What Should We Preserve? The Question for Heritage Libraries in a Digital World

Margaret E. Phillips

Abstract

A primary role of national libraries is to document the published output of their respective countries. Traditionally, this has meant collecting, describing, and preserving for future generations at least one copy of every item published in print, including books, serials, newspapers, maps, music, posters, and pamphlets. In the last decade, online publishing has had a revolutionary impact on the creation, publication (dissemination), and use of information. This has presented libraries, particularly national (deposit) libraries and other cultural collecting institutions, with the daunting task of collecting, storing, describing, managing, and preserving the vast quantities of information that are being produced online.

A key question to be asked when embarking on this task is, "What should be collected and preserved?" National libraries have responded to this question in different ways. Some, including the National Library of Australia, have taken a selective approach, while others have engaged in whole domain harvesting, or a "comprehensive" approach. This article discusses the advantages and disadvantages of each of these approaches and looks in some detail at the selective approach as exemplified by PANDORA, Australia's Web Archive.

Introduction

A primary role of national libraries and other deposit libraries is to document the published output of their jurisdictions. Traditionally this meant collecting, describing, preserving, and providing access to library materials for current and future generations. Library materials have included printed books, serials, newspapers, maps, posters, music, and pamphlets.

LIBRARY TRENDS, Vol. 54, No. 1, Summer 2005 ("Digital Preservation: Finding Balance," edited by Deborah Woodyard-Robinson), pp. 57–71

Subsequently the definition of "library materials" was extended to include information stored on other physical carriers such as microfilm, film of various types, audio cassette tapes, video tapes, computer disks, CD-ROMS, and DVDs. These have all presented challenges to libraries because of the need for special equipment to display items in these formats, obsolescence of this equipment and/or the formats themselves, and the need to preserve the information contained on sometimes fragile storage media.

With the development of the World Wide Web in 1993, which opened up online publishing as an easily available, ubiquitous, and relatively inexpensive means of creating and distributing information, national and other deposit libraries accepted that, once again, they must expand their roles to encompass this new form of publishing and all that its collection, description, storage, management, preservation, and provision of access entailed. There are additional challenges to face over and above those inherent in the formats that they already collected. The volume of online publishing is huge. Almost anyone can set themselves up as a publisher, meaning that issues of quality and authority of information need to be addressed, as well as a wide range of competence (or otherwise) in using publishing software and compliance in applying standards. In addition, many of these items are complex Web objects—for instance, Web sites that contain a number of different file formats—and this makes strategies for preservation particularly difficult to formulate and undertake.

What Should Be Collected and Preserved?

While national and other deposit libraries have largely accepted responsibility for collecting and preserving online publications, at least in principle, those that have embarked on the task have responded to it in different ways. They have assessed the task before them in relation to the resources available and have made different decisions about what "finding the balance" is in their particular situation.

Some have argued that, because national and other deposit libraries are typically comprehensive in collecting the published output of their jurisdiction, this same approach should prevail with online publishing. As far as humanly possible, all online publishing must be collected and preserved. Others have argued that, because online publishing is a completely different paradigm from print and other physical format publishing and a different order of magnitude, then a different, selective, approach is necessary and acceptable, and perhaps even desirable. This has led to two broad national approaches to collecting and preserving online publications—the whole domain or comprehensive approach, and the selective approach.

In the mid- to late 1990s a small number of national libraries began archiving programs and exploring different approaches to archiving national documentary heritage online. It is interesting to note that, within five or six years of embarking on a chosen course, most of those libraries seemed to be at a crossroads with regard to planning their future directions for digital archiving (Gatenby, 2002). Whether they were engaged in whole domain (comprehensive) harvesting or selective archiving, each was recognizing the limitations of their chosen approach. There are a number of approaches that national libraries are currently employing to build archives of their countries' publications, which are discussed below.

Selective Archiving of Static Web Resources

The National Libraries of Denmark and Canada have been the principal exponents of this approach. Resources that are like print publications and that do not change or contain interactive or dynamic elements are archived on a selective basis, with library staff making the selection decisions.

Selective Archiving of Static and Dynamic Web Resources

Australia is the only known country with an established program for archiving dynamic as well as static publications and Web sites on a selective basis, once again with a high degree of intellectual input from library staff.

Whole Domain Harvesting

Libraries attempt to harvest automatically the entire Web domain of their respective countries using harvesting robots and a minimum of human intervention for identifying resources. This involves harvesting not only all the resources in the specific country domain but also identifying those of country origin or subject matter in .com and other generic domains. The National Libraries of Sweden, Finland, Iceland, Norway, and more recently Austria have been pursuing this approach. The Internet Archive, a public, nonprofit organization in the United States, attempts to archive the whole Web every two months.

Combination of the Selective and Whole Domain Approaches

The Bibliothèque nationale de France is involved in a research project to program a robot to archive both automatically and selectively those resources likely to be of research value. Researchers there have asked the question, "Is it possible to define relevant and automatically computed parameters to focus a robot on only that part of the Web we want to archive?" (Masanès, 2002).

Thematic Approach

The Library of Congress undertakes some selective archiving and, in conjunction with partners, including the Internet Archive, also builds thematic collections that are the result of harvesting as much as possible on a given subject, such as the 2002 election and the events of September 11, 2001 (Kresh et al., 2004). Supplementing its online deposit collection, in 2001 and 2002 the Royal Library of Denmark worked with the State and University Library, Aarhus, and the Centre for Internet Research at the University of Aarhus to test the viability of the thematic approach (event-based archiving) through the Netarchive.dk project (Royal Library, Denmark, 2003). The Royal Library and the State and University Library together have gone on to incorporate event-based archiving into a three-pronged approach to Web archiving, including automatic snapshot harvesting and selective harvesting (Royal Library, Denmark, 2004).

Archiving Based on Collaborative Agreements with Selected Commercial Publishers

The National Library of the Netherlands has taken a different approach altogether, responding to a particular situation where 30 percent of all scientific publications in the world occur in that country. It has focused on commercial publications and, in association with IBM, has developed technical infrastructure and organizational relationships with a small number of commercial publishers, including Elsevier Science and Kluwer Academic, to archive, preserve, and provide limited access to the whole digital output of the publishers concerned (National Library of the Netherlands, 2004). It takes in large volumes of online publications from a small number of publishers. Collaborative agreements with publishers also work well under the selective model, and the National Library of Australia and the Commonwealth Scientific and Industrial Research Organisation (CSIRO) have recently reached an agreement whereby the library will archive all of CSIRO's online commercial publications.

The Selective Approach to Archiving

Advantages

A selective approach to archiving enables libraries to achieve six important objectives:

- Each item in the archive is quality assessed and functional to the fullest extent permitted by current technical capabilities.
- A gathering schedule can be individually tailored for each selected title, taking into account its publication schedule or the frequency with which the Web site changes, thus enabling the content gathered to be as complete as possible.
- Each item in the archive can be fully catalogued and therefore can become part of the national bibliography.
- Each item in the archive can be made accessible via the Web to readers immediately because permission to do so can be negotiated with publishers.
- The "significant properties" of individual resources and classes of resources within the archive can be analyzed and determined. (These are the attributes that convey the full meaning and intellectual content of an item and enable it to be experienced as the creator intended.)

This enhances our knowledge of preservation requirements and enables risk assessments and preservation strategies to be put in place.
- Sites that are inaccessible to harvesting robots can be identified and archived using other methods, by arrangement with the publisher.

Disadvantages

In selecting titles for the archive, libraries are making subjective judgements about the value of resources and what researchers of the future are likely to find useful. Librarians have always made these collection development decisions. However, the print environment has been a more established, structured, stable, and predictable environment in which to make such decisions.

Dissemination of information online is still in its infancy, and the way that researchers will want to access, use, and apply the potential of the Web is also still under development. Though we believe that we are selecting titles based on sound professional experience and judgement, do we really know what will be important for future researchers? Selection is largely based on a traditional understanding of the concept of "publication." Perhaps in the future this will not be as relevant, or, perhaps more likely, something in addition to this traditional approach will also be required.

The extent of a selective archive is very limited in comparison with the large volume of material in a country's domain. While it is valid to argue that a lot of this material is of no future research value, it is also certain that resources that do have research value are being missed. The selective approach is very labor-intensive, and the unit cost per item is therefore high. The amount of material that can be archived at any one time is heavily dependent on and proportional to the number of staff that can be allocated to the activity. In a time of contracting funding for staff, the amount of archiving that can be carried out also contracts, unless increased sophistication of the technical infrastructure supporting the archive can be brought to bear to counteract it. The selective approach takes a resource out of context and often does not include other resources to which it is linked. Contextual meaning is therefore lost, and this will be more critical for some resources and research requirements than others. The value of "sampling" is as yet unproven. Will this approach satisfy the majority of research needs for these kinds of resources in the future?

Whole Domain Harvesting

Advantages

In theory, the obvious advantage of the whole domain harvesting approach is that the whole domain is captured automatically at periodic intervals with minimal human intervention and therefore comparatively low staff cost per item gathered. The whole domain is available to future researchers, and resources can be seen in their broader context, with links to other documents retained.

Disadvantages

In practice this ideal is a long way from being the reality. Because whole domain harvests are demanding in relation to computer time and storage space, they are usually run at intervals of at least a couple of months. Any publications that come into being and disappear in the interim are missed. Any changes to existing sites that are made and overwritten in that period will also be missed.

Because of the huge volume of publications involved, quality control checks cannot be made on more than a very small sample of titles. The experience of the National Library of Australia would suggest that approximately 40 percent of harvested titles could be incomplete or defective in some way. Nationally significant material is likely to be missing, and the archive administration will not be aware of it. It is possible that, in time, improved intelligence of harvesting software and reliable quality-checking software may increase the accuracy of automatic archiving and therefore ameliorate this disadvantage.

While staff costs per item are low in comparison to the selective approach, the whole domain approach is expensive in terms of costs to download and store data. With the current level of system reliability, there can also be the need for a staff member to monitor the harvesting process twenty-four hours a day, seven days a week, and to intervene when problems arise.

Commercial sites that employ passwords or other inhibitors (a part of the Web referred to as the "Deep Web" to access) will not be accessible to a harvesting robot and therefore will not be gathered. Some of the most important digital heritage is found on commercial sites.

Whole domain archives still have major drawbacks from the point of view of resource discovery and access, although it is likely that these problems will be resolved in time through improved methods of gathering and organizing descriptive metadata. The Nordic Web Archive has done groundbreaking work in the area of indexing and free-text searching across the contents of diverse archives (Nordic Web Archive, 2002). For copyright reasons, access to whole domain archives is usually strictly limited, and at best the contents may be available within the library building. The Swedish National Library has made a major gain in the area of access through ministerial support and a government decree that authorizes the library not only to collect Swedish Web sites but to allow public access to them on library premises (National Library of Sweden, 2002). Appropriate legal deposit legislation enacted by national governments could ease this limitation of the whole domain approach by permitting unfettered access to freely available, noncommercial publications and Web sites archived by national libraries.

The only example of a whole domain archive that is readily available for evaluation is the Internet Archive, which attempts to capture the whole Web every two months. Valuable though this resource is, having commenced its work in 1996 and now having amassed a considerable volume of historical data, it does have limitations of concern to agencies looking for completeness and version control of documentary heritage.

Hybrid Approaches

All of the approaches discussed so far have disadvantages—the selective approach misses material that may be of future value, the whole domain model is not as comprehensive as its name would suggest, and collaborative agreements with publishers to date exclude the majority of publishers and a lot of freely available resources. A multipronged approach that combines a periodic snapshot of a country's domain, supplemented by selective archiving of nationally significant, authoritative publications of long-term research value and provision for deposit of publications by agreement with specific publishers, would be ideal. As already described, the Royal Library of Denmark and the State and University Library, Aarhus, have recently embarked on a hybrid approach (Royal Library, Denmark, 2004). Funding is an issue, however, with each approach having its own technical infrastructure and staff support costs. Most libraries are struggling to support just one approach.

PANDORA, Australia's Web Archive

This section examines in more detail a particular approach to selective archiving as implemented at the National Library of Australia through PANDORA, Australia's Web Archive (National Library of Australia, 2004). In late 1995 the National Library of Australia accepted that it needed to collect and preserve online publications. It recognized that there was information about Australia and by Australians appearing on the Internet that was not available in any other format. It was the content, rather than the format, that was the determining factor for collection. Under the National Library Act of 1960, the National Library has a mandate to collect, preserve, and provide access to a collection of documentary resources that supports in-depth research on all matters relating to Australia, its history, its culture, and its peoples. These resources come on a wide variety of media, including whalebone, stone, gum leaf, and glass, as well as the more usual media of paper, film, and disk. The Internet is just another medium, after all, and while it undoubtedly poses unique challenges, there was no reason to exclude it on this basis.

The National Library realized from the very beginning that, in order to develop an archive of sufficient breadth and depth, collaborative effort among collecting institutions with similar objectives would be essential. It developed policies, procedures, a template for selection guidelines, and technical infrastructure to support a collaborative Web archiving program and invited the state libraries and other collecting agencies to join it. The first partner, the State Library of Victoria, joined in 1998, and by 2004 there were ten partners in all, including five state libraries and the Northern Territory Library and Information Service. These partners are all deposit libraries, like the National Library, and have similar responsibilities in collecting and preserving the published outputs of their jurisdictions. Three other partners—ScreenSound Australia (The National Screen and Sound Archive), the Australian War Memorial, and the Australian Institute of Aboriginal and Torres Strait Islander Studies—bring important subject expertise, which enriches and deepens the scope of the archive. Each of the partners develops its own selection guidelines, which are published on the PANDORA Web site,[1] and thus takes responsibility for a stated portion of the national published output online.

The original impetus for this work came from the collection development area of the National Library and from librarians with all the traditional library skills. These origins for the Web archiving program have strongly influenced the approach taken—one that can be seen as largely an extension of traditional library practices with emphasis on balanced and rounded collection development, quality and authority of resources collected, and description of resources according to full MARC records and inclusion in the national bibliography. All of these collection-building activities relating to PANDORA are now carried out by the Digital Archiving Branch in the Collections Management Division.

The Library's Information Technology Division has provided strong support from the beginning and has developed and maintained the technical infrastructure, including the PANDORA Digital Archiving System (PANDAS), Web-based software that supports the collaborative Web archiving program. The Preservation Services Branch was also involved in defining policies and procedures from the beginning, and it takes responsibility for ensuring that the library can provide long-term access to the contents of the archive as the hardware and software on which it is dependent for display changes over time.

The selective approach to Web archiving enabled the staff of the National Library of Australia to start in a small way and to learn how to manage this completely new task as we went along. It enabled us to make a start in a practical way, without being overwhelmed by the enormity of the task. It also enabled us to tailor the activity to the staff resources available and to do something rather than nothing. Since the library and its partners have at no stage received additional funding for this costly activity, it has been important to focus our activity and to collect and preserve those publications and Web sites considered most likely to be of long-term value to researchers. Sometimes "finding the balance" comes down to something as practical as this.

The first task, when we started this work in 1996, was to decide what we would collect, and this resulted in the publication of our first selection guidelines. As we implemented these guidelines and learned more about what was in the Australian domain, and as the Australian domain grew and developed, the selection guidelines were modified to accommodate new categories or to clarify our approach to them. We have always applied them very flexibly, being more inclined to include, rather than exclude, an item that was borderline.

The essentials of the selection guidelines, then and now, are the following:

- To be selected for national preservation, a significant proportion of a work should be about Australia;[2] be on a subject of social, political, cultural, religious, scientific, or economic significance and relevance to Australia and be written by an Australian author;[3] or be written by an Australian author of recognized authority and constitute a contribution to international knowledge.
- It may be located on either an Australian or an overseas server. Australian authorship or editorship alone is insufficient grounds for selection and archiving. In the case of online publications, content is the pre-eminent factor determining selection.
- When a title is available both in print and online, the print version only will be collected. The online version will be collected only if it has significant additional information or value. This is a policy that the library adopts only out of necessity. Especially in relation to government publications, the library would like to be able to collect both the print and online versions because of the access advantages. However, the staff resources needed to collect both versions are not available.
- Highest priority is given to authoritative publications with long-term research value, and these are selected and archived as comprehensively as possible. In addition, the library seeks to include in the PANDORA Archive examples of the different types of online publications on a wide range of subjects to document Australian society as it is represented on the Internet.
- The library does not attempt to collect all versions/editions of all changing sites but sets a manageable gathering schedule for each title based on its content, publication pattern, long-term research value, and the stability of the site.
- Links to external resources from a selected publication are not archived.
- Content is the pre-eminent criterion for selecting online publications and Web sites. Static publications, dynamic Web sites, and databases will all be selected for archiving if the content is within scope of the guidelines. In practice, technical limitations at a given point in time may inhibit our ability to actually archive a publication or Web site. However,

> we do our best to solve the problems presented by a publication or class of publications. For instance, our desire to archive Deep Web sites, including databases, has led us to embark on a research project in conjunction with the International Internet Preservation Consortium (IIPC) to find or create tools and methods for collecting and preserving information presented in these dynamic formats.

In 2003, after seven years of selecting and archiving online publications and Web sites, the library conducted a major review of its selection guidelines to ascertain whether they remained relevant and flexible enough to encompass new categories of Web resources that had appeared in the intervening period. This review also considered whether the selective approach to archiving was still the most valid approach for the library. It concluded that, under ideal circumstances of adequate funding, the library would like to undertake periodic Australian Web domain harvests to supplement the selective archive. However, this is currently beyond the means of the library. Living within our means, the advantages of the selective approach still outweighed the disadvantages, and it remained the most viable for this library for the time being.

In embarking on this review of the selection guidelines, it was anticipated that new categories of online publications would be identified for collecting and that recommendations would be made to expand the scope of the selection guidelines. The review did, in fact, highlight that there were types of resources that were not being collected, which would likely have long-term research value. It also identified significant gaps in our collecting of some categories already included in the selection guidelines.

When the library commenced Web archiving in 1996, the volume of online publishing was much lower than it is today. For instance, there was relatively little Commonwealth government publishing in online formats only before 2000. Since then the volume of material has mushroomed, but the resources available to deal with it have not. The volume of online publications that meets the selection guidelines is much greater than the staff available for the activity can manage.

Facing the reality that the library was unable to archive everything that it would like to, some hard decisions had to be made. The review recommended that, rather than expanding the selection guidelines, it should prioritize its collecting of online publications currently within scope to focus on six categories. The choice was between collecting a broader range of publications superficially or focusing the collection activity and archiving defined areas in some depth.

The six categories to receive priority are as follows:

- Commonwealth and Australian Capital Territory (ACT) government publications (the state library partners take responsibility for state government publications)

- Publications of tertiary education institutions
- Conference proceedings
- E-journals
- Items referred by indexing and abstracting agencies (which are frequently from the first four categories but also include items with print versions)
- Sites in nominated subject areas (specified in an appendix to the selection guidelines) on a rolling three-year basis, and sites documenting key issues of current social or political interest, such as election sites, the Sydney 2000 Olympics, the Canberra bushfires, etc.

Even these six categories cannot be archived comprehensively. For instance, conference proceedings are further limited:

- Only sites that contain the full text of a substantial number of papers presented at a conference will be archived. Powerpoint presentations alone do not contain sufficient information to warrant archiving.
- Priority will be given to conferences held by Commonwealth and ACT government bodies, professional associations and institutes, and academic disciplines.
- The proceedings of international conferences that are affiliated with an Australian body and that are held in Australia may be selected for archiving. The relevance of the content to Australia will be a factor in influencing selection.
- Preference will be given to major conferences over small seminars such as those held by a university department.

Categories that had not been collected prior to 2003 and that would continue to be excluded were

- datasets
- online daily newspapers
- news sites
- discussion lists, chat rooms, bulletin boards, and news groups
- web cams
- blogs (except those that support the tertiary institutions publications category)
- portals
- games.

Except for portals and games, which were excluded for good reasons, the remainder was excluded reluctantly. There is much content in these other categories of research value, and they were excluded at this stage largely because of resource constraints.

In some ways this outcome for the review was disappointing. It was apparent that radical innovation would be required to empower the library and its partners to collect online publications at an adequate level. The National Library was already planning ways to increase dramatically its intake of online publications, especially government publications.

Government publications are a very important category for collecting, both for the National Library and its state library partners. Having collected Commonwealth government publications comprehensively in print, the National Library now found itself in a position where it could collect only a small fraction of government online publications, an increasing number of which are available in no other format. Even had the staff available for Web archiving done nothing else but archive government publications, they would still only manage to archive a fraction of the available publications. The state library partners were under similar stress.

At the beginning of 2003 the National Library launched the Commonwealth Government Metadata Pilot Project (later renamed the Australian Government Metadata Project). The aim of this project was to (1) increase significantly the coverage of government publications in the National Bibliographic Database (NBD), which is made available through the Kinetica[4] service; and (2) batch load this metadata into the PANDORA Digital Archiving System to trigger automatic harvesting and archiving of these publications.

The first part of the strategy for doing this was to work initially with seven government agencies of different sizes to identify work flows for creating metadata for government publications and contributing it to the NBD. As a result, a small number of models for contributing metadata to the NBD would be formulated, which other agencies could then use. Some metadata is now being routinely added to the NBD as a result. Another seven agencies joined the project in 2004. The second part of the strategy is to enhance the PANDORA Digital Archiving System to enable it to receive batch-loaded metadata and to harvest and archive publications with as little human intervention as possible. This enhancement is underway. This development will mean that, for government publications, the national library will relinquish control over the selection of titles for inclusion in the PANDORA Archive. The government agencies will define what is to be collected by including descriptions for it in their metadata sets, which are made available to the NBD.

Moving Forward

Thus far this article has looked at the approaches being taken by national libraries to collecting their online documentary heritage and has examined in detail a particular implementation of the selective approach, the PANDORA Archive. It is possible that in the future we will look back on this early period of Web archiving and see it as the exploratory phase, when national libraries individually sought their own solutions. Only a handful took up the challenge in the mid- to late 1990s, the early days of the World Wide Web.

Those national libraries that were active in the field have developed a lot of knowledge, expertise, systems, and software to manage the activity. Most of this knowledge and these systems have been developed in isolation, with sharing of information taking place at conferences and workshops, through visits of staff from one organization to another, and through publication in professional journals and on organizational Web sites.

As mentioned earlier in this article, around 2002 those libraries with well-established Web archiving programs were beginning to question whether their chosen models were meeting all their needs. The formation of the IIPC in 2003 was an indication that libraries had realized the limitations of working on their own and the value of collaboration and shared effort and infrastructure.

The foundation members of the IIPC are the Bibliothèque nationale de France (coordinator); the National Library of Italy; the Royal Library, Denmark; Helsinki University Library, the National Library of Finland; the Internet Archive; the Royal Library, National Library of Sweden; National and University Library of Iceland; Library and Archives Canada; the National Library of Norway; the National Library of Australia; the British Library; and the Library of Congress. "The mission of the IIPC is to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations" (International Internet Preservation Consortium, 2004a). Its goals are as follows:

- "To enable the collection of a rich body of Internet content from around the world to be preserved in a way that it can be archived, secured and accessed over time.
- To foster the development and use of common tools, techniques and standards that enable the creation of international archives.
- To encourage and support national libraries everywhere to address Internet archiving and preservation" (International Internet Preservation Consortium, 2004b).

The IIPC is achieving its goals through members' active participation in six working groups. Through these working groups the IIPC plans to create a shared technical basis for Web archiving activities; to develop procedures and tools for providing immediate and long-term access to Internet material; to devise means for evaluating coverage and performance of Web archiving programs; and to identify strategies and produce tools for archiving content that is inaccessible to crawlers.

The question "What should we preserve?" is directly addressed by one of these working groups, the Researchers Requirements Working Group. It recognizes that, because of the huge volume of material on the Internet, it is inevitable that not everything can be collected. This means that the decisions that we make about what to collect now will have an enduring impact on what is available to researchers of the future. This working group, which consists not only of members but also of invited researchers in the area of Internet studies, is aiming to define a common vision of what needs to be collected (International Internet Preservation Consortium, 2004c).

The methods, procedures, and tools developed by the IIPC will be available to all members, as well as to other national libraries with Web archiving programs. It is to be hoped that this shared infrastructure will enable national libraries to expand significantly what they can collect and preserve. In a few years time our national Web archives may be very different from what they are today.

Conclusion

The question "What should we preserve?" has been addressed in different ways by national libraries with responsibility for collecting and preserving online publications and Web sites. At this stage, whether the whole domain or selective approach has been adopted, or collaborative arrangements have been entered into, or hybrid approaches put in place, it is likely that all would in part answer this question with "More than we are currently able to do." In this first decade of the Web, programs by national libraries to archive Web resources have necessarily been experimental, and what we have collected for long-term preservation has often been determined as much by our limited (though developing) technical capability as by our judgment of what is likely to have research value in the future. We know, for instance, that there is a huge volume of valuable information in the Deep Web that has been beyond our reach.

However, it has been important to do what we could with what we have available. In doing so we have learned a lot, are still learning fast, and are developing increasingly sophisticated methods and tools for dealing with different collection scenarios and with different types of publishing formats. The formation of the IIPC and the collaboration and shared effort that that stands for will be a springboard into collecting and preserving the second decade of the Web. During that period our national archives will become very different beasts.

Notes

1. The selection guidelines of all partners are available at http://pandora.nla.gov.au/selectionguidelinesallpartners.html.
2. This is defined as "dealing with Australians, with Australia, or with a State, Territory or any other subdivision of Australia" (National Library of Australia, 2003).
3. An Australian author is one who was born and has resided in Australia; who has continued to be recognized as Australian although residence in Australia has not been continuous, or who, although not born in Australia, has been identified through work and residence in Australia as an Australian.
4. Information about Kinetica is available at http://www.nla.gov.au/kinetica/.

References

Gatenby, P. (2002). *Report on Senior Executive Fellowship to research digital archiving in national libraries.* Canberra: National Library of Australia. Retrieved February 7, 2005, from http://www.nla.gov.au/nla/staffpaper/2002/elect.html.

International Internet Preservation Consortium. (2004a). *Netpreserve.org.* Retrieved February 7, 2005, from http://netpreserve.org/about/index.php.

———. (2004b). *Mission.* Retrieved February 7, 2005, from http://netpreserve.org/about/mission.php.

———. (2004c). *Researcher Requirements Working Group.* Retrieved February 7, 2005, from http://netpreserve.org/about/researchers.php.

Kresh, D., Ammen, C., Thomas, D., Grotke, A., Hayes, A., & Guenther, R. (2004). *Oh, the places you'll go!: Recommending in the digital environment* [Powerpoint presentation]. Washington, DC: Library of Congress. Retrieved February 7, 2005, from http://www.loc.gov/minerva/presentations/DigFutures.ppt.

Masanès, J. (2002). Towards continuous Web archiving: First results and an agenda for the future. *D-Lib Magazine, 8*(12). Retrieved February 7, 2005, from http://www.dlib.org/dlib/december02/masanes/12masanes.html.

National Library of Australia. (2003). *Online Australian publications: Selection guidelines for archiving and preservation by the National Library of Australia.* Retrieved August 15, 2005, from http://pandora.nla.gov.au/selectionguidelines.html.

———. (2004). *PANDORA, Australia's Web Archive.* Retrieved February 7, 2005, from http://pandora.nla.gov.au/index.html.

National Library of Sweden. (2002). *New decree for Kulturarw3* [Press release]. Retrieved February 7, 2005, from http://www.kb.se/Info/Pressmed/Arkiv/2002/020605_eng.htm.

National Library of the Netherlands. (2004). *The archiving system for electronic publications: The e-Depot.* Retrieved February 7, 2005, from http://www.kb.nl/kb/dnp/e-depot/dm/dm-en.html.

Nordic Web Archive. (2002). *Final report for the Nordunet2 Secretariat.* Retrieved February 7, 2005, from http://folk.uio.no/mdahl/17-FR.pdf.

Royal Library, Denmark. (2003). *Experience and conclusions from a pilot study: Web archiving of the district and county elections 2001: Final report for the pilot project "netarkivet.dk"* [English version]. Retrieved February 7, 2005, from http://www.netarkivet.dk/rap/webark-final-rapport-2003.pdf.

———. (2004). *About netarchive.dk: 2nd phase.* Retrieved February 7, 2005, from http://netarchive.dk.

Margaret E. Phillips, Director of Digital Archiving, National Library of Australia, Parkes Place, Canberra ACT 2600, Australia, mphillips@nla.gov.au. Margaret Phillips is Director of Digital Archiving at the National Library of Australia. She has worked in libraries since 1976 and joined the staff of the National Library of Australia in 1987. In the mid-1990s, as manager of Acquisitions, she increasingly dealt with electronic materials and from 1996 began to devote full-time attention to online publications in her capacity as manager of the unit that builds PANDORA, Australia's Web Archive. She has been closely involved in establishing policy, procedures, and infrastructure for ensuring long-term access to Australian Internet publications.