



What are some of the things that
the ‘Feds’ are doing about
Digital ‘Stuff’ ?

Presentation to
NSW State Library

David Pearson

15 February 2011

Collecting framework for Digital Collections

The National Library Act 1960:

requires us to maintain and develop “...a comprehensive collection of library material relating to Australia and the Australian people”

The Collection Development Policy:

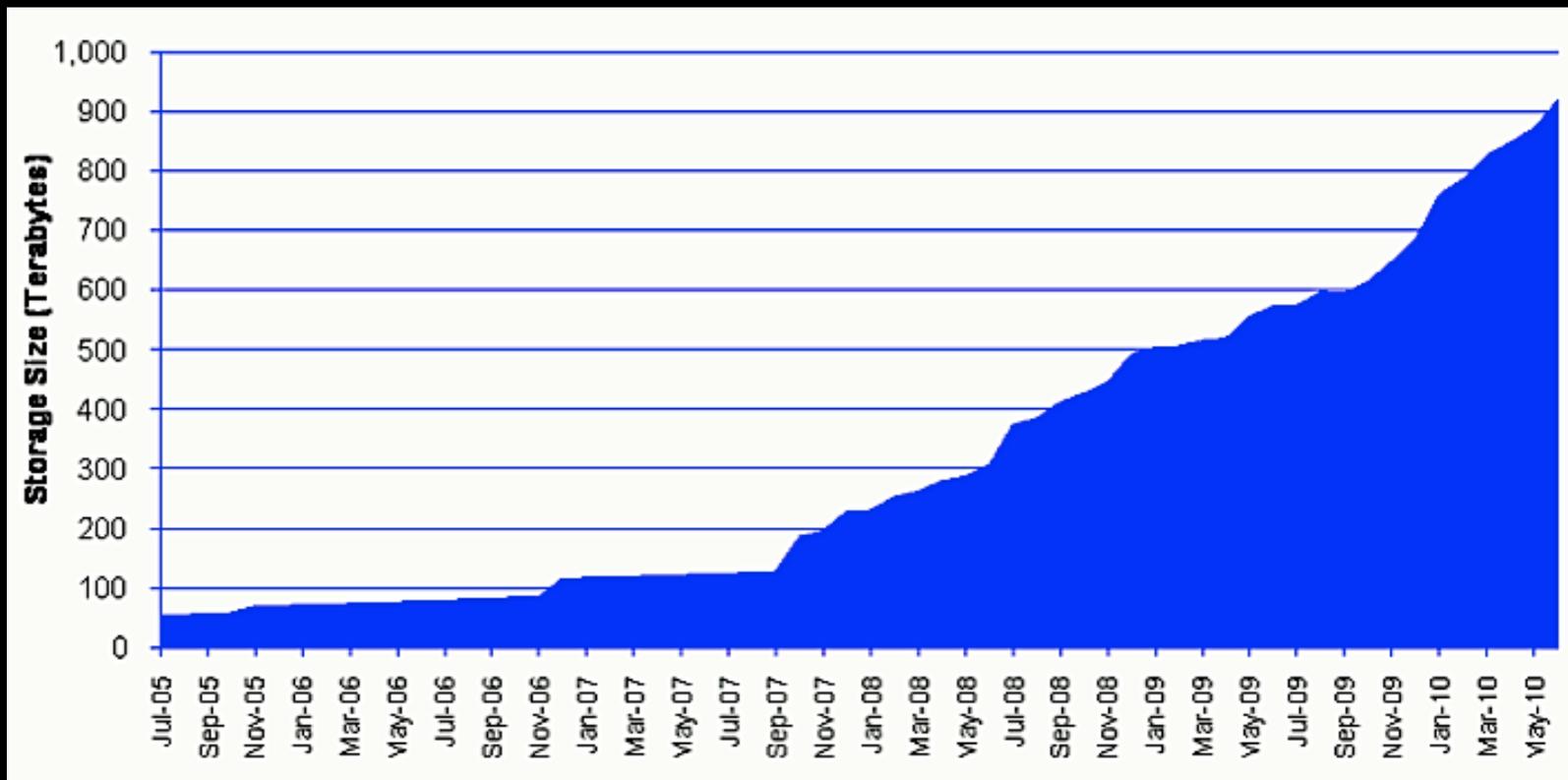
outlines collecting intentions for print and digital resources of all kinds – e.g., pictures, maps, publications ...

How do we get it?

The NLA collects digital ‘stuff’ in many forms and through a number of different ingest mechanisms. Generally this can be categorised as:

- Internally generated;
- External sourced (purchase, donation and future legal deposit).

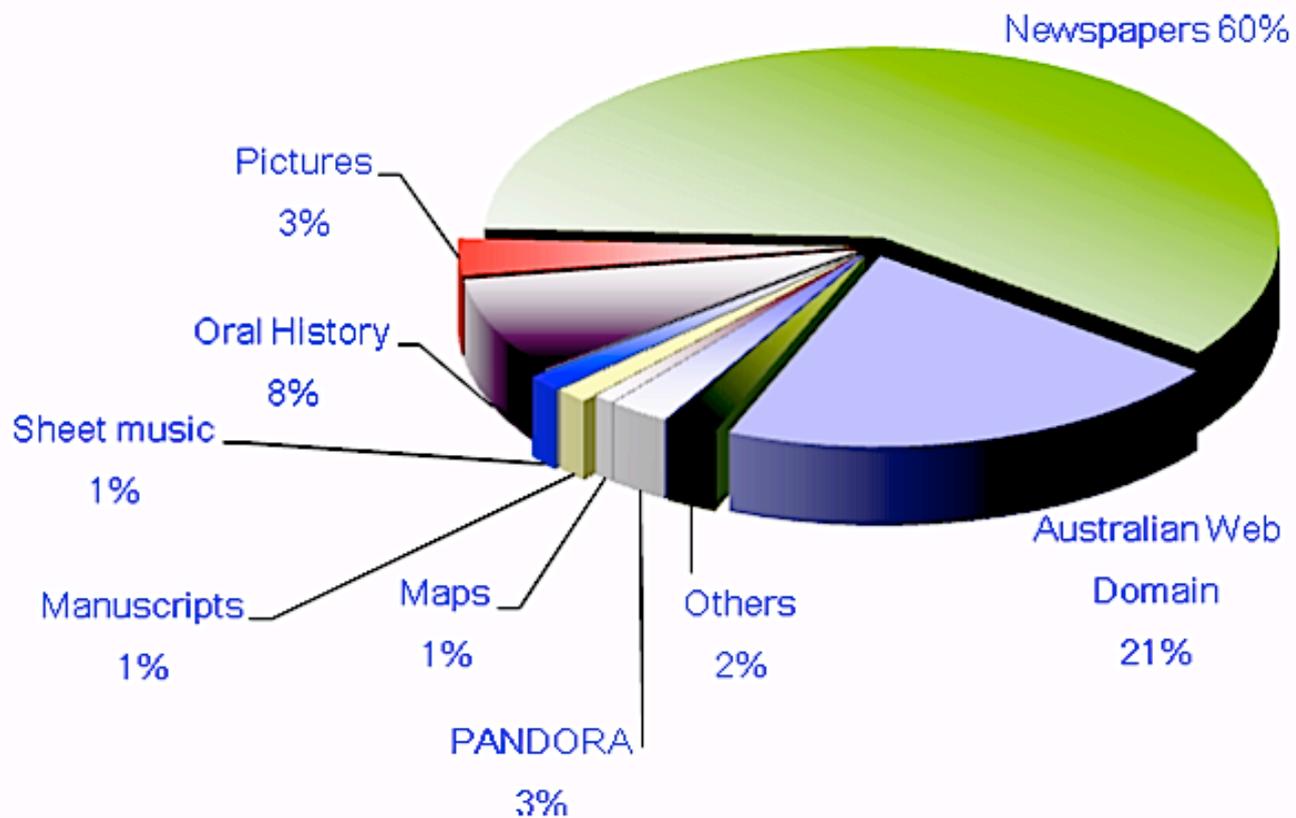
Growth in digital collections stored July 2005 to June 2010



1,000 Terabytes = 1 Petabyte

(= 1,000,000 Gigabytes or 250,000 x 4 GB USB memory sticks)

Diversity of digital 'stuff' stored by material type - 2009/10



Who is responsible for digital 'stuff' ?

At the NLA, Digital Preservation is not just the name of a business unit, it is the act of maintaining access to collection materials.

Therefore everyone at the NLA is responsible for digital preservation in some way:

- Collections Owners;
- Digital Preservation; and
- IT.

Digital Preservation is about maintaining access to the collection

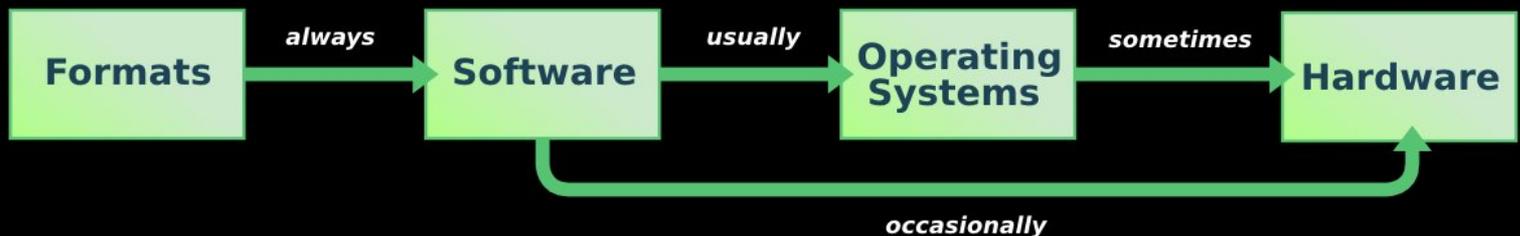
But this is an ongoing process as ‘digital stuff’ is dependent on technology at all stages:

- Creation/capture;
- Storage; and
- Access.

For some of us, the scale of the problem is exacerbated by the scale of our digital ‘stuff’ .

The general preservation problem

The digital ecosystem is inherently fragile in that our ability to access the dependences for a file change over time (sometimes rapidly, sometimes more slowly).



Entropy

- Loss of availability or degradation of the technology over time.
- Loss of skills and understanding of the technology
- This is a constant process



Entropy effects every part of the ecosystem, from the data to the mechanisms to access the data.

There are the things that we know.

There are the things we don't know.

Then there are the things we know we don't know.

And things we don't know that we don't know (D. Rumsfeld)

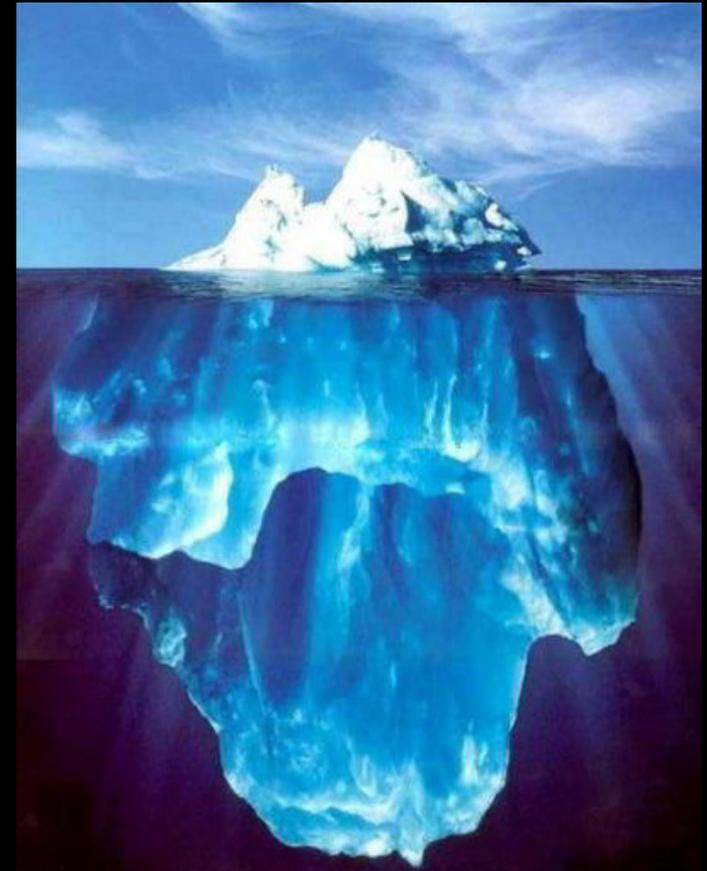


Image from <http://media.photobucket.com/image/image%20of%20iceberg%20cross%20section%20water/unenergy/Black%20Beach/iceberg-below-water.jpg>

Maintaining accessibility: The three headed dog

There are three main facets to our preservation problem:

- Maintaining access to the bit-stream;
- Maintaining access to content;
- Maintaining access to meaning.



Image from <https://supportforums.cisco.com/message/3269168>

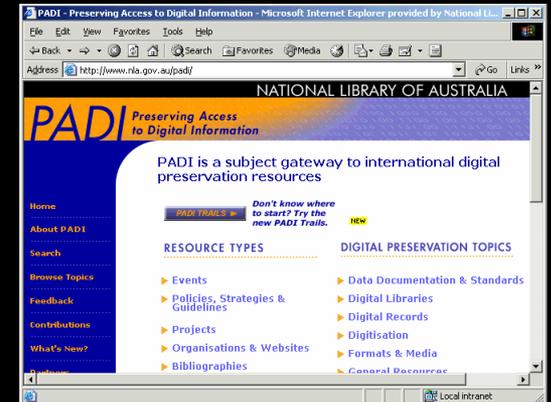
Access to content and meaning

Levels of Representation

```
0011001111010101010100100101001010010010010  
111110101010010101010010101010010100000010  
11010101001001001001001011111111110100101  
010101010100000001010101111010010101010101  
10100101010010001001010111101111010010101  
11111110101111010010101001001000010111111  
10101010101111001010101010101010101010101  
101010010101000100101001001001110101010101
```

Bits

```
<BODY BGCOLOR="#FFFFFF" LEFTMARGIN="0" TOPMARGIN="0" M  
MARGINHEIGHT="0" LINK="#6666FF" VLINK="#FF9900" />  
  
<TABLE WIDTH="100%" BORDER="0" CELSPACING="0" C  
<TR>  
<TD BGCOLOR="#000000" ALIGN="right"> <  
HREF="http://www.nla.gov.au"><IM  
HEIGHT="20" ALT="National Librar  
BORDER="0"></A></TD>  
</TR>  
</TABLE>  
<TABLE WIDTH="100%" BORDER="0" CELSPACING="0" C  
BACKGROUND="/padi/images/new_background.gif">  
<TR BACKGROUND="/padi/images/new_backgrou  
<TD WIDTH="433"><IMG SRC="/padi/images/  
HEIGHT="54"></TD>  
<TD COLSPAN="3" ALIGN="right" VALIGN="t  
</TR>  
</TABLE>  
<TABLE WIDTH="80%" BORDER="0" CELSPACING="0" C  
<TR VALIGN="TOP">
```



Representations
of content

? Meaning

? Meaning



Bits

```

0011001111010101010100100101001010010010010
111110101010010101010010101010010100000010
11010101001001001001001011111111110100101
010101010100000001010101111010010101010101
10100101010010001001010111101111010010101
111111101011111010010101001001000010111111
10101010101111001010101010101010101010101
101010010101000100101001001001110101010101

```

Item from an NLA Manuscripts collection

```

±
Fy@T-2yyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyy#mmmmmmmmmmmmmmmmmm
A. XXXX
: MONTHLY OUTGOINGS mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
-----
INSURANCE ACCT      †18.20 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
LIFE & CITS         3.04 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
EQUITABLE          150.00 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
INSURANCE           23.85 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
9.62 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
29.60 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
152.50 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
-----
TOTAL:
386.81 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
mmMULTI BROADCAST (TV & VCR) 15.99 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
13.95 mmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmmm
CONSUMERS' ASSN²

```

```

*****: MONTHLY OUTGOINGS
-----
DSS INSURANCE ACCT      £18.20
MUTUAL LIFE & CITS      3.04
SCOTTISH EQUITABLE     150.00
PROVINCIAL INSURANCE   23.85
ALLIED DUNBAR           9.62
GEN. ACCIDENT LIFE     29.60
ROYAL BANK (HIP)       152.50
-----
MONTHLY TOTAL:         386.81

plus

MULTI BROADCAST (TV & VCR) 15.99
13.95

CONSUMERS' ASSN

```

opened with notepad

opened with other software

? Meaning



Thus, we have to collect digital 'stuff', we have lots of it, and we need to maintain access to the majority of it for extended periods of time.

Simple, hey?



NLA Image



Google Images

The Digital Preservation Business

At the NLA in relation to digital ‘stuff’ , we have to:

- Understand what we have;
- Understand what we want to do with it;
- Understand our current level of support and anticipate the impact of change;
- Plan and take appropriate actions on a scale appropriate with the size of the target.



Understanding our collections

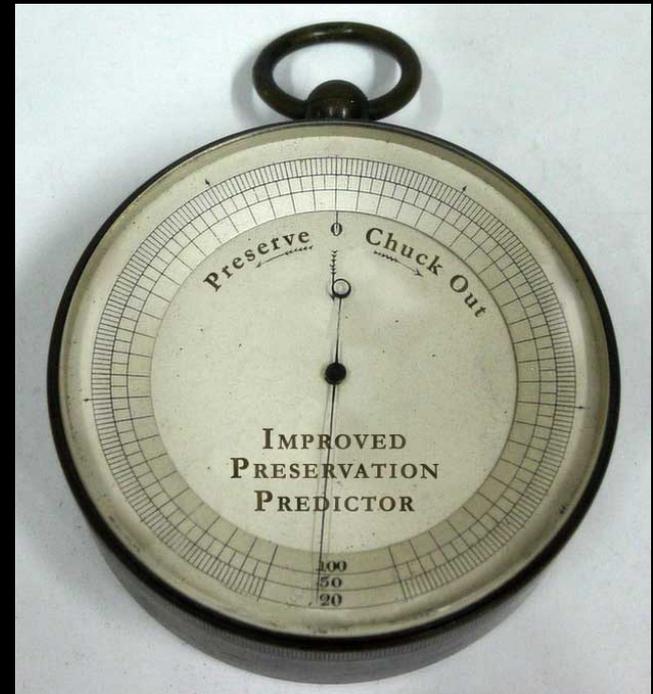
For all digital objects in the collection we ultimately want to be able to get ‘live’, reliable information from collection metadata about:

- format types and versions,
- file numbers
- file size
- file functions
- business owner
- and other digital object metadata

Understanding the preservation intent for all digital 'stuff'

The NLA needs to be able to group all files in the repository into a preservation profile so that we have an agreed understanding of what we wish to maintain over time. For example:

- What group do these files belong to (could be grouped by genre, format or function etc.)?
- Do we want to preserve them? If so, for how long?
- In what way?



NLA Image

Manage the level of support that we wish to maintain

In order to understand the current level of support for formats and more particularly content, the NLA needs to know:

- What software do we have that can access this content?
- Are there any restrictions?
- Based on this how long are we likely to be able to adequately access the file formats .

Plan and take action

Thus, we require certain consistent information to be able to preserve a digital object. Based on this information we can then:

- Work out the best preservation action to maintain the preservation intent.
- Plan and test appropriate actions (this could be simple or complex)
- Schedule actions and carry them out
- Repeat as needed

We understand what we have to do ...



http://www.starstore.com/acatalog/ROTK_Mount_Doom_poster_L.jpg

... doing it is another story

Current NLA Strategies

Current digital preservation strategies include:

- making information available about the level of support that the NLA can provide for digital materials;
- helping to process digital content from physical carriers to managed storage for bit level preservation;
- working towards building a holistic understanding of the NLA collection;
- Building knowledgebases about formats, software and dependencies;
- Influencing IT infrastructure:
 - ie. DLIR (Digital Library Infrastructure Redevelopment project)



Helping to process digital content from physical carriers to managed storage for bit level preservation – i.e. Prometheus

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

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

- Processing material that we already have
- Processing new materials as they come in



Select Media Type

Please contact Digital Preservation if your media is not listed here.

 CD-ROM or DVD-ROM	 External Hard Disk or USB	 Floppy Disk (3.5 inch)	 File
--	--	---	---

Cancel

Select Media Type

 CD-ROM or DVD-ROM	 External Hard Disk or USB	 Floppy Disk (3.5 inch)	 File
---	---	--	--

Available only to Digital Preservation

 Floppy Disk (5.25 inch)	 CD-ROM Image	 DVD Image	 Block Image
--	---	--	--

Cancel

Working towards building a holistic understanding of the NLA collections

Dashboard > Web Archiving & Digital Preservation > ... > Collection Info > DCM

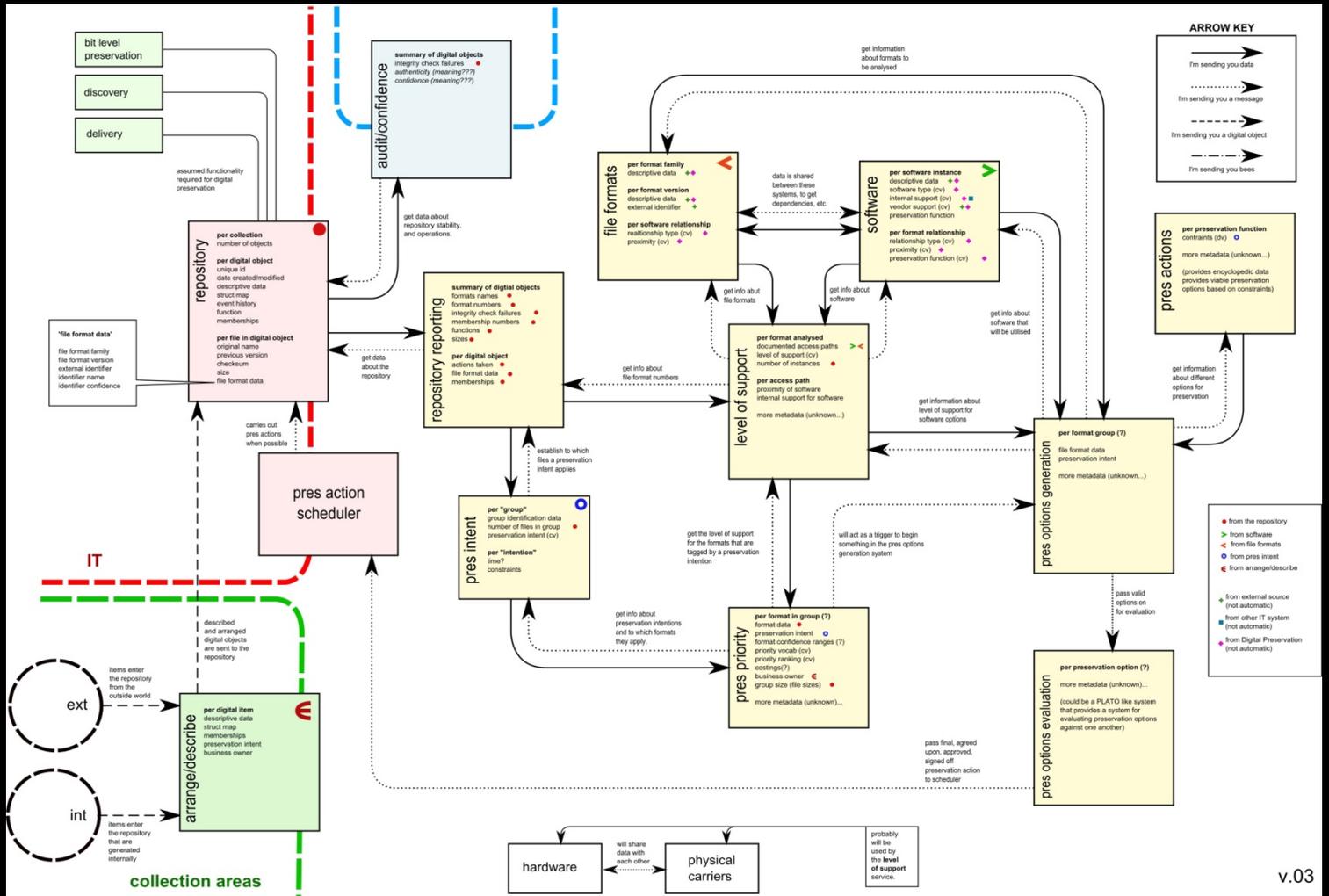
ourWiki DCM

This page presents summary information on the DCM collection content. This information was obtained via an IT Cube Report

IT Identified Collection	File Format	Copyrole	No. of Files	Size	Unit	Source/Date	Pres Evaluation	Pres Intent
nla.aus (Australian Books & Serials)	JPEG file	examination	15,112	2.72	GB		DNP	Unknown
		thumbnails	16,070	112.57	MB		DNP	Unknown
		view	15,113	1.63	GB		DNP	Unknown
			46,295	4.47	GB			
	MRSID	special delivery	224	616.35	MB		DNP	Unknown
	PDF file	print	957	26.36	GB		DNP	Unknown
	TIFF file							
		co-master	4,257	107.73	GB		P	Unknown
		derivative master	16,057	52.90	GB		P	Unknown
		master	15,648	677.54	GB		P	Unknown
			35,962	838.16	GB			
	XML file	finding aid	6	0.62	MB			Unknown
Total			83,685	869.59	GB	DCM Cube 11/08/10		
nla.con (Conservation)								Collection Self Managed?
	JPEG file							
	derivative master		774	294.34	MB		n/a	Unknown

This report for DCM gives us most of this data. But not all NLA systems can provide this data.

Build knowledgebases





File Format - Digital Preservation Knowledge Base - OurWiki - Microsoft Internet Explorer provided by National Library Of Austr

http://ourweb.nla.gov.au/apps/wiki/display/DPKB/File+Format

File Edit View Favorites Tools Help

Google Search Translate Define Wikipedia Translate page Tools Callro... a 18°C dapeerso

File Format - Digital Preservation Knowledge Base - O...

Dashboard > Digital Preservation Knowledge Base > Knowledgebase Home > File Format

Browse David Pearson Search

All File Formats

Family Name	Name	Type	Date modified
	BMP		2/11/10 11:09 AM
BMP	BMP - 1	image	1/11/10 9:59 AM
BMP	BMP - 2	image	1/11/10 10:05 AM
BMP	BMP - 3	image	1/11/10 10:06 AM
BMP	BMP - 4	image	1/11/10 10:07 AM
BMP	BMP - 5	image	1/11/10 10:08 AM
	BWF		2/11/10 11:11 AM
BWF	BWF - 1	audio	1/11/10 10:43 AM
	DNG		2/11/10 11:11 AM
DNG	DNG - 1.0.0.0	image	1/11/10 10:48 AM
DNG	DNG - 1.1.0.0	image	1/11/10 10:48 AM
DNG	DNG - 1.2.0.0	image	1/11/10 10:49 AM
DNG	DNG - 1.3.0.0	image	1/11/10 10:50 AM
	DOC		2/11/10 11:12 AM
DOC	DOC - 95	document	28/09/10 10:48 AM
DOC	DOC - 2003	document	28/09/10 11:05 AM
DOC	DOC - 2007	document	28/09/10 11:06 AM
	GeoTIFF		2/11/10 11:12 AM
GeoTIFF	GeoTIFF - 1	image	1/11/10 11:03 AM
	GIF		2/11/10 11:13 AM
GIF	GIF - 87a	image	29/10/10 11:31 AM
GIF	GIF - 89a	image	29/10/10 11:32 AM
	H.264		3/11/10 11:43 AM
H.264	H.264 - Version 1	video_codec	3/11/10 11:30 AM

Atlassian Confluence 3.2.1_01, the Enterprise Wiki: Intranet software for documentation and knowledge management | Report a bug | Atlassian News

Done, but with errors on page. Local intranet 100% 3:12 PM

Start Inbox - Microsoft Outlook File Format - Digital P...



NATIONAL LIBRARY OF AUSTRALIA



File Format - BMP - Digital Preservation Knowledge Base - OurWiki - Microsoft Internet Explorer provided by National Library Of

http://ourweb.nla.gov.au/apps/wiki/display/DPKB/File+Format+-BMP

File Edit View Favorites Tools Help

Google Search Translate Define Wikipedia Translate page Tools Califo 18°C

File Format - BMP - Digital Preservation Knowledge Bas...

Dashboard > Digital Preservation Knowledge Base > ... > File Format > File Format -BMP

Browse David Pearson Search

ourWiki File Format -BMP

Added by Nicholas DelPoza, last edited by Nicholas DelPoza on Nov 02, 2010 (view change)

Edit Add Tools

Click on field label for more info.

File Format family

family name	
general information	BMP is a contraction of Bitmap. BMPs are very simple containers for image based information. Even though they are practically ubiquitous, it can be difficult to pin point particular versions and release dates. Tentatively we can say the format was first released by Microsoft probably around 1986. Subsequent versions seem to have followed the release of new Windows operating systems.
format type	image
extension	.BMP
vendor	Microsoft
technical capacity	BMP is a raster image format that can support a variety of bit-depths, but only in the RGB colour space. It holds enough metadata to describe the version of the format being used, and the dimensions and bit-depth of the image, but no more than this. BMP is an uncompressed format.
conditions of use	The specification for BMP is widely available and appears free of restrictions.
adoption	Support for BMP is fairly ubiquitous. This probably has a lot to do with the specification being widely and easily available, and totally unrestricted in usage. It also probably has a lot to do with this being the de-facto choice of image format for many early Windows installations. However, usage of BMP is currently very limited, especially in preference of GIF or PNG.
simplicity	BMP is a very simple format. Given that it does not contain any superfluous metadata, and does not have support for compression schemas, it should be considered even more simple than other bitmap style raster images, such as TIFF, or PNG.
stability	BMP is a highly stable format, from the perspective that it has not seen any significant changes in some time, and is unlikely to undergo any drastic changes in the foreseeable future.
specification	http://www.fileformat.info/format/bmp/spec/e27073c25463436f8a64fa789c886d9c/view.htm

Atlassian Confluence 3.2.1_01, the Enterprise Wiki: Intranet software for documentation and knowledge management | Report a bug | Atlassian News

Done Start Inbox - Microsoft Outlook File Format - BMP - Di... Local intranet 100% 3:11 PM





File Format - BMP - 1 - Digital Preservation Knowledge Base - OurWiki - Microsoft Internet Explorer provided by National Library

http://ourweb.nla.gov.au/apps/wiki/display/DPKB/File+Format+-BMP+--+1

File Edit View Favorites Tools Help

Google Search Share Bookmarks Translate AutoFill dapearso

Translate Define Wikipedia Translate page Tools

File Format - BMP - 1 - Digital Preservation Knowledge Base

Dashboard > Digital Preservation Knowledge Base > ... > File Format -BMP > File Format -BMP - 1

Browse David Pearson Search

ourWiki File Format -BMP - 1

Added by Nicholas DelPozo, last edited by Nicholas DelPozo on Nov 01, 2010 (view change)

File Format family version

version name

general information	Released circa 1986.
---------------------	----------------------

Click on field label for more info.

External identifiers

source	PRONOM
identifier	fmt/114

Software - File Format relationships

Relationship	Proximity	Software
open	third party relative	Software - Photoshop CS5

Edit file_format_version

Add Comment

Atlassian Confluence 3.2.1_01, the Enterprise Wiki: Intranet software for documentation and knowledge management | Report a bug | Atlassian News

Done Local intranet 100%

Start Inbox - Microsoft Outlook File Format - BMP - 1 - ... 3:10 PM

- Add new Software item
- Add new Hardware item
- Add new Physical Carrier item
- Add new File Format item

- File Format
 - File Format -BMP
 - File Format -BMP - 1
 - File Format -BMP - 2
 - File Format -BMP - 3
 - File Format -BMP - 4
 - File Format -BMP - 5
 - File Format -BWF
 - File Format -DNG
 - File Format -DOC
 - File Format -GeoTIFF
 - File Format -GIF
 - File Format -H.264
 - File Format -HTML
 - File Format -ISO 9660
 - File Format -JP2
 - File Format -JPEG
 - File Format -JPEG 2000
 - File Format -JPEG File Interchange Format
 - File Format -JPX
 - File Format -Matroska
 - File Format -MP3
 - File Format -MrsID
 - File Format -PDF
 - File Format -Photo CD
 - File Format -PNG
 - File Format -PPT



Software - Photoshop CS5 - Digital Preservation Knowledge Base - OurWiki - Microsoft Internet Explorer provided by National Lib

http://ourweb.nla.gov.au/apps/wiki/display/DPKB/Software++Photoshop+CS5

File Edit View Favorites Tools Help

Google Search Share Bookmarks Check Translate AutoFill

Translate Define Wikipedia Translate page Tools

Calif... 18°C

Software - Photoshop CS5 - Digital Preservation Know...

Dashboard > Digital Preservation Knowledge Base > ... > Software > Software - Photoshop CS5

Browse David Pearson Search

ourWiki Software - Photoshop CS5

Added by Nicholas DelPozo, last edited by David Pearson on Nov 25, 2010 (view change)

[Edit](#) [Add](#) [Tools](#)

Click on field label for more info.

Software

software name	
general information	Photoshop is a very popular image manipulation and creation program developed by Adobe. The program itself is very complex, and provides support for a very broad range of functionality and uses. For example, it can be used as a tool to re-touch photos, or to create a new image entirely from scratch. It has a very broad user base, particularly for those who work in the graphics or photographic industries. The default working format for Photoshop, PSD, is practically an industry standard, and can be opened in a variety of programs.
software type	Application
current version	Release
highest level of internal support	Localised
supersedes	
superseded by	

Requirements

operating system	Microsoft® Windows® XP with Service Pack 3; Windows Vista® Home Premium, Business, Ultimate, or Enterprise with Service Pack 1 (Service Pack 2 recommended); or Windows 7
computer and processor	Intel® Pentium® 4 or AMD Athlon® 64 processor
memory	1GB of RAM
hard disk	1GB of available hard-disk space for installation; additional free space required during installation (cannot install on removable flash-based storage devices)
drive	DVD-ROM drive
display	1024x768 display (1280x800 recommended) with qualified hardware-accelerated

Atlassian Confluence 3.2.1_01, the Enterprise Wiki: Intranet software for documentation and knowledge management | Report a bug | Atlassian News

Done, but with errors on page. Local intranet 100%

Start Inbox - Microsoft Outlook Software - Photoshop...

EN 3:12 PM





Software - Photoshop CS5 - Digital Preservation Knowledge Base - OurWiki - Microsoft Internet Explorer provided by National Lib

http://ourweb.nla.gov.au/apps/wiki/display/DPKB/Software++Photoshop+CS5

File Edit View Favorites Tools Help

Google Search Translate Define Wikipedia Translate page Tools

Software - Photoshop CS5 - Digital Preservation Know... Page Safety Tools

Dashboard > Digital Preservation Knowledge Base > ... > Software > Software - Photoshop CS5

name: Wikipedia
url: http://en.wikipedia.org/wiki/Adobe_Photoshop

Software - File Format relationships

Relationship	Proximity	File format
edit-save	native absolute	File Format - PSD - CS5
open	third party relative	File Format - PNG - 1.2
		File Format - PNG - 1.1
		File Format - PNG - 1.0
		File Format - PNG
		File Format - Photo CD - 1
		File Format - Photo CD
		File Format - JPX - 1
		File Format - JPX
		File Format - JPEG - 4
		File Format - JPEG - 3
		File Format - JPEG - 2
		File Format - JPEG - 1
		File Format - JPEG File Interchange Format - 1994
		File Format - JPEG File Interchange Format
		File Format - JPEG 2000 - 1
		File Format - JPEG 2000
File Format - JPEG		
File Format - JP2 - 1		
File Format - JP2		
File Format - BMP - 5		
File Format - BMP - 4		
File Format - BMP - 3		
File Format - BMP - 2		
File Format - BMP - 1		
File Format - BMP		
open	Native Relative	File Format - TIFF - 6

Atlassian Confluence 3.2.1_01, the Enterprise Wiki Intranet software for documentation and knowledge management | Report a bug | Atlassian News





File format - Software - Digital Preservation Knowledge Base - OurWiki - Microsoft Internet Explorer provided by National Libra

http://ourweb.nla.gov.au/apps/wiki/display/DPKB/File+format+-+Software

File Edit View Favorites Tools Help

Google Search Translate Define Wikipedia Translate page Tools Califo a 18°C

File format - Software - Digital Preservation Knowledg...

Dashboard > Digital Preservation Knowledge Base > ... > Relationships > File format - Software

File format - Software

ourWiki File format - Software

Added by Brendon McKinley, last edited by Brendon McKinley on Nov 26, 2010 (View change)

Software	Relationship	Proximity	File format version
Software - Excel 2007	open	native absolute	File Format -XLS - BIFF8
Software - Excel 2007	edit-save	native absolute	File Format -XLS - BIFF8
Software - Excel 2007	open	Native Relative	File Format -XLS - BIFF7 File Format -XLS - BIFF5 File Format -XLS - BIFF4 File Format -XLS - BIFF3 File Format -XLS - BIFF2
Software - Excel 2007	edit-save	Native Relative	File Format -XLS - BIFF7 File Format -XLS - BIFF5 File Format -XLS - BIFF4 File Format -XLS - BIFF3 File Format -XLS - BIFF2
Software - Photoshop CS5	open	third party relative	File Format -PNG - 1.2 File Format -PNG - 1.1 File Format -PNG - 1.0 File Format -PNG File Format -Photo CD - 1 File Format -Photo CD File Format -JPX - 1 File Format -JPX File Format -JPEG - 4 File Format -JPEG - 3 File Format -JPEG - 2 File Format -JPEG - 1 File Format -JPEG File Interchange Format - 1994 File Format -JPEG File Interchange Format File Format -JPEG 2000 - 1 File Format -JPEG 2000 File Format -JPEG File Format -JP2 - 1 File Format -JP2

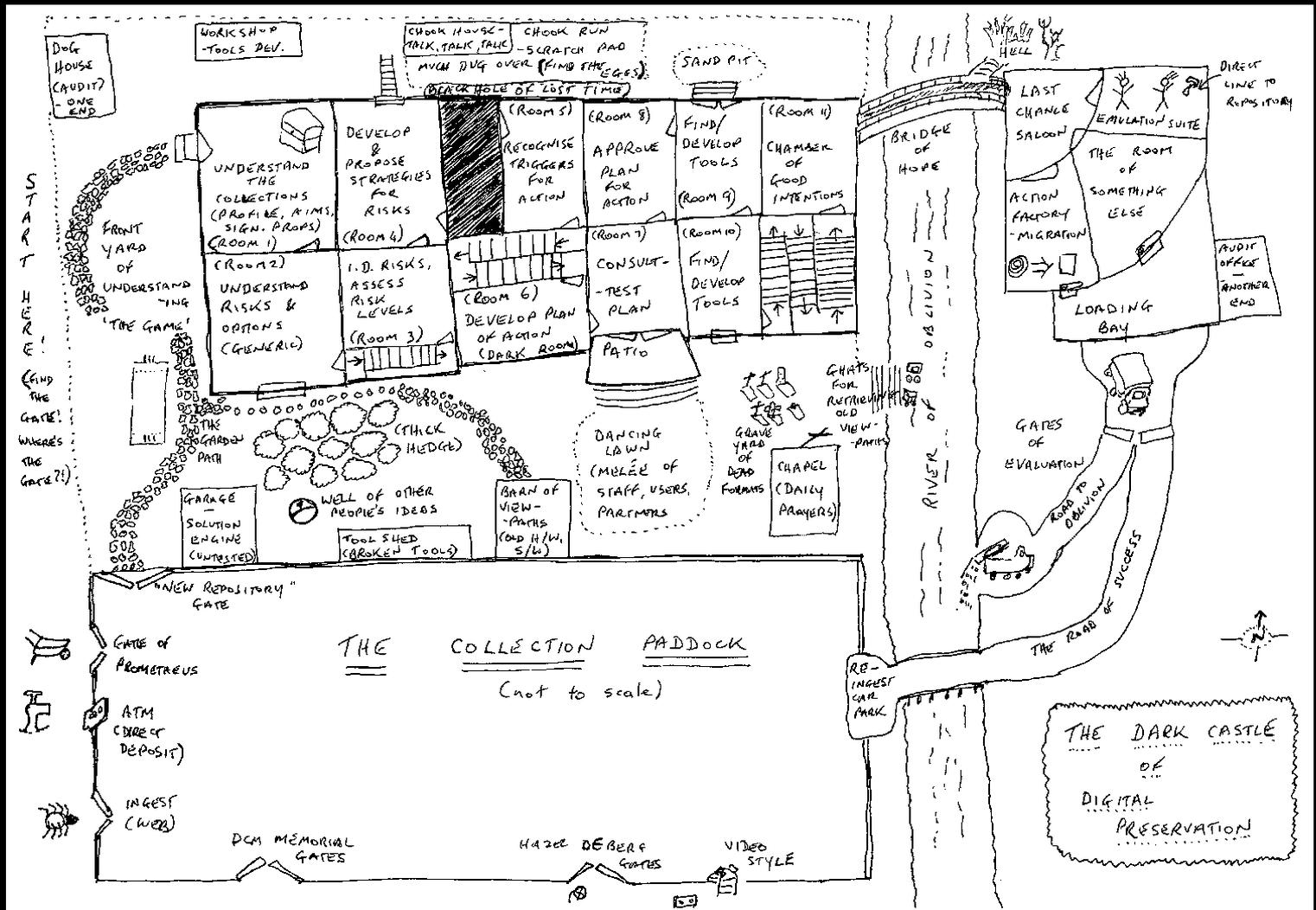
Atlassian Confluence 3.2.1_01, the Enterprise Wiki: Intranet software for documentation and knowledge management | Report a bug | Atlassian News

Done, but with errors on page. Local intranet 100% 3:17 PM

Start Inbox - Microsoft Outlook File format - Software...



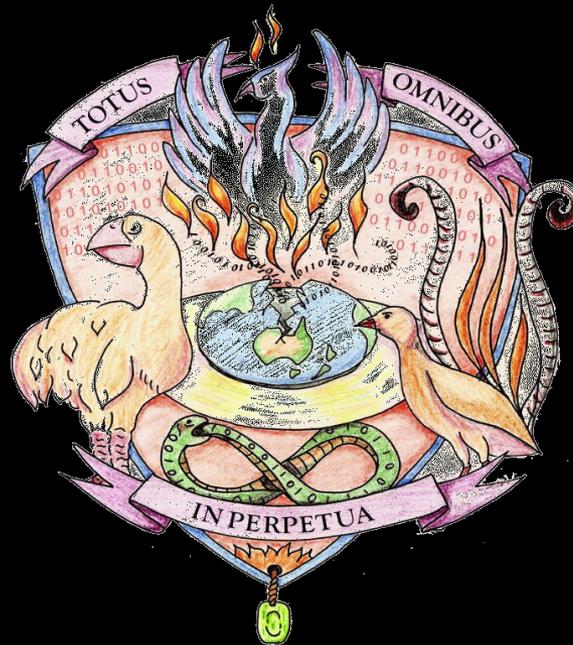
Influencing IT infrastructure development



In conclusion:

- The NLA is responsible for a lot of digital ‘stuff’ and we need to actively understand what it is and what we want to do with it in the future;
- 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.

Thank You



*Everything, for Everyone
Forever*