Document 3

18th CDNLAO Annual Meeting, April 13, 2010
National Diet Library, Japan

CDNLAO Questionnaire Survey on Web Archiving (Q2-Q7)

Q2	Do you have a legal deposit or other legal system for Web archiving? If not, please let us know the reason and future prospect.
Australia	Under current Australian Legal Deposit legislation digital materials are not included. Therefore web archiving cannot be performed under Legal Deposit, and all archiving for PANDORA has to be conducted with individual publisher's permission. An amendment to the copyright law is being sought which will allow for digital materials to be included in Legal Deposit, but at this stage we do not know when that will be.
China	Now we do not have a legal deposit or system for Web Archive. Now Web Archiving is in experimental stage in China, and we will focus on that later.
Japan	In July 2009, the Act for Partially Revising the National Diet Library Law was enacted to come into force in April 2010. This law enables the National Diet Library (NDL) to copy and collect Internet information produced by government and government-related institutions, such as government departments, local authorities, independent administrative agencies and national universities. The legal framework is different from the legal deposit.
Korea	We added a new article, 'acquisition of online materials' (Article 20, 2) in which prescribes legal deposit for worth preserving online materials, to the existing Library Act. This new Library Act was amended and promulgated 25th March 2009, and is currently under process of amending Enforcement Decree and Enforcement Rule.
New Zealand	Yes we have legal deposit. The 2003 National Library Act extended New Zealand's legal deposit to cover born digital material (including websites) and this extension came into force in August 2006.
Singapore	The National Library Board (NLB) Act (Cap. 197) has in section 10, a legal requirement for all published works in Singapore to be compulsorily deposited with the National Library Board. Our legislative amendments to the NLB Act to include the legal deposit of digital and online publications are currently being considered. As such, we have been undertaking the process of getting the written consent of digital publishers [including online publications] to deposit their digital media works with the National Library Board.

Q3	Do you or will you have a Web archiving project? (If you do not, please go to Q6).
3a.	If you have, please describe it briefly, mentioning the name of the project and special features.
Australia	1. PANDORA (http://pandora.nla.gov.au) PANDORA is a selective archive of Australia web publications. Begun in 1996 it contains web publications harvested with publisher's permission. 2. Whole Domain Harvests

	Since 2005 the National Library of Australia has commissioned the Internet Archive (http://www.archive.org) to conduct annual whole domain harvests of the .au domain.
China	We have a Web Archiving project called WICP (Web Information Collection and Preservation). This project started in 2003.
Japan	Launched in April 2004 as a pilot project, WARP（Web Archiving Project）has been in full-scale operation since July 2007.
Korea	The name of the project is 'OASIS' (Online Archiving & Searching Internet Sources), which is mainly collecting web documents including research reports, policy database, and statistics. It also provides users with the opportunity to recommend or contribute materials on the OASIS homepage. Those materials are collected and preserved after evaluating their value as a collection.
New Zealand	We had a web archiving project based around the development of the Web Curator Tool. The project has now become part of business as usual. Another project took place in 2008 when the National Library of New Zealand harvested the .nz domain.
Singapore	The National Library Board (NLB) currently undertakes a limited web archiving project. The name of the project is Web Archive Singapore (WAS).
3b.	Please describe the contents, mentioning acquisition policy and scope of acquisition.
Australia	1. PANDORA (http://pandora.nla.gov.au) As PANDORA is a selective archive of web publications the contents within it are those that conform to our selection criteria. The criteria stipulate that archived publications should be produced by or for Australians, be about Australia or Australians and have long term research or cultural value. Approximately 50% of the contents of PANDORA are government publications; the rest contains websites that give a broad outline of Australian intellectual, social and cultural life. PANDORA Statistics as at 27 May 2009 Number of archived titles 22,104 Number of archived instances* 45,839 Number of files 67,796,218 Data size 2.94 TB * An archived instance is a single snapshot or copy of a title that has been added to the archive. Many titles are copied into the archive more than once to capture changing content, for example, when serial titles add new issues. 2. Whole Domain Harvests The whole domain harvests are not selective; the scope is the whole of the .au domain. The size of the harvests is dependent on the duration of the crawls, which in turn depends on the budget available for the project each year. Whole Domain Harvest statistics Year Number of unique documents/files crawled/harvested 2005　185,549,662

	2006 596,238,990 2007 516,064,820 2008 1,028,604,399
China	We make the archiving policy that we collect the web pages about the events that have great influence on the society, economy and so on, and the sites in 'gov.cn' domain.
Japan	We collect and provide online periodicals, websites of governments and institutions cooperating in this project based on permission from each copyright holder. (From April 2010, we will collect without permission Internet information produced by government and government-related institutions, such as government departments, local authorities, independent administrative agencies and national universities.)
Korea	We collect materials on the principle deliberating utility for present and future demand, reputation of the author, uniqueness, scholarly contents, up-to-date information, update frequency, and accessibility. We also have a policy to collect materials related to Korea or created by Korean authors. Online digital resources of the government, university publication, conference documents, electronic journals, contributed or recommended resources, online digital resources related to recent issues, and domestic web sites take precedence in acquisition. We exclude chatting sites, press sites, boards and news groups, and other resources which are not able to be collected or considered of little worth as our collection.
New Zealand	Selective web archiving: Includes special events such as the New Zealand general elections; topical areas, predominantly in the social sciences – Arts, music, community, Maori, Pasifika, education, but also environmental issues and health. Acquisition policy: Selective within a topical area; we look at a range of factors before acquiring the website: we assess the website's content for research value; risk of loss from the Internet, and technical ability to archive the website to a quality standard. Events: in political events we aim to collect websites across the political spectrum – all party, candidate blogs and websites when possible; election pages from major news sites; lobby group blogs and websites. Whole of Domain: .nz because it's in scope for Legal Deposit. Some additional URLs (e.g. .com sites that were registered in N.Z.) were added that had been identified as being in scope for Legal Deposit by Turnbull Library staff.
Singapore	We have both thematic (selective) and whole domain web archiving. For selective archiving, the National Library Board with the written consent of the website owners, archive about 1000 sites three times a year. For whole domain archiving, we get the list of .SG-registered domains from the local domain registrar for websites to archive. We plan to do whole domain archiving once or twice a year, depending on our resources.
3c.	Please describe the tools you use for the Web archiving system (harvesting,

	preservation and access), mentioning the name of the system and the software and the size of the system.
Australia	1. PANDORA (http://pandora.nla.gov.au) The harvesting for PANDORA is conducted using HTTrack (http://httrack.com). The archival management system we use is PANDAS (http://pandora.nla.gov.au/pandas.html) and was developed in-house by the National Library of Australia. 2. Whole Domain Harvests The harvests are conducted by the Internet Archive using Heritrix; the files are stored in WARC format and delivered via the Internet Archives' Wayback delivery system. In both cases, we are currently investigating our options for preservation management and action, and are likely to use a combination of in-house workflows and tools and externally sourced tools.
China	We have built a Web Archiving system with the free software IIPC recommends, which are Heritrix & Nutchwax & Wayback.
Japan	[Name of the system] WARP(Web ARchiving Project) [Software structure] All the functions of the present system, including collection, preservation and provision, were developed using scratch development. As a crawler, Wget1.8.2, a tool for downloading files on the websites, is used. For full text search engine, a commercial product of Accela Technology Corporation, Accela BizSearch is used. For a new system which will be working from 2010, we adopted Web Curator Tool (Ver1.4.0) for collection function (included crawler is Heritrix). Functions of preservation and provision was developed using scratch development. For full text search engine, we use Apache Solr. [Size of the system] It has a capacity for approx. 28TB of archived contents (including two sets: one for the original collected contents and the other for provision).
Korea	We use a web robot system titled 'WebBee' which is developed by the National Library of Korea cooperated with software company. We apply international standard Dublin Core for metadata.
New Zealand	Selective web archiving: we use Web Curator Tool and Heritrix harvesting software. (http://webcurator.sourceforge.net/) About 3600 websites have been archived in the New Zealand Digital Heritage Archive (NDHA). Preservation of selectively archived websites: The content acquired during selective web archiving activities is stored in the NDHA's preservation repository (currently operational but still under development and scheduled for completion in March 2010). The repository is being designed and implemented in partnership with Ex Libris Corp. and is built around the Rosetta repository application. Web archives are deposited through the Web Curator Tool and undergo a validation process that includes checksum generation for fixity, virus

	scanning, format identification, and technical metadata extraction. The repository adheres with the PREMIS definitions for intellectual entities, representations, files, and bitstreams, and packages data according to the METS schema. The repository will provide format risk assessment, preservation planning, and migration/transformation capabilities. Whole of Domain harvest: 105 million URLs harvested (3 terabytes). The Internet Archive ran this for the National Library of New Zealand. It is stored on a server (not in NDHA) in the National Library of New Zealand.
Singapore	The National Library Board are using tools as recommended by IIPC for web archiving, such as Heritrix, Nutchwax and WERA.

Q4	What are the challenges or obstacles to implementing the Web archiving project? Do you have legal or technical problems or others?
Australia	The National Library of Australia has been archiving from the web for many years, the initial challenges were the lack of basic archival tools for gathering, managing and displaying content. These obstacles were resolved by the National Library investing heavily in creating its own systems for web archiving. There are however still a number of outstanding challenges relating to web archiving, these are primarily: 1. Lack of Legal Deposit legislation. Seeking archival permission for titles in PANDORA is time consuming and costly, while lack of permission for the Whole Domain Harvests means they are inaccessible to the public. 2. Inadequate resources. Selective web archiving requires a large degree of administrative work and technical manual intervention. The Australian web presence is very large and there are not sufficient resources to archive more than a small sample. The Whole Domain Harvests do not require as much human intervention but have a high initial cost to commission. 3. Preservation. There is currently no identifiable single solution to preserve and provide access to web archived materials in the long term; research and development in this field remain critical challenges.
China	Surely we have many challenges and obstacles when doing Web Archiving, such as legal problems and indexing on mass data storage. Everything we collected has something to do with the intellectual property. So we have the legal problems to harvest and publish them on the internet. And we have tried to do full-text indexing on 1TB data, but ran into the java leaking problems.
Japan	[legal problems] As collection of Internet resources affects copyright, it is necessary to have the permission of copyright holders or other legal measures. The Legal Deposit System Council points out that the permanent preservation of Internet resources could wither people's expression of opinions. [technical problems] Realization of techniques of duplication reduction is needed for an effective collection.

Korea	The current web archiving project is not applicable to some web pages containing high technology such as Flash or JAVA script.
New Zealand	Technical challenges – the more sophisticated the website the more difficult it is to harvest, archive and view properly – especially Web2.0 sites. Development of harvesting/viewing tools need to keep pace with current technology and requires resources. Currently javascript/Flash is a major obstacle to archiving some websites. Legal challenges: Legal Deposit in NZ has been extended to e-publications (including websites) but the legislation is limited in scope. New publishing paradigms relating to Web 2.0 and third party hosting have emerged that are not explicitly covered in the legislation. New Zealanders are publishing their content on overseas third party hosted sites. It is difficult to get permission from third party hosts to harvest New Zealand material. Collection development challenges: the legal/technical issues we face means that there are important websites that we have been unable to archive.
Singapore	The National Library Board from time to time, face some technical challenges – such as crawl traps, which often incur down time for investigation and rectification and we remain very interested in sharing access to web archives.

Q5	What access do you provide to web archive contents? Are you able to provide access to the contents to the level you would like? What constraints are there on access?　What other ways do you use or would like to use web archive contents?
Australia	1. PANDORA (http://pandora.nla.gov.au) The PANDORA Archive is freely available online. There are some titles which are restricted due to publisher's request or due to the nature of their content. The restricted material is however less than 1 per cent of the Archive's contents. We would like to improve search interfaces to improve the experience of using the Archive. 2. Whole Domain Harvests The Whole Domain Harvests are all restricted from public view due to the lack of Legal Deposit legislation. Access has been granted to a small number of Australian academics to conduct research. The National Librray of Australia is intending to make the Harvests accessible once legislation is in place.
China	This year we plan a project on our archives about the important events happen in the period of 2005 to 2008. We make some organizations of the metadata of the events on the higher level for browsing, and in detail pages we will give out the link to Wayback pages about these events. Now the web archive content is only available in intranet in National Library of China. The patrons who are onsite at the NLC can also access the content using the intranet of NLC. We would like to provide a full-text index to the content. But this is very difficult. It is so too when we want to do metadata for the archived pages.
Japan	WARP is available on the website: http://warp.ndl.go.jp/. Main texts can be

	searched in addition to bibliographic information such as titles, publishers and subjects. However, some contents cannot be provided on the Internet because of a condition of the permission. For Internet resources collected based on the legislation from 2010, they are provided basically only inside the NDL. Only those for which we get permission will be made available on the website.
Korea	We provide access to OASIS on the web site (http://www.oasis.go.kr). Using keywords or subject index, users are not only able to access the information but also check recent and frequent materials. Materials shown in this process are limited to resources which have permission of copyright holders for their use.
New Zealand	Currently researchers access published born digital materials (including harvested web sites) through the Library's online catalogue (available via the NLNZ web site). They are also able to discover this material through the Library's new metasearch facility called FIND which you will be able to access soon through the NLNZ main web site. They are able to view born-digital content by selecting the link from the cataloguing record. We are hoping to provide a mechanism to enable researchers to search for websites that are in the NDHA or captured during the Whole of Domain harvest. We would also like to greatly improve the visibility of the published born digital collections by having, for instance, the ability to display published digital items to the public to highlight what is held in the collections, e.g. bringing together and displaying digital books, maps and websites relating to Anzac Day; or websites covering the 2008 general elections.
Singapore	We "meta tag" websites before making them available on the Web Archive Singapore website (http://was.nl.sg), which is an archive service open to public access.

Q6	Does your country have a framework or any plans 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?
Australia	The National Library of Australia is responsible for the archiving, display and preservation of Australian web archiving. However PANDAS is a distributed system which allows other Australian agencies (currently 9, made up of 6 State/Territory Libraries and 3 national cultural institutions) to contribute archived content to PANDORA. It is not possible to extend this type of cooperation internationally, but the National Library of Australia is supportive of working in other ways with web archiving partners in the Asia and Oceania regions. The National Library is active in all international web archiving organizations and works with its partner libraries under the auspices of the International Internet Preservation Consortium (IIPC) and IFLA, (via IFLA PAC (http://www.nla.gov.au/initiatives/internat/iflapac.html). The National Library of Australia has sought to build strong links with, amongst others, the National Diet Library of Japan and the National Library of Korea, sharing information and conducting mutual visits to broaden understanding.

China	As far as we know, no. In China there are 3 institutions involving in Web Archiving fields, National Library of China, Peking University and National Science Library. But there is no national framework about Web Archiving because every institution works separately. We are working on it and in the future we may have one.
Japan	No.
Korea	Sharing resources with KISTI (Korea Institute of Science and Technology Information), KADO (Korea Agency for Digital Opportunity & Promotion), and Ministry of Culture, Sports and Tourism, the National Library of Korea provides those resources on the Dibrary Portal (http://www.dibrary.net). We are making an effort to extend organizations and institutions in cooperating in future.
New Zealand	Web Archiving and web preservation is undertaken by the National Library of New Zealand as part of legal deposit requirements. We collaborate with Archives New Zealand who work with the public sector to maintain and archive their web records under the Digital Continuity Strategy. We work with the International Internet Preservation Consortium (IIPC) to develop web archiving tools; web preservation standards; and discuss curatorial issues with other web curators. On a broader level there is the NZ Digital Content Strategy (http://www.natlib.govt.nz/about-us/current-initiatives/nz-digital-content-strategy) which provides information about accessing and creating digital content. NDHA is the library's preservation initiative (http://ndha-wiki.natlib.govt.nz/ndha/)
Singapore	The National Library Board, Singapore is the only Singapore Government agency empowered by the law to preserve published cultural heritage, including website contents as a record of our nation's publishing output.

Q7	What level or role of cooperation in Web archiving do you currently see as possible, practical and useful among the CDNLAO members? What do you expect from other CDNLAO members in cooperating in Web archiving?
Australia	The National Library of Australia seeks to continue to work cooperatively with other libraries in CDNLAO. The National Library continues to conduct testing and research on many aspects of web archiving and digital preservation, this information is made available through its website, through academic publications and international forums. Web archiving is an activity that transcends national boundaries and the solutions on how best to archive and preserve content on the web will need international solutions. The National Library believes that working through the IIPC is one potentially effective way for all libraries to share information, systems and tools for web archiving.
China	We think we can do cooperation on techniques of Web Archiving. It will be great that we can have more cooperation partners.
Japan	The NDL became a member of the International Internet Preservation Consortium (IIPC) in 2008. As IIPC's tools for web archiving are based on software developed for web archiving for Western languages, we think that there

	are several issues to be solved for adopting them in countries where non-Western languages are used. Accordingly, we hope to share with CDNLAO countries information about NDL's web archiving which starts to use tools developed by IIPC from 2009 and to investigate technical issues and their solutions for archiving contents in non-Western languages. We would also like to exchange opinions about issues under study in IIPC, such as systems for long-term preservation of web archives, and other technical and operational issues which are not discussed in IIPC.
Korea	We would like to cooperate with other CDNLAO members at the level of sharing relevant knowledge and materials.
New Zealand	It would help if each institution in CDNLAO was also a member of the IIPC since that is where most of our web archiving discussions and development take place. (http://netpreserve.org/)
Singapore	Sharing of experiences on issues/challenges/practices; Sharing archives that concern the members of CDNLAO like for e.g. the global financial crisis; H1N1; etc - for all to learn and refine work processes in capturing these events.