

## **For the Record: Assessing the Impact of Archiving on the Archived**

### **Introduction**

This article examines whether Internet archiving affects the page views or commercial success of live web published items. It also asks how publishers feel about Internet archiving and whether the process of archiving affects the content or longevity of publications in the knowledge that they will now be preserved online forever.

PANDORA Australia's Web Archive at the National Library of Australia has been archiving web based publications for 10 years in conjunction with participants at the Australian State Libraries and other cultural organisations including the Australian War Memorial, National Film and Sound Archive and the Australian Institute of Aboriginal and Torres Strait Islander Studies. There are approximately 12,000 titles within the Archive, each title may be a single discrete document or a whole government website containing thousands of pages.

Many studies and articles have emanated from PANDORA, examining archival practice and policy. None however have attempted to gauge the effect of archiving on the archived - that is the publishers and their publications.

This study of publisher behaviour and attitudes relied on research conducted in three parts. An online survey was placed on the National Library of Australia website and 4920 emails were sent out inviting people who had given permission for PANDORA archiving of a resource between 1996 and May 2005 to complete it. The cut-off date of May 2005 was used so that information from those who had been archived for more than one year was received, as it was thought that only after a reasonable amount of time would the effect of archiving (if any) be noticeable. To complement this survey a selected range of archived publications was examined to discover publication patterns pre and post archiving. A small range of electronic resources that were not archived by PANDORA (or have been archived in the Library's Whole of Domain Harvest or by the Internet Archive) were compared with items archived by PANDORA. In this way a sample of knowingly archived and unknowingly archived items was available for comparison. An analysis of published comments appearing on archived websites was also undertaken.

There are a number of Internet archiving projects currently gathering websites for preservation. Most of these and the largest – the Internet Archive - do so in general without the express consent or knowledge of the web publisher, As such, the web publisher does not automatically know that a copy of their publication is in existence

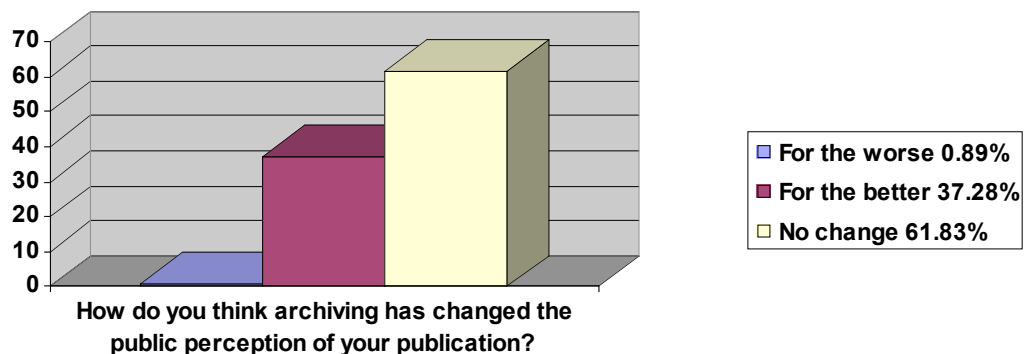
elsewhere and what they are producing will potentially have a much longer life than may have been intended. However this is not the case with PANDORA as it is one of the few archiving projects which explicitly seeks permission from publishers before archiving and notifies them post-archiving. This study then queries only those knowingly archived publishers.

### The PANDORA publisher's survey

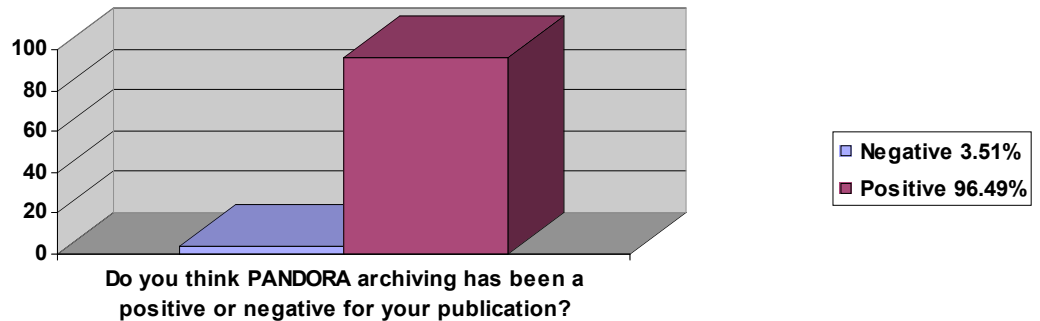
Material produced for the Internet is generally still not afforded the respect that is garnered by traditional print publishing, it is often not subject to peer review, is frequently perceived to contain unreliable information when compared to print publications and is often hard to rank. Usage statistics are one means of defining quality and usefulness but popularity does not always indicate quality, reliability or link stability.

One way that Australian Internet publications can receive permanence and recognition is by being invited to be archived in PANDORA.

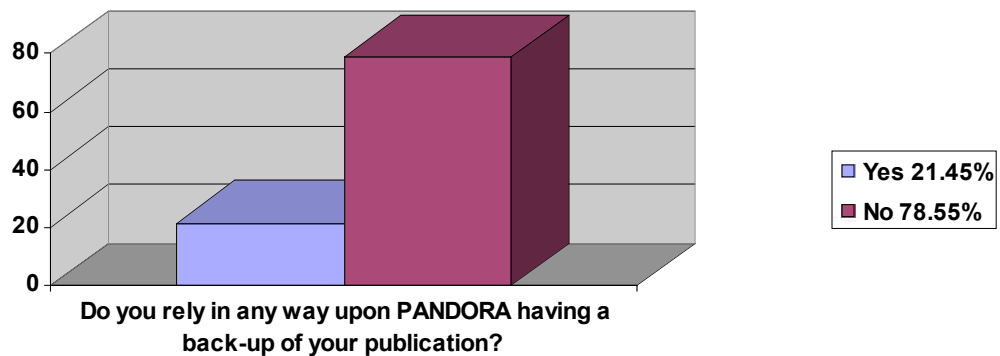
PANDORA is a selective archive and as such web publications within are those that have been selected by staff using selection criteria, giving many the impression that the items archived are in some way different and more significant than those not archived. The PANDORA form letter which is sent out to publishers when initially asking for archival permission includes the claim that the desired item has both 'lasting cultural value' and 'national significance'. These compliments clearly resonate with many publishers and seemingly give many publishers a perception of recognition and even acclaim. Some publishers have even chosen to repeat some of our letter in their publications, seeking to make their audience aware of the Library's estimation of their publication. One publisher of an online novel has even used the excerpted sentence to make it seem like a positive review.



Many publishers are therefore very happy to be archived. When asked in the survey whether PANDORA archiving was worthwhile 97% said that they thought it was. 96% also thought that archiving had been a positive thing for their publication. However, conversely the survey also showed that prior to our first contact when requesting archival permission just over 52% of publishers had not heard of the Archive. And once aware of the Archive only 35% had ever used it to view any other website. Interestingly 29% of publishers also believe, contrary to what we plan for, that it is improbable that PANDORA will preserve their publications in the long term.

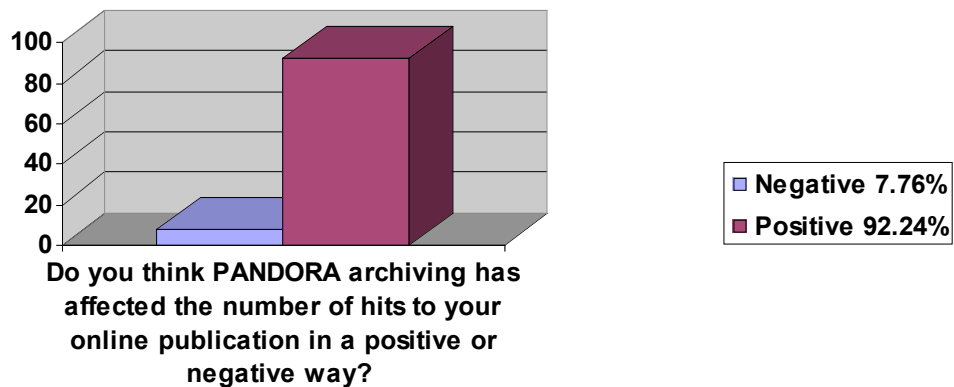


The majority of publishers also did not appear to have any reliance on the Archive as a back-up of their publications. Indicating that they either do not give importance to back-ups or that they are in organisations that have risk management strategies in hand. On occasion however publishers have been assisted when they have suffered serious problems with their computers or ISPs and have lost the content to their own websites, and we were able to give copies back to them. Another service that PANDORA provides for some publishers of websites and online journals is the ability for them to point to our Archive for past issues of their publication, so that they do not have to host them themselves, thus saving presumably on their storage and hosting costs.



## Survey findings

It would generally be expected that making available to the public an archived exact copy of a website would entail a loss of page-views for live websites as users would now be given a choice of access points, something about which web publishers would naturally be concerned about. The PANDORA Archive does receive a relatively large number of hits; the usage 2004-05 was 5,390,459 page views. These views and thus users could conceivably have gone to the live websites. PANDORA page views are continuing to grow, however, although most of the sites in the Archive are still available the highest ranking site in PANDORA is almost invariably a website that is no longer available on the live web.



The PANDORA survey asked web publishers if they believed that archiving had affected the number of hits to their publications, 65% said that it had not. Of the 34% who said archiving had had an effect 92% said that it had been positive. Caution however should be taken in extrapolating these results to all archives. PANDORA may lead to increased hits and usage of live websites only because it actively attempts to do this as a reciprocal gesture for publishers to allow them to archive. Every title within PANDORA has a Title Entry Page (TEP), which serves as the first point of entry, on this page is a link to the live site. PANDORA also uses a pop-up which informs users that they are entering an archive and not the live site. This pop-up appears when users enter into the Archive from a link from a page that is not within the National Library web domain. Robots exclusions have also been used to direct search engines so that they deliver up the TEP in their search results rather than a direct link to lower level pages. These activities especially the active link from the National Library of Australia, our metadata on the TEP and the individual catalogue records created for each title on Libraries Australia all raise the visibility of the live resource to a marked degree. Web archives that do not take measures such as this will be unlikely to have such a beneficial effect on web publishers live websites, but

given that most other archives are not directly available to search engines, they should not have a negative impact either.

## Blogs

The National Library began a concerted effort to archive blogs in May 2005. It had previously selected and archived the blogs of some notable Australians who were politicians or journalists. This archiving was not done because the medium was a blog, but rather because they were high profile individuals and we wanted to capture the online only adjuncts to their traditional media output.

The blogs chosen to be archived from mid 2005 were not produced by people otherwise involved in the media or politics. They were instead archived to show a representative sample of the use of blogging as a popular means of communication and personal publishing. Bloggers in particular were very enthusiastic about being archived and consequently many wrote up the experience on their blogs. One young man wrote an online letter for future youth, another considered the effect on their publication habits thus:

“One of the weird things is that I’m going to need to resist the urge to be more self-conscious, now that I know that my words here will be preserved in this manner. I feel kind of inspired to keep this blog going and may end up trying to improve the quality and quantity of my posts, which is a good thing ... I guess ;)”<sup>i</sup>

Another was less impressed writing that, for posterity

“you’ll still be able to get your fix of ill-informed commentary, shilling for gambling sites and hot Asian chicks. And it’s all thanks to you, the Australian taxpayer. I would have preferred they just gave me one of those \$50,000 grants, but beggars can’t be choosers.”<sup>ii</sup>

The purpose of an archive is to record material exactly as it appeared for the benefit of a future audience. There is therefore some justified fear that informing a content provider that what they produce will be recorded and made available for the long-term may tend to influence what they produce. Thus we might create the ‘observer effect’ whereby things are changed merely by the fact of observing them. Happily, from a comparison of blogs both before and after archiving by PANDORA, it is possible to see that archiving does not appear to have affected the content. Bloggers, although they may consider it in the short term after initial archiving (and have commented thus) do not appear to continue self-consciously writing for a possible future audience, but concentrate on the immediate and the quotidian. From a brief textual analysis there appeared to be no evidence that archived bloggers censor themselves any more than they did prior to being archived, a result which was confirmed by the general survey responses. The bloggers who discuss sexual, political and personal information continue to do so, and where there is no

illegality, the Library takes no role except to restrict some archived websites and blogs to adult researchers.



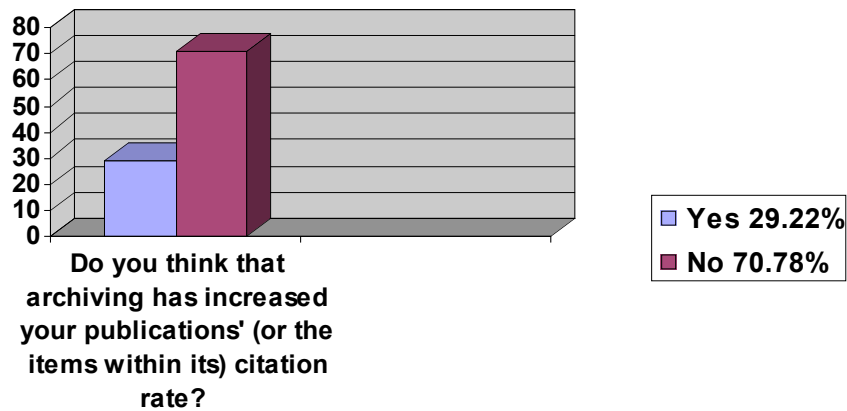
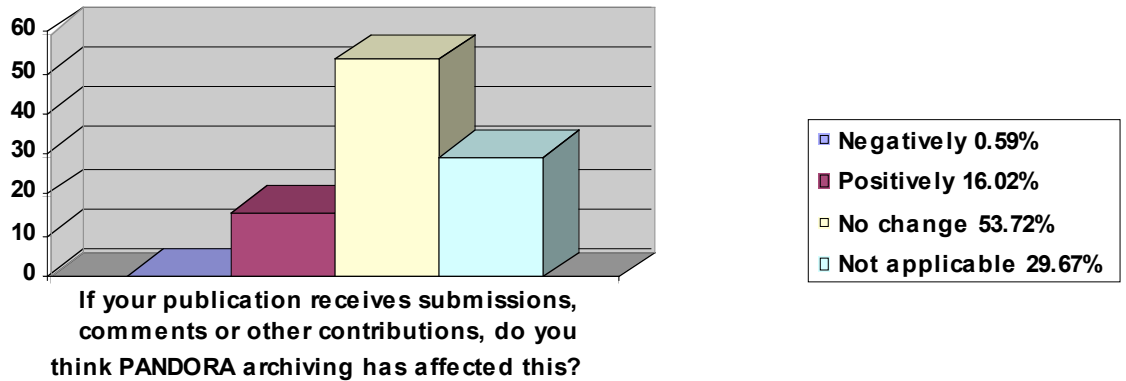
From studying a set of both archived and un-archived blogs there also does not appear to be any changes in blogs' longevity by being archived. The author's personal circumstances, available time and 'having something to say' seem to be far greater determinants of a blogs' longevity than any effect of archiving.

### E-journals

E-journals are journals that are published only in online form; they can emanate from any source, but in Australia are most widely used by government and academic, rather than by commercial publishers.

The National Library and PANDORA are often in at the outset of these publications as frequently the first task of a newly created serial publication is to apply to the Library for an ISSN. Built into the ISSN application form is a PANDORA archival notification clause. PANDORA therefore is often able to archive many Australian online journals from inception.

That archiving in PANDORA significantly improves quality or maintains publishing life cannot be proved from an analysis of archived serials. The survey asked whether archiving increased submissions, comments or other contributions. Publishers mostly reported that they had experienced no change, but, where there was change it was predominately positive. There also appeared to be no evidence that archiving prolonged publishing life.



There was some indication that some serials had increased citation rates from being archived. Their perception of the usefulness of our creation of Persistent Uniform Resource Indicators (PIs) for their publications was very low however. Only 14% of survey respondents believed that our creation of PIs for their publication had any benefit. We are however aware that a very large number of links to the Archive is made by indexing agencies using PIs as PANDORA has ongoing relationships with a number of indexing agencies (who actively advise us on the selection decisions in their specialist areas). The lack of knowledge of PI usage is therefore possibly due to the fact that the links point to the Archive and not the live resource.

### Commercial websites

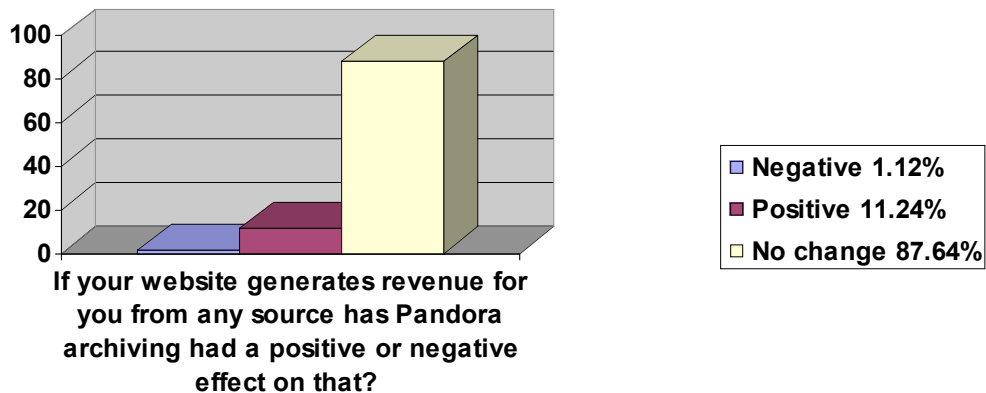
Most websites within the Archive exist to disseminate information and many publishers welcome PANDORA's role in promoting that further. Commercial websites differ in that they exist to gain revenue from users. Diffusing their customers has the potential then to lose them revenue. It was interesting to note from the survey results, that publishers who

use their websites to conduct a commercial enterprise say that they do not appear to be detrimentally affected by being archived.

There are three major ways in which an archive could have an economically damaging role for a live website. The archive may take away hits, leading to less page view based advertising revenue and possible click-through revenue. Archiving may also display materials that normally are only viewable at a cost, such as a commercial subscription only online journal. The other main problem could be user confusion, whereby a user unknowingly accesses the archive rather than the live website and attempts and fails to complete a purchase, leading to dissatisfaction with the company and loss of potential revenue for it.

PANDORA has made efforts to not interfere with live websites commercial activities. If it archives commercially available material it makes prior negotiation with publishers to restrict access for a publisher specified period. PANDORA also tries to make sure that users are aware that they are in the Archive, and restricts sales confusion by not archiving or allowing any transaction pages or functions and using re-directs to point to the live website.

Consequently PANDORA has limited effects on commercial activity. The survey results showed that this policy seems to have worked as only 1% of commercial publishers believe archiving has had a negative impact.



## Conclusion

This study does not seek to give an opinion on how publishers perceive all Internet archiving projects. However it can be seen that if a known openly searchable archive such as PANDORA has few problems with publishers then a less searchable (deep web) whole domain or larger archive should present even fewer problems for publishers – as

has been shown by the relatively few requests for take-downs that have been received by the Internet Archive.

The National Library has a statutory duty to collect and preserve Australia's documentary history, and born digital publications are no exception. To successfully create an archive that encompasses the broadest range of publications requires the ongoing consent of publishers. While there are archived copies and live websites it will remain the archives' responsibility to make sure that their activity does not adversely affect online publications – both their content and their commercial value. In the long term, most non-institutional publisher's websites will no longer be available outside of the archives. However, the archived publications will for many years still be protected by copyright and so the Library will need to continue to have publisher's consent for them to be made accessible.

The results of the study show that PANDORA archiving has thus far not had a detrimental effect on publications, and is in fact mostly benign and in many cases beneficial. It is to be hoped that the knowledge that Internet archiving does not necessitate any conflict between archivists and publishers will assist in guiding future negotiations.

Edgar Crook  
ecrook@nla.gov.au  
Digital Archiving Section  
National Library of Australia  
June 2006

---

<sup>i</sup> Wilson, Morgan, *explodelibrary.info*, [http://www.explodedlibrary.info/2005/07/the\\_national\\_li.html](http://www.explodedlibrary.info/2005/07/the_national_li.html) accessed 14 June 2006

<sup>ii</sup> Ward, Sam, *A yobbos view*, <http://www.gravett.org/yobbo/?m=200508>, accessed 14 June 2006