration and reconstruction of the affected areas and for disaster prevention measures. In the aftermath of the Great East Japan Earthquake, WARP has been proactively collecting content from the websites of the affected municipal governments, public institutions, NPO, NGO, volunteer aid, and other related organizations. The metadata for the collected websites can also be searched via HINAGIKU.

WARP is used not just at the NDL but by the very public agencies that are targeted by WARP, who refer their users to WARP in order to assure access to old information that is no longer available on their websites. Seen in this light, WARP not only collects and preserves the websites of other agencies, it also helps them keep the volume of their websites at a more manageable level.

5. Toward further utilization of web archiving in Japan

WARP plays an important role as a national web archive on behalf of Japanese libraries, which have only limited resources. Thus it behooves the NDL to publicize WARP and promote its resource-sharing programs to libraries without the resources to do web-archiving on their own, especially for public libraries in Japan.

Public libraries in Japan do an excellent job in both acquiring large quantities of materials published in their local area, and grasping information that is produced there. The NDL supports public libraries by providing training programs and publishing guidelines for digitizing materials, collecting digital data locally, and building digital archives. As mentioned above, the NDL cooperates with public libraries in providing integrated search services such as NDL Search and HINAGIKU, thereby enhancing public awareness of this digital data. Such programs fall under the heading of digital resource sharing.

In contrast, the NDL is still looking to promote utilization of digital data. Increased availability of our digitized materials has stimulated their overall utilization. For example, NDL efforts since 2000 at copyright clearance have enabled us to make about 490 thousand items available on the Internet as of April 2015. Moreover, about 1.4 million out-of-print items have been available on the Digitized Contents Transmission Service for Libraries (in Japanese) since 2014, which enables public and university libraries in Japan to provide their users with on-site access to materials digitized by the NDL. On the other hand, websites archived in WARP are not fully utilized by public libraries, except those that use them for reference services on local administration. In Japan, large-scale mergers of municipalities in the 2000s has reduced the number of municipal governments by 46%, or 1,500 in total. Although there actually are municipal government websites which diminished before they could be archived in WARP, the ones that are available are clearly a significant source of information for learning about and getting to know those areas.

It does not seem easy for most Japanese public libraries to conduct web archiving by themselves. Therefore, the NDL will take a further step to provide even better services by sharing and utilizing resources in WARP, including the framework of ongoing digital resource sharing programs led by the NDL.

As the national library of Japan, the NDL has an important role to play in constructing and operating a platform and network for digital resource sharing to be used for disseminating

information to the general public as well as conveying this information to future generations as records and memorabilia.

We would be pleased if our experiences and lessons learned could be of help to those who are currently in charge of web archiving or those who are planning to launch web archiving in the future, regardless of the size or scope of your project.

Acknowledgments

I express my gratitude to Mr. Yoshiyuki Kanematsu and my colleagues for their kind co-operation and encouragement which help me in completion of this paper.

References

Akiyama, T. (2014). Struggles of the National Diet Library in Collecting Online Publications in Japan. Paper presented at: IFLA WLIC 2014 - Lyon - Libraries, Citizens, Societies: Confluence for Knowledge in Session 87 - Information Technology with Preservation and Conservation and National Libraries. In: IFLA WLIC 2014, 16-22 August 2014, Lyon, France. http://library.ifla.org/id/eprint/886, (accessed 2015-05-01).

Maeda, N. (2013). 10 Years of Web Archiving Project (WARP). Paper presented at: Monthly meeting of the Information Organization Research Group, Nippon Association for Librarianship. In: Monthly meeting of the Information Organization Research Group, Nippon Association for Librarianship, 18 May, 2013, Osaka, Japan. http://warp.ndl.go.jp/warp10years.pdf, (accessed 2015-05-01).

Sato, T. (2009). Archiving of web information at the National Diet Library. Paper presented at: The 28th mutual visit program between the National Diet Library and National Library of China. In: The 28th mutual visit program between the National Diet Library and National Library of China, 24 November - 1 December 2009, Tokyo, Japan. http://www.ndl.go.jp/jp/aboutus/cooperation/pdf/theme1_sato.pdf, (accessed 2015-05-01).

Shimura, T. (2013). Current status of Web archiving of the National Diet Library, Japan. Paper presented at: 2013 General Assembly of the International Internet Preservation Consortium. In: 2013 General Assembly of the International Internet Preservation Consortium, 22-26 April, 2013, Ljubljana, Slovenia. http://netpreserve.org/resources/current-status-web-archiving-national-diet-library-japan, (accessed 2015-05-01).