PANDORA : Collecting in a Digital World – Where is the artefact?

A paper presented at The ALIA Acquisitions and The Bob Hawke Prime Ministerial Library Symposium: The Acquisition of Cultural Artefacts
October 10, 2007, Bradley Forum, Hawke Building, University of South Australia

Paul Koerbin
Manager Web Archiving
National Library of Australia

The focus of this symposium is the acquisition of cultural artefacts. In talking about the PANDORA Archive and about collecting in a digital world, I am therefore extending our discussion to the idea of 'digital' cultural artefacts.

The first question to ask is whether this is a valid concept – does it make sense? Is it something we can usefully speak of, specifically as practitioners in collecting and preservation? If so, what then do we mean by 'digital cultural artefacts'? Or put another way, where is the artefact?

An artefact, a material product of human culture, implies, I think, something static, representative, probably discrete and of itself (having dimensions); and ultimately deserving of preservation. This is a point that is critical to understand in considering the World Wide Web: the object must surely exist in some persistent way in order to be an artefact. This may seem somewhat counter-intuitive in regard to the Web which persists in the now but is dynamic, interactive and evolving. So, is it relevant or useful to think of digital culture in terms of 'artefact"? Well, I think it probably is and my talk will give my reason why.

The consideration of the Web as an entity of sorts that should be considered as a digital cultural artefact is not new. Brewster Kahle and Peter Lyman founding directors of the Internet Archive considered this issue in an article in D-Lib Magazine in July/August 1998[1]. At that time they stopped short of formally defining digital cultural artefacts considering it something "still being shaped by experimentation and practice". That situation has never changed. The Web is more than ever a medium of experimentation and practice. Instead Lyman and Kahle preferred to describe some of the significant differentiations of digital cultures from other cultural artefacts. Among these were:

- Things occupy spaces while digital documents are "electronic signals with local storage but global range"
- Digital documents are both ubiquitous (because of their global range) and a universal medium by virtue of their "size and ability to represent culture expressions in all other media"
- Digital documents are at once tangible (the "representation in code") and intangible (the code is meaningless unless transmitted and represented)
- Digital objects are not the property of cultural elites for the "medium is profoundly democratic".

I think we can probably agree that the World Wide Web does represent a cultural artefact of some description or other: it is a material construct (even in its virtuality) representing human culture. The problem of course arises, and is the nub of the issue in respect to collection and preservation (which is my concern as a practitioner) when we seek to locate what the artefactual element is.

[1] Lyman, Peter and Kahle, Brewster. Archiving Digital Cultural Artifacts: Organizing and Agenda for Action. D-Lib Magazine, July/August 1998. http://www.dlib.org/dlib/july98/07lyman.html (viewed on 12 September 2007)

On the one hand we can of course approach the web as a massive publishing medium and we can deconstruct it in such ways as we find appropriate or practical to collect, describe and preserve discrete publications. This is very much the model we have pursued for the PANDORA Archive over the past decade, since we established it at the National Library in 1996. Discrete publications of course exist on the web; they have done from the outset and still occupy a significant place. Here I am thinking of things such as print-like reports in PDF format[2]. But the Web is much more than this of course. It is a mass of hyperlinked complex objects presenting information often with dependencies on other resources and without clear demarcation[3]; and evermore interactive with the distinction between the creator and the user of information less delineated. It is possible of course to make arbitrary determinations of scope to define a publication of convenience for collecting. Again this has been something we have done for PANDORA. So a "title" entity in PANDORA can be anything from a single PDF file containing a print like publication to an entire website and anything in between.

In a sense this is to recognise and indeed give precedence to the content element of the web, especially in as much as it resembles the traditional concept of (print) publication models. Lyman and Kahle were able to say about the web in 1998 that "its guiding metaphors are derived from print publication"[4]. Think of text, pictures, and some multimedia. Even today this holds true, to some degree, although it is important to recognise that this may be in the browser of the beholder. To remain with that metaphor however is to miss the point that the medium is fundamentally different than print. The hypertextual, interactive democratic nature of the medium means that it is contextualised in the performance of the person engaged in some way with the web. As an example, consider how an issue is discussed in the blogophere: the publication, a blog, consisting perhaps of a discourse and commentary while linking discussion to other blogs and sources, and perhaps combining multimedia (let's say a video on YouTube, possibly embedded transparently in the blog page); it crosses across simply defined entities and may only be defined in the performance of someone engaged with this discourse in a particular way at a particular time. Researchers are indeed concerned with how to map this, and in how to archive for future reference this stream of interaction[5]. And we can go further; the value of preserving, for example, search query logs has also been raised as being valuable records of the expression of the collective human consciousness[6]. Are these too digital cultures we should consider as artefacts?

To phrase this another way, there is a sociology about the web that we as curators concerned with collection and preservation should remain aware of. It is not only the content but the web as an object of culture and its uses that is of interest to researchers[7]. That is, the manner in which it is used; the manner in which more and more its real-time construction, growth and change become part of the cultural and political process. A current example is the Prime Minister's use of YouTube recently for policy announcements[8]. Is it perhaps in these aspects of the web that

[2] Anecdotal evidence and responses to the PANDORA User Survey form suggest specific documents are something people are regularly looking for.

[3] For example a politician's personal website that presents media releases or speeches delivered from the party or parliamentary websites

[4] Lyman, P. and Kahle, B. op. cit.

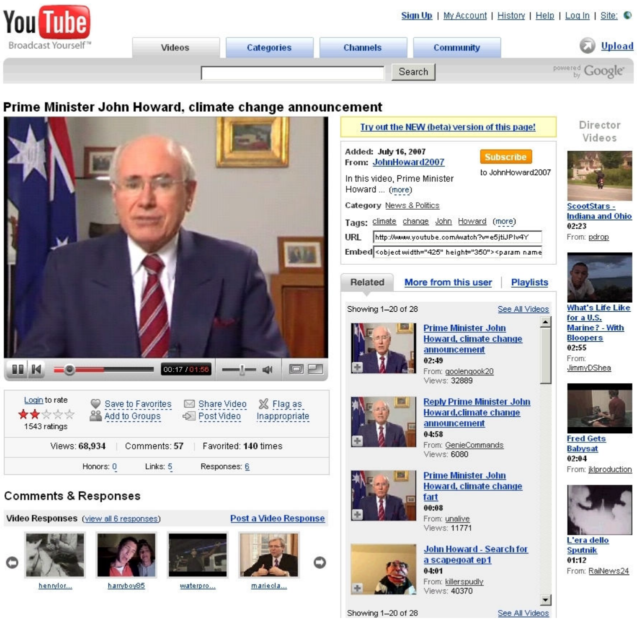
[5] See Bruns, Axel. Methodologies for Mapping the Political Blogosphere: An Exploration Using the IssueCrawler Research Tool. First Monday, vol. 12, no. 5 (7 May 2007). http://www.firstmonday.org/issues/issue12_5/bruns/index.html (viewed on 8 October 2007).

[6] Jansen, Bernard J. *Preserving the Collective Expressions of the Human Consciousness*, 16th International World Wide Web Conference, 2007. http://www2007.org/workshops/paper_58.pdf (viewed 8 October 2007)

[7] See for example Schneider, Steven and Foot, Kirsten A. The Web as an Object of Study. New Media & Society, vol. 6, no. 1 (Feb. 2004) pp 114-122.

[8] Probably the first political leader to seriously use a website to promote his election campaign in a personal way in Australia was Victorian Premier Jeff Kennett in 1999 with his site jeff.com. This was not able to be archived

we really should seek the digital cultural artefact? That is, in the use and context of the content, not solely in the content per se? To illustrate, here we have a snapshot of the Prime Minister's climate change announcement as it appeared on YouTube:



Prime Minister John Howard, climate change announcement

Added: July 16, 2007
From: JohnHoward2007
Subscribe to JohnHoward2007
In this video, Prime Minister Howard ... (more)
Category News & Politics
Tags: climate change John Howard (more)
URL http://www.youtube.com/watch?v=e5jtiJPlv4Y
Embed <object width="425" height="350"><param name

Login to rate
1543 ratings

Views: 68,934 | Comments: 57 | Favorited: 140 times

Honors: 0 Links: 5 Responses: 6

Comments & Responses

Video Responses (view all 6 responses)

henrylor... harryboy85 waterpro... mariecla...

Showing 1–20 of 28 See All Videos

Prime Minister John Howard, climate change announcement
02:49
From: goolengook20
Views: 32889

Reply Prime Minister John Howard,climate change announcement
04:58
From: GenieCommands
Views: 6080

Prime Minister John Howard, climate change fart
00:08
From: unalive
Views: 11771

John Howard - Search for a scapegoat ep1
04:01
From: killerspudly
Views: 40370

Showing 1–20 of 28 See All Videos

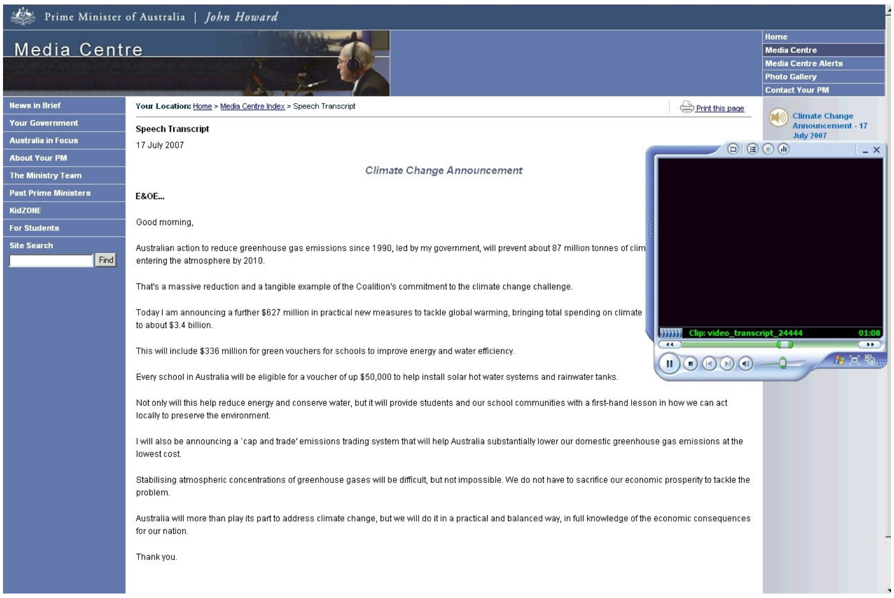
ScootStars - Indiana and Ohio
02:23
From: pdrop

What's Life Like for a U.S. Marine? - With Bloopers
02:55
From: JimmyDShea

Fred Gets Babysat
02:04
From: jklproduction

L'era dello Sputnik
01:12
From: RaiNews24

This was not able to be captured for PANDORA as permission has not been able to be obtained from YouTube to re-use content for the Archive. However, we were able to capture the same content from the Prime Minister's official website for which we do have archiving permission. So we get this:



Your Location: Home > Media Centre Index > Speech Transcript

Speech Transcript

17 July 2007

Climate Change Announcement

E&OE...

Good morning,

Australian action to reduce greenhouse gas emissions since 1990, led by my government, will prevent about 87 million tonnes of clim entering the atmosphere by 2010.

That's a massive reduction and a tangible example of the Coalition's commitment to the climate change challenge.

Today I am announcing a further $627 million in practical new measures to tackle global warming, bringing total spending on climate to about $3.4 billion.

This will include $336 million for green vouchers for schools to improve energy and water efficiency.

Every school in Australia will be eligible for a voucher of up $50,000 to help install solar hot water systems and rainwater tanks.

Not only will this help reduce energy and conserve water, but it will provide students and our school communities with a first-hand lesson in how we can act locally to preserve the environment.

I will also be announcing a `cap and trade' emissions trading system that will help Australia substantially lower our domestic greenhouse gas emissions at the lowest cost.

Stabilising atmospheric concentrations of greenhouse gases will be difficult, but not impossible. We do not have to sacrifice our economic prosperity to tackle the problem.

Australia will more than play its part to address climate change, but we will do it in a practical and balanced way, in full knowledge of the economic consequences for our nation.

Thank you.

While we do have the content captured for preservation purposes, clearly an important part of the context, the artefactual element is lost.

in PANDORA as permission to include it in the Archive was not able to be obtained. The site produced by Stephen Mayne in response to it however was captured. Mayne asserts that the site played a major part in Kennett's defeat owning to the publication of his substantial "treatise" about Kennett's administration that was included on the site. See http://nla.gov.au/nla.arc-15033. Another interesting site was that of John Schumann for his campaign against Foreign Minister Alexander Downer for the seat of Mayo in the 1998 Federal Election, which resulted in the Foreign Minister's near defeat. This site was also captured in PANDORA, see http://nla.gov.au/nla.arc-10222.

So, what exactly are we dealing with when we attempt to archive web material? There is the code, the un-interpreted digital files that are the digital building blocks, transmitted and retained on physical media; then we have the intellectual content, the text, image, sound, multimedia, stylesheets; and then we have the performance, the hyperlinking, the self-referencing and the browser and browsing users that animate the web and define its ontology. Where in this, then, is the artefact?

To deal with the code, briefly. There is certainly a case to be made that this, being the building blocks of what is delivered as a web experience is an artefact of some kind. Although, at the point of publication this is complicated by the manner in which content is commonly delivered, utilising many elements to make up that experience from database content, HTML mark-up, stylesheets, JavaScript, binary media and so one. So, once again, we can ask: what is the artefact in this?

It may be a simpler case to consider milestone technological innovations like the first web page as a digital cultural artefact, and I mean the code and all. That first page we don't have, CERN did not preserve it, but the URL was http://infor.cern.ch/hypertext/WWW/The Project.html[9]. After that, it starts to get very complicated! I don't intend to go beyond this in this brief presentation in regard to the code, other than to say that in preservation terms, the integrity of this is critical, as is the ability to interpret that code over time into anything meaningful. And because of technological obsolescence, to do this we will need to take actions that may transform that original thing in order to render it with meaning and integrity – an apparent contradiction, perhaps. But this is indeed the central issue and concern of digital preservation. Unless you can fulfil (or seriously intend to achieve) this preservation goal of persistent access, there is little point in collecting in the first place and the artefact would not exist in any useful way.

To bring this presentation towards some more practical considerations, I think the act of web archiving may be seen as a process creating an artefact of digital culture.

Collecting significant sites that document aspects of our cultural heritage, expressions and events may be one way to look at this artefactual aspect of digital culture. Let us take, by way of example, the Official Sydney 2000 Olympic Games site. The Sydney Games were the first summer Olympic Games to have widespread web coverage and to be subjected to targeted web archiving in the lead up to and during the actual Games[10]. The official Sydney 2000 site was one of something in the order of 150 websites we collected in respect to the 2000 Olympic and Paralympic Games for the PANDORA Archive. So what then constitutes the artefact? Is it the content? Is it the look and presentation? Does the functionality of the site matter? And what do we make of the changes to the site over time and over the period of the Olympics – where capturing snapshots over days gives us a slow shutter like representation of the site? Since this is all that remains of the site from the time it was active[11] there is, I think it is fair to claim, an artefactual quality to this archival resource. And I think the fact that we did our best to collect the content as completely as possible; to capture it every day of the Games; and to retain the look and style of the site all adds up to create an identifiable digital cultural artefact and one that

[9] Mentioned in Jansen, op. cit.

[10] The 1996 Atlanta Games were covered on the Web and some content has been picked up by the early web crawling and archiving activity of Alexa and the Internet Archive in October and December 1996, more than two months after the Games finished. See http://web.archive.org/web/1996*/http://www.atlanta.olympic.org/

[11] The Internet Archive did not capture the site during the period of the running of the Games. It was crawled apparently on 6 July 2000 and next on 1 October and then 18 and 19 October 2000. However the bit they seem to have picked up appears to date from around 19 October 2000. See http://web.archive.org/web/*/http://www.sydney.olympic.org/

was in a sense very deliberately formed (and could have been done differently). It has defined dimensions in extent (no external links were harvested) and chronology and is discrete in its isolation from the original context of the Web at the time it was captured.

If we look at another end of the scale: in the second half of 2005 the NLA contracted the Internet Archive to do a scoped large scale crawl and harvest of the Australian web domain. We have repeated this crawl again in 2006 and in 2007. This represents a somewhat different approach to the archiving we have been routinely doing for PANDORA since 1996. Rather than being highly selective it is a scoped crawl of linked content, so it retains something of the quality of a snapshot of the Australian web domain – albeit a long exposure if we are to extend the photographic simile, since these crawls take around five weeks to complete. This archival collection does not function in the same way as the live web of course. While it is interactive in as much as links can be followed and media activated, it does remain an artefact, a static representation of the web at a certain (5 week) point in time. It is however, clearly more than a publication. It does give evidence and a representation of the ontology of the web, or the part represented by the archival collection, though it is just as clearly not exactly the same thing. Within its defined scope, content remains in context with its hypertext linkages intact as far as possible. This may represent somewhat better what might be regarded as an artefact of the sociology of the web although specific content – and thus discrete digital cultural artefacts – may or may not be collected in an acceptable or ideal manner. This sort of collecting relies on the machine to a much greater extent than the 'hand crafted' selective archiving undertaken for the PANDORA Archive.

I will conclude by highlighting some points in respect to web archiving and the concept of the digital cultural artefact:

1) The act of web archiving is of itself the act of creating a digital cultural artefact. This could be a specific site that has genuine cultural significance (such as Sydney 2000 Olympics official website) or a temporal snapshot of a scoped part of the Web domain.
2) Cultural artefacts represent something created by humans engaged in the culture in some way. In this the Web presents major challenges for collecting since the nature of the web is dynamic and interactive. How to adequately collect that performative aspect poses a great challenge for us. For the present we do our best to preserve what we can of our digital culture through our web archiving activities even though the digital artefacts we create in doing this do not represent the full experience of the Web.

Thinking of the web as a medium of digital cultural artefacts challenges us, those of us engaged in collecting and preserving the documents and artefacts of culture, to think beyond the metaphors of print publications and to try and address digital culture on its own terms. Capturing the content is fundamental and important; the challenge remains to capture more than the content.

Recognising that collecting web resources is in many respects a process of identifying and in effect creating digital cultural artefacts helps us to consider critically what we are in fact doing in our web archiving pursuits. The web provides a particularly complex publishing medium and cultural space to challenge the assumptions and approaches we take to collecting and preserving our cultural heritage. Where is the artefact – well, arguably, some are in the PANDORA Archive and may also exist as web domain snapshots!