Digital Preservation
The Saga Continues
Agenda

• 10:15 – 11:00
  – Introduction and group formation

• 11:00 – 12:30
  – Filling your box with tools

• 12:30 – 13:45
  – Lunch
Agenda

• 13:45 – 15:15
  – Using your toolkit effectively.

• 15:15 – 15:30
  – Coffee

• 15:30 – 16:45
  – Can one tool do the job of 10?

• 16:45 – 17:30
  – Panel discussion and wrap up.
‘Digital Preservation: what I wish I knew before I started’

• It won’t do itself

• It won’t go away

• Don’t wait for perfection
Digital preservation makes bleak reading ...
Let’s restate the problem …

• Digital stuff has value. It is an asset.
• It has potential and creates new opportunities.
• Use gives rise to direct and indirect outcomes.

…but...

• Deployment depends on software, hardware and people.
• Software, hardware and people change.

...therefore...

• Access is not guaranteed without (some) action
• Value, opportunity, impact not guaranteed
Key responses

1. Migration
Changing the format of a file to ensure the information content can be read

2. Emulation
Intervening in the operating system to ensure that old software can function and information content can be read

3. Hardware preservation
Maintaining access to data and processes by maintaining the physical computing environment including hardware and peripherals.

4. etc
Research and development field, new solutions and new approaches continue to emerge, e.g. virtualisation for preservation
Access and long term use depends on the configuration of hardware and software and the capacity of the operator.

Change is not a bug.
Technology continues to change creating the conditions for obsolescence.

Need to become a learning institution
Storage media have a short life and storage devices are subject to obsolescence.

Be mobile and format neutral
Digital preservation systems are subject to the same obsolescence as the objects they safeguard.

Standards and modularity
Digital resources are intolerant of gaps in preservation.

Ongoing process
The problems are more subtle than we realised a decade ago...

*e.g. file format obsolescence*

Changing file formats? Conformant containers? Units of information?
How to pick a winner ...

beyond and potentially over-writing the criteria ... repository managers should align the recognition and weighting of criteria with a clear preservation strategy that articulates the purpose of the repository and the needs of its designated community;

How to pick a winner

You ain’t seen nothing yet

Data growth on 3 axes
- volume
- complexity
- expectation

... it’s not going to be about obsolescence so much as workflow and capacity
Digital Preservation as a ‘discipline’

Daunting challenge
Decade of research and development
Replete with jargon and acronyms
Turf war between professions?
A whole new barrier

The last decade has shown definitively that using fancy words are not the same as solving problems
How much does it cost?

Lifecycle costs of digital objects vs Lifecycle costs of books vs Lifecycle costs of museum objects vs Lifecycle costs of archives vs Lifecycles costs of historic environment
digital preservation

technology

organization

resources

Scape
The reality?

You don’t need to understand or do all of this.

... and it doesn’t all have to exist at the same time
The reality?

Get started now
not later
Preservation Lifecycle

Identification

Characterisation

Risk Assessment

Planning

Action

DROID, FIDO, FILE, FITS, TIKA...

JHOVE, JPYLEZER, exiftool, FITS...

Knowledge + Policy + Risk = Continue

Plato

Migration, Emulation
This Training

Identification

- DROID
- FITS
- FIDO
- TIKA
- FILE

Characterisation

- JHOVE
- JPYLYZER
- Exiftool
- FITS
So you have dug a hole?
Stage 2

• What did you find?

• Is it worth preserving?

• What are the problems?
Aim of Training
Agenda

• 10:15 – 11:00
  – Introduction and group formation

• 11:00 – 12:30
  – Filling your box with tools

• 12:30 – 13:45
  – Lunch
Agenda

• 13:45 – 15:15
  – Using your toolkit effectively.

• 15:15 – 15:30
  – Coffee

• 15:30 – 16:45
  – Can one tool do the job of 10?

• 16:45 – 17:30
  – Panel discussion and wrap up.
Time to get married

1. Luis Bravo
2. Jose Casanova
3. Vitor Fernandes
4. Sebastien Leroux
5. Joao Pereira
6. Rui Rodrigues
7. Carlos Velentim
8. Jose Carvalho (Papiro)
9. Omar Coelho

1. Jose Carvalho (SDUM)
2. Carlos Duarte
3. Luis Ferreira
4. Cristiana Freitas
5. Claire Johnson
6. Anthony Laerdahl
7. Helena Medeiros
8. Antonio Rodrigues
9. Cidalia Ferreira

Column 1

Column 2
Getting Started (1)

- Wifi Network = SMS, password = Sarmento1881127
- Download Virtualbox (if you don’t have it)
- Start Virtualbox
- Plug-in USB memory key
- Open the memory key folder and double click the extension pack file to install it (follow instructions at this point)
- Return to virtual box:
  - From the main menu (*file*), select “Import Appliance”
  - Browse to the memory key and select the only file
  - Wait for this to import
  - Once done you can safely remove the key.
Getting Started (2)

- Once done, click the machine and press the settings button (maybe in right click)
- Click shared folders
- Click add
- Add a shared folder (e.g. your desktop or downloads folder)
- Tick auto-mount!
- Click OK to return to the main screen
- Start the machine
- Wait..
Getting Started (3)

• Password is training.
• Ignore update manager if it appears
• Press the top left ubuntu home button and type terminal (select and run the app)
• Type: cd /media/sf_Desktop (where Desktop is the folder you shared previously) and press enter
• Type: fido *
Bundle or Not?

• Pros
  – Single Input/Output
  – Consistent
  – Easy

• Cons
  – Out of date
  – Doesn’t Scale
Questions (1)

• What tool would you use?
Keeping Control - Scalable Environments for Identification and Characterisation
Aims

This training course will cover elements dealing with scalable identification, characterisation and validation of large collections of varying file types. Users will be introduced to a number of tools designed for each of these purposes and involved in problem solving scenarios. Further, users will be required to evaluate the use of scalable and cloud based technologies in developing solutions for given scenarios.
Learning Outcomes (1)

• Distinguish between different file types and identify the requirements for characterising each.

• Carry out a number of identification, characterisation, and duplication detection experiments on example files.
Learning Outcomes (2)

- Critically evaluate characterisation and identification tools and assess their advantages and disadvantages when used in different scenarios.
Learning Outcomes (3)

- Conduct an in-depth analysis of large volumes of identification and characterisation data and find representative sample records suitable for preservation planning experiments.
Learning Outcomes (4)

- Compare and contrast the differences in running characterisation and identification tools both stand-alone and within workflows.

- Envisage a system that combines workflows with identification, characterisation and validation tools to suit a variety of scenarios.
Our Last Commitment

Slides will be available Monday!
Thank You

Franz San Galli
Thank-You
Thank You
Next Time...

Building Applications Infrastructures for Action Services

London, September 2013 (wet)
Then....

Critical Path: Effective Evidence Based Preservation Planning

Denmark, November 2013
(cold)
Tonight

www.goo.gl/q6wKB

7:15pm Eleven Bar
@ Hotel Fundador

Free Beer*

Our Table

* 1 Free Beer subject to completion of online survey!