# Web Resources as Cultural Heritage

International Symposium on Web Archiving

January 30, 2002

National Diet Library, Japan

# Contents

## Profile

### Dr. Brewster Kahle

Brewster Kahle founded Alexa Internet with Bruce Gilliat in April 1996. In June 1999, Alexa Internet became a wholly-owned subsidiary of Amazon.com. At Alexa Internet, Kahle helped build the free Alexa service that provides information about web sites and about products on Web pages. Alexa Internet's services are bundled into more than 80% of Web browsers, and the full Alexa service is available as a companion toolbar.

As an Internet pioneer, Kahle invented the WAIS (Wide Area Information Server) system and in 1989, founded WAIS Inc., a pioneering electronic publishing company that was sold to America Online in 1995. Kahle also helped start Thinking Machines, a parallel supercomputer maker in 1983, serving there as lead engineer for six years.

Kahle earned a B.S. from the Massachusetts Institute of Technology (MIT) in 1982. As a student, he studied artificial intelligence with Marvin Minsky and W. Daniel Hillis. He is profiled in Digerati: Encounters with the Cyber Elite (HardWired, 1996). He was selected as a member of the Upside 100 in 1997, Micro Times 100 in 1996 and 1997, and Computer Week 100 in 1995.

### Cassy Ammen

Team Leader of the MINERVA (Mapping the Internet: Electronic Resources Virtual Archive) Web Preservation Project Team, joined the reference staff in the Main Reading Room in June of 1987. She is a subject specialist and recommending officer for Library Automation, and provides automation support in the Humanities and Social Sciences Division in the areas of publications development, staff training, Windows NT network administration, Integrated Library System Onsite Expert, and serves on the Humanities and Social Sciences' Automation Board and Rep-OPAC team, the latter being a team that answers electronic reference queries. Ammen currently is a member of the Library of Congress' Technology Users Group Planning Team and the Integrated Library System implementation team for the OPAC user interface. During her career she has participated in the development of the Machine Readable Collections Reading Room and the Library's first cd-rom network, served on the CD-ROM Reference Plan Working Group, assisted in the implementation of the BANYAN local area network, worked on the Integrated Library System implementation teams for the OPAC user interface, OPAC testing, staff training and staff workstations, and served on the Collaborative Digital Reference Service (CDRS) Tech Team. Ammen received a MLS from the University of Maryland College of Library and Information Science in 1987 and holds a Bachelor of Science in Music Education from Radford College.

### Margaret E. Phillips

Margaret Phillips has worked in libraries since 1976 and joined the staff of the National Library of Australia in 1987. As manager of Acquisitions she increasingly dealt with electronic materials and from 1996 began to devote full-time attention to online publications as manager of the unit that builds the National Collection of Australian Online Publications (PANDORA Archive). She has been closely involved in establishing policy, procedures and infrastructure for ensuring long-term access to Australian Internet publications. In 2000 she led a team that formed a consortium of State, Territory and National libraries for purchase of access to commercial electronic resources. She is currently the manager of Digital Archiving at the National Library.

### Birgit N. Henriksen

Graduated as cand. mag. in history and computer science from University of Copenhagen, 1987. Worked nine years (1987-1996) in the telecomunication sector as IT consultant and system developer. Works at the Royal Library (since 1996). The past 4.5 years as Head of the Digitisation and Web Department – one of 2 IT departments at the library - with responsibility for (among other things):

-> web archiving projects
-> long time preservation of digital material
-> digitisation processes
-> web publishing and RL's website

### Machiko Nakai

Ms. Machiko Nakai has worked for the National Diet Library since 1975. She has extensive experience with the public service of periodicals, the cataloging of Japanese books and periodicals, JAPAN/MARC and classification. In April 2000, she assumed the position of Director of the Electronic Library Development Office.

# Opening Address

Masao Tobari
Librarian, National Diet Library, Japan

Before starting our International Symposium on Web Archiving - Web Resources as Cultural Heritage, I'd like to say a few words of greeting on behalf of the organizer.

As you know, our Library functions under Japan's National Diet, or parliament. As such, we have performed our mission to collect and preserve an enormous amount of literary materials based on the legal deposit system and make it available to all areas of government as well as to the public. But times change and so have the materials handled by libraries in general. Electronic publications such as the CD-ROM have appeared in addition to the conventional printed word. To accommodate this, two years ago our Library revised its deposit system to include offline electronic publications. On the other hand, the number of publications provided online without a physical medium has sharply increased, and the recent spread of the Internet has expanded the amount of information, particularly as sent via the World Wide Web, at an astonishing rate and includes not only electronic replacements of printed periodicals but also provisional information having various objective values.

Libraries, which traditionally have engaged in collecting, preserving and making available printed publications, essentially books, now find themselves faced with the problem of how to deal as quickly as possible with the recent flood of electronic information. Though belatedly, our Library has begun to resolve the issue and at present we are pouring a great deal of energy into finding solutions.

As part of our effort, we have invited lecturers from countries already engaged in coping with online publications to this symposium, with its theme "Web Archiving," our objective being to share their knowledge on the subject and exchange views on global cooperation. When announced, our plan drew enthusiastic response from both home and abroad, and it delighted us that such widespread interest existed.

We have invited four lecturers. One is Dr. Brewster Kahle, president of Alexa Internet in the United States, which has built a pioneering archive. Another is Ms. Cassy Ammen from Library of Congress; a reference librarian in its human sciences department, she implements plans and practices regarding the archiving of web information. We also have Ms. Margaret E. Phillips, manager of digital archiving at the National Library of Australia, and last but not least Ms. Birgit N. Henriksen, head of digitization and the web department of The Royal Library of Denmark. It is also gives me great pleasure that Dr. Winston Tabb from Library of Congress is here to join our symposium.

I wish to express my sincere thanks to our lecturers for taking the time to be with us here today, and to our many participants. I look forward to active discussions and to rewarding results.

## Keynote Speech

# Public Access to Digital Materials
# Roles, Rights, and Responsibilities of Libraries

Dr. Brewster Kahle
Director, Internet Archive

It is a great honor to be here at the National Diet Library. I have enjoyed my visits to Japan in the past, and this is no different. The combination of a deep understanding and carrying of cultural heritage, as well as the best of the highest technologies make it thrilling for me to visit Japan. I think the combination of those two will also make Japan a very strong player in building a digital library. The cultural interest as well as the high technology can, and I believe will, raise Japan to be one of the pre-eminent digital library countries in the world. It is for this reason that I am very happy to be here.

The opportunity, as we see it, is to offer universal access to our cultural heritage, which is something that has never been possible before. The idea of having a deep collection of our cultural heritage and to be able to make it available to anyone is now possible because of the technologies that are available to us. The storage technology, as we all know, has been going phenomenally up, up, up, up. The costs have been going down, so the storage is capable of being done. Another aspect is the communications, to be able to make it so that any child walking into any library in the world can have access to the best collections anywhere else. This is a fantastic opportunity. But there is a third piece that we have as well that is necessary, beyond the ability to store digitally and to distribute using the Internet. We also have the political will to have an open society. To have a society where access to information is not only allowed, but it is encouraged. It is seen as the important step towards having a successful knowledge economy and a successful information economy. So I believe that these opportunities combine to make something quite unique and it is a wonderful time to be alive now, to be in the library world, and to be in the computer world as they start to come together.

In this talk, I will describe some examples of just one project that is going on. There are many, many projects and some are here, but are many in other places, so do not think that this is even a very important project, but just ones that can show an example of how we have dealt with the technology issues and the rights issues; the issues of what can and should be done, from a law and society perspective. Those two aspects are what I hope to give some experience of what we have done in the United States.

So in this talk, I am going to suggest what are the methods of making public access to digital materials. This involves either materials that used to be on paper or film, and digitized, or things that were born digital, things that started digital, like World Wide Web or email, and the difficulties there. I am going to give a few examples of how a library might work, but I think I should start with an issue that has come up again and again. Some people think that libraries are no longer needed, that it is over. They feel there is no need for physical places, of places to bring things together. No, no, no. They think the publishers will do that for us. The idea that the publishers will go and keep their materials and make them accessible to people, and there is no real need for libraries any longer.

I think that this is wrong, and a way of illustrating this is a particular example in the United States. This is a page from an electronic book. It is Alice in Wonderland, which is a famous children's book that is out of copyright. It was keyed into a computer by volunteers in a Project Gutenburg and made publicly available. A company called Adobe used this material to make an electronic book. If you download this book as a sample, and click to see what permission you have, you have very little. In the bottom, they indicate under copying that no text is allowed to be copied. Can you print it? No, you are not allowed to print it. Can you lend it to someone? No, you are not allowed to lend it to someone. Can you give it to someone? No, you are not allowed to give this to someone. Can you read it aloud? No, you are not allowed to read this book aloud. You are not allowed to read this children's book to your child. This makes no sense, and I think we have seen, at least in the United States, a swing to holding on to information property too tightly. There is a role for libraries to do their traditional roles of preservation and access. Libraries serve the public; companies serve their private interests. Asking one to do the other is wrong. So I believe, and I think we all hold, that there is a role for libraries in preserving our cultural heritage, which may not be profitable, but it is still important.

Here I will give some examples of how we might think of the roles and responsibilities of libraries. Therefore, if these are the roles and responsibilities in the future for libraries, then the right to do these things is part of our society. Here are some examples. The first, which is the subject of this, is a web collection, and I will spend some time on that idea of how to deal with untraditional materials. These are materials that are not classically published. They are made available on the World Wide Web. Another is experience on digitizing a small collection of archival materials, of paper, to also deal with some of the issues of the technology and the rights issues. Another example is movies, and using this as an example of a donation of materials to the public. These are privately held donations to public library institutions. Why would someone do this and how can we encourage more?

Then there are the other aspects that are key to the workings of libraries, which are interlibrary loans. So can one library lend it to another? If one were to walk into a library anywhere else, could they get access to the materials that are in the National Diet Library, the Library of Congress, or any other libraries? We think so. The last topic is loaning of materials. How does one loan digital materials? Is this important to try to preserve, or not? These are some examples of some systems that use these concepts.

I will start with web collections. The Internet Archive has been collecting World Wide Web in collaboration with other companies and groups. The Alexa Internet has been donating a copy of the World Wide Web to the Internet Archive now for five years. The collection is over 100 terabytes (TB) in size, so it is a large collection. There are over 16 million different websites, with over 10 billion pages in this collection. It spans about five years. What has been interesting to me is that it is not very expensive to do. If you are good at saving money, it can be done very cost effectively on a very large scale. This wall of computers stores about 30 TB. So how big is 30 TB? A book is approximately 1 megabyte (MB). If you just take the text in the book, it is about 1 MB. The Library of Congress, I am told, has 26 million books. So 26 million MB is 26 TB. So if you just take the text in the Library of Congress, not the images and the movies and other things that are more difficult to quantify, it is 26 TB in the Library of Congress. The Web Collection is now about 100 TB, so it is larger. I do not say it is higher in quality, but it is a very large collection, and growing larger very quickly. The Web Collection grows at about 10 TB each month, so we buy more and more and more of these.

The cost of these sorts of machines is quite low; it is about US$3,000 for every TB. So if you put it on computer so you have both computer and storage, you find that computer is very important for data mining. It is not very expensive to build even large collections such as 100 TB.

You do not want to just store it, you want to organize it. There are a couple of things I will show on how we have tried to organize a large collection such as this. One is to do automatic cataloging. This cataloging is not as good as a person would do, but it is much less expensive and on a very large number of sites. The idea is to suggest who is behind a website, where is it located, when was it made available, and if there are any reviews or comments from others, that you can have access to these. There is a service from Alexa that does this for free, and it is part of the Netscape and the Internet Explorer (IE) browser in the “What’s Related?” area. The toolbar is a little bit better, but here is an example of the National Diet Library and site information about it. How popular the website is, how many people have gone there, and how many people are linked to it are all indications about this website; it is meta-data or information about this website.

Alexa has also attempted to do subject indexing of websites, so these are related links. They are other websites that are similar or related to this website. Much like on Amazon, it says, “People who bought this book also bought these books”. This is, “People who looked at this website also looked at these websites”. And that is how this is used. It is done by users and passed through the net, and here is a list off of the links off of the National Diet Library’s Japanese page. If one went to the English version, the links would be different, because people who look at the English page of the National Diet Library, look to other websites. Does this make sense? So this is not just at one level, it is for every page inside websites. We have cataloged over 80 million different areas of the World Wide Web using these technologies. It is used by millions of people every month, but it is just a mechanism of trying to have a card catalog for the Internet.

We also tried something else to give access to such a large collection. One is to be able to look at past websites. We started this in 1996 in conjunction with the Smithsonian museum in Washington D.C., where we archived some of the 1996 Presidential election websites. They have used it as a kiosk, as a stand-alone computer, in the museum for people to access these old websites. This work has continued and expanded greatly in cooperation with the Library of Congress, Compaq Computer, and Alexis Internet to create a much larger collection based on the year 2000 Presidential election. I would like to emphasize that these are not just the official sites, but also news sites and other related sites. The idea was to collect it all and to be able to make it available to researchers, historians, and scholars on the Internet.

The approach on the rights issues was to proactively crawl these websites. Without asking specific permission from each one; there are too many to do that. So the idea was to go out there and collect it, but if there were indications that the website did not want to be archived, or did not want to be crawled, then we did not. The idea was to do it like the search engines of Google or Altavista. It is a case of going and doing it, except where people ask to not do it. This has worked very well over the last five years, and where many people will write and say, “Why are you doing this? You are downloading my whole site, everything, my images, everything.” And we explain what it is that we are doing, and most of the time, 90% of the time, people are happy. The say, “Okay, you are crazy, but okay.” But sometimes, they say, “We do not want to be a part of it.” And we say, “That is fine.” Then we take them out and we tell them about robot exclusions. And this has worked out very well over the last five years. We do miss some key sites, and maybe working with the Library of Congress, that they would be more proactive, but when are acting independently, this is what it is we have done.

We took it one step further than just making access to those special collections, to make a Wayback Machine. A Wayback Machine is being able to look backwards in time. The idea is to surf the World Wide Web as it used to be and to turn time back so that you look at the World Wide Web in 1996, or 1997, or 1998, and be able to click around and see the World Wide Web. We do not have copies of the deep web, but if you could click on it before, you can see it again. We think that this is extremely important for preserving our history and to give accountability to our institutions. The World Wide Web is now so important in the United States that it is the primary materials that journalists, businessmen, and students use for their work. We need, as a culture, we believe, to save these materials, so that if they were available before, they can be seen again. If we depend on these materials, we should live up to our society’s need for the preservation of our cultural heritage. The amount of publishing on the World Wide Web is huge. There are over 10 million people’s voices—the writings of 10 million people—available on the World Wide Web today. It is a flourishing of publishing that we libraries should take, I suggest, very seriously.

So the Wayback Machine was made available about three months ago, and we wondered whether people would like it or not like it. And the answer is that people have liked it very, very much. It has been much more popular than we thought it would be. People have used it extensively, mostly to see their own sites. We were very much surprised, by looking at the usage, of where the users coming from. After the United States, the lead country that was using this was Japan. I do not know why, but there have been many, many Japanese that have used this service, and consistently. For instance, on the Alexis service, the other country is Korea. Why? I do not know. But in the Wayback Machine, it is the Japanese. I would like today, to present a gift to the National Diet Library from the Internet Archive, of a small sample of the Japanese websites that have been collected by the Internet Archive over the years. The collection is on a machine that is over here, a desktop computer, and it is storing 20 million webpages from Japan, from 1996, 1997 and 1998.

This is a web browser that is talking to that machine. If I click on one of the letters, this is the letter G; these are the sites in Japan that began with the letter G. It goes on and on; there are many. We can just click one. Here we have this from 1998 and we have two copies. Some have images and some do not, but we worked very hard to get the texts. I thought I would show kantei.go.jp. We have several copies. These are old websites from Japan. Another is Yahoo. So you can type www.yahoo.co.jp and see Yahoo from five years ago. You can visit Yahoo, but also to the websites that it points to if they are within the Japanese domain. So this is a collection that we hope will be useful in starting the Web Collection of the National Diet Library.

So that is what we have done, based on the World Wide Web, and the idea is to go beyond just collecting the World Wide Web. So I realize this a conference about that, but I hope that we can start thinking beyond, because of the opportunities that are with us, is to make many other collections in bulk, available on the World Wide Web. One we have tried using is just taking and scanning pages to get them online inexpensively. Sometimes, for valuable works, it is done expensively and very, very well. This was not to do that. This was to do a lot, inexpensively. We found that at most, it is US$0.1 per page to be able to scan these materials. We also tried to understand the rights issues.

We took a collection of 5,000 pages about the Advanced Research Projects Agency Network (ARPANET) that were used to research a book; classic research materials. But the question is, can you put them on the World Wide Web? There is a bit of question. So after we did the scanning, I talked informally, so this is not on the record, with the head of the Copyright Office in the United States and asked, “What do you do? Can we make these available?” All of these works are copyrighted. Libraries hold copyrighted materials. Can you make them available on the Internet? And her reply, as I remember it, was that it is possible for people to sue you for copyright, but usually they send a letter that says, “Please take it down. Please remove it from the Internet”. Then you have a choice. If you remove it, they often go away happy. If you do not remove it, then there is an issue. So this seemed like a reasonable idea. We would make it available, and if people asked to take it down, we would take it down. So we made it available three years ago, and we have not received any letters to take it down. In fact, we have not received any letters at all. So we do not know if it has been very popular, or unpopular, but the lesson for us was to try and make it available, and see what happens.

Here is another experiment in making an intellectual property preserve. Like a national park, it is a place that is without private ownership. It belongs to everyone. Sometimes these parks are donated or bought by the public from private interests. In the United States, there are tax incentives for people to donate to the public. This is an example of one of these. It is a collection of 1,000 movies that are part of the Prelinger Archives, a commercial company that sells access to their films so that pieces can be used in other films. He donated the right to make these publicly available at high resolution, DVD quality, on the Internet for free. He is looking for a tax benefit for doing this. He also likes the advertising so that this is now a more famous collection, and other people will use his collection for more and more things. It was very inexpensive to make this available.

These films are typically 15 minutes long, they do not have copyright issues, they are older, they are government films, educational films, industrial films that are for one reason or another out of copyright, or he has the rights to use them for this purpose. So he donated those. We scanned the films for about US$200 each to make it available on the Internet for people to do what they want. We have been again very surprised at how people have used these. Many students have been using them to make their own films. So they take pieces and tell their own stories using these materials. One of the great uses of libraries is to help other people learn their own lessons and create something that is theirs. So this is an example of the idea of donation of materials to a public.

As for interlibrary loans, I suggest the interlibrary loan law is very important and can be extremely valuable to us in the digital age. I am not familiar with how this works in Japan, but in the United States, libraries may loan things freely from one library to another. It is possible, then, for someone to walk into a library and have access, theoretically, to it all. In practice, in the physical world, it is very, very slow and very expensive. In the digital world, it does not need to be this way.

Imagine a child walking into a library in Uganda and being able to have access to the best materials. Say this child is HIV-positive, has Acquired Immune Deficiency Syndrome (AIDS), and wants to get at the newest medical journals or the lectures by researchers in this field. This child will be very interested to find out about these materials, if they are available. I think it is our opportunity to make these available to people in every library in the world. This trade-off of restricting it to in-library use can keep it from damaging the publishers excessively. To have universal access to those that cannot pay, or the under-served, or the researchers and scholars, while preserving the financial benefit for those that want to bring it home, or buy it at home.

So interlibrary loan, I suggest, is a mechanism that we can use to cause all of our small libraries throughout the world or within our countries to become world-class libraries. They all become part of the National Diet Library and they all become part of the Library of Congress, if we were to aggressively use interlibrary loan as a mechanism of moving digital materials to those that enter a library.

We do not have any experience yet in interlibrary loan programs. I am sorry to say. But we have tried loaning materials. Loaning materials is the bedrock; it is the fundamental mechanism that libraries use. They buy one copy and they loan it out. A limited number of copies can be loaned to patrons. There have been a couple of commercial attempts at this in the United States, one is called NetLibrary, and it has not been successful. So in the commercial sector, they tried to get explicit agreement with publishers, they ended up with a very small collection, and the company is now bankrupt. Loaning materials, though, I believe in the public sector, non-profit libraries, can succeed.

There is a particular law in the United States, part of the copyright law, that allows for the archiving and the lending of television news. This was done because a library, Vanderbilt University, tried to do this and CBS network said no. They went to a judge and the judge said, "Yes, you may do this." We have used this to make a particular collection available. We took television news about the September 11 events. So it is September 11 to September 18, seven days, 24 hours a day; just that one week. We just took the news from 20 channels from around the world. This is Russian television, Chinese television, NHK from Japan, CBS, NBC, ABC, CNN, BBC, Iraq Television—television from Iraq, very interesting—and Palestinian television. So 20 channels from around the world for one week, and allowed people to comment on it and have scholars use it to understand how the world saw this event that is so important to us. It has been used somewhat by scholars and historians, and has been very positively received. I would not say this a competition for the networks, it is more like a library in providing materials at lower resolution for archival and scholarly work.

That concludes my talk. These are several different approaches towards the technology and the rights issues around building digital libraries. The goal that gets me very excited, and many people very excited, is the opportunity to offer universal access to human knowledge. This is the line from Raj Reddy, from Carnegie Mellon, who is involved in digitizing 1 million books in collaboration with India and China. So he has coined this phrase, and I think it is a good one. "It is our opportunity to do something that will be remembered for many generations". Thank you very much.

Report

# MINERVA
# Mapping the INternet Electronic Resources Virtual Archives

Ms. Cassy Ammen
Reference Librarian, Library of Congress

Good afternoon. I want to express my thanks to all of you and especially the staff of the National Diet Library for organizing this most interesting gathering of people. This is my first trip to Asia, my first trip to Japan, and so it is quite an honor to represent my National Library at your National Library. So thank you very much.

I would also like to acknowledge Winston Tabb who is here with me. Winston is the Associate Librarian for Library Services and he is my boss, so it is important that I recognize him. He is seated here in the front. But he has placed great faith in me to represent my library well, and I hope I do it justice.

The image you see on the screen of Minerva is a very important image to the Library of Congress. In our oldest building, which was erected and opened in 1897, the Minerva icon is the focal point in the great hall of that building, which is a ceremonial hall where many prestigious events are held and where the public is allowed to roam and admire its beauty and its architecture. Minerva stands approximately 10 feet tall and is a mosaic. It is a mosaic as she represents all fields of knowledge, so she is an iconic figure for the Library of Congress that we are intended to preserve knowledge of all humans and to preserve it for future generations. So we took the name of Minerva to try to bridge the past to the present, and to make it a name of our project as we move into a digital realm. So I wanted you to understand the imagery.

Today we can only imagine the content of and audience reaction to the lost plays of Aeschylus. We do not know how Mozart sounded when performing his own music. We can have no direct experience of David Garrick on stage. Nor can we fully appreciate the power of Patrick Henry's oratory. Will future generations be able to encounter a Mikhail Baryshnikov ballet, a Barbara Jordan speech, a Walter Cronkite newscast, or an Ella Fitzgerald scat on an Ellington tune?

This prophetic call was issued in 1996 by the Commission on Preservation and Access and Research Libraries Group in its report, *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* This statement is heavily Western and heavily American in its perspective, but its underlying truth can be applied to any world culture. It speaks representatively to those aspects of our world’s cultural history that have been lost due to lack of technologies to collect and preserve them in their original forms. With a reported 93% of the world’s information existing today in digital formats, our challenge to preserve our digital heritage is formidable. Terry Kuny, a noted researcher in digital studies, proclaimed the following in a 1997 International Federation of Library Associations (IFLA) report: “As we move into the electronic era of digital objects, it is important to know that there are new barbarians at the gate and that we are moving into an era where much of what we know today, much of what is coded and written electronically, will be lost forever. We are, to my mind, librarians and archivists to hold to the tradition which reveres history and the published heritage of our times.” That report was: *A Digital Dark Ages? Challenges in the Preservation of Electronic Information*.

This is an image from the photo archives of the Library of Congress taken in 1897 when the original Thomas Jefferson building was opened and the copyright registrations were moved into the building and scattered about because they were disorganized. The Library of Congress, as well as your own libraries, have always faced such challenges of organizing information—from the preservation of books, maps, journals and manuscripts, to photographs, sound recordings, film, radio and television broadcasting, to machine readable files, geographic information systems (GIS), portable document format (PDF) files, digital trading cards and other electronic publications to the Internet. This slide is a representation is of Internet traffic that was prepared from statistical graphical information in 1999. How we deal with this mass of digital information, disseminated via the Internet, is our focus today.

The Library of Congress’s mission statement is on the screen now and I would like to read it for you. It is a very important mission statement. The Library’s mission is to make its resources available and useful to the Congress and the American people and to sustain and preserve a universal collection of knowledge and creativity for future generations.

For our current time, the time that I work at the library, we have begun to develop programs to add digital content to our research holdings. Part of that endeavor is to collect and preserve open access materials from the World Wide Web. To that end, a multidisciplinary team was created in the summer of 2000 to study and evaluate methods of web preservation. Members of that multidisciplinary team came from our cataloging field, our reference and collection development field, and our information technology field. We were originally known as the Web Preservation Project, but as our pilot project developed, we took on the name of MINERVA, which is of Mapping the Internet Electronic Resources Virtual Archive.

The copyright issues remain, and I do not want to say troubling, but it is an area of study that the Library has to very carefully examine, as the Copyright Office is a part of the Library of Congress. Work is under way at the Library to interpret the Library’s authority that it has already been granted by statutory authority, to move what we have done in our analog forms into the digital context. Interpretations will be consistent with the Library’s established practices for non-digital materials, including our regular safeguards for rights holders’ interests.

Some of the areas of the copyright law that are being studied at present are the Mandatory Deposit authorities, applications under the American Television and Radio Archives (ATRA), which grants the Library authority to capture broadcast materials. A study is also being done to see if the capturing of web materials can be based on the fact that they may be broadcast. We have Preservation Authority granted to us to care for the cultural heritage of the United States. Whatever we do, we will acknowledge that the public has access rights, including interlibrary loan in the digital realm.

In our first explorations into web archiving, we came up with some definitions and we chose our points of study from very early on. We decided in our first pilot projects to take an open access approach, that is that we would select open access web materials, both in a very selective approach and in a bulk approach. It was not until later would we do a closed type of access where we work with publishers. So I will be talking, first and foremost today about open access collection.

The Library had two projects that ran simultaneously in the first studies. One was entirely devoted to library staff taking on the functions of doing the web archiving. We were to select, collect, and catalog websites and then build a prototype access system to test and develop procedures for a production system. Our other pilot was with the Internet Archive, Brewster Kahle's organization, and to work with them and their experience to help us to find sets of digital resources for which the Library will assume long-term curatorial responsibility. We also strove to work with them to gain experience in harvesting and archiving US-based websites, and to use the experience gained to create appropriate selection policies for digital materials. It has been very much a productive, collaborative work with the Internet Archive and I have personally have greatly enjoyed working with the staff of Alexa Internet and Internet Archive.

We are now in a new phase of a pilot, to work with a group that is newly formed known as WebArchivist.org. This is a very small group of two academics, their names are Doctors Kirsten Foot and Steven Schneider, in the fields of communication and political science. The focus of their research is in the Internet and how the Internet affects culture in society. They have been studying how to collect the web, and actually have done collections for themselves for their own research, and they are now beginning to share what they have learned with the Library of Congress. So we will be using some of their web based tools to help us learn how to select web resources and to guide collections decisions such as what is the depth of a website and how frequently do you want to capture it. It will also help us practice an experiment with collecting different type and creating different types of metadata to describe websites, and it will improve access to those collections by a query of a metadata database that describes the websites.

In our earliest attempts at beginning the selection process of what you decide to collect, because the pilot was very small we asked a small group of recommending officers in the Library of Congress to forward to our attention websites of importance in their own field. We asked them not to choose very large or technically complex sites, but ones that were of scholarly content. We asked that they ask themselves if researchers would be glad that the library selected this particular website. We also requested that they focus on timely websites or "at risk" websites, ones that might disappear if they were not collected. And we would study then whether to do a selective or bulk collection of those websites.

For the library pilot, we focused on approximately 30 websites and we focused on three for a much more detailed study. As you recall, the library pilot was done entirely by the staff from deciding what to collect, using software to download the sites, cataloging the sites, and then to provide access to them. So we focused on the campaign sites for the two presidential candidates in the fall of 2000 in the United States and on the White House. In our Internet archive pilot, this was for a set of materials based on a theme, which was the Presidential Election in the United States in the fall of 2000. We began with a list of 150 websites that were reviewed and monitored by our political science recommending officers. As the collection commenced—it began in August of 2000 and went until January of 2001—the collection grew to over of 800 websites, approximately. These websites were collected on a daily basis, sometimes more, sometimes several times a day.

This of course was an interesting election in the United States because of the conflict from the election process. As the emphasis changed after the election, then we began to collect different types of sites that no one had anticipated would be active in the web sphere. So it was an interesting time for us to be doing a pilot about our own electoral process.

Our understanding of collecting websites is very basic. A website is downloaded using a mirroring program. A snapshot is stored in an archive and additional snapshots are made at selected time intervals. Some of the challenges that we faced and some of the things that we discovered, you will find repeated throughout talks you hear about web archiving, these are standard problems.

One of the problems, once you download a website, is that you have imbedded links, you have executable programs that you do not know what to do with, and you have variant languages presented. The architecture of the website sometimes causes strange things to happen when you download a copy. There are problems with persistent naming and date stamping. There are problems with actually defining what is a website. If a link on a website points to another universal resource locator (URL) that is outside the boundaries that you have decided upon, is that a new website to be organized, to be described, or is it part of the first? This is an intellectual decision to decide this.

There are problems with redirection, such as links or movement to another website that you were not anticipating, or to another web address that then your software does not know what to do with that particular move. There is change in content when a website becomes so different from when you first observed it, even though the URL remains the same and its location remains the same, but its content is so very different. How does that impact your description of that website? We found that the administration of managing a system like this is very costly. To go through all these steps, to select, to collect it multiple times, to organize it by cataloging it, and then to provide access through a web based prototype system is an expensive process.

For our cataloging, we made some decisions. Some sites would receive individual cataloging and item level cataloging, while some of the collections would receive a collection level cataloging record that might describe hundreds of websites. We are using a system known as CORC, which stands for Cooperative Online Resource Catalog. This is part of the Online Computer Library Center (OCLC) cataloging system. It is used to catalog Internet resources. Our computer files cataloging specialists would generate a core level machine-readable record cataloging (MARC) using the CORC interface. That would then be imported into the cataloging module of the Library of Congress's Online Public Access Catalog (OPAC) system, and then we would add Library of Congress (LC) classification system, subject headings, a 583 note field and two 856 fields—one for the URL for the current website and then we would also add a "handle" persistent identifier that would point to the archive snapshots. Once the catalog record was completed, it would migrate into the public catalog where people can find it and have information resource discovery of archived websites.

For access, we developed a web based prototype system very much mirrored on the Preserving and Accessing Networked Documentary Resources of Australia (PANDORA) Project. We decided not to reinvent the wheel. This is a very fine access system and we modeled ours after the PANDORA system. I will be showing you some screen shots from that system in just a moment, but I wanted to describe it in brief for you first. It has a section on information and issues. It has search methods that where you can search for websites by title, by subject, and by URL. They library is presently testing a search engine Inktomi, which we hope to use against our MINERVA website.

Each website receives what is known as a title resource page and on that page, the title of that website is listed and the collecting objective, that is it could be collected daily, weekly, monthly, or once. It also contains links back into the catalog record for you to see the bibliographic record. It has links to the active website, if it is available, and then to the archived versions.

This is the web address for the MINERVA prototype, but it is only available from the Library campus. It is not available from off sight because it is still a prototype and not a fully functional system. This is the opening screen. As you can see, there are searches for title, subject and URL, and some basic information about the project. There is an alphabetical listing by title. There are applications made into the broad subject areas of the LC Classification scheme, so that if you linked from political science, you would receive websites that were political science websites. The last search method by URL, which was done in a key word and context format so that there would be entries for Al Gore under "A" for Al, also under "G" for Gore.

This slide shows what we call a title resource page. There is the title is at the top, which is taken from the title of the web page as determined by the cataloging specialist; a link to the LC MARC catalog record; a link to a Dublin Core View which used the Dublin Core identifiers; and a link to the live web, if it is currently available. This one "algore2000" is no longer available on the web. Then there is the list, by date, that they were archived.

This is a screen shot from the Al Gore website. This was soon after Lieberman was joined into the race as the vice-presidential candidate. This, as you can see in the upper left hand corner, is a snapshot taken on August 30. It is very much a website that describes a campaign in progress, talking about one of the major issues for Al Gore, to improve health care. Another screen shot taken after the election was held on November 7. This one was actually taken on November 28 shows a change in emphasis in the website. It is now, instead of being a campaign to win an election, is a site to try to secure monetary contributions to make the fights in the court. So its content has changed even though the URL remained the same.

This next screen shot. If you can read up in the location bar, this is what we call a redirect. Even though we tried to capture algore200.com, we were redirected to another server, to another URL, the goreliebermanrecount.com, which is something our software could not handle, and so we did not actually capture it because it went outside the boundaries of the website that we wanted to capture. This goes to show that even if you decide to capture websites on a regular basis, monitoring must be done to make certain that you capture what you think you are trying to capture.

This is the CORC record that I referred to earlier from the Cooperative Online Resource Catalog. It shows, down the left-hand side, the identifiers. It is just another way of looking at a bibliographic record for a website.

This is the same record in MARC display in our OPAC, the Library of Congress OPAC. If you can notice down here in the 583 states that this is archived at the Library of Congress, DLC is the code for the Library of Congress. And down here in the 856 fields, you can see the URL links to the live website and then the second 856 is a “handle” or persistent identifier assigned to this particular title so that if it were to get moved around on the servers at the Library of Congress, we would still be able to find it. It is a naming convention that helps us maintain where things are stored.

The next concept in working with archived web materials is the concept of long-term preservation. I have been speaking about the MINERVA Project now for a little over a year and this is the first time, in a public presentation, that I have had something to say about preservation. Most of the time this screen has been blank because we do not have direction as yet on how we would pursue preservation of born-digital objects, but I am pleased to say that we are moving in that direction.

The Library of Congress has been appointed the lead agency among several government agencies to develop the National Digital Information Infrastructure and Preservation Program, otherwise known as NDIIPP. This program is intended to develop a program on the national scale to preserve born-digital information. There are many groups working at the Library of Congress and at other agencies to come up with a plan. Several of those groups have been meeting at the Library of Congress, and I had the privilege of serving on two of them—the Preservation Policy Group and the Preservation Metadata Group—where we began to study these issues on what it means to preserve digital objects in their original context and how to think about either emulating them or migrating them, as standards change. Those reports will become public, I think, at some time in the future.

Another activity that the Library is beginning to participate in is a joint project through the Online Computer Library Corporation (OCLC) and the Government Printing Office (GPO). It is a pilot program to capture web documents, not websites, but web documents in an archival program that will be monitored by those two agencies. Libraries will be able to know that a government document is preserved in an archival form and you can link to it or receive a copy of it and add it to their own digital collection. So this is a program that the Library will be studying.

The Library is also pleased to have a consultant, Dr. Margaret Hedstrom, who will be working with the Library of Congress in studying preservation needs for digital content. She is going to specifically look at preserving of websites and geographic information systems and make proposals to the Library of Congress on how to maintain and build a preservation program for digital content.

Now I want to talk a little bit about projects that we have done in cooperation with the Internet Archive. There are two. The first was the Election 2000 collection, which Brewster talked about briefly in his presentation this morning. It is available on the web at the address that you see there. It was a partnership with the Internet Archive and the Library of Congress. The Internet Archive then worked with the Compaq Computer Corporation for their crawling technology. The Internet Archive staff provided quality assurance and project coordination in working with the Library of Congress. Alexa Internet built the Wayback machine, which provides access into the collection.

This is the catalog record. Now this is a catalog record for a collection level, as opposed to an item level. They are very similar, but the link down at the bottom goes to the link outside of the Library, to it being hosted at the Alexa Corporation; it is a full MARCed catalog record.

These are some screen shots from the Election 2000 collection. The access is by using the Wayback machine to type in the URL that you might know of or might like to see if it is in the collection. Or there is an alphabetical listing, also by broad categories of the 800 and some sites that are available in this collection. This is the result of the search for www.georgewbush.com. It shows the daily impressions made from early in August until January 21, the period of time that we collected this material. Once you had hit any of those links, you would go into the archive copy at the Internet Archive.

Our second collection is known as the September 11 web collection. The tragic events of September 11, 2001 prompted web creators around the world to respond. The September 11 web archive preserves these web expressions of individuals, groups, the press, and institutions in the United States and from around the world in the aftermath of the attacks in the United States on September 11, 2001. Memorial sites, tribute sites, and survivor registries were created. In other words, new types of content on the web emerged as a result of September 11. The list of survivors was very heartwarming and the list of the dead was sad. Corporations and non-profits began to solicit money and large amounts of money were collected through the web. This had, to my knowledge, had never been done before using the web as a vehicle for such mass scale charitable giving in the United States. New sites form countless countries dedicated their resources to reporting the disaster in its aftermath, and US government sites sought to inform and reassure the people.

Our record of this event in the United States would not be complete without our having captured this material about September 11. A call went out from the Library of Congress to the Internet Archive on September 12. On September 11, we were evacuated, but when we returned to work on September 12, we immediately began the process of identifying websites to collect. This went on until December 1. So for almost a three-month period, we collected thousands of sites on a daily basis.

It was a joint effort between the Internet Archive, WebArchivist.org, the Pew Internet & American Life Project, and the Library of Congress. For the first time we used mass input from many different parties to try to decide what to collect. The Library of Congress has a program of over 300 staff members who are responsible for building the collections. These recommending officers were asked to examine websites in their subject area, or their format area, and then forward those sites to me so that they could be included in the collection. WebArchivist.org and staff people at Internet Archive also contributed sites. A public appeal was sent out through a Listserv asking for input from the general public and we received many recommendations in that manner.

The collection itself is at present about five terabytes (TB) of data and it includes all kinds of websites. It is available at the website that was on the previous screen for you to see. The cataloging record again, created by the Library of Congress, is in our online catalog so that people can discover this rich resource of research information.

This is a screen shot of the opening page as it sat during the collection period. A user that would come to this site could contribute a site to the archive, they could recommend a site, they could search the archive, search it, or they could visit a random site. The image you see over on the right-hand side is regenerated every time someone hits the refresh button or comes into the archive. They receive a new, fresh image they can just go and explore that one if they were just curious of what was in there. There is some analysis done by the Pew Internet & American Life Project.

We are now beginning some interesting work with WebArchivist.org to apply metadata to this collection so that we will be able to not only search using the Wayback machine, but also search metadata created that describes each of the individual websites. This work will be ongoing and this information added to this site through next summer.

What will we do in the future? The Library has taken on the concept of what we call fanatic collections as an appropriate way for us to continue collecting. We hope to do an Olympic games collection, the Olympic games are in the Unites States in just a few days, and we hope to collect some website information about the Olympic games. We are planning another election collection, this one far greater than the Presidential Election because it will focus on each district House of

Representatives seat and each Senatorial seat that is open. We are looking at a magnitude of a large collection of probably 2,000-3,000 sites. We are also continuing our aftermath of September 11 with the Homeland Security War on Terrorism-type web collection. The Library of Congress has its own Portal to the World that our subject specialists from the area studies sections of the Library have created. We are considering archiving these very important resources that have been selected by the Library of Congress specialists. We are also thinking about Independent Film and how to reach that market that is appearing on the Internet as its only form of distribution.

As for future explorations of what we need to consider in our growing project, we want to continue to study our copyright and access issues. We want to establish selection criteria and recommended procedures for selecting both individual sites and for thematic collections. I think, at some point, the library will consider if we need to do non-selective bulk collection of the US domain, although that is a very frightening concept to me. We need to automate many of the systems. Our information technologies department is presently working on a traffic manager to help in the workflow of once a site is selected, analyzed, shipped on to cataloging, shipped on to be added to our access system MINERVA. All of these things need to be as automated as possible for the workflow to be functional, so we will be working on all of these.

I think we will also experiment with various approaches to indexing and cataloging, automatic indexing, continuing MARC, and Dublin Core. Something we have not yet done, but have tossed around, is to use the Encoded Archival Description (EAD) method of describing. It is presently used to describe manuscript collections, but we think there might be an application to describe large masses of information based on a subject.

We need to continue our planning for preservation. We need to continue to develop partnerships with the library, archival and museum communities on a national and international scale. And, like many times you will hear today, the concept of the deep web, the archiving of databases, is one that looms before us.

In the area of international cooperation, there are some things that I am keeping my eye on and wanted to share with you. The Electronic Resource Preservation and Access NETwork (ERANET) in Europe is a European Union endeavor, which is trying to preserve scientific information. This might be a source of information for us in the library field. UNESCO will be discussing, in its next budgetary year, preserving our digital heritage so that it might become a program in UNESCO. Next fall in Rome, the European Conference on Digital Libraries will be held. This past year in September, at this conference there was a workshop on preserving digital heritage that several of us attended. I attended and one of the next speakers attended. We think that there will be another workshop, so some of you may wish to travel to Rome to participate in that conference. IFLA and the IPA, the International Publishers Association, have issued a joint draft statement on the archiving and preserving of digital information. This is something we need to keep eyes on and see how IFLA might become a supportive institution in preserving digital information.

If you like to subscribe to Listservs, there is a Listserv specifically devoted to web archiving. It is hosted by the French National Library, and the address is there if you would like to subscribe.

There is another organization that we have been introduced to by wedarchivist.org known as the Association of Internet Researchers. This is a group of academics from around the world who have studied the impact of the Internet on society. They are holding a conference in the Netherlands next fall, and they have an interesting website where they talk about their conferences. This is an organization that I think might be a benefit to the web archiving community.

If we are effectively to preserve for future generations the portion of this rapidly expanding corpus of information in digital form that represents our cultural record, we need to understand the costs of doing so and we need to commit ourselves technically, legally, economically and organizationally to the full dimensions of the task. Failure to look for trusted means and methods of digital preservation will certainly exact a stiff, long-term cultural penalty. Thank you very much for your time.

Report

# Archiving the Web:<br>The National Collection of Australian Online Publications

Ms. Margaret E. Phillips
Manager, Digital Archiving, National Library of Australia

Thank you. It is a great privilege for me to be here with you this afternoon. This is my first visit to Japan, as well, and it is a wonderful country from what I have seen already. It is a wonderful experience for me to be talking to you on a topic about which I have a lot of enthusiasm. I have been working in this area for five years now and when I think back to when I was training as a librarian 25 years ago, I thought that I was choosing an interesting and a rewarding profession. Little did I dream that I was choosing a very exciting one and that I would end up working in an area of such key challenges for the library profession. So this afternoon I would like to talk to you about what we are doing at the National Library of Australia in this area of web archiving.

Recognizing that online publications are an intrinsic part of the documentary heritage, the National Library of Australia, together with a number of partners, is building the National Collection of Australian Online Publications. The purpose of the National Collection is to ensure that Australians of the future will be able to access today's significant Australian information resources. This paper describes what we are doing, why we are doing it, and how. It also discusses the lessons we have learned along the way, and aspects of web archiving that we still have to address adequately.

I do not need to tell this audience about the challenges to libraries in collecting, storing and conserving the digital materials for which they are responsible. You know these challenges and you are facing them. That is one reason why we are all here today, to learn from each other. So I will go straight on to the situation in Australia and what we are doing to address these challenges.

In Australia there are three layers of government—national, state and local—and the national and state libraries are responsible for collecting and preserving the documentary heritage of their respective jurisdictions. At the national level, the National Library is responsible for the published output of the nation as a whole. The State Libraries are responsible for building collections of works published in their respective states, and there is scope for duplication here. Each jurisdiction has legal deposit legislation, some of which currently includes electronic publications, and some of which does not. I will speak about the national legislation further on.

The print collections of the National and State Libraries have necessarily remained quite separate, though resource discovery has been integrated through the National Bibliographic Database. This is a national union catalog that records the collections and holdings of 1,100 Australian libraries.

In the online environment, the National Library quickly saw not only the possibility of, but also the necessity for, collaborating with the State Libraries to build one national collection of Australian online publications. Especially in the case of freely available Internet publications, it is possible to collect just one copy of a title in order to provide access to all Australians, and indeed, to everyone in the world. Since the management of electronic resources is such an expensive business, sharing the work and the cost would enable a collection of greater scope and depth to be developed.

In Australia at the national level, legal deposit provisions are contained in the Copyright Act 1968, which is currently under review. In 1968 we were not thinking about electronic publications, and so the Act does not include provision for them. In a joint submission of the Copyright Law Review Committee by the National Library of Australia and the National Film and Sound Archive, the following was recommended:

"…that the scope of publications to be covered by the legal deposit provisions of the revised Copyright Act be extended to include microforms, audio-visual materials of all kinds, and electronic publications, both networked and artifactual, for instance CD-ROM, and all formats yet to be developed, in addition to the print-based publications that are currently included."

This approach to extending legislation is supported by the International Conference on National Bibliographic Services held in Copenhagen in November 1998, which reaffirmed the value of legal deposit and recommended that states should, as a matter of urgency, examine existing deposit legislation and consider its provisions in relation to present and future requirements and, where necessary, existing legislation should be revised.

A key issue for legal deposit legislation and the libraries that administer it is the definition of "publication", a concept that has become blurred in the online environment. Are all websites necessarily publications? Which documents on a government site or other organizational websites are publications to be collected by libraries? Which are organizational records to be collected and preserved by archives?

While it will be advantageous to have legal deposit provisions for online publications, it will not be a panacea for all our problems. The Library wants to archive any publication it deems to be of national significance, but it does not want to be obliged to archive all Australian online publications and websites. It wants the right to remain selective. It is likely that the Library will find itself called to account for its selection decisions more than has been the case in the print environment, and we will have to be able to justify these decisions. A clear collection development policy will be more important than ever.

Legal deposit legislation will not remove the necessity for liaising with many publishers to ensure that publications are complete and accessible. We will need to continue to have their assistance with archiving certain types of files that are difficult for us to download. In the case of commercial publishers, we will still need to negotiate access to publications that are secured in some way, for instance, by password, and we will still need to negotiate periods of restriction so as not to jeopardize their commercial interests.

At the National Library of Australia, in the relatively early years of the web we accepted that online publications are important social, intellectual and cultural resources and that, with or without legal deposit legislation, our statutory responsibilities in relation to this material were no less than for the material we traditionally collected. Ensuring long-term access can be seen as a two-step process. First, the materials have to be identified, collected and made accessible in their current, or native formats. That is the archiving process. Second, the materials have to be managed in such a way that they remain accessible as technology changes—this is the preservation process.

The National Library has always recognized the importance of both these steps, and their interdependence. For practical reasons, the first step received most attention when we set up the PANDORA Archive in 1996. The preservation process has, however, been the growing focus of the Library's efforts as the archive has moved beyond the proof of concept stage to the operational National Collection.

The name PANDORA is derived from its goal of Preserving and Accessing Networked DOcumentary Resources of Australia. In English, that creates the acronym of PANDORA. I do not know how that works out in Japanese. For a number of years now the work has been a mainstream operational activity, and we refer to it as the National Collection of Australian Online Publications, although the name PANDORA is still used interchangeably, and with some affection I might add, for it.

The National Collection of Australian Online Publications is an extremely selective one, containing to date only 2000 websites. In comparison with the volume that Dr. Kahle was talking about to us this morning, it is a tiny archive. Nevertheless, it already constitutes a strongly representative sample of Australian web publishing by academic, government and commercial publishers, as well as community organizations that are publishing online. A number of the websites captured in the archive, including the official website for the Sydney Olympic Games, have already disappeared from the live Internet. Moreover, about one-third of the sites have been captured on multiple occasions, allowing the gathering of successive issues of serials, and enabling the collection of a sequence of snapshots, which demonstrate how some sites have changed over time.

The Collection now comprises almost 11 million files, and uses 320 gigabytes (GB) of storage. It is growing at about 500 new titles each year. The number of titles we can add each year grows with the increasing efficiency of our staff, and we are also adding about 400 re-gatherers every year.

The gathering of titles is undertaken by the National Library as well as its partners. Our partners are the State Libraries of Victoria, South Australia, and New South Wales; the Library and Information Services of Western Australia, the Northern Territory Library and Information Service, and Screen Sound Australia, the national film and sound archive. In addition, the National

Library cooperates closely with the State Library of Tasmania and its independent web archiving venture, Our Digital Island.

A primary goal is to provide immediate as well as long-term access to readers both within the buildings of the National Library and our partner libraries, as well as to readers anywhere else in Australia. To this end, the permission of the publisher is sought and received prior to the site being included in the Collection.

I have already referred to the fact that we do not yet have legal deposit for online publications, for electronic publications, so we have to first seek access permission from the publisher to make a copy into the archive, but even when we have legal deposit legislation, it is likely that we will still have to seek permission to provide the broad networked access that we like to be able to provide. We want to be able to provide access beyond just the reading room of the library.

All partners gather websites and publications for the National Collection using the Digital Archiving System, which was developed by the National Library, and which I will describe in more detail later in this paper. All material, including that gathered by other partners, is currently stored on the National Library Server and its associated storage systems. At some time in the future, we envisage that our partner libraries will want to store their files on their own servers, but at this point in time, the National Library is providing that support.

The gathered version of every publication is subjected to quality checking to ensure that all files have been correctly captured. Each website or publication is fully cataloged, and the catalog entries are included the Library's own catalog, as well as the National Bibliographic Database. This policy is based on the Library's firm view that access to all information resources, digital and traditional, should be integrated. This approach has been endorsed by the International Federation of Library Associations (IFLA), which supports the treatment of online publications as part of the national imprint, and incorporated into the National Bibliography.

As an alternative access pathway, titles in the National Collection are accessible via alphabetical and subject lists on the PANDORA homepage. We are also in the process of building a search engine that will give access to full text over the PANDORA archive, and we are hoping to be able to implement that new facility within the first half of this year.

Within the National Library, the work of building the Collection is now a routine part of the selection, acquisition and collection management processes of the Library. Organizationally, it is managed by the Electronic Unit, which forms a part of the Technical Services branch in the Collections Management Division of the Library.

The Library and its partners do not attempt to capture everything published online that relates to Australia, or is published by an Australian. A policy of selective acquisition has been adopted, partly dictated by resource issues, but also by the determination that everything in the Collection should be accessible. In order to make that possible, we need to negotiate with publishers, and that is a time-consuming and expensive process.

To govern the selection of online publications and websites for which it takes responsibility, the Library has developed a set of guidelines. They can be found on our website. These guidelines take account of characteristics such as Australian content and Australian authorship, whether the publication is indexed by a recognized indexing service, and stress the authority of the author, the research value, and the quality of the publication. Publications are selected for their research value in their own right, but also as part of their ability to form thematic collections, which provide a broader sense of what the Internet is like in Australia at a given point in time. What are the issues that Australians are concerned about and expressing opinions on?

Up until now, the selection guidelines have excluded publications with print equivalents, since in the early years of our work we needed to limit its scope to manageable proportions. We estimate that including publications with print equivalents would double the number of publications that we would need to archive. However, we are constantly reviewing the selection guidelines and we aim to be as inclusive as limited resources permit. As we gain in experience and efficiency, we can expand our selection guidelines and take in more material.

The policy of excluding those publications with print equivalents has already been relaxed to a certain degree, when two years ago we invited indexing and abstracting agencies to enter into partnership with us. These services were already finding that some uniform resource locators (URLs) cited in their references to electronic publications were defunct. The Library now has arrangements with six services, whereby they notify us of publications that they are referring to, and we archive them, even if they have print equivalents.

The National Library of Australia has deliberately pursued the selective approach to archiving for its advantages of quality control, and permission from publishers to archive and provide access. However, we have remained interested in the work of other agencies such as the Internet Archive and the Royal Library of Sweden, which are aiming to collect and preserve entire domains. We recognize that there is value in a comprehensive approach, and we have had initial discussions with the Internet Archive to ascertain whether cooperation with it would enable us to extend our collecting of the Australian domain.

To facilitate the building of the National Collection, a digital archiving system has been developed by the Library, with the latest version implemented in June 2001. Specifications for the system were devised by the Library's Electronic Unit, and the system was built by contract staff using WebObjects as an application development environment. The system is designed to support the following functions: manage the metadata about titles that have been selected for inclusion in the Archive or rejected. For instance, this information includes the title, the URL, publisher details, gathering schedule, and whether access conditions apply. Other functions include: initiate gathers of titles to be archived; manage the quality checking and problem fixing process; prepare the item for public display and generate a title entry page; manage access restrictions; and provide management reports.

The digital archiving system is a web-based tool that uses Internet Explorer 5.5 to enable the partners to achieve a more efficient workflow without the need for special desktop software. Currently all titles, including those archived by partners, are stored and delivered from the National Library's server. The system will support distributed content as an option for the future, so each of the partner libraries could have their own digital archives, but still use this digital archiving system software.

Publications stored in the National Collection are to be preserved for posterity, and they may be cited in many places, including other web documents, scholarly articles, library catalogs and databases, indexes and abstracts, and bookmark files within browsers. They therefore need to be accessible indefinitely, and require persistent identification.

In the year 2000, the Library engaged a consultant to provide advice on the direction it should take in relation to persistent identification of digital objects. We need to be able to identify not just those objects in the PANDORA archive, but  all the objects in our digital collections, including those that we create as a result of digitizing our traditional library materials. The report produced by the consultant has been valuable in assisting the Library to determine its policy on naming standards, and its procedures for managing digital information resources for persistent access. In the absence of a global operational system of persistent identifiers that meets the National Library's needs, the Library has implemented the guidelines for persistent identification recommended by the consultant and they are reproduced in the appendix to this paper.

The resulting persistent identifiers are automatically assigned to titles and component files in the PANDORA collection by the Digital Archiving System. They are not dependent on a domain name to provide unique identification of the resource, but could be used as the namespace specific string in the context of any naming system implemented at a later date.  In the future, therefore, if a global scheme becomes available and viable for us to use, we could use this existing persistent name that we have adopted as part of that persistent identifier of the future. This standard aims to ensure that each file will have a persistent, unique identifier, that there will be sufficient granularity of identification to support preservation activities, and that there will be sufficient intelligence to enable a measure of grouping and relating of versions.

Since June 2001, when we implemented the latest version of the digital archiving software, we have had the ability to cite the persistent identifier on the title entry page for every title in the PANDORA archive. While this has been an important step forward in the promulgation of the use of persistent identifiers, it does not assist those researchers or indexing agencies that need to refer to part of a publication, for instance an article in an e-journal. The new version of the digital archiving software that is coming out in March will provide the persistent identifier for every file in the Archive.

Access restrictions are applied to a number of titles in the Collection, and are managed by the digital archiving system. Restrictions may be applied for commercial, privacy or cultural reasons, or as part of a policy decision that we have taken for certain categories of material. Items can be restricted for use only within the buildings of partner agencies for a specified period of time, or can be password-protected so that only designated researchers can obtain access.

To date there is little commercial publishing on the web in Australia, with most of the publications in the Collection being freely available. However, there are a growing number of commercial publications in the Collection, and we need to be able to cater for them. When a commercial title is selected, the Library negotiates with the publisher to determine access conditions that will not undermine the publisher's commercial interest. Our preferred model is for the publisher to allow use within the reading rooms of the partner agencies for an agreed period of time, after which the title would become freely available to external users. External users might be in Australia or anywhere in the world, for that matter. Periods of restriction negotiated to date range from three months to five years, and agreements of this type have now been reached with publishers of about 80 commercial e-books, e-journals and other publications.

The Digital Archiving System manages all of these restrictions, and allows only those users in designated locations, based on IP address or with the required password, to gain access to archived versions. As time periods for restrictions expire, the system automatically updates the title entry pages to indicate the changed access conditions. The system checks every title each night to see whether those access conditions have expired.

The ultimate purpose of the National Collection is to ensure that the Australians of the future will be able to have access to significant Australian information resources currently being published on the web, which is their digital heritage. As mentioned above, the first step in managing an archive of online publications involves creating and managing a safe place where digital resources can be stored. The second and vital step is to ensure that the resources remain viable and accessible in the long-term.

Given the diverse collections and the formats for which it is responsible, the Library will use a combination of preservation methods for the foreseeable future, including some technology preservation, such as maintenance of software, and even some hardware; negotiating with publishers to supply stable source files of some streaming or dynamic formats; migration strategies for those file formats which correspond to compatible new formats, and which are amenable to mass conversion; the use of emulators if they can be found or developed, for some file formats; and simply keeping and refreshing some titles not amenable at this stage to migration or emulation, in the hope that a suitable access pathway will emerge.

The Library stores and manages at least two copies of archived files, maintaining the original copy unchanged in its native formats as the preservation copy, and creating a second copy that may have to be manipulated in some way for access. When preservation action is required, it will take place on yet another copy, which will be created for that purpose. This procedure will safeguard the collection in the situation where, for example, a series of migrations leads to a dead end. We will always have the original files to go back to.

In 2001, a very small trial migration of 127 files was successfully undertaken. It was a start and it was successful. In preparation for this, the Library had identified file formats held in the Collection, with a view to establishing an effective preservation path for each format. It had also identified dead and deprecated hypertext markup language (HTML) tags in the Collection, which may not be recognized by future browsers.

This analysis found 127 tags already dead in HTML 4.0, which is why we migrated them. There were 7 million tags due to be made non-standard in later versions of HTML, and 14 million with some deprecated attributes. Clearly, our future migration tasks are going to be much larger.

Another preservation initiative is a working group of staff involved in digital archiving, with the task of identifying significant properties of various types of digital objects, including those stored in the PANDORA archive. We are not only looking at PANDORA files, but also at files in other digital collections within the Library. This group is attempting to answer the question, in relation to any given digital object or class of objects, "What exactly is it that we want to preserve?" Is it just the content or is it the full experience, the look and feel, of this digital object?

The Cedars Project Team in the United Kingdom has also grappled with this matter and defines significant properties as the level of content and functionality retained by the archiving and preservation process. The Cedars Project report notes that the "significant properties of an object, as they have been agreed, may alter the costs associated with preserving those specific properties. The preservation of an object's full functionality may prove more costly than just a bare-bones preservation of the basic intellectual content. The question that has to be asked is whether the object's long-term value is worth the long-term expense of preserving the 'bells and whistles'".

Defining significant properties of different kinds of digital objects in the Collection will help the Library to do a number of things such as understanding the full scope of its preservation task better; setting up quality control criteria for judging whether preservation action has been successful; allocating resources to preservation work that is really needed; and determining the level of metadata required to support the management of digital objects.

Ultimately, decisions about significant properties will be implemented automatically to whole classes of material, in order to manage the Collection cost-effectively. Defining the significant properties of classes of material will be an important step forward.

Significant properties are closely related to the next matter—preservation metadata. Determining the significant properties will dictate the amount of metadata that must be recorded and stored. A successful preservation process relies on an accurate record of the types of files to be preserved, the nature of any preservation activity carried out on them, and the result of such activity.

What information are we likely to need? In addressing this situation, the Library developed and published a draft set of digital preservation metadata requirements in 1999. In the year 2000, the Research Libraries Group (RLG) and the Online Computer Library Center (OCLC) invited the Library and a number of others to join an international working group charged with the task of proposing a draft international standard for preservation metadata.

The working group has taken as its starting point, the approaches to preservation metadata of the NEDLIB Project in Europe, the Cedars Project in the UK, the Harvard University Library's Data Repository Service, as well as the National Library of Australia's draft. A White Paper issued in January 2001 compared these approaches as a springboard for defining a comprehensive, structured preservation metadata framework, applicable to a broad range of digital object types, archival processes and institutions. The working group's recommendations for a preservation metadata standard are expected in early 2002, so it should be out soon.

Now I would like to talk about cooperation with publishers. In the online environment, legal deposit legislation will not succeed without cooperation between libraries and publishers. Accordingly, the National Library has been working with the Australian Publishers Association to develop a Code of Practice for archiving, preserving and providing access to commercial publications. It has been important to establish mutual trust and to overcome the suspicion that has existed between publishers and libraries in relation to access to electronic publications. We have had excellent relations with the publishers regarding open-access material, the freely available material. Publishers, by and large, have been very enthusiastic about including them in the Archive. But understandably, the commercial publishers have been concerned about their commercial interests. Publishers have feared that control over their intellectual property would be lost, and their livelihoods undermined. Libraries need to demonstrate to publishers that they can control access in agreed ways, and that there are advantages to publishers, as well, in legal deposit, through wider knowledge of their works, and also the long-term care and maintenance of them.

The Australian Publishers Association and the National Library have recently reached agreement on the Code, and will now ask a small number of publishers to trial it. The Code recognizes that safeguarding Australia's published cultural heritage is a concern shared by publishers and the National Library alike. It outlines the conditions and responsibilities that each partner agrees to observe in order to ensure Australian online publications remain available for use into the future.

I only have ten more minutes, so I would like to move on and share with you some of the lessons that we have learned in the last five years. We have learned a lot about the technical, legal and organizational aspects of the work, and the kind of infrastructure needed to support it. We reported on what we learned in conference papers, journal articles, and documents on our website. But there are some aspects that are not commonly discussed, and I would like to share a few of these more informally with you.

We found that in the digital environment, it is advantageous to liaise and cooperate with a wider range of parties than has been necessary for print collection development. Active working relationships are required with publishers, and indexing and abstracting agencies. Collaborative relationships with other libraries and collecting institutions, both in Australia and overseas, are beneficial for exchange of information, and to share the high costs. We have been cooperating and liaising, not only with other libraries, but also with archives and museums in Australia.

Taking a practical approach and learning by doing has worked well for us and continues to do so as we investigate new aspects of this work. Once we start work on a particular task, instead of just considering it theoretically, we sort out many issues as we go along, and gain confidence in our ability to tackle the task at hand, no matter how hard. There are still sometimes problems that we cannot solve immediately, but at least we know what they are.

A team-based approach to devising and implementing policy and procedures enables us to draw on the expertise of a wide range of staff within the Library, and helps to motivate their commitment to the work. Starting small has also worked well for us. This approach kept the workload manageable while we built up efficiency and expertise. It also enabled us to address issues like quality control in detail, and gave us time to learn.

Deciding not to exclude a website or publication on the basis of its file formats meant that we learned a great deal about web publications and what can be done with them, even the very complex ones. That does not mean we have always succeeded in gathering them, and databases are one of the major categories we have yet to come to grips with. We have also found that staff undertaking this work needs certain qualities. For the staff in the electronic unit who deal with these materials day after day, we have found the need for a willingness to learn, adaptability, persistence, initiative, ability to tolerate frustration, and creative thinking, as well as the usual skills and knowledge of librarianship, collection development and management ability.

And, just for the record, there are some issues that we still have to resolve, and it is a great opportunity to be here and talk to colleagues who are undertaking similar work, so that we can share these issues and learn from each other.

Digital archiving is an area of work where the learning and development never stop. There are still a number of issues that the National Library has to solve, and given the nature of the digital environment, as soon as we find solutions for these, I am sure that others will emerge.

As already mentioned, the National Library still lacks legal deposit for electronic publications, and has to resolve many of the issues that go with that. Our Digital Archiving System, although successive versions of it give us increasing efficiency, does not yet deal adequately with a number of significant matters. Some of these include gathering and storing the software and plug-ins that publications and websites are dependent on for display. We know that the Danish system manages that, and we would love to know how. Clearly indicating to users that they are viewing an archived object, rather than a current version on the website, is an issue. If you are knowledgeable and you look up in the location box, you will see this is an archived version in the PANDORA archive, but if you are an untrained user, you will not be immediately aware of that. There are also a number of other authenticity issues.

Although we know in theory what preservation metadata we want to record, at this stage we do not have a mechanism for recording all of it. The Digital Archiving System keeps some of it. This deficiency is being addressed through the development of a Digital Object Management System, which will contain, among other things, the data to support migration and other preservation activities aimed at providing long-term access to digital objects. The Library has developed a specification for its Digital Object Management System, and is currently developing part of this system in-house, known as a Special Collections Manager, after failing to identify a sufficiently affordable and low-risk solution in the marketplace. Completion of the Special Collections Manager, which will provide for preservation metadata, is planned for late this year.

Another issue still requiring resolution relates to the archiving of databases, as I have mentioned. We have some plans to try to address, in a small project, a possible way of dealing with this matter. We would like to do that this year.

In adopting the selective approach to archiving, we know that we are archiving only a very small proportion of the Australian domain. We would like to supplement this selective collection with periodic snapshots of the entire Australian domain. A short-term consultancy last year identified the associated issues for us. While we have developed a system of persistent identification for the Library's digital objects, we see a need for a national system. A proposal for an Australian Digital Resource Identifier has been endorsed in principle by the Council of Australian State Libraries, but this initiative requires further practical development, and the implementation of a national agency for allocation of identifiers.

RLG/OCLC are in the process of defining the attributes of a trusted digital repository with their report due out early this year. The Library will assess where any deficiency exists in our own system, and work towards meeting the standard. Like every other digital archive in the world, we still have a host of preservation issues to resolve. We will keep abreast of research,  contribute a little of our own, and work in cooperation with others to find solutions.

So in conclusion, as a national deposit library, the National Library of Australia has clear responsibilities for collecting and preserving the documentary heritage of Australia, now appearing in digital formats. After five years experience with archiving web publications, building the National Collection of Australian Online Publications has become a routine part of the Library's collection development activity.

While there are significant challenges ahead, we believe we are establishing a good foundation for meeting them. Thank you very much for your time.

## APPENDIX:
## PANDORA PERSISTENT IDENTIFIER STANDARD

A recent consultancy recommended the following form of identifier for files stored in PANDORA:

<collection id>-<work identifier>-<archive date>-<publisher's URI>-<generation code>

where

<collection id>: for the archival collections the collection ID will be "nla.arc"
<work identifier>: a unique number within the digital archival collections assigned to the parent work of which the resource is a component

<archive date> the date the file was archived in the format YYYYMMDD

<publisher's URI> currently the host name, path name and file name of the resource on the publisher's site

<generation code>: a two digit code representing the version of a resource which has been migrated from its original format.

Report

# Danish Legal Deposit on the Internet

Ms. Birgit N. Henriksen
Head of Digitization and Web Department, The Royal Library, Denmark

Good afternoon, ladies and gentlemen. This is the first time that I have ever been to Asia and to Japan, and I am very happy to be able to be here. As a representative for the Royal Library, it is my pleasure to talk about legal deposit on the Internet in Denmark. The Royal Library is a similar institution to the National Library. It is the Danish National Library.

In Denmark, we have two legal deposit libraries. One of them is the State and University Library in Aarhus and the Royal Library is the other one. The Royal Library, which I represent, is responsible for many things and also the legal deposit of net materials. First, I will talk what effect the changes of the legal deposit law in 1997 had had on the archiving of electronic material from the Internet in Denmark. And secondly, I will focus on some of the initiatives going on right now in Denmark in this field. Finally, I will give my point of view to the future strategy for web archiving in Denmark.

Where is Denmark? Well, this is the smallest country in Scandinavia. We cannot trace our civilization back as long as you can, but at least for the last 1,200 years Denmark has been a royal monarchy. We are 5 million people and I have learned that this is half of the people living in this inner city of Tokyo. We live in an area like Hokkaido in Japan. We are a very small country with our own language, and we are very much concerned about preserving our cultural heritage. As for Internet access, 55% of the people in Denmark have access from home and 72% have access to the Internet from home or from work. So, most Danes use the Internet very frequently.

In more than 300 years, Denmark has had a legal deposit law and it had been changed time to time. I will talk about two of the changes here because they are important to understand this. The 1902 law was the most extensive and it was passed just as the Danish Industrial Revolution had changed the printing industry and thus increased the amount of printed matters deposited immensely. In 1997, the Danish legislation on legal deposit was modernized and updated in order to match the information society. Working on a definition of what was to be deposited, the text ended up with two very important keywords—work and published. There was also a very important point, which was "regardless of medium". Work was defined as being a delimited quantity of information, which must be considered a final and independent unit. Published was defined as one or many number of copies of the work that have been placed on sale or otherwise distributed to the public. So, the modernized law only covers selective collection and archiving of Internet material in Denmark.

What does that mean in practice? When the law was passed, the governmental instructions were also made and it was during this stage that we developed a concept called “static and dynamic”. Static publications in the Danish model are monographs, as we know them, reports, books and periodicals, if they are not changed too often. We have guidelines that say that if it changed more than a month then it is included in the law but, if it is less than a month, then it is too dynamic. The other thing we defined was that dynamic publication included databases, as we have heard earlier today, and homepages. The law excludes them. The consequences of these instructions were that, for instance, news and media, which are normally very important for the library, are not included in the law and then excluded from the legal deposit. A homepage can have an area with publications that are included, but the homepage as such is not included.

To support the law, a website containing information about the law and a system for retrieving, archiving and viewing and interaction with the rendered form of the archive material was developed in the beginning of 1998. I think that it is because we did not have a good acronym for it that you have never heard about it. Because it has never been a project actually, it has been in production just from the start. So, since everything that we have done is only documented in Danish and we do not have this brand name for it, I think that maybe that is why nobody has heard about it.

The formulation of the existing law, which was only required for the portion of the content of the net should be deposited, makes it very difficult for us to create a fully automatic model by which all relevant material is harvested and registered. So until now, our system has been based on a model where the notification starts a download of the publication. The person in charge of the technical completion of the digital copy is the one responsible for the notification.

This approach has the disadvantage that far from every producer of the work covered by the Legal D Law is aware of it. Earlier it was the printers that were responsible and they have known the laws for 300 years. Now, we are talking to a totally new group of people and they have never heard about legal deposit. Mail campaigns and advertising in newspapers have speeded up notifications, but as you shall see later on, the figures are still very low.

Notification is done by filling out a form at the legal deposit website. They just have to go there and fill in a form and then we will take over. As soon as this is done, it must be done as soon as the publication is available on the net, this is the Royal Library’s responsibility to download it. The law states that we have in practice three months to do it, but we try to do it as quickly as we can. Once we send notification back to the publishers saying that we now have archived this document, they are not allowed to change it and are not allowed to remove it. Of course, we try to do it as fast as we can. If they change it, they have to notify us again and we will take a new version back to be archived.

Notification forms for monographs and periodicals have been developed to promote the use of Dublin Core (DC) metadata. We added an extra notification form for monographs containing DC metadata. Publishers who include the required metadata will have a far easier time notifying us than others. Normally, publishers have to supply us with a lot of information, and you shall see the list in a minute, but if there is notification form for metadata used, you simply just add a few things—your email address, name, phone number, and the uniform resource locator (URL) where we have to start the download—and then a program will extract as many of the fields into the notification form as possible. And the extracted metadata is reused for cataloging purposes when possible.

Now I will try to see if it is possible for me here. I have not made any screen dumps. I have only made one in my presentation because I thought there is no value in it because everything is in Danish. I do not expect any of you to understand it. So, this is a Danish notification form. You can see here, it has e-mail address here, it requests name, institution, phone number, and then URL. I will try and see if it works online and it does. It expands the notification form and fills in what it could find in the Dublin Core metadata, as much as it could.

So, what has to be added now is to tell us here which formats are represented. Is it a PDF document? We would like to know the data format in advance. If any special program is going to be used, we would like that. If it is a commercial program, the Library has to buy it. If it is freeware, we would like to know which one. We also have to know if we can access it just from the net or if we have to have user ID and passwords. The law says that we are allowed to go behind user ID and passwords. If it fulfills the requirements of static publications we are to go to the archives. This part here is just if they had some remarks, say, sometimes they say, "Call us because we do not know how to fill in all of the things." This last one here, if I can point it out, this at the bottom says, "Commit the information." As soon as this is committed, we will start looking at the library.

The library staff at the Royal Danish Department now receives the notification and they go and inspect this to see if it is a publication covered by the law. Of course, the law is not so clear to people so often also things that should not be inside the archive is notified. It could be a whole website and sometimes it means that we nearly track down the Danish sub-domain. That would be nice for us, but we are not allowed to. We just inspect and normally it is good enough, and then we start the download with our own program. We have built our own harvesters to do this. We download only the plug-ins and put information into the system that this plug-in is used in connection with this work.

The work is verified. We see that we have all of the items and the images that should be in there and we verify that all hyperlinks are valid. If it is not, we sometimes call the publisher back. If we can go and see it in the browser, then there is something wrong with our tools. But if things do not work in a browser either, and it very often does not, then we call back and tell the publisher that there is something wrong with the work. There are some links, for instance, that do not work. Normally, they will say, "Thank you very much." They fix up the link for the material, and then we just archive the revised version as if it was the first version.

The work is now cataloged and classified in the Online Public Access Catalog (OPAC) and I will come back to that later because we have made some restrictions now so that means we only classify periodicals. Finally, it is transferred to the archival server. The archival server is mirrored every night to the other legal deposit library in the State and University Library in Aarhus, which is on the other side of the country. This mirror is a part of a back-up and security strategy with two physical copies placed far away from each other in different, independent systems. Each legal deposit library gives access to its own copy of the archive and this is also exactly what we normally do with printed materials. It is exactly matched.

When a publication is downloaded, the Danish Bibliography Centre is notified by email. It is not the Royal Library who makes the whole National Bibliography. The Danish Bibliographic Centre (DBC) makes machine-readable record catalog (MARC) records of that part of the works that has to be included in the National Bibliography. Some of the notified publications are not covered by the National Bibliography and earlier the Royal Library makes MARC records of all the remainder part. This is because Denmark and the Royal Library is the Danish International Standard Serial Number (ISSN) Center and is responsible for the registration of periodicals in Denmark.

When the project started, MARC records were made for all notified publications, but this has changed. Instead, we have supplemented access by searching the OPAC, access by searching directly in the data provided by the publishers. A full text search in the archived material through a web index is also being prepared and will be there very soon. The intention is to give access to this material the same way as when it was online on the net and drop or minimize the access through the OPAC, which is very different from what we have heard.

Due to the copyright legislation, we are not allowed to give access over the net. The archived net publications can only be viewed at the reading rooms in the legal deposit libraries. There is one PC at each library and it is free for all, but there are no possibilities to make electronic copies from the material. Only print paper printouts for personal use are allowed, and this is stated in the law.

What does that mean in practice? No access. In practice, no one uses this facility. Last year, approximately 20 people used the archive this way. Such a restriction is nearly the same as no restriction, in my point of view. This restrictive access seems strange to most Danes, since parts of many Danish websites now are accessible on the archivist.org website. We feel it is very odd that we are only allowed to archive a little bit and we are not allowed to give only very restricted access and the same material is archived and accessible from a server outside of Denmark.

How much have we archived up until now? We have approximately 1,000 Danish sub-domains represented in the archive. In Denmark, 352,000 sub-domains are registered and we guess that around 200,000 are really active. That means that we have less than 0.1% of the domains represented in the archive. In four years, we have collected 10,500 net publications, which are constituted by nearly 700,000 files and a total volume of only 23 gigabytes (GB). Compared to what we have heard today, this is a very small collection. It is about 10% of the Australian collection.

Two-thirds of the publications are periodicals and the reason for this is that the staff at the Royal Library is very much concerned about this stuff and wanted to be as complete as possible. At the moment, it is very complete. That means that they actually, if they can see something missing, notify the publications themselves in order to have it all in the library. Two-thirds of the publications are done by public publishers such as the government or universities, and only one-third comes from private publishers. The material collected mainly consists of working papers, reports, scientific reports, guides, periodicals and newsletters. The minority of this material also exists in printed versions, and therefore they are also coming to the Library in these printed versions. In general, I will say that the closer we come to the individual citizen and the private concerns, the less the citizens are represented in the archives at the moment.

How expensive is this selective web archiving for the Danish National Library? We do not need much hardware, so nearly all the costs are manpower, which is why I just showed you the staff resources. At the moment, we spend a little over 1.3 man years to do this work. It is more than we actually have for it because we have only granted one man year for it. We always look for ways to do things a little more quickly and easily. Of these 1.3 man years, 0.5 half man year goes for the technical stuff—running the server, developing the system, correcting errors, and helping the librarian staff when they need something that is really tough on the net. And 0.8 man years go for the librarian part of verifying, securing that everything is okay, making the MARC records, and so on.

You can see that the first year was very special because we did the development of the system, and therefore it cannot be countable. The next year, we used a little more than one hour per publication and that was the phase where we classified and cataloged all publications. When we ended that year, we decided that we had to do something dramatic in order to decrease the amount of work, so we skipped cataloging except for periodicals. You can see that we halved the time spent per publication. We only use 35 minutes per publication now to archive it.

So, the experience from our system is that building these select collections is expensive and we have to be careful what material we select for this expensive treatment.  Selections ought therefore, from our point of view, be directed to a greater degree towards types of material that we wish to preserve, but which we cannot catch via the more mechanical methods. This is, for example, interactive works, streamed material, or collections based on themes.

One of the duties of the Royal Library is to collect, store and make the files available now and in the future. This could be problematic, but if you look at these figures, you will see that only 1% of the files we have inside our library are not in well-known and widespread formats, which we may reasonably expect to be maintained and remain available in some form in the future.

That dynamic and interactive publications are not to be archived does, of course, skew the distribution of data formats. But the figures I have put up in the other column from Sweden show that the picture does not change dramatically merely by changing from selective to bulk collections. As long as you harvest, we do not see this as a problem. If other methods of collection are to be used, such as actual deliveries of the technically more difficult material, then the picture could easily change.

The technical problems we have run into in connection with our current system are very much like those we know from large harvesting projects. We have errors in HTML standard documents, which makes it very hard to harvest. And similarly, problems with downloading and later accessing documents that use flash or java, client-side elements like java script or other types on inline code. Document retrievement, depending on cookies and providing user id and passwords, can also provide a problem depending on how the features are implemented on the websites. Sometimes it works, sometimes it does not.

What was impossible yesterday may be possible tomorrow. The global boom in portals offering search engine services, both highly specialized and very general, indicates that search engine developers are faced with many of the same problems regarding document decoding, as people in the archiving. This boom implies that there are now many more brains and heads working on the problem than just a few archivists. I am sure that solutions will be found.

Let me quickly summarize the situation in Denmark right now. We have practiced selective archiving for the last four years. We do not get a representative part of the net since most of the net is dynamic and most of the material we get is published by public publishers such as the government. We do not get the most advanced part of the net because our download is based on harvesting. We only get static publications, many of them also available in printed version. But what we get is validated and plug-ins are downloaded and most of it is cataloged. The selective approach, based on notification, is labor intensive and therefore cataloging is partly replaced by search facilities in the metadata. Due to copyright legislation, it is only possible to give restricted access to the archive.

Like you, we had an international conference. This took place in Denmark last June and the Danish Electronic Research Library sponsored it. The purpose of this conference was to clarify the user expectations to web archiving in Denmark. Scholars and scientists put different aspects such as the content and the context and the evidence of use of the material into focus. Especially archiving of news and media on the web with very high frequency was important. But it was also felt that archiving materials like chats and more interactive and process-oriented materials should be included.

Most archivists found that different archiving approaches were needed to meet all these requirements. The budgets for making snapshots and selective collections seemed to be comparable and that means that budgets will have to be doubled up, for instance in institutions like mine, if we should go and support both things. Methods for archiving interactive material are not available at the moment and should be developed, if such material is to be included.

What are the arguments for going into bulk collections instead of selective collections? The material collected as a consequence of the changes in the 1902 law is material that researchers use today when history of firms is written or research if done on the breakthrough of industrialization.

In 2000, 168,000 printed items of the type called ephemera were archived. Of the corresponding electronic publications, we have nothing at all except for a very few annual reports from a firm. In line with rising cost connected to printing and distribution, more and more of this material will only exist in electronic versions. An example of what we already have missed could be advertising campaigns using banners on the net and short films on the Internet for product promotion.

If we are going to collect the matching electronic material in Denmark, we have to use techniques such as harvesting the entire Danish web space in bulk collections. We cannot see any other ways to do it. That will also mean that we will have a better coverage of both the private and public publisher's things on the net and material about Danes, as well as materials that interest the Danes, not only more official sorts of papers.

We will also get a better coverage of Denmark outside the public sphere, but we will be also be able to catch new trends in function, and in contents and design on the net as soon as they appear; if they are not too technically advanced, I must say.

The last thing is that now we only get the first version of a publication. If we go into harvesting, we will be sure that we will have the next and the next and the next version by itself. People are not going to remember to notify us over and over again. It will be possible for us to make accumulating harvesting of news and media. At the moment, we are not able to do that and the law does not cover it. Of course, we want to do harvesting once or twice a day of, for instance, newspaper sites.

Why is harvesting not enough? When exclusively harvesting the entire .dk domain, it is more difficult to secure archiving of all the necessary plug-ins, especially because most of the plug-ins are placed in the .com domain. They are therefore excluded by the actual configuration of the harvester in order to stop the robot harvesting the whole Internet.

Some material available to us as a user simply is not available for a harvester. This is, for instance, streamed material and webcasted material but it also includes applications that are made with flash applications. They are hard stuff for us at least in Copenhagen. Interactive sites and sites with a lot of interactive material cannot be harvested using existing methods. Sites that adapt their content and design to the current user's preference would also be difficult to archive as a whole, and likewise with material depending on a program running on the web server.

Services like travel route planners, online maps, OPACs, home banking systems, auctions, games, e-commerce, portals and such services are not archived by harvesting. This slide tries to demonstrate the web services. Common for many of these services is that they are based on detailed data that is often placed in databases with a service program on top. We are not actually interested in retaining the many detailed data for posterity, but merely want to document that they were there, and show how they were used.

We have been practicing now for four years and are ready to start looking at new methods. The Danish Electronic Research Library sponsors two different web archiving projects at the moment and both terminate this summer. The Royal Library is participating in the Nordic Web Archiving project, a project with focus on developing software that gives access to a web archive consisting of standardized documents by searching and navigating. We could call it a Google, a search facility like Google, but with a possibility to search and navigate not only in space, but also in time. The project partners are the five Nordic National Libraries. At present, a vendor for the search engine has been chosen and a virtual software team, with participants from the different Nordic countries, has been set up.

The other project has another focus. The first one has access and the second has archiving strategies. The State and University Library of Aarhus, Center for Internet Research at Aarhus University and the Royal Library are partners in the other project, which we call netarchive.dk. If you go in that way, it has also another name—we have a name with the same meaning in Danish—but if you go in that way, the content will be English and you will be able to follow the projects. Every time we make a report, we try to make up a summary in English and put it up there.

The scope of the project is to test different archival approaches and subsequent usability of the archived material for research purposes. The case is the Danish municipal elections last November and this was the first time that the two legal libraries did event-based archiving.

If we try to describe the material on the net that has to be archived in a two dimensional model, it could look like this: a vertical axis representing the publishing frequency, with real time dialogue such as chat having the shortest lifetime, over news in the middle to static publications and scientific reports with the longest lifetime. If we draw a horizontal axis representing the interactivity, with the material with static HTML pages or static web forms, for instance, at one extreme, over database publishing and e-learning in the middle, to services such as net auctions and web games at the other end, the other side. The nearer you come to the bottom of this figure, the less libraries have normally been involved in collecting and giving access. These are some of the questions that I think that go on now in Denmark. How far shall we go in our archiving efforts?

You will find that existing projects are concerned in the uppermost left corner. And as a rule, it is for material in this area that archiving tools are available. Netarchive.dk, as a project, investigates this corner, too. One of the test cases we will do now is, we will harvest an online newspaper and, at the same time, get the published article pushed by the publisher in order to compare the quality of the two methods of collection with respect to completeness. This spring, we will begin investigating the possibilities for archiving that part of the net where the processes are more interesting than the data itself. This includes processes like surfing the Internet and paying our bills or reserving a book in OPAC or a seat on a flight.

We are in the middle of a project period and the material that has not been covered by the existing concept, like web sites, discussion rooms, portals, chat, and conferences, are now being archived. Since we have not been involved in event-based archiving before, one of the biggest experiences for us was how hard it was to find the relevant new URLs to be harvested during the event. When the things go on, where do we go? We cannot use the search engines—they are too slow, they are three weeks back, we need it now. How do we find it? We found out we still have something to learn there.

Since Danish law only covers static material on Danish websites, and the event that was to be archived was not restricted to this, contracts or agreements had to be signed with the publishers in order to archive dynamic publications. We sent proposals for agreements to news, media and political parties and named our user agent for our harvester with the web address of our project. But only a few publishers reacted to this proposal and, until now, only three out of 44 agreements have been signed. It was the experience that knowledge of an agreement does not spread out sufficiently in a top-down organization as we had assumed it would. So although we have already approached these organizations, we still have strong reactions from individual departments who have never heard about it. Our proposal for an agreement was focused on the possibility of harvesting and giving access later on. But some of the publishers, even the big ones, were also concerned, or more concerned, about technical questions such as the time of the day when the harvesting went on, or whether we would be able to limit the harvester to the relevant area. Such requirements turned out to be hard to control with a harvester designed mainly for a snapshot oriented archiving strategy, which was used most of the time.

We used three different harvesters during the harvesting period, and basically none of them could handle bad HTML code with errors or interpret, for instance, javascript as well as browsers. Browsers are updated all the time and if we should be able to work as well as we could, harvesters should be upgraded at the same speed. As a consequence of this, the project netarchive.dk has decided to do an extra test, harvesting some of the difficult websites with a browser like Mozilla as a layer in front of the harvester, archiving the HTML code interpreted by the browser.

Furthermore, it turned out that there was an extensive use of redirects from servers included in the archiving to sites not included. This meant that part of the archiving failed because it was impossible to manually find out fast enough whether the new site should be archived. Do we have an agreement? Could we do it or not?

What can we do with process-oriented materials? One of the obvious choices is to do the same as we normally do when we want to document processes in our time, which is to capture the process on a film, which is a traditional container with a well-know preservation strategy, but we will lose all functionality as a consequence.

We think there is another possibility. We will try to film the net through a browser, defining as archiving a chronological series of displayed web pages in a standardized form, for instance, as images with a mouse-over functionality telling which options you have if you want to proceed in the image stack. You would no longer need to preserve information about required plug-ins, link structure, how to interpret HTML code, and so on since the experience is frozen in the images. This is only meant to be used on works that we cannot catch because of a high degree of interactivity. We will not use it in widespread or normal HTML pages. Software from business intelligence tools and tools used in usability laboratories might be taken into consideration when we have to go into this area this coming spring.

What shall then be the strategy in the future? It is clear in Denmark that we work very hard to get dynamic material included in the law. But we do not imagine that we can and will archive the whole Internet, but we wish to archive a broad coverage of the type of material that is to be collected as well as minimize the cost connected to it. Harvesting the entire subdomain is suitable for this purpose and should be used to gather net material. But harvesting cannot solve all problems. We still need the possibility to collect selectively for various purposes and need to use different archiving methods in order to document interactive or very technically advanced material.

Therefore, we urge that the present Danish Legal Deposit law be amended by rules that will make it possible for the national legal deposit institutions to harvest those sections of the Internet that the institutions deem essential for documenting the national heritage and at the intervals that will best serve this purpose. At the moment, access to the archived material is only given at the Legal Deposit Library. I hope that in the future it will be possible to find a solution where material could be freely accessible from the Internet, or at least after a period, for instance, five years.

If free access to the archive cannot be granted through law, it should be considered whether access might be granted to material where the copyright owners have granted free access at the time when it was collected. This could be done by adding a simple tag to web pages if you, for instance, allow us to archive it—that could be one tag—or if you want us to give access—that could be another tag. So at the moment, we go for our next strategy in Denmark with respect to collection methods and a less restrictive access policy, but most of all, it should be covered by the law. Thank you for your attention.

Report

# Collecting and Archiving of Information Resources on the Internet and the National Diet Library

Machiko Nakai
Director, Electronic Library Development Office, National Diet Library, Japan

## Introduction

My name is Machiko Nakai and I head the Electronic Library Development Office of our National Diet Library. First, I'd like to thank you for joining us here today. This symposium resulted from discussions in our office, which is planning toward the gathering of information resources via the Internet as part of our Electronic Library Project. We find it most gratifying that this symposium has become a reality and that we are privileged to make a report.

## 1. Background

The National Diet Library was established in 1948 attached to our National Diet, or parliament, and it remains today Japan's only national library. Under the National Diet Library Law, we engage in the collection of materials based on the legal deposit system for domestic publications, and prepare and provide information such as the "National Bibliography of Japan." At the same time, we furnish library services to Diet members, the government and the public.

2002 marks an epochal and, in a sense, thrilling year in our history of over five decades, as the Kansai branch of the National Diet Library will open in Kansai Science City of Kyoto Prefecture and start providing services this October. Accordingly, we plan to upgrade our Electronic Library function while publishing the bibliographic information stored to date, its electronic materials, and the like.

The concept of our Kansai branch goes back to the 1980s. Its basic functions as envisioned at the time were to furnish literary documents with no geographic or collection restrictions and to make electronic publications available. It was also to have a large-scale storage facility for bibliographic materials.

However, the present status of electronic publications evidently is clearly very different from that of fifteen years ago. An enormous volume of electronic information is now open on the Internet; it is used, updated and disappears. It would seem that unknown sources are challenging libraries, whose task has always been to store knowledge in the form of the printed word and make it available to the public. Of the three projects I plan to describe today concerning our National Diet Library's Web resources, each has satisfactorily coped with the challenges.

What, then, is the present status of our Library? The truth is we are merely at the starting line.

As part of our Electronic Library Project, which began in 2000 as a three-year plan aiming at 2002 when our Kansai branch would open, we scheduled the collection and archiving of Web resources, and the system remains in the development stage.

Let me briefly explain our Electronic Library Project. In 1998, we developed what we call our "Concept of an NDL Electronic Library" and positioned it as one of our National Diet Library's future projects. At present, we are proceeding with a plan to digitize books published in the Meiji Era, which lasted from 1868 to 1911. In 1999, an "Electronic Library Development Office" was formed within our administration department's planning section, and ever since, it has advanced the project. This April, we will install a "Digital Library Section" in our Kansai branch to accelerate electronic library work.

In reality, though, we regret that this plan to archive Web resources and develop its means of access may have been too much to undertake as part of our Electronic Library Project, as it affects traditional library operations like collecting, cataloging and making materials available. We regard it as an undertaking that demands new and unique approaches and support systems.

I shall treat the challenges later. First, however, I shall outline the plan and its basic policies.

## 2. Basic policies

### (1) Collection

The basic policies for the plan cover three major points. The first is collection.

To base collection on selective acquisition, not legal deposit as at present, is very important. Following the April 2000 revision of the National Diet Library Law, in October of that year, we started collecting offline electronic publications such as CD-ROMs and the like based on the legal deposit system. A report by the Legal Deposit System Research Council in February 1999, which formed the basis for the revision, recommended not including online electronic publications in the deposit system at least for the time being, and that the Library should engage in selective collection. Thus, owing to the enormous amounts of electronic information available and the difficulty of imposing a legal deposit on publishers while converting it to physical form, etc., online information lacking physical media was excluded from the current deposit system.

Accordingly, this plan is grounded on the "Guidelines for Collecting Materials" as revised in October 2000. That means only online electronic publications sent within Japan and deemed necessary or useful by the Library will be selectively gathered. Incidentally, our library defines "Online electronic publications" as "Characters, images, sounds or programs made public via telecommunication lines", and interprets this as including all electronic information available on the Web and Internet.

That brings us to what we should selectively collect. For this, we can refer to our "Guidelines for online electronic publications," as compiled in March 2001 after consultation within the Library, the answer being at present to gather administrative and scientific information. Specifically, we gather information available on the websites of government agencies, academic societies and associations, like white papers, research reports, statistics, publicity materials, bulletins, etc… in other words, statistics similar to traditional materials. To preparation for it, this February we conducted a survey of electronic information resources on the Internet among libraries and such organizations, and 2,300 institutions responded.

(2) Access methods
The second point is a policy concerning preparation of bibliographic information which can be accessed. For the standard of bibliographic information created for the information amassed, we adopted the Dublin Core Metadata Element Set while observing moves to standardize metadata concerning electronic information.

We formed a working group within our section to collate a bibliography. Based on our studies, in March 2001 we prepared an "NDL Metadata Element Set" and fixed a qualifier for more detailed descriptions, in addition to the fifteen elements cited by Dublin Core to enable mapping with JAPAN/MARC bibliographic information as prepared by our Library for book and offline publication bibliographies.

Information for bibliographies is not entirely what we collect and archive. For instance, electronic library contents of library or museum archives are excluded from selection inasmuch as they are stored and preserved by each institution. Also, dynamic information like that stored in databases will not figure among the information we collect for technical reasons. Nevertheless, in many cases this information is more useful than that obtainable. Thus, to collect information of this type, bibliographic data is prepared and made accessible by linking them.

Accordingly, our plan envisions a system capable of handling information resources collected and archived as well as those linked and navigated by compiling bibliographic data even when in exterior sites. Put another way, we hope to retrieve archived bibliographic data and external information integrally by both preparing bibliographic information and retrieving it for materials archived at other libraries.

(3) Business models
Lastly, there is a need to create a new business model. This plan will establish for our Library a new type of business, including collection, storage, management, cataloging, and provision of our information resources. To this end, we must develop a system to efficiently engage in various tasks and build a model for that purpose. Since March 2000, an online electronic publications-related system as a component of our electronic library infrastructure subsystem has been in development under a three- year plan. Even before March 2000, we made a metadata cataloging prototype for bibliography preparation and a prototype for acquisition is now in development. But as I mentioned earlier, since we included everything we could think of, it has been a long trial-and-error process.

3. Overview

The title of our project is still tentative but we have nicknamed it the Web Archive Program (WARP). Figure 5-35 shows what it looks like.

For the collection aspect, we request each publisher to cooperate in the collection of information made public on its website and, after obtaining permission, we gather the information. As means of acquisition, four ways are available:

[1] Acquisition from the publisher using web information harvesting software (Web robot)
[2] File transfer from the publisher to our Library
[3] E-mail delivery from the publisher to our Library
[4] Delivery via an archiving medium from the publisher to our Library

According to the results of our survey, publishers understand our intention to collect information and about 60% of them expressed their willingness to cooperate. As a means of collection, using software, namely the Web robot, is most apt to attract publisher cooperation. This being the case, our program centers on collection using the robot.

We register the knowledge resources gathered and create metadata as bibliographic information. Though its name has yet to be determined, a database for retrieval where metadata is assigned will be made public. Our metadata will have a new URL for archiving and the original URL for the original site, both of which link to each type of resource. We also foresee preserving collected information in such physical media as CD-Rs, etc.

4. Workflow and system

Having given you an overview, I shall now explain the workflow and system. While I am using the screen being developed, please understand that the system at this time is not in actual operation.

There are two types of workflow: new acquisition and reacquisition. Since Web information has the salient characteristic of continual updating, it is seldom complete when collected, requiring update management, hence a flow for reacquisition.

With respect to new acquisition, I'll now describe our plan.

(1) Primary survey

New acquisition starts with a primary survey. Which sites do we archive? Even if sites are seemingly desirable, their structure, the part that can be archived and the means to do so, merit first consideration.

(2) Negotiation for permission and contract

Primary survey completed, we must contact the publisher and negotiate to secure permission for acquisition and conclude a contract.

This is the screen we expect to use for entering into a contract (Fig. 5-36). When preparing new contracts, various types of information as affecting the publisher will be input, and the publisher's conditions for collection or for providing services, etc. will be set, registered and maintained as contract stipulations.

(3) Secondary survey

The secondary survey naturally follows the first. In this, a unit of information for use as a unit for actual collection is determined. Unlike books and periodicals, web information has no physical form; it is an aggregate of files, hence it lacks a definite unit to identify the information resource. We call it "granularity" and plan to judge which scale we can regard as a collective entity.

(4) Creating preliminary metadata

The fourth step calls for creating metadata for information resources whose unit is determined during the secondary survey. Various items exist but initially, as this is for collection, nothing more than the title, URL, and the like will be registered, but only the minimum data necessary (Fig. 5-37).

(5) Setting harvesting (acquisition and reacquisition) conditions

The fifth step is to set the conditions for harvesting (acquisition and reacquisition) (Fig. 5-38). Since the robot does the harvesting, it requires a URL as the starting point. As the robot will start from the URL and automatically track the link to HTML files, that is when we determine harvesting depth and how far in the link hierarchy harvesting will be made. If the hierarchy depth is too shallow, harvesting will be incomplete. On the other hand, if set indefinitely deep, the robot will find it hard to return. Although the web information appears single-screen based, in many cases it has numerous images attached, and thus is by no means what it seems to be. When web information consisting of multi-layered links is ready for harvesting, various problems may arise even though the robot is excellent. In this plan, a freeware software called "wget" is used, but it becomes hard to perform perfect harvesting.

The frequency of re-harvesting for reacquisition on a regular basis to prepare for updating of information resources will also appear in this screen, with harvesting carried out accordingly. We consider that there should be two ways of harvesting, one accomplished at night, the other, instantly.

(6) Harvesting

The screen to check the conditions of harvesting has also been prepared (Fig. 5-39).

(7) Trimming and registration of individual objects

The next step following collection is trimming and registering individual objects (Fig. 5-40). Trimming, a WARP language term, is a process to organize collected information like shaping a plant by pruning excess branches. The robot cannot judge which area constitutes information, but by designating the depth of the hierarchy link for harvesting, the robot will retrieve all files at that depth, which naturally includes unnecessary files. Unnecessary files should be purged and arranged for registration, but doing so is quite tedious. Accordingly, we are conducting a study on this sort of trimming to store collected files in hierarchical form and eliminate the check box.

Registration of individual objects comes next. We use “individual objects” because various versions will result for one metadata when updated information is re-harvested.

Information like years and months represents “individual objects.”

(8) Creation of metadata
This step registers archived information through a flow such as this and develops metadata based on the criteria I referred to earlier:

Various functions are examined for reacquisition including the method of update check, but as this becomes rather complicated, I won’t explain them now. Instead, I’ll stick to the main issues.

5. Challenges
(1) Characteristics of Web resources and WARP
As I mentioned earlier, WARP is a project to create a business model aimed at making use as library resources of Web information different from traditional library materials. We intend to cope with Web information characteristics as follows:

- Indefinite granularity that cannot clarify from where to where a unit is scoped
- Easy to update, change and delete
- No tidy hierarchy owing to hyper-links
- Dynamically generated Web pages such as CGI and scripts

The first issue will necessitate defining the standard of granularity and continuing practical experiments based on the required type of unit harvesting. The system is one in which harvesting can be premised on either site or file units, but we are considering the need to solidify the standard through experimentation. Obviously, to set granularity finer will demand greater manpower for harvesting and bibliographical preparation, thus boosting the cost of labor to where it becomes a concern. Harvesting grounded on site units conflicts with the conventional policy of collection, and causes a problem with data capacity.

Next, as for the problem of updating Web information, we plan toward a flow of re-harvesting as a mechanism enabling update management. However, it also relates to issues such as harvesting frequency, manpower and data capacity.

With respect to harvesting information with a non-hierarchy structure, it presents a difficulty in that the robot function will not work, as I mentioned when explaining the setting of harvesting conditions. It also concerns how far we should seek perfection.

In the case of harvesting dynamic Web information, we cannot deal with it for the time being. Moreover, if Web information is stored in a database, and no longer shows as remaining in deep Web, we have no way to collect it. But, as I said at the beginning, there is a concept of navigation, and we plan to deal with it based on navigation to external sites by metadata. To build a gateway-like database for information provided on the Web, we are now preparing a list.

There is one more issue. When trying to obtain permission for acquisition, we will encounter various difficulties. While in some cases those who prepare information become publishers themselves, in other cases, they do not. As seen in these instances, complicated rights when securing permission also present issues as revealed through our survey and on-site research.

(2) Legal deposit system and WARP

Now, to the legal deposit system. I have already mentioned "selective collection" as the basic policy. At present, there is a movement afoot in our National Diet Library concerning the legal deposit system, and the Legal Deposit System Council is scheduled to launch discussions on the acquisition of online electronic publications in early 2003. Consequently, a major objective of this symposium is to learn the progress toward revision of the legal deposit system occurring in each country's library as reference for the discussion.

The projected discussion will address issues other than those associated with traditional library materials, such as whether a country's legal deposit system should adopt inclusive collection for online and offline information instead of the selective type, if the concept of "publications" should be the current interpretation, how reproduction and modification required for archiving should be adjusted with copyrights, etc.

Conclusion

Having heard various reports during this symposium, I feel that the role of our National Diet Library with respect to Web information resources is indeed wide-ranging and the issues we face range from technical subjects to those related to systems. At present, though, WARP has certain basic policies such as for collection, since so far we lack a policy for collection such as Pandora at the Australian National Library. Nor do we have the experience of pursuing many projects while having a legal deposit system like the Danish Royal Library. And we don't have a partnership arrangement with the American Library of Congress like the International Archive. Considering this, Dr. Kahl of Alexa has made us a surprising gift of Japan's Website Collection, which we certainly did not expect. It may mean that we now face another issue: how our National Diet Library should make use of it.

We have much to do in clarifying uncertain factors. To that end, we wish to start by attending to collection on a project basis while proceeding with necessary research and seeking cooperation from external organizations.

I think this symposium is a unique opportunity to learn from the experiences of libraries already advancing and modernizing their practices.  I sincerely hope that we can to report on the results of our project on the occasion of our next gathering.

That's it regarding our effort at the National Diet Library.  Thank you.