

# Preserve or Preserve Not, There is No Try: some dilemmas relating to Personal Digital Archiving

Luke: All right, I'll give it a try.

Yoda: No. Try not. Do... or do not. There is no try.

Star Wars: Episode V - The Empire Strikes Back (1980)

*David Pearson  
Manager –  
Digital Preservation Section  
National Library of Australia*



# Outline

What I am going to talk to you about today is some of my experience in dealing with, and the consequences of not looking after digital materials.

Part 1: The current environment, some observations

Part 2: Monster, what monster?

Part 3: Some personal experiences 2002-2009

Part 4: What can 'I' do about it?

Part 5: Conclusion

# Part 1: The Current Environment

So essentially we know that:

- There is a lot of ‘digital stuff’ in various personal and institution archives and it is growing;
- loss of access to it (or to much of it) will occur unless action taken;
- The process of keeping it accessible and usable needs to be managed; and
- Any ‘solutions’ have to be scalable, reliable and automated and in the long term sustainable by a community.

Not managing 'it' is not an option. Collectively and or individually we, need to:

- Understand our 'digital stuff' & associated risks;
- Provide safe storage, ensure immutability and authenticity;
- Ensure access over time as technology changes;
- Develop & implement preservation workflows, skills, standards, & strategies for ongoing access; and
- Enable content to be shared and used in different ways in the future.



There are a number of problems which we face during the life-cycle of digital material:

- There appears to be a massive difference in the way that people create and use data, and the way in which collecting bodies would like to receive data in order to preserve it.

The challenge is that either:

- We need to be able to handle any sorts of data that comes to us (impossible!); or
- Find a way that the data that we receive works for us, without affecting how the user creates and use their data. The data needs to be capable of being identified, preserved and accessed.

# Part 2: Monster, what monster?

The challenging part is keeping these rich information resources available.

One of our greatest problems is knowing that we have a problem or will have a problem.

We can ignore the monster at our peril, but sooner or later it will bite 'someone' on the ...!



Yet, the game keeps changing!

The digital environment is a dynamic entity. This means that in relation to digital materials, some problems relating to collecting archives are:

- They are constantly being created;
- People are submitting more and institutions are actively looking for content;
- The ways in which it can be submitted are changing;
- The way in which the content is presented and the files types keep changing;
- The accompanying metadata that comes to us is generally not consistent, if there at all (standard change); and
- In many cases we only know what was sent to use when we open it up!

# Act or risk losing it

‘Digital stuff’ is dependent on technology at all stages:

- Creation/capture
- Storage
- Access

Because Technology changes - sometimes rapidly, sometimes more slowly, but eventually over time, we will lose access. Thus software, hardware, media, file formats, operating systems will become obsolete.

Unless managed deterioration can occur rapidly (e.g. data can be corrupted or lost in storage or transfer process).



# Different flavours of obsolescence



Differential preservation of technology which is driven by:

- The availability of the technology
  - Vendor designed and ultimately consumer driven obsolescence;
  - Disruptive and or destructive technologies (destroys stability and content); and
  - Increase in the capabilities of technologies.

# It's 'only MOSTLY dead'



Miracle Max: Whoo-hoo-hoo, look who knows so much. It just so happens that your friend here is only MOSTLY dead. There's a big difference between mostly dead and all dead. Mostly dead is slightly alive. With all dead, well, with all dead there's usually only one thing you can do.

Inigo Montoya: What's that?

Miracle Max: Go through his clothes and look for loose change.

The Princess Bride 1987

## Entropy:

- Loss of availability or degradation of the technology; and
- Loss of skills and understanding of the technology.



# An example of MOSTLY dead

An example: If you did not know already, you may have forgotten that to access a 5 ¼ inch Floppy Disk (in one particular configuration) you might need the following:

A working 5 ¼ inch Floppy Disk



A working 5 ¼ inch Drive



A working 5 ¼ inch floppy cable

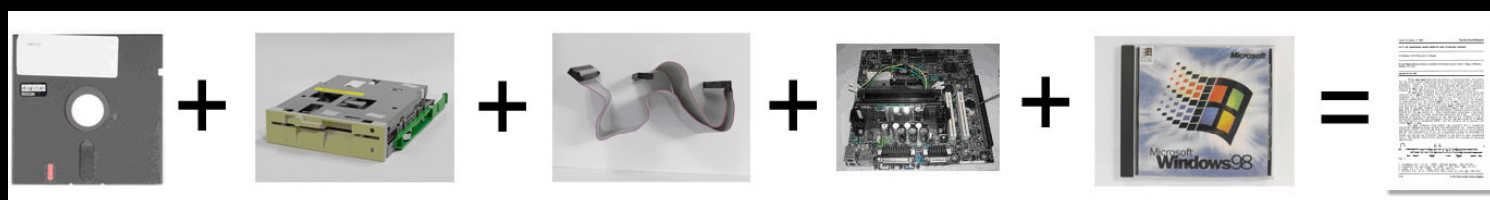


A mother board that can connect to a 5 ¼ inch floppy cable and has a compatible BIOS for this drive type



A working version of a compatible OS







# Part 3: Some Personal Experiences 2002-2009

## The National Archives of Australia.

Dealing with the Digital Preservation of Government Manuscript materials.

For example, data recovery of 300 data carriers containing Royal Commissions records from 1970-1995. Data

Recovery from:

- 9 Track ½ Magnetic Tape
- 8 inch Floppy Disks
- 5 ¼ inch Floppy Disks
- 3 ½ inch Floppy Disks



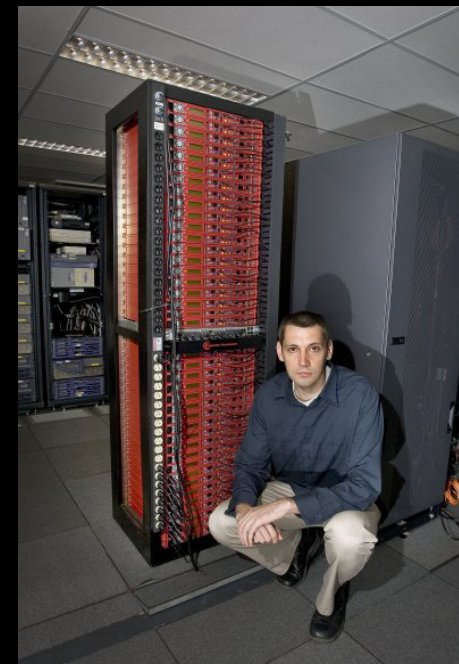
# Some Experiences 2007-present



## National Library of Australia.

Dealing with Digital Preservation of published (serial, monograph), unpublished personal manuscripts, large scale digitisation of newspapers and web sites (around 350TB of data).

For example, The PANDORA Web Archive:



National Library of Australia

nlc.int-nl39551-ls2-v

# NAA Part 1

I have learnt the some of the following during my 8 years in Digital Preservation.

NAA (data recovery from carriers):

- Problematic with sensitive or security classified data;
- External Data Recovery Companies don't like recording the types or descriptive data which we like to record. Is the detailed metadata recorded about series, media and data objects necessary for future audit and authentication? What will a researcher require in the future when they are looking at some derivative data object?
- Very expensive. In this case costing about \$277 per carrier (300 carriers) for about 8Gb of data (about \$35 per Mb);
- Surprisingly, recovered data as follows:
  - 245 (81.9%) carriers with 100% data recovery
  - 20 (6.7%) system or known duplicate data
  - 14 (4.7%) with partial recovery
  - 14 (4.7%) found to be blank
  - 6 (2%) failed the process completely.



# NAA Part 2

- Although the Data recovery process was first concerned with the media carriers, there are in fact two distinct problems:
  - accessing the physical media (hardware/software issue)
  - accessing the file system by rendering/performing/converting the file (software problem).
- Because of potential media degradation, caution or conservative action should be taken when accessing legacy media;
- The media is only the record carrier (such as the box in the analogue world);
- Manifest of the content is essential;
- There is a vast variety of media types and file formats used over the last 40 years which exacerbates the legacy problem;



# NAA Part 3

## Arrangement and Description (A&D):

- Is continued A&D required at the conclusion of each step of the data recovery process?
- As the custody model at the time of initial digital transfers of these records was distributed, it is assumed that the controlling agency also might have provided with paper copies.
- Ideally when conducting A&D on these series, cross-referencing all digital files with paper records of the same series. Looking for a direct correlation between both.
- Which has primacy, duplicates of digital or paper records? Are the paper printouts of digital records copies? Does digital make more sense from a storage perspective?

# NLA Part 1

In relation to Web Archives, the following are problematic:

- Knowing what you have;
- Measuring numbers of files;
- Dealing with imbedded digital objects;
- Processing and identifying 'unusual' file formats;
- Maintaining links between objects;
- Dilemmas about migration & emulation of at 'risk files'; and
- Browsers are so tolerant of poor conforming code that migration results can be variable.

# Part 4: Solutions

No one has solved this yet. We all have small parts of the puzzle. As a community we need to concentrate on services, tools application that work together. We need to:

- Avoid duplication of effort;
- Build tools that work not only for your own business but could work for others;
- If a product is good, concentrate effort rather than start yet another project;
- Use standardized APIs;
- Make vendors partly responsible for their own file formats and hardware; and
- Don't lock ourselves into a situation where there is no way out 'except pain'.

# Ok, I'll do it

This is easier said than done!

As we all have:

- Different businesses (to varying degrees. Although we all want to preserve data).
- Different internal environments (are usually dependant on applications and services which are customised to our business)
- Different common understanding of the problem and what is required to fix the problem; and
- Different levels of resources and sustainability models.



# Part 4: What can 'I' do about it?

I have been working on a number of projects to assist in the acquisition, ingesting, and ensuring access to digital content (these are):

Mediapeda (soon to be a community web-base service for carrier identification and knowledge) aims to:

- be a curated repository for business knowledge and 'trusted' sources;
- enable the identification with a low level of audience knowledge;
- cater for both general and specialists audiences;
- contain additional information about dependencies (such as hardware, connectors, interfaces, software, etc.) and access paths for usage which are not currently represented in other sources such as Wikipedia; and
- be a web based service which can be integrated internally or externally to other services.

# Mediapeda



**Holotypes**

HomeContactsCalendarProjectsGenresCandidates**Holotypes**Variants



# H000007Candidate

Holotype DetailsList ViewTable ViewDocumentation

Basic InfoGenre and NotesImage Management

Data

5 1/4" Floppy Diskmagnetic disk sleeved



HolotypeID: H000007

Carrier Type: disk sleeved

Process Type: magnetic

Medium Type:

Name: 5 1/4" Floppy Disk

Standards: EBU: FD5

Dimensional Info

Height:133 mm

Width:133 mm

Diameter:

Depth:1 mm

Running Length (Min>Max):

Holotype Status: **Candidate**

Add Line Item

Stretch window to view more line items

| ID       | Name of Variant Product                | Product Code | Manufacturer | Prod. Start > Stop | Required Device                         |   |
|----------|--|--------------|--------------|--------------------|---|---|
| V0000176 | 5.25" diskette digital data disk, reel |              |              |                    |   | X |
| V0000177 | Floppy Disk 5.25" SS                   |              | various      | 1976               |   | X |
| V0000178 | Floppy Disk 5.25" DSSD                 |              | various      | 1978               |   | X |
| V0000179 | Floppy Disk 5.25" SSDD                 |              | various      | 1978               |   | X |
| V0000180 | Floppy Disk 5.25" DSDD                 |              | various      | 1978               |   | X |
| V0000181 | Floppy Disk 5.25" DSDD                 |              | various      | 1978               |   | X |
| V0000182 | Floppy Disk 5.25" MD2-HD               |              | Verbatim     | 1982               |   | X |
| V0000183 | 5 1/4" Floppy Disk RX50K               |              | Digital      | 1970               | 1980s for use in Digital Micro PDP-11 & | X |

Created By: Digital PreservationDate:Modified By: Digital PreservationDate: 17/02/09

© Leonard French

# Prometheus

The NLA built a application called Prometheus to transfer digital content off common media carriers into managed storage system in a systematic manner.

The NLA had to designed a system and process which was:

- semi-automatic;
- modular;
- scalable (able to deal with multiple carrier types); and
- capable of automatically harvesting and generating appropriate metadata for future access.


The acquisition of digital materials on carriers is a constantly growing problem. Therefore, the NLA had to address:


- the backlog
- adding to the backlog



# Prometheus



**PROMETHEUS**  
digital preservation workbench

**NATIONAL LIBRARY OF AUSTRALIA**

[Logout](#) [About](#) [Help](#)

Search

My Jobs

My Jobs [Create or Find Job](#)

Assigned

| Name   | Priority | Last Update         |
|--|----------|---------------------|
| <a href="#">Victorian government gazette 1862</a>            | High     | 14 Oct 08, 01:54 PM |
| <a href="#">Victorian government gazette 1877</a>            | High     | 14 Oct 08, 12:21 PM |
| <a href="#">Victorian Post Office directory 1888 (Wise)</a>  | High     | 14 Oct 08, 11:45 AM |
| <a href="#">Victoria police gazette compendium 1901-1905</a> | High     | 13 Oct 08, 04:09 PM |
| <a href="#">Solo piano 2</a>                                 | Medium   | 13 Oct 08, 03:25 PM |

Working

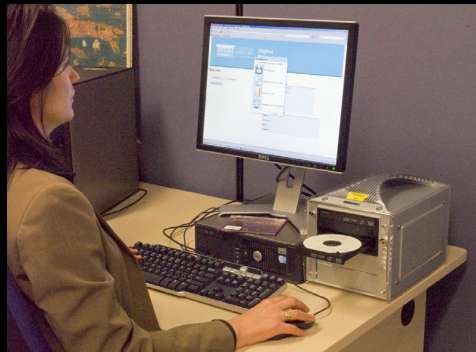
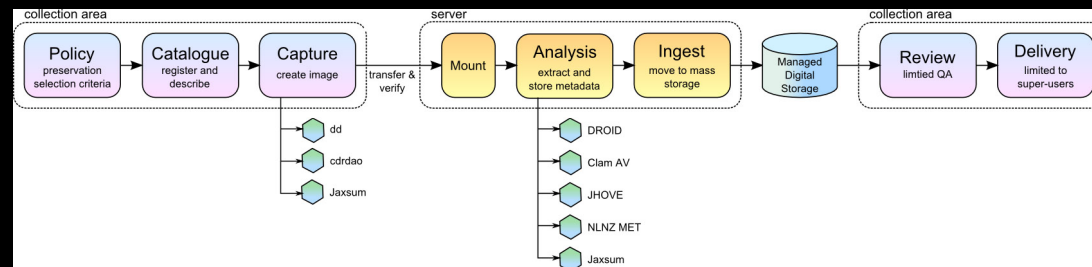
| Name  | Status  | Priority | Last Update         |
|---|---------|----------|---------------------|
| <a href="#">West Tanami</a>   | Working | Medium   | 13 Oct 08, 10:45 AM |
| <a href="#">Choices: stories of young women's experiences with binge drinking: a short film</a> | Working | High     | 13 Oct 08, 09:29 AM |
| <a href="#">Then we were three: building a stronger, healthier relationship</a>                 | Working | High     | 13 Oct 08, 09:20 AM |
| <a href="#">Instant families: building a stronger, healthier relationship</a>                   | Working | High     | 13 Oct 08, 09:04 AM |
| <a href="#">Taking the first step: building a stronger, healthier relationship</a>              | Working | High     | 13 Oct 08, 08:59 AM |

Finished

| Name  | Status   | Priority | Last Update         |
|---|----------|----------|---------------------|
| <a href="#">Nature conservation (estuarine crocodile) conservation plan 2007 and management program 2007 - 2017</a> | Finished | High     | 13 Oct 08, 08:44 AM |
| <a href="#">Queensland Police Gazette: compendium 1916-1920</a>   | Finished | High     | 10 Oct 08, 03:31 PM |
| <a href="#">Little Billabong: Hume Highway duplication</a>  | Finished | High     | 10 Oct 08, 02:48 PM |



# Prometheus

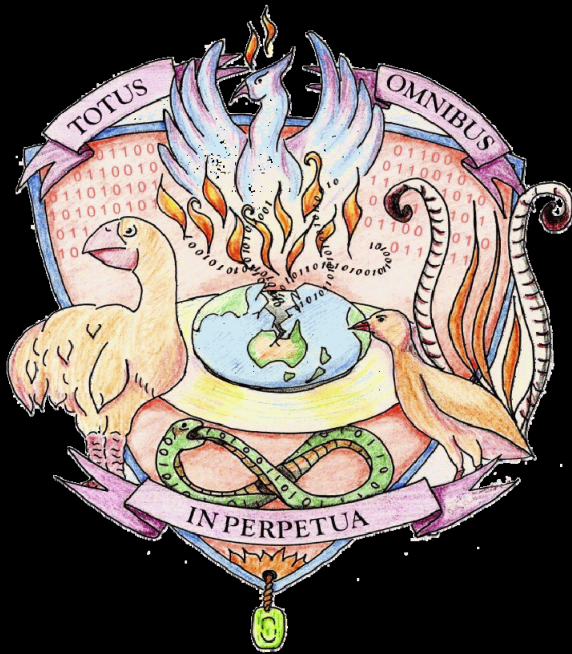


# Conclusion

In conclusion:

- We are all responsible for a lot of “digital stuff”;
- If we simply collect and store it, it will become unusable in a relatively short time as technologies change;
- Maintaining the ability to access it requires a lot of good management, planning, & dedicated resources; and
- We have to find and use solutions that can be applied automatically and reliably to billions of digital files.

# References



Everything, for Everyone  
Forever

## Websites

NLA Digital Preservation Section  
<http://www.nla.gov.au/preserve/digipres/index.html>

Mediapeda Information  
<http://www.nla.gov.au/mediapeda/>

Prometheus Sourceforge Website:  
<http://prometheus-digi.sourceforge.net/>



## Some Recent Articles:

Elford, D., del Pozo, N., Mihajlovic, S., Pearson, D., Clifton, G. and Webb, C. 2008. Media Matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia. In World Library and Information Congress: 74th IFLA General Conference and Council 10-14 August 2008, Québec, Canada  
[www.ifla.org/IV/ifla74/papers/084-Webb-en.pdf](http://www.ifla.org/IV/ifla74/papers/084-Webb-en.pdf).

Pearce, J., Pearson, D., Williams, M. and Yeadon, S. 2008. 'Australian METS Profile: A Journey about Metadata' *D-Lib Magazine* March/April 2008, Vol. 14 No.3/4, <http://www.dlib.org/dlib/march08/pearce/03pearce.html>

Pearson, D. 2008. Titans in the Library: Prometheus Unbinds At-risk Data. In Gateways Dec 2008.  
<http://www.nla.gov.au/pub/gateways/issues/96/story02.htm>.

Pearson, D. and Webb, C. 2008. 'Defining File Format Obsolescence: A Risky Journey', *The International Journal of Digital Curation* (IJDC), Issue 1, Volume 3 (July 2008), pp.89-106. <http://www.ijdc.net/ijdc/article/view/76/78>