“What is it and what the hell are we supposed to do with it?”

A happy-go-lucky walk through preservation functionality

Kevin De Vorsey and Peter McKinney

*International Perspectives on Digital Preservation*

2 February 2009
Discussing…?

Main purpose: to describe some of the preservation functionality being developed here

by

Posing some thoughts on how we should/would/will/have dealt with objects coming into the archive

We are aware of the accents….
Let’s play the ‘Preserve the digital heritage’ Game.
Purpose of the Game

To preserve indefinitely the documentary heritage of New Zealand

Main restriction
- To do so in a fashion that maintains their value and is a true representation of what was collected.
Rules of the game

• We must accept content in any format
• We must preserve everything indefinitely
• We must preserve every version
• People can give you anything at any time
• Migration is allowed
• Fixed sum game
For the game today, we’ll use 100 files that we know nothing about, from lots of different people (producers).
Begin
Step 1 – what is it?

Validation Stack Tools

Validation Stack

- Fixity
- Virus
- Validation
- Characterisation

TOOLS

- CRC32
- SHA-1
- MD5
- CLAM AV
- DROID & PRONOM
- JHOVE & NLNZ ET

Baseline for monitoring bitstream change
Virus free?
What format and version?
What properties (e.g. colour encoding)? Is the format well-formed?
ID the files

25 definitively identified objects
(spanning 5 formats)

39 generic matches
(spanning 3 formats)

20 extension mismatches
(spanning 4 different formats)

15 unidentifiable

1 error
Download and:
1. Use another tool to identify
2. Use info sources (forums, info repositories, experts) to explore possibilities
3. Look through the eyes of a hex-editor
DROID fails to identify an RTF file.

The specification describes an order for declaring the format.
Examination shows the character declaration is misplaced.

And it is easily fixed.
And DROID is successful.
Results (perfect world)

100 identified objects
0 probable matches
0 extension mismatches
0 unidentifiable
0 errors

45% not well-formed
So...we have:

It means we have:
30 TIFF images (of dubious version -- rules)
25 JPEG images (of known versioning)
17 Word 97 documents
15 PDF documents (of known versioning)
5 Word Perfect Documents
4 XML documents
2 MP3 files
1 AVI
1 .mus file
What does this mean?

We know what formats our content is in.

IMPRESSIVE
What next?

Safe in the knowledge we know what we have, eat a preservation banana and plan the next steps.