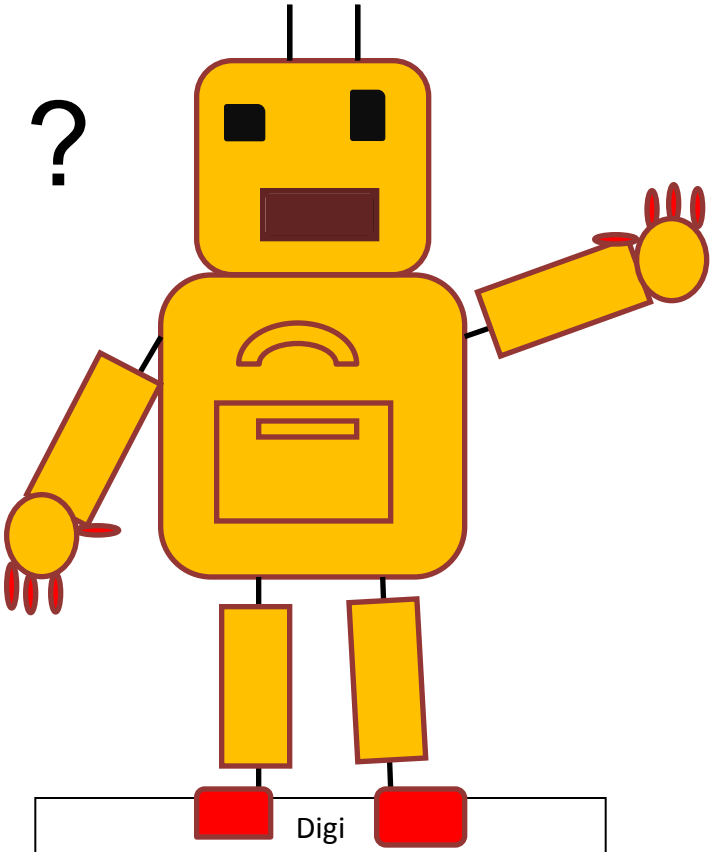
Those Mad Men from the Antipodes:

Presentation Intent at the National Library of Australia

David Pearson

What do we need to do?

?

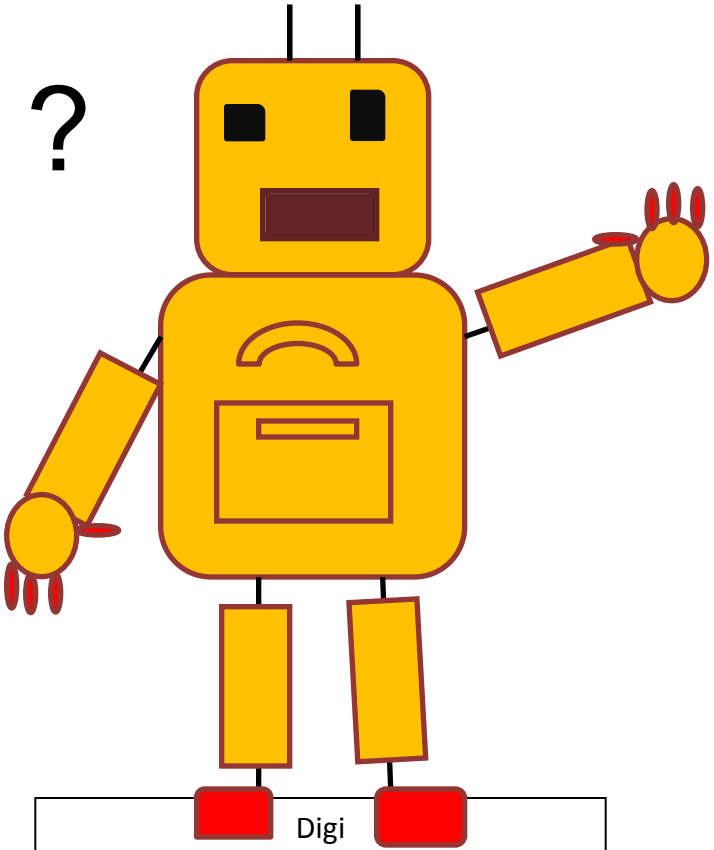


Digi
By Imogene Pearson (7 years)
(March 2012)

A range of preservation tools and methodologies!:
Google Images (2012)

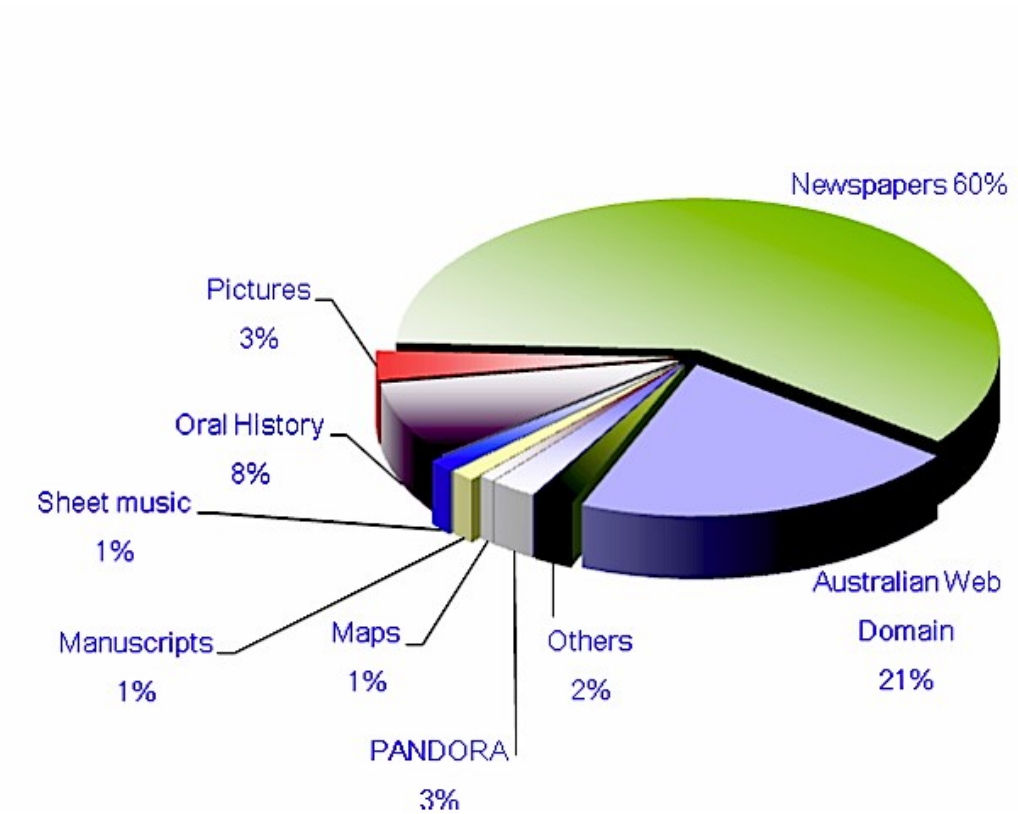
What do they want?

Rosetta Stone
British Museum:
Google Images (2012)



Digi
By Imogene Pearson (7 years)
(March 2012)

What is in the box?



NLA Collections

Preservation Intent - Asian Collections and Overseas Collections Management — Version 1.0

Preservation Intent - Australian Books and Serials — Version 1.0

Preservation Intent - Dance — Version 1.0

Preservation Intent - Manuscripts — Draft

Preservation Intent - Maps — Version 1.0

Preservation Intent - Music — Draft

Preservation Intent - Newspaper Digitisation — Version 1.0

Preservation Intent - Oral History — Version 1.0

Preservation Intent - Pictures — Version 1.0

Preservation Intent - Selective Web Harvesting — Version 1.0

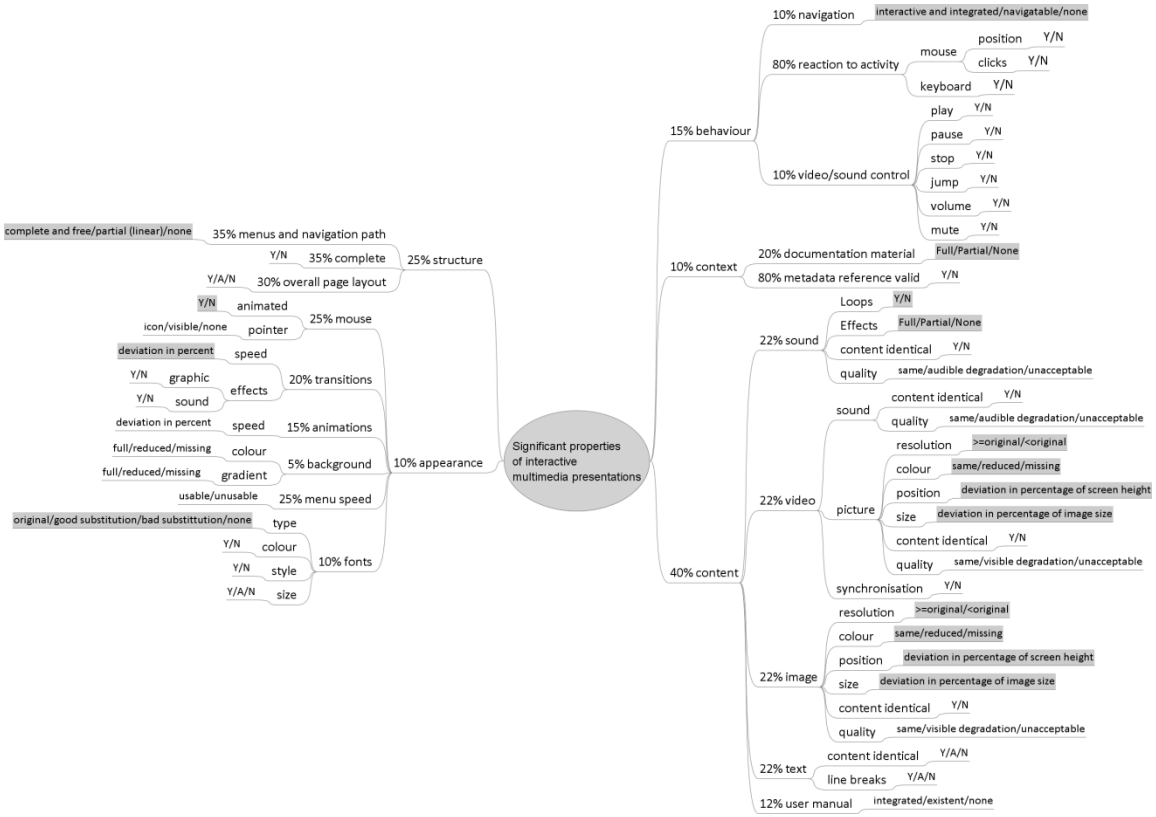
Preservation Intent - Web Domain Harvests — Version 1.0

Preservation Intent – Australian Government Web Archives — In progress (Dec 2012)

These statements are designed around questions of such as:

- How can we describe and classify a collection, which makes sense to the collection managers? (splinter or lumper)
- Who is responsible for authorising what happens to any specific collection materials?
- Do we need to keep these materials accessible? If so, for how long?
- In general terms, what would an adequate level of accessibility look like (such as, integrity or fidelity of bits, viewing functionality, editing functionality, navigating, and the ability to manipulate content)?
- Who is responsible for preservation actions (i.e. digital preservation specialists, curators, ICT staff or another party)?
- Can we identify any issues, including operational or non-collecting issues, which may hinder preservation efforts in the future?

What about Significant Properties?



Significant Properties
PLATO
Google Images (2012)

An Example: Pandora Web Archive

The NLA's web collections fall into three broad categories.

- Selectively gathered 'titles' which make up PANDORA, Australia's Web Archive.
- 'Whole Domain Harvests' which aim to capture a broad periodic snapshot of the Australian Web Domain.
- Coherent bulk collections, such as '.gov.au' seedlist derived collections.

General Description

The NLA's selective web harvesting activity currently consists of the PANDORA Archive, which contains a selective collection of web publications and websites relating to Australia and Australians. PANDORA was established by the NLA in 1996 and contains historical online materials harvested from 1996 to the current period. Online materials, ranging from discrete publications to complete websites, are selected for inclusion in the collection with the purpose of providing long-term and persistent access to them.

Who is Responsible?

The Web Archiving Section (the relevant collecting and curatorial agent) intends that:

General Intent

1. All PANDORA digital preservation masters, including all associated metadata (currently known as the 'preservation master,' the 'display master' and the 'metadata master'), should be retained in perpetuity. All technical properties should be maintained to the fullest extent possible.
2. Content, connections and context are of primary importance. How it is ultimately presented to a user is a secondary consideration.
3. The original harvested copy, that is the 'preservation master' that represents the initial and untouched collection of files gathered by the harvest robot, is of less importance than the 'display master' which includes the results of quality assurance and curatorial work. However, as the impact of curatorial and quality assurance work upon harvested instances may not be known, the 'preservation master' (although a lesser version in terms of completeness) should be retained at least at the bit level.
4. The 'derivative copy' of the 'display master', which is created for display and access, should be maintained only for as long as it is useful. A new derivative version may be generated according to future access requirements.

Important Aspects

Preservation of Different Versions

Some of the Collecting Issues

The NLA understands that web archives pose several problems.

1. The NLA has no control over the creation of the original content and its format, standards or quality. The NLA can only currently harvest what is delivered in a single published form through a browser/server request (for example, the original data in the publisher's databases are not collected).
2. Current methods for collecting and rendering are also not ideal in ensuring the complete capture of all files or retaining full functionality in the version delivered from the archive.
3. We are only taking time specific and time limited snapshots of web content.

Therefore, the NLA accepts that what is to be preserved is not a mirror representation of the web nor of a website but, rather, a snapshot of content that was once arranged and published as a website, with only limited functionality of the original. The archived artefact is formed out of the collecting process which is inevitably lossy. Our aim is to define and control this loss. In addition, the way in which the content is collected and displayed places a significant limitation on the presentation of the archived artefact as an authentic record of the publisher's original data or of the version of that data originally published on the web.

Preservation Issues .

The NLA's intention is long-term access for all users. However, over time access to certain content may be available only on-site due to technical constraints in supporting remote access to all possible operating environments.

The harvested web content, being composed of complex objects, is contained in either a compressed package (tarball) or a container file (WARC file). While the tarball retains the directory structure of the original harvested website, the WARC file may contain random collections of files plus metadata, which are managed and located by indexes.

The PANDORA collection, having begun in 1996, includes content collected through various methods with a growing legacy of inconsistency regarding Uniform Resource Identifiers (URIs), metadata and quality assurance interventions. Processes are underway to move the content to a consistent archival format (WARC), although the underlying legacy variations may not necessarily be removed in this process.

Preservation Issues ..

The PANDORA collection can be broadly categorised as consisting of about 75% text, 20% images (JPEG, GIF, PNG) and 5% multimedia and style elements (Java script, CSS files, Flash and so on), including linkages. Because of the variable nature of the collected entities (understood as the PANDORA 'title' and its archived 'instances') ranging from simple documents to complex multi-file objects, there are some parts of the collection where style elements are more important, and some parts where this is less so. Style elements are problematic from the outset since they are sometimes difficult to harvest and often remain impossible to render. Because content is harvested through a browser-type request on a server, in many cases only a subset of possible style element files are delivered (those required for the browser request). Moreover, harvesters are not able to thoroughly parse complex JavaScript which may also result in the collecting process not identifying and missing many style elements (JS, XML and so on).

Preservation Issues ...

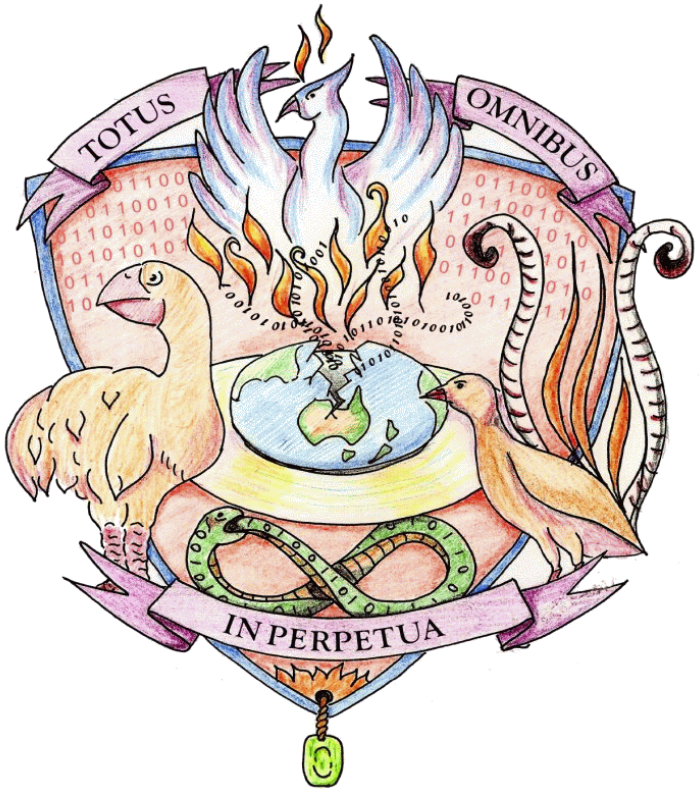
Contemporary browsers are fairly tolerant for accessing both current and legacy web content. However, due to the variability of this content (collected from 1996 to the present day) and factors such as being poorly formed (no standards) can mean that viewing content as it was at the time of creation can be problematic.

The status of the visual accuracy of the harvested copy of a site has not been systematically documented (although implied in QA workflows). Thus, the look of the original may only be surmised from the content collected, the context of embedded links, tags and file types and the context of technologies known to exist at the time of harvesting. The NLA's objective is not to misrepresent the material in any way that would compromise its legal warrant to collect, preserve and make accessible the archival content. Thus particular care in retaining the integrity of the intellectual content including embedded links and domain-related image material is a priority.

This collection is currently in a state of transition in how it is stored, described and understood via technical metadata."

For more information see:

Webb, C., Pearson, D. and Koerbin, P. (in press) ‘ “Oh, you wanted us to preserve that?!” Statements of Preservation Intent for the National Library of Australia’s Digital Collections’, in D-Lib Magazine.



NLA Digital Preservation Crest:
Totus, Omnibus, in Perpetua
(Everything, for Everyone Forever)