Date submitted: 17/05/2010

Collecting online newspapers: the National Library of Australia experience with archiving the *Sydney Morning Herald* (smh.com.au)

Pam Gatenby
Assistant Director General, Collections Management
National Library of Australia
Canberra ACT, Australia
(pgatenby@nla.gov.au)

[This paper draws heavily on an internal project review document prepared by Paul Koerbin, Manager of Web Archiving at the National Library of Australia. His contribution to this paper is acknowledged.]

Meeting: **102. Newspapers**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract

The National Library of Australia commenced archiving selected online newspapers during 2009 for access via its PANDORA web archive. The undertaking has raised many issues, confirming that newspapers, especially in online form, are complex entities. The presentation will demonstrate and explain the approach we are taking with reference to the Sydney Morning Herald (http://nla.gov.au/nla.arc-101523) and will discuss the main issues we have encountered so far. These relate to definitions and scope; the special features and characteristics of online newspapers; curatorial decisions about the value of what can be collected; technical and system constraints; and resource implications. The presentation will also present the outcome of a review of our current approach that will be undertaken mid-way through 2010.

1. Background

The National Library of Australia commenced selective archiving of "born digital" web publications in 1996 when it established PANDORA: Australia's web archive (http://pandora.nla.gov.au/). PANDORA selection criteria give priority to web resources of national importance reflecting all aspects of life. Online publications with a print equivalent are for the most part not collected on the assumption that they mainly duplicate content available in print form. Until recently this selection criterion was used to exclude online newspapers but it is now quite evident that the nature of online newspapers has changed considerably.

Following a review of our web archiving activity in 2008-2009, we decided we needed to be more active in seeking to collect online newspapers. Up to this point some newspaper type content had been collected in an ad hoc manner, including specialist non-mainstream online news and opinion sites like Crikey (http://pandora.nla.gov.au/tep/13027) and the blogs of some political commentators.

2. What is an online newspaper?

A key consideration for us in deciding to collect online newspapers was to understand their characteristics and whether the traditional concept of a newspaper remains relevant in the online context. Online newspapers are in a period of quite dynamic change and they seem to be trying to be a number of things. Some common features are:

- they include not only news reporting but also substantial online content in the form of opinion pieces which might not be replicated in the print version, and blogs and contributor comment;
- they serve as complex portals to a range of services and information and use new web technologies to present the content in a dynamic way;
- their content changes regularly; and
- they can be very large in terms of file size.

The rapid development of online newspapers suggests they will develop further in ways we cannot foresee and that therefore attempting to deal with them in the same way as print titles will not be useful. It may be more appropriate and practical for collecting purposes to target specific content they publish such as the blogs, comments and 'front pages", or whatever aspects have significance for an institution's collecting policy. In making this decision, practical considerations such as the frequency with which content on the sites changes need to be taken into account as this can pose challenges for timely and efficient capture, depending on the digital archiving infrastructure in place.

3. What we are collecting

In the second half of 2009 the National Library of Australia began systematic action to build a collection of online newspaper content. First, ethnic community newspapers were selected in support of the Library's broader collection development focus on multicultural publishing. For the most part these are not complex to collect. They are much smaller than mainstream dailies, it is more straightforward to negotiate permission to archive, and content is more static. Permission was also obtained at this time from the Fairfax Company to archive content from the online *Sydney Morning Herald* (smh.com.au) which provided us with the opportunity to begin dealing with the scoping, and procedural and technical issues associated with genuinely complex and dynamic major online newspapers.

This presentation will focus on our experience with the *Sydney Morning Herald* and the issues identified after one year of archiving the title.

We now have around 45 titles listed under the "Newspapers" subject category in PANDORA. (See http://pandora.nla.gov.au/subject/221 for a list of the titles.) These include 22 individual born digital ethnic community papers, four titles for which newspapers are the subject interest (that is, they are not themselves newspapers) and an assortment of other special interest newspaper type publications.

4. Sydney Morning Herald (smh.com.au)

The *Sydney Morning Herald* is one of Australia's longest running and reputable newspaper titles. It has been owned and published by the Fairfax Company since 1842. The online newspaper is a complex site with articles and blogs covering many topics. Following discussions with the Editor-in-Chief of the online version, which clarified its composition and the editorial decisions that determine content, and taking into account the complex, dynamic nature of the site, the Library decided to limit collecting to manageable components that we consider have particular cultural and historic value. The prominence given to opinion writing and blogs is an obvious aspect of online newspapers that we decided should be preserved in this case. In addition, we decided to collect and archive the "front page"* of each online issue as this captures the most important stories of the moment and has significance in its own right. The "pages" convey editorial decisions on what would have particular appeal to the online readership and decisions on how to present the content to capture audience attention. There is usually something for everyone, ranging from more serious political and business news to popular appeal stories on all aspects of life.

* [The concept of the "front page" of online newspaper titles is not necessarily straightforward. It is obviously the main page (or first screen) but could also include, for the purposes of archiving, one level of content linked from the main page. The headline and opening paragraph of a story will appear on the main page but capturing the "front page" would include the linked full story.]

Our approach to collecting the SMH

The *Sydney Morning Herald* has been collected on a daily basis, seven days a week, since June 2009. A daily archiving schedule has been set up in PANDAS, the gathering and content management system that supports PANDORA, which captures the content in the early hours of the morning, normally sometime between 12 midnight and 2am. This instance is retained for weekend captures but during the working week a harvest is manually initiated at 10am. Providing the 10am harvest is completed successfully this is retained for the archive and the instance harvested in the early hours is deleted from the working area and not retained for the archive. The rationale of this approach is to capture a mid-morning prime reading time version of the newspaper but to have the version available between 12 midnight and 2am as a backup copy if required.

The harvest is scoped to capture the 'front page' of the newspaper. As discussed above, this is more than an image snapshot of the first screen as it includes the content available from one additional 'click'. So the content directly linked from the first page is also captured meaning that, in most cases, not only the headline text is captured but also the substance of the main articles.

The size of the daily harvest is around 50 Mb and includes something in the order of 4,500 files and takes around 10 to 15 minutes to collect. While it is difficult to say what an 'average' size site is for harvesting purposes, given that everything from a single PDF file to very substantial sites are collected for PANDORA, this could be considered, in isolation, a relatively small site. However, as that constitutes nearly 18 GB over a year it might also be considered a very substantial harvest (given that the most common archiving schedules are yearly and half-yearly).

In terms of staff time involved, harvesting the "front page" includes around 5 minutes at 9:55am preparing to send off the manual harvest request through PANDAS, which can also involve pausing harvests currently running so as to allow the *Sydney Morning Herald* harvest to run, as only four concurrent harvests can be run in PANDAS.

The quality assurance (QA) process takes around 10 minutes per day and involves copying a number of pre-edited files to the harvested instances to fix display problems; then doing a spot check on the instance to see that it works and that there are no new format changes to the site. The instance is archived and then published to title entry page. The 'midnight' harvest is deleted after the 10am archiving is completed. On Mondays and after public holidays there are the weekend or holiday instances to check, fix, archive and publish.

The display of collected content in PANDORA is based on a title page for the content which conveys information such as the number of instances collected, the Persistent Identifier for the content and whether on not it is still available from the live Internet. (See http://pandora.nla.gov.au/tep/99834 for an example of a title page.) For the *Sydney Morning Herald* a new title is created on PANDAS each month so we have titles such as *Sydney Morning Herald* (June 2009), *Sydney Morning Herald* (July 2009), for example. This is to avoid having hundreds of links on very long title entry pages. These multiple titles are listed under the "Newspapers" subject listing on PANDORA and in order to bring them more usefully together a PANDORA "collection" has been created that provides a collection title entry page linking to all the individual monthly title entry pages. (See http://pandora.nla.gov.au/col/10281)

The opinion, analysis and commentary pieces associated with the Fairfax Network of online newspapers including the *Sydney Morning Herald*, which have been contributed by columnists and the public, are issued as a separate online entity, the *National times.* This collection is archived every 3 months. See http://pandora.nla.gov.au/tep/100981 for the PANDORA title page for this title and also for a link to the publisher's detailed general conditions of use statement.

Issues and constraints

The main issues and constraints that follow relate to our experience with archiving the SMH and are consequently influenced by the approach we took and the systems environment in which we operate.

The technical aspects of the archiving process are not fundamentally different from any other title collected for PANDORA. Like most complex sites collected for PANDORA each instance does require quality checking and some manual fixing. The main issues with collecting the front page of the SMH arise from its daily harvesting.

While PANDAS was already setup with a daily harvest schedule capability this has generally been used for short and finite periods of daily harvesting, as in the case of some content during election campaigns. The current PANDAS workflow and technology does have some limitations for the ongoing daily harvesting for newspapers.

Scheduling

PANDAS includes a pre-set daily schedule however this cannot be configured to a specific time. Like all scheduled harvests, the daily harvest schedule is initiated in the early hours of the morning. This is a good low impact time for harvesting but not necessarily a good time for harvesting content that is time focused such as an online newspaper. PANDAS does allow more than one harvest to be done in a day, but these have to be manually initiated. So, currently, in order to collect a high value version of the content a manual harvest needs to be initiated by a staff member mid-morning; and this obviously requires additional staff management to ensure someone is available to do this. In order to collect weekend and holiday harvest – and as a backup in case the mid-morning harvest cannot be run – the daily schedule needs to be maintained. This means that five days a week a 50 MB harvest is run and discarded. An option would be to keep this harvest – although the value of that has not been considered at this time.

Maintaining quality

Newspapers sites are complex and require manual fixing to ensure the formatting is retained. Given the objective of collecting the daily front page view of the newspaper, the formatting is considered particularly important. Some efficiency have been achieved by identifying the style elements that need editing and keeping copies of these fixed files and copying them into each harvest daily. The issue with this is that the staff member needs to keep an eye on any changes to formatting and consequent changes required to these 'pre-edited' files. This has already happened at least twice during the course of collecting the SMH.

Display

As discussed above, the PANDORA display does not lend itself to displaying a title collected daily for any length of time. This would result in a very long page – obviously in just one year of harvesting it would have a vertical listing of 365 links. To deal with this a decision was made to create a new title each month which means that a new title entry page is created for each month. This keeps the size of the title entry page down to a functional level it means that there are numerous title entry pages emerging for the one title which results in a non-chronological listing in the "Newspapers" subject listing. The display issue and the decision to create a new title each month also obviously adds additional work since a new title on PANDAS has to be created, links set up on the title entry pages to link earlier and later titles and copying the pre-edited files to the new instance directory.

Conclusion

While harvesting the *Sydney Morning Herald* daily is managed within existing PANDORA workflows, the infrastructure is not particularly well suited to the requirements of ongoing daily archiving in respect to scheduling and managing the presentation. The complexity of the target site and the decision to capture the content at a particular time also require a high degree of staff attention to respond to any changes to the site in order to maintain the consistency and objectives of the archiving. This has been found to be readily manageable for one major online newspaper using the current infrastructure and would probably be manageable for an addition one or two, but it remains a very small scale approach. To expand collecting to a number of newspapers, it would be necessary to address the current limitations in our existing infrastructure.

Brief biographical information for Pam Gatenby, Assistant Director General. Collections Management, National Library of Australia

Pam has worked at the National Library of Australia since the late 1970s in a range of management positions. Her career started in cataloguing and she has maintained a special interest in ways of expediting access to collections through metadata creation and using digital technologies. She currently holds the position of Assistant Director General, Collections Management Division. In that position she is responsible for collecting Australian and overseas publications, including publications in digital form, and for cataloguing, digitisation and preservation of all formats of material in the Library's collections. Specific responsibilities include PANDORA: Australia's web archive; the newspapers digitisation program; and the Australian Newspaper Plan (ANPLAN).

Pam serves on a number of national and international committees including the Committee of Principals that oversees the development of the new cataloguing rules RDA; and ICADS (the IFLA CDNL Alliance for Digital Strategies).