**World Library and Information Congress:
70th IFLA General Conference and Council**

**22-27 August 2004
Buenos Aires, Argentina**

Programme: http://www.ifla.org/IV/ifla70/prog04.htm

Code Number: 026-E
Meeting: 148. IFLA-CDNL Alliance for Bibliographic Standards (ICABS)
Simultaneous Interpretation: -

Collecting and managing web resources for long-term access: web harvesting and guidelines to support preservation (ICABS Actions 3.3 and 3.4)

Pam Gatenby
Assistant Director General
Collection Management
National Library of Australia
pgatenby@nla.gov.au

Abstract:

The National Library of Australia is contributing to the ICABS work plan through Goal 3 – to advance understanding of issues related to long-term archiving of digital resources. It is committed to actions to advance collaboration in the areas of web archiving (Action 3.3) and preservation of digital resources (Action 3.4), drawing on its own experience in these areas. Specific actions include leading the Deep Web Working Group of the International Internet Preservation Consortium (IIPC), holding an international conference on web archiving, working on automating the deposit and archiving of online government publications, promoting information on approaches to web archiving, and surveying information available to guide digital preservation decision-making. The aim is to facilitate the availability of information relating to these areas of interest, and to make the information readily accessible through the PADI (http://www.nla.gov.au/padi/) subject gateway.

1. Introduction

The National Library of Australia is pleased to be a participant in the ICABS alliance as we are committed to collaboration on standards development and strategies intended to support cost-effective access to national heritage collections.

Our main contribution to the ICABS work program is though Goal 3 – To Advance understanding of issues related to long-term archiving of digital resources.

We chose to contribute to this particular goal because we have experience in digital archiving and we are aware that many national libraries are seeking guidance in this area, being unsure how to proceed. Also, in order to keep moving forward ourselves, we recognise that it is necessary to work in practical partnership with others as there are too many issues to solve alone.

ICABS has provided a stimulus for us to think carefully about how we can share our knowledge and experience more effectively and how we can more usefully contribute to the goal of long-term access to digital resources of national significance.

2. Background to digital archiving and preservation activity at the National Library of Australia

Digital archiving

The National Library of Australia has been collecting significant Australian web sites and managing them in PANDORA: Australia's web archive (http://www.nla.gov.au/pandora/) since 1996. This is now a routine collection development activity carried out by a team of 5 staff. Titles are collected with the permission of their publishers as Australian Commonwealth legal deposit law does not yet apply to electronic publications.

The archive currently includes around 6,000 titles and over10, 000 regatherings (or instances) of web sites and is over half a terabyte in size. Development of PANDORA is a collaborative undertaking involving the Australian state libraries and 3 other national collecting institutions. Selection of sites is based on detailed criteria which give emphasis to research and information value as well as cultural significance, and which also provide scope for subject based selection. Each web site archived is quality assessed for content and functionality.

To manage and support its digital archiving responsibilities, the Library developed its own Digital Archiving System know as PANDAS (Pandora Digital Archiving System.) The system supports the following functions:

- managing metadata about titles selected and rejected
- initiating gathers of selected titles
- managing the quality checking process
- preparing items for public display in the Archive
- managing access restrictions
- managing persistent identification
- providing management reports

Several collecting institutions around the world have enquired about using PANDAS as there are no comparable commercial systems available yet to support collecting and managing web resources. In response to the enquiries, the Library has made PANDAS available on request for evaluation purposes. We are also prepared to make the software available to others for implementation on the understanding that we will not provide support services.

Digital Preservation

Preservation considerations and strategies have been integral from the outset to the Library's development of digital collections. The objective of collecting web resources that are archived in PANDORA is to manage them over time in order to provide on-going public access to them. This requires both immediate and long-term planning and active day-to-day management of the archived resources. A Digital Preservation Policy (http://www.nla.gov.au/policy/digpres.html) and action plan outline the key directions and activities that the Library is pursuing. Emphasis is given to:

- data management (refreshing, backing-up, etc)
- identifying and documenting the characteristics of archived resources
- identifying and managing the risks and threats associated with different file formats
- developing preservation strategies for each file format
- using digital preservation metadata
- contingency planning

Collaboration is also central to the Library's digital preservation approach. To facilitate collaboration the Library maintains the PADI (Preserving Access to Digital Information) subject gateway (http://www.nla.gov.au/padi). PADI provides links to a wide range of information related to digital preservation and also provides contextual information about key topics. It is central to our contribution to ICABS, enabling us to maximise information sharing.

3. The National Library's involvement in ICABS

The National Library of Australia is contributing to the ICABS Action Plan in two action areas - Web harvesting (action 3.3) and Preservation of digital materials (action 3.4)

Web harvesting (ICABS Action 3.3)

The objective of our activities in this area is to explore and promote approaches to collecting and archiving web publications and to identify the associated issues. The purpose is to make this information widely available and easily accessible through PADI with the particular aim of assisting those institutions that are planning or reviewing digital archiving and preservation programs.

Key activities are to:

- lead the work of the International Internet Preservation Consortium (IIPC) Working Group on the Deep Web;
- hold an international conference on archiving web resources;
- automate deposit and archiving of online government resources; and
- Explore and promote methods to archive web-based publications collected by web harvesting.

(i) Lead the work of the IIPC Working Group on the Deep Web

The International Internet Preservation Consortium (IIPC) was established in 2003 in order to facilitate collaboration among national libraries on developing standards and tools to support archiving and preserving web resources. Eleven national libraries plus the Internet Archive are members and the consortium is coordinated by the Bibliotheque nationale de France. The focus of the consortium is practical. Participating libraries pay a membership fee and are obliged to contribute actively to the work program of the group.

The National Library of Australia is leading the work of the IIPC Deep Web Archiving Working Group which is investigating the identification, acquisition, storage and display of publications and web sites that are database driven. This category of web resource is of concern to us as an increasing number of important resources are issued in this form. For technical reasons, we are currently unable to collect them.

The objectives of the Deep Web Group are to:

- identify submission information required from publishers for deposited databases;
- develop a data model and schema for archiving the contents of a database as XML; and
- develop an online access tool to search and navigate structure archived databases.

Software that is developed will be open domain and available for use by anyone with an interest in it.

Other IIPC working groups are working on issues such as defining architecture requirements for archives; identifying the needs of researchers in using web archives; and developing access tools and web crawling software.

(ii) International conference on archiving web resources
(http://www.nla.gov.au/webarchiving/)

An international conference on archiving web resources and issues for collecting institutions will be held at the National Library in Canberra from November 9-11 this year. The following events will be held in conjunction with the conference:

- a meeting of the IIPC Steering Committee
- an Australia Day on the 8th at which current Australian projects relating to digital resources generally will be presented
- an Information Day on the 12th at which web archiving tools and methodologies from around the world will be presented.

The main objective of the conference is to identify significant issues facing cultural heritage institutions in collecting web resources and to explore how the issues are being addressed internationally. Particular emphasis will be given to reviewing the approaches collecting institutions are currently taking and to identifying research projects underway to support the different approaches. The conference is not intended to be technical, rather to concentrate on the business and strategic needs of collecting institutions.

Information gathered via the conference that is relevant to the ICABS action relating to web archiving will be used to augment information that is already available through PADI.

(iii) Automate deposit and archiving of online government resources

In early 2003, the Library launched the Commonwealth Metadata Pilot Project which has two main objectives:

- to improve national discovery of and access to online Australian Commonwealth Government publications through the National Bibliographic Database (NBD) available through the service (http://www.nla.gov.au/kinetica), and
- to streamline identification and archiving of these publications in PANDORA or another public archive to ensure on-going access to them via the NBD.

An underlying objective is to review metadata standards required and to develop and promulgate guidelines to encourage participating agencies to create good-quality metadata.

The Pilot Project involves a number of government agencies and it will extend until 2005. It consists of the following stages:

- exploring options (including harvesting) for acquiring metadata from the partner agencies;
- converting gathered metadata from Dublin Core to MARC21 via MODS for loading onto the NBD;
- automating the harvesting, conversion, loading and updating of metadata; and
- enhancing the PANDAS software to trigger automated harvesting and archiving of the web resources associated with the metadata (this is expected to occur during 2005).

(iv) Explore and promote approaches to archiving web-based publications collected by harvesting.

Useful information on this topic is already available through PADI and as mentioned above, the conference on web archiving will provide a vehicle for exploring further current practices and recent developments. In addition, a survey of current activities by a range of cultural agencies will be undertaken to ensure these are also represented in PADI.

In order to make the information available via PADI more easily accessible, PADI will be restructured in 2005. The restructure will address some current issues with the system that affect resource discovery. It will aim to make PADI more useful by reorganising key resources according to popular topics – including Digital Archiving – and making these topics (to be called Trails) accessible from the PADI Home Page.

Preservation of digital materials (ICABS Action 3.4)

The objective of our activities in this action area is to survey existing standards, guidelines and codes for preservation of digital materials, in cooperation with IFLA's Preservation and Conservation Section.

Development of programs for preserving digital materials is a daunting challenge for many libraries. The purposes of our activities in this area are to identify factors that influence the readiness of libraries to get involved, and to identify forms of guidance that libraries can use to improve their readiness or to make their existing programs more effective.

Use of PADI to identify existing guidance documents and to make these more easily accessible, is central to this action. Significant gaps will also be identified and brought into PADI by mid-2004. As mentioned above, PADI structures and interfaces are being restructured to enable information on key topics such as standards and guidelines, to be more easily accessible.

The Library's relationship with other libraries in the South East Asian and Pacific regions will also provide an excellent opportunity for assessing the form of guidance that would be most useful in a large, diverse region. The Library serves as the Regional Centre for Oceania and South East Asia for the IFLA Preservation and Conservation Program and has working ties with the UNESCO Information Society Division's campaign for the preservation of digital heritage.

In November 2002, UNESCO held a Regional Consultation on digital preservation in association with the preparation of guidelines which the National Library of Australia was commissioned to write. Based on the information gathered, in February 2004 we prepared a checklist of readiness factors and guidance on improving readiness, for a UNESCO Memory of the World nomination workshop held in Manila, The Philippines. The readiness factors include for instance, an existing collection of digital resources; access to Information Technology knowledge and skills; and dedicated staff resources to manage the digital collection.

These documents will be revised in the light of feedback from the workshop and prepared for wider circulation by June 2004. They will of course be accessible via PADI.