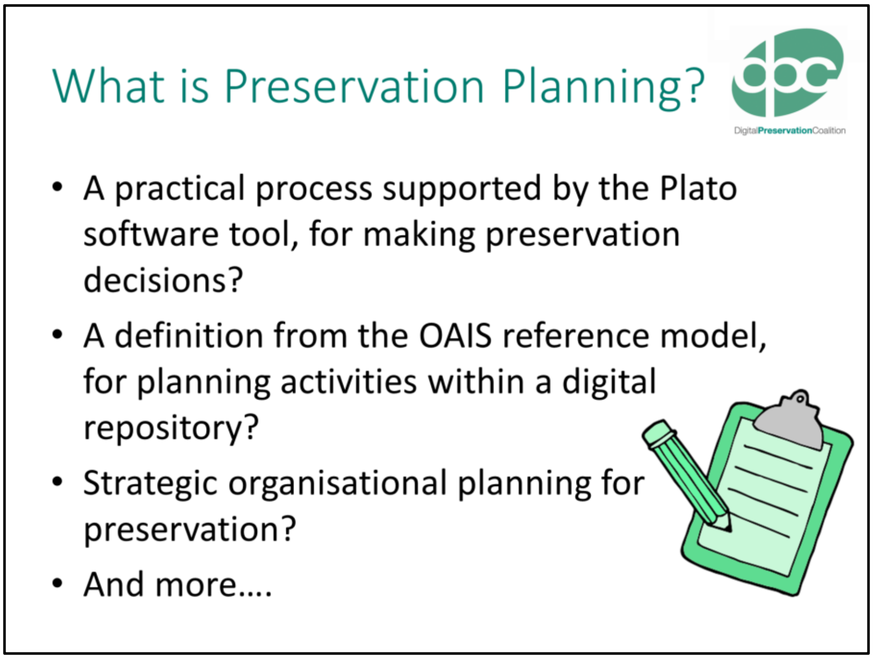
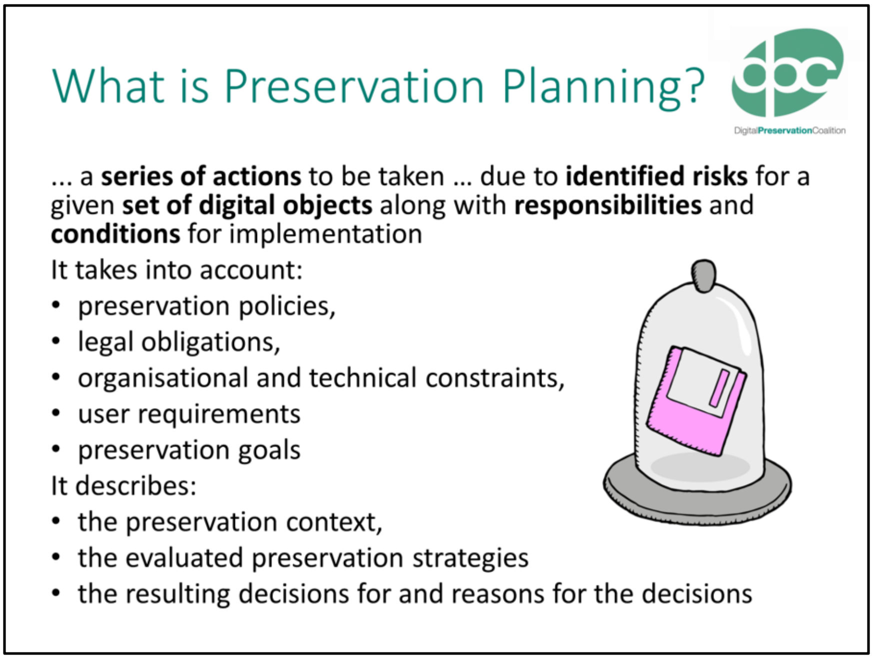
Practical Preservation Workshop

Digital Preservation Coalition



What is Preservation Planning?

- A practical process supported by the Plato software tool, for making preservation decisions?
- A definition from the OAIS reference model, for planning activities within a digital repository?
- Strategic organisational planning for preservation?
- And more….



What is Preservation Planning?

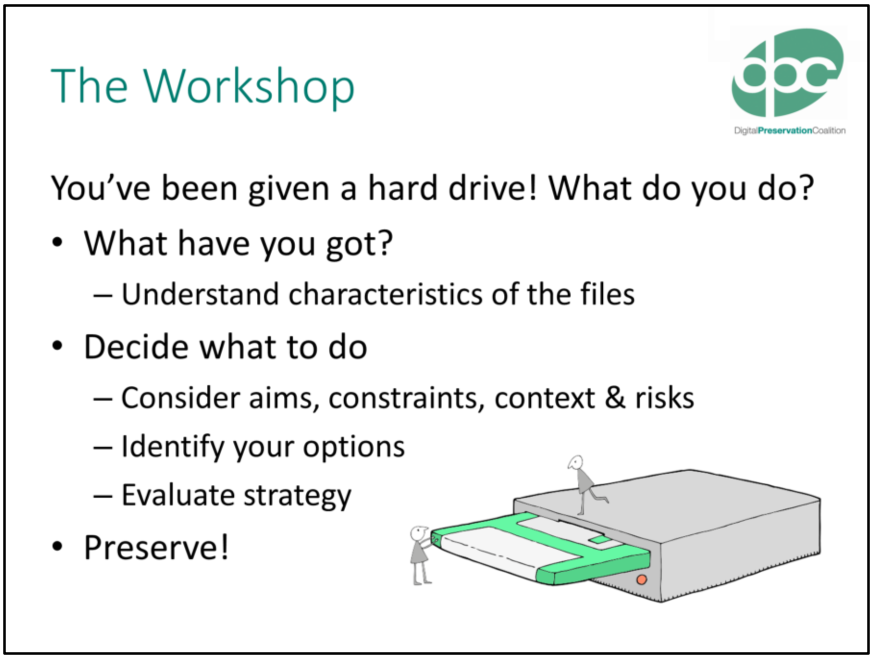
... a **series of actions** to be taken ... due to **identified risks** for a given **set of digital objects** along with **responsibilities** and **conditions** for implementation

It takes into account:

- preservation policies,
- legal obligations,
- organisational and technical constraints,
- user requirements
- preservation goals

It describes:

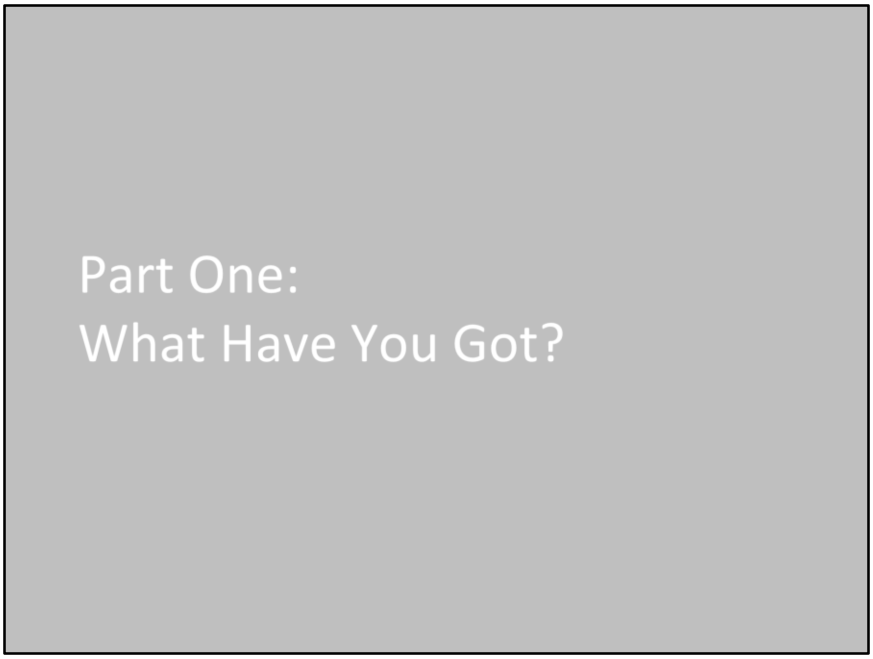
- the preservation context,
- the evaluated preservation strategies
- the resulting decisions for and reasons for the decisions



The Workshop

You've been given a hard drive! What do you do?

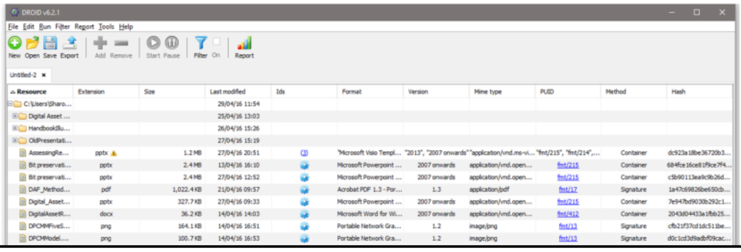
- What have you got?
  - – Understand characteristics of the files
- Decide what to do
  - – Consider aims, constraints, context & risks
  - – Identify your options
  - – Evaluate strategy
- Preserve!



Part One:
What Have You Got?

Basic Characterisation: DROID

- Works with PRONOM file format registry
- Analyses contents of folder(s)
- Captures information such as:
  - File name, location, file size, last edited, format, version, PRONOM ID, checksum
- Outputs raw data or a variety of reports



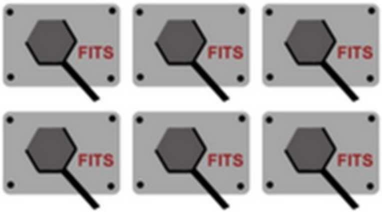
Resource	Extension	Size	Last modified	Ids	Format	Version	Mime type	PUID	Method	Hash
C:\Users\Sharo...			29/04/16 11:54							
Digital Asset ...			25/04/16 13:03							
HandbookIllu...			26/04/16 15:26							
OldPresentati...			27/04/16 15:19							
AssessingRe...	pptx ⚠	1.2 MB	27/04/16 20:51	(3)	"Microsoft Visio Templ...	"2013", "2007 onwards"	"application/vnd.ms-vi...	"fmt/215", "fmt/214",...	Container	dc923a18be36720b3...
Bit preservati...	pptx	2.4 MB	13/04/16 16:10	✱	Microsoft Powerpoint ...	2007 onwards	application/vnd.open...	fmt/215	Container	684fce16ce81f9ce7f4...
Bit preservati...	pptx	2.4 MB	27/04/16 12:52	✱	Microsoft Powerpoint ...	2007 onwards	application/vnd.open...	fmt/215	Container	c5b90113ea9c9b26d...
DAF_Method...	pdf	1,022.4 KB	21/04/16 09:57	✱	Acrobat PDF 1.3 - Por...	1.3	application/pdf	fmt/17	Signature	1a47c69826be650cb...
Digital_Asset...	pptx	327.7 KB	27/04/16 09:33	✱	Microsoft Powerpoint ...	2007 onwards	application/vnd.open...	fmt/215	Container	7e947bd9030b292c1...
DigitalAssetR...	docx	36.2 KB	14/04/16 14:03	✱	Microsoft Word for Wi...	2007 onwards	application/vnd.open...	fmt/412	Container	2043d04433a1fbb25...
DPCMMFiveS...	png	164.1 KB	14/04/16 16:51	✱	Portable Network Gra...	1.2	image/png	fmt/13	Signature	cfb21f37cd1dc511be...
DPCMModel....	png	100.7 KB	14/04/16 16:53	✱	Portable Network Gra...	1.2	image/png	fmt/13	Signature	d0c1cd3d9adbf09cac...

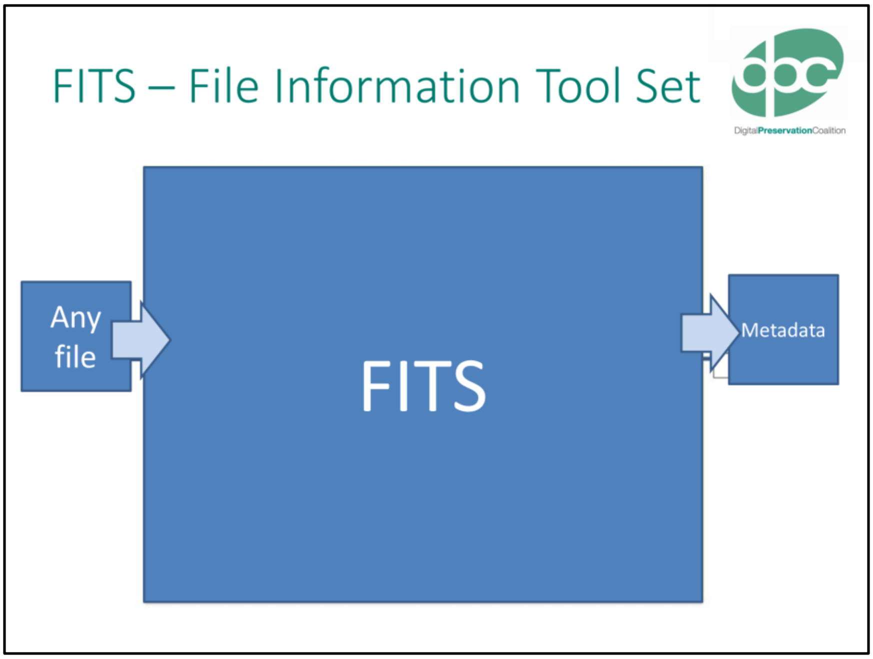
Delving Deeper: FITS

http://projects.iq.harvard.edu/fits

- Wraps together a selection of open-source tools
- Identifies, validates and extracts technical metadata
- Command line operation
- Consolidates info

  into an XML file



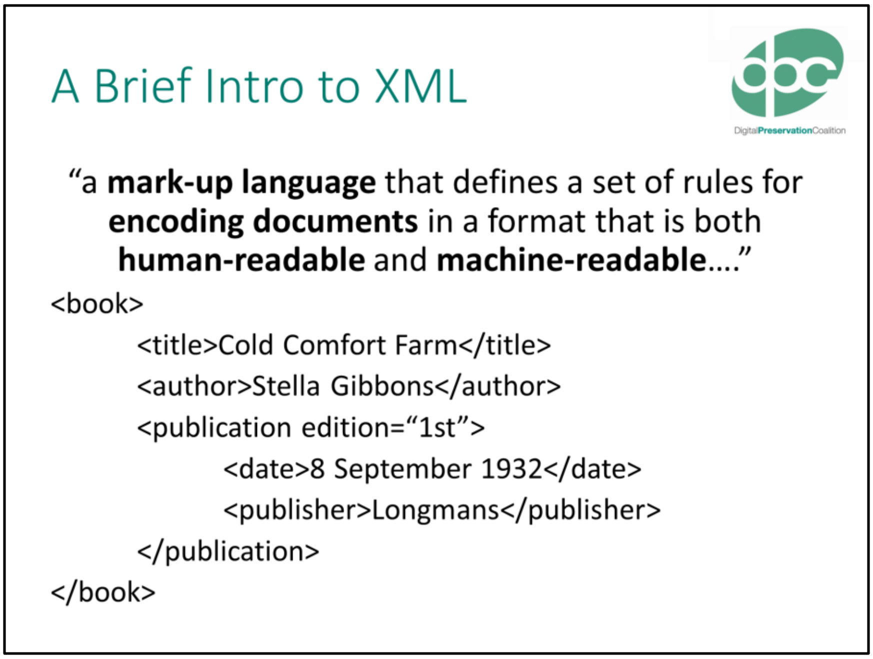


FITS – File Information Tool Set

Any file

FITS

Metadata

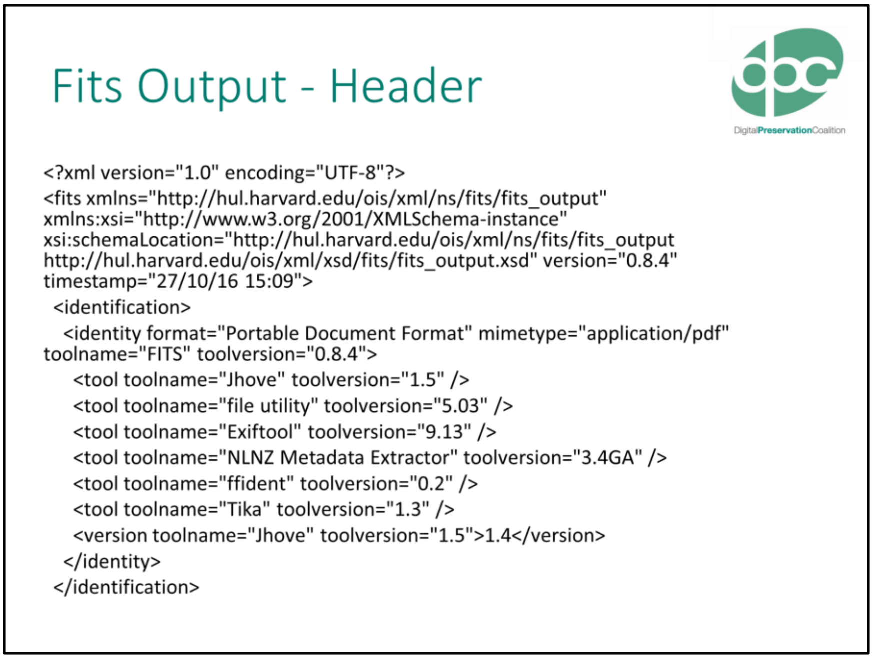


A Brief Intro to XML

“a **mark-up language** that defines a set of rules for **encoding documents** in a format that is both **human-readable** and **machine-readable**….”

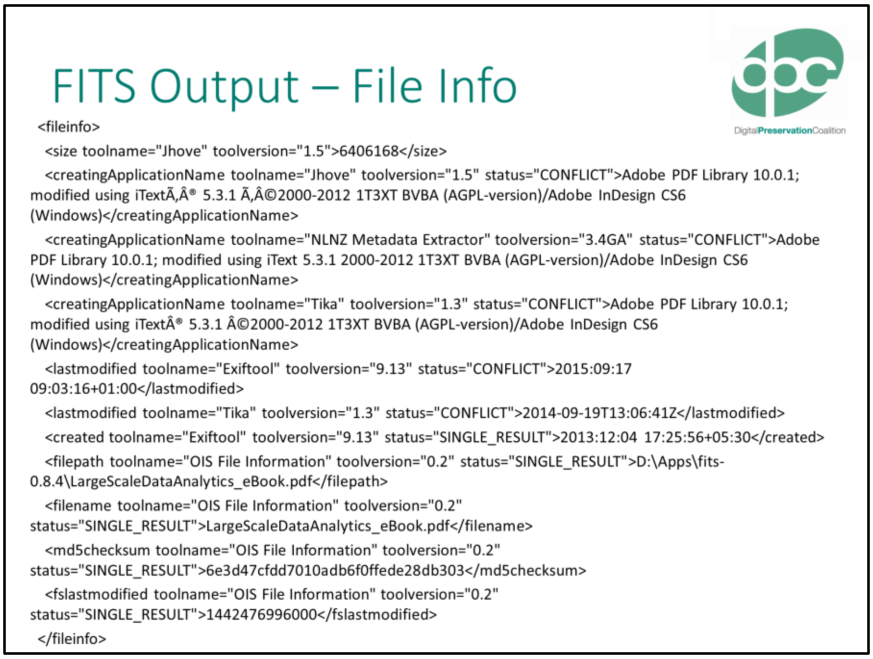
```
<book>
        <title>Cold Comfort Farm</title>
        <author>Stella Gibbons</author>
        <publication edition=“1st”>
                <date>8 September 1932</date>
                <publisher>Longmans</publisher>
        </publication>
</book>
```

Attributes



Fits Output - Header

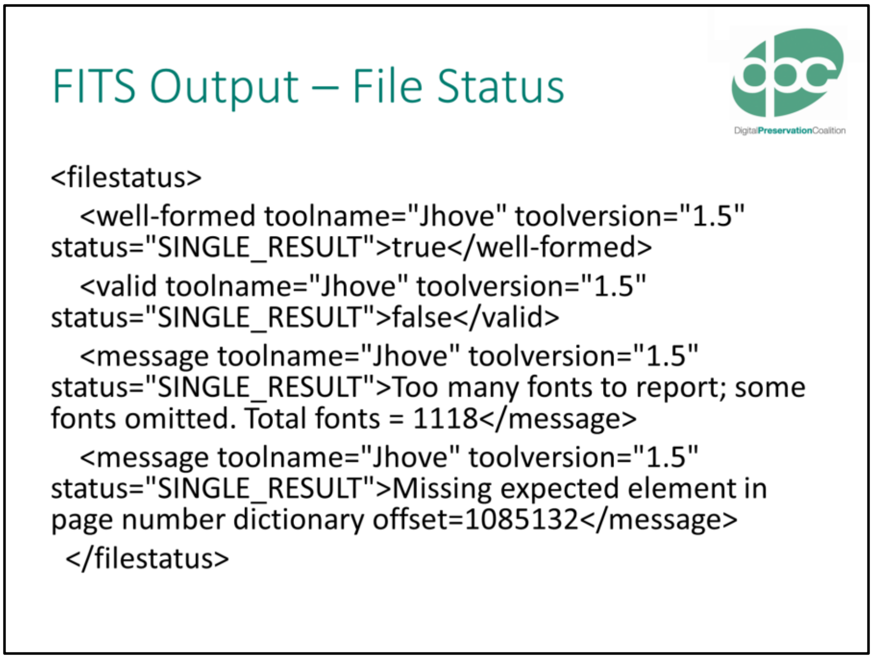
```
<?xml version="1.0" encoding="UTF-8"?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/fits/fits_output
http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd" version="0.8.4"
timestamp="27/10/16 15:09">
 <identification>
  <identity format="Portable Document Format" mimetype="application/pdf"
toolname="FITS" toolversion="0.8.4">
   <tool toolname="Jhove" toolversion="1.5" />
   <tool toolname="file utility" toolversion="5.03" />
   <tool toolname="Exiftool" toolversion="9.13" />
   <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />
   <tool toolname="ffident" toolversion="0.2" />
   <tool toolname="Tika" toolversion="1.3" />
   <version toolname="Jhove" toolversion="1.5">1.4</version>
  </identity>
 </identification>
```



FITS Output – File Info

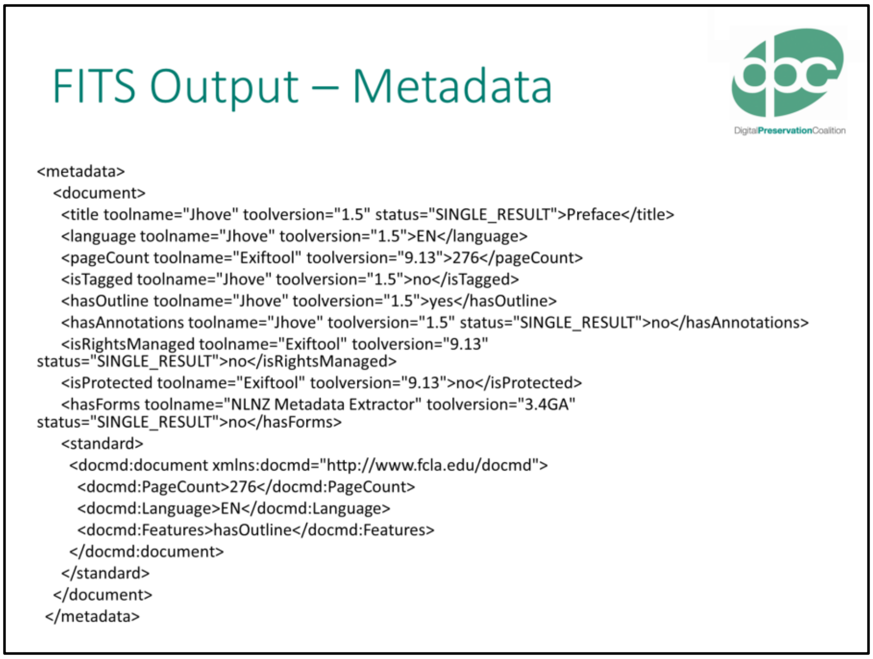
```
<fileinfo>
  <size toolname="Jhove" toolversion="1.5">6406168</size>
  <creatingApplicationName toolname="Jhove" toolversion="1.5" status="CONFLICT">Adobe PDF Library 10.0.1;
modified using iTextÃ,Â® 5.3.1 Ã,Â©2000-2012 1T3XT BVBA (AGPL-version)/Adobe InDesign CS6
(Windows)</creatingApplicationName>
  <creatingApplicationName toolname="NLNZ Metadata Extractor" toolversion="3.4GA" status="CONFLICT">Adobe
PDF Library 10.0.1; modified using iText 5.3.1 2000-2012 1T3XT BVBA (AGPL-version)/Adobe InDesign CS6
(Windows)</creatingApplicationName>
  <creatingApplicationName toolname="Tika" toolversion="1.3" status="CONFLICT">Adobe PDF Library 10.0.1;
modified using iTextÂ® 5.3.1 Â©2000-2012 1T3XT BVBA (AGPL-version)/Adobe InDesign CS6
(Windows)</creatingApplicationName>
  <lastmodified toolname="Exiftool" toolversion="9.13" status="CONFLICT">2015:09:17
09:03:16+01:00</lastmodified>
  <lastmodified toolname="Tika" toolversion="1.3" status="CONFLICT">2014-09-19T13:06:41Z</lastmodified>
  <created toolname="Exiftool" toolversion="9.13" status="SINGLE_RESULT">2013:12:04 17:25:56+05:30</created>
  <filepath toolname="OIS File Information" toolversion="0.2" status="SINGLE_RESULT">D:\Apps\fits-
0.8.4\LargeScaleDataAnalytics_eBook.pdf</filepath>
  <filename toolname="OIS File Information" toolversion="0.2"
status="SINGLE_RESULT">LargeScaleDataAnalytics_eBook.pdf</filename>
  <md5checksum toolname="OIS File Information" toolversion="0.2"
status="SINGLE_RESULT">6e3d47cfdd7010adb6f0ffede28db303</md5checksum>
  <fslastmodified toolname="OIS File Information" toolversion="0.2"
status="SINGLE_RESULT">1442476996000</fslastmodified>
</fileinfo>
```

DigitalPreservationCoalition



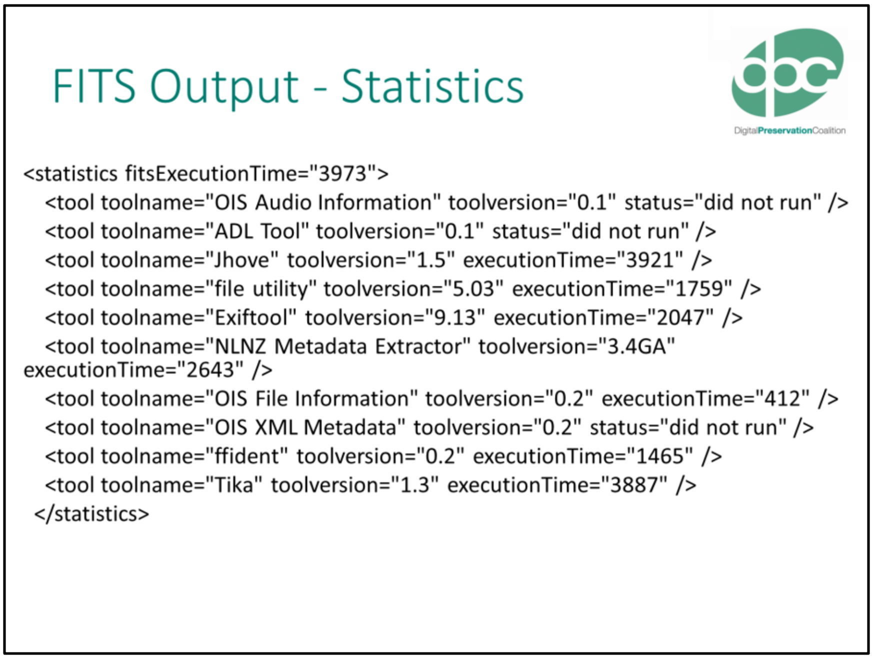
FITS Output – File Status

```
<filestatus>
   <well-formed toolname="Jhove" toolversion="1.5"
status="SINGLE_RESULT">true</well-formed>
   <valid toolname="Jhove" toolversion="1.5"
status="SINGLE_RESULT">false</valid>
   <message toolname="Jhove" toolversion="1.5"
status="SINGLE_RESULT">Too many fonts to report; some
fonts omitted. Total fonts = 1118</message>
   <message toolname="Jhove" toolversion="1.5"
status="SINGLE_RESULT">Missing expected element in
page number dictionary offset=1085132</message>
  </filestatus>
```



FITS Output – Metadata

```
<metadata>
  <document>
   <title toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">Preface</title>
   <language toolname="Jhove" toolversion="1.5">EN</language>
   <pageCount toolname="Exiftool" toolversion="9.13">276</pageCount>
   <isTagged toolname="Jhove" toolversion="1.5">no</isTagged>
   <hasOutline toolname="Jhove" toolversion="1.5">yes</hasOutline>
   <hasAnnotations toolname="Jhove" toolversion="1.5" status="SINGLE_RESULT">no</hasAnnotations>
   <isRightsManaged toolname="Exiftool" toolversion="9.13"
status="SINGLE_RESULT">no</isRightsManaged>
   <isProtected toolname="Exiftool" toolversion="9.13">no</isProtected>
   <hasForms toolname="NLNZ Metadata Extractor" toolversion="3.4GA"
status="SINGLE_RESULT">no</hasForms>
   <standard>
    <docmd:document xmlns:docmd="http://www.fcla.edu/docmd">
     <docmd:PageCount>276</docmd:PageCount>
     <docmd:Language>EN</docmd:Language>
     <docmd:Features>hasOutline</docmd:Features>
    </docmd:document>
   </standard>
  </document>
 </metadata>
```

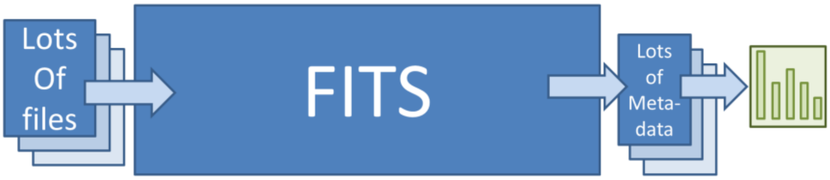


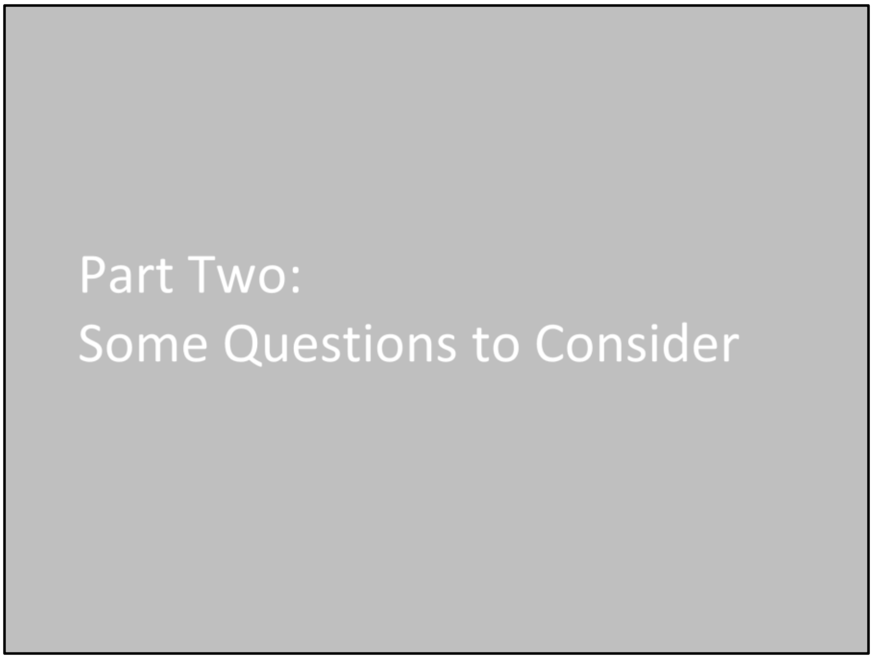
FITS Output - Statistics

```
<statistics fitsExecutionTime="3973">
  <tool toolname="OIS Audio Information" toolversion="0.1" status="did not run" />
  <tool toolname="ADL Tool" toolversion="0.1" status="did not run" />
  <tool toolname="Jhove" toolversion="1.5" executionTime="3921" />
  <tool toolname="file utility" toolversion="5.03" executionTime="1759" />
  <tool toolname="Exiftool" toolversion="9.13" executionTime="2047" />
  <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA"
executionTime="2643" />
  <tool toolname="OIS File Information" toolversion="0.2" executionTime="412" />
  <tool toolname="OIS XML Metadata" toolversion="0.2" status="did not run" />
  <tool toolname="ffident" toolversion="0.2" executionTime="1465" />
  <tool toolname="Tika" toolversion="1.3" executionTime="3887" />
 </statistics>
```

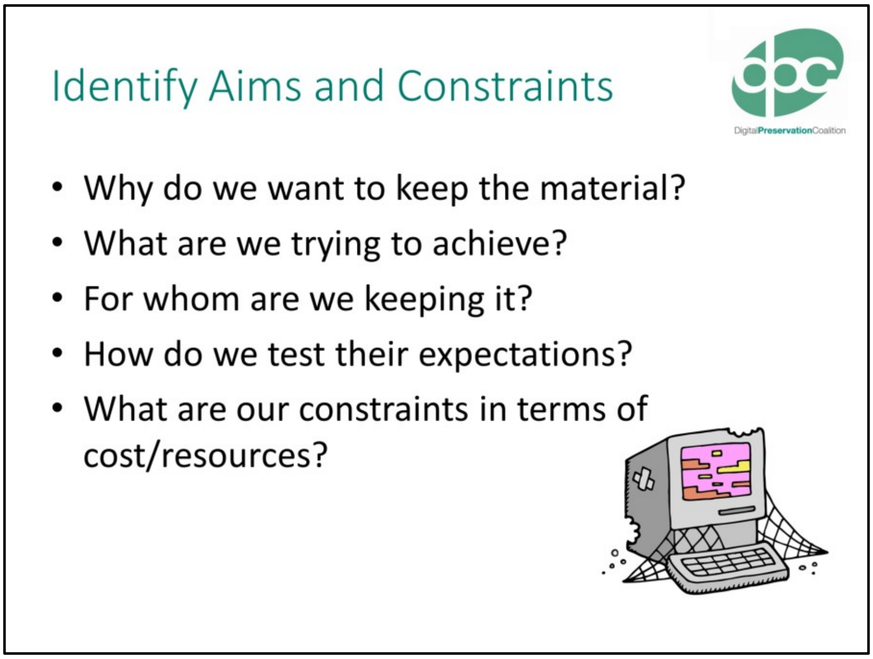
C3PO

- FITS: Content profiling tool
- Works with FITS to visualize FITS output metadata
- Setup can be a little more taxing





Part Two:
Some Questions to Consider



Identify Aims and Constraints

- Why do we want to keep the material?
- What are we trying to achieve?
- For whom are we keeping it?
- How do we test their expectations?
- What are our constraints in terms of cost/resources?

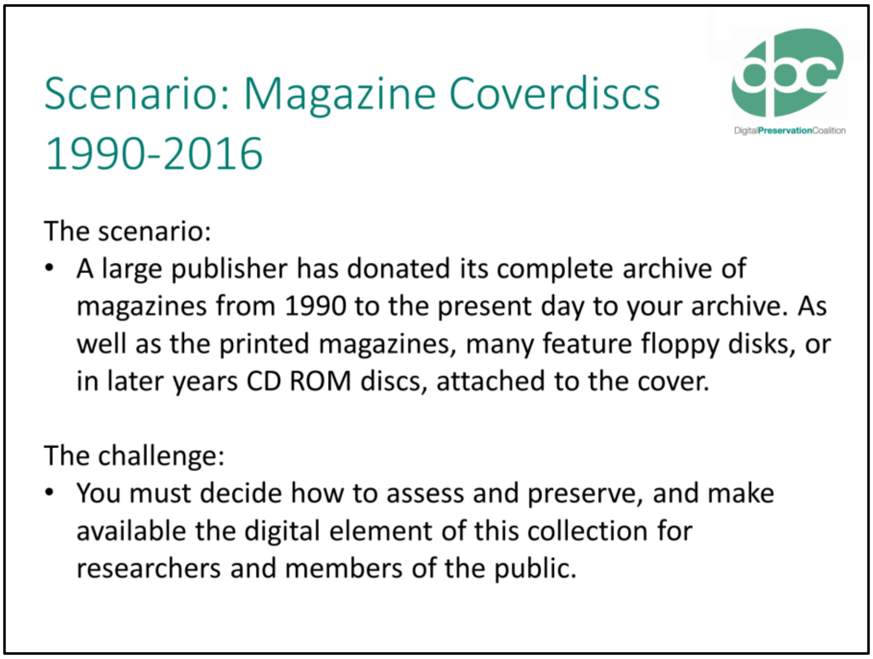
Why do we want to keep this stuff?
- Collecting/Acquisition Policy
- Organisational Goals
- User Needs/Appetite
- Value – historical, financial etc.

What are we trying to achieve?
- Create a record of an event/organisation/individual
- Provide access
- Meet legal requirements

For whom are we keeping it?
- Users now or in the future
- What skills will they have – cognitive and technical

How do we test their expectations?

What are our constraints in terms of cost/resources?



Scenario: Magazine Coverdiscs 1990-2016

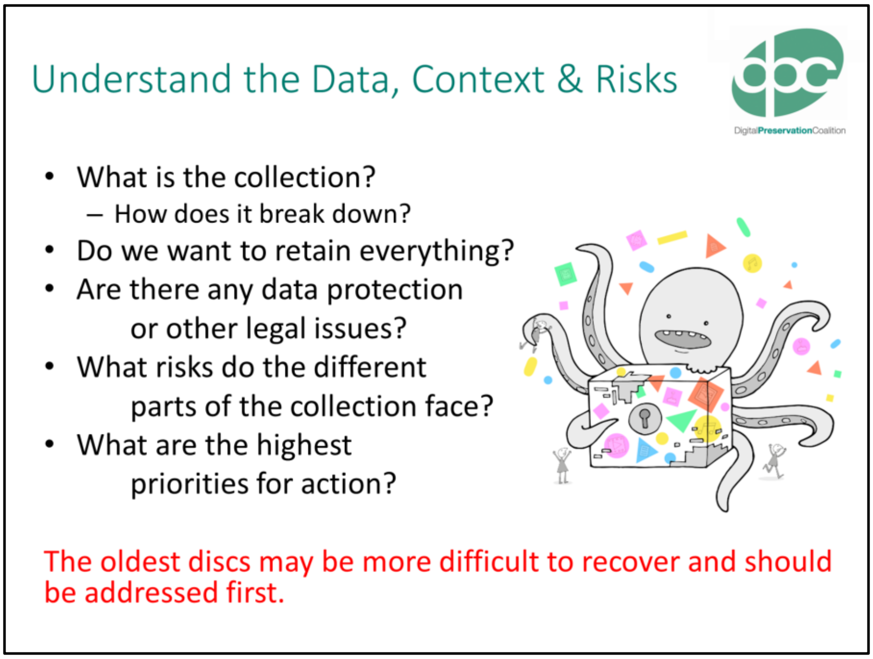
The scenario:

- A large publisher has donated its complete archive of magazines from 1990 to the present day to your archive. As well as the printed magazines, many feature floppy disks, or in later years CD ROM discs, attached to the cover.

The challenge:

- You must decide how to assess and preserve, and make available the digital element of this collection for researchers and members of the public.

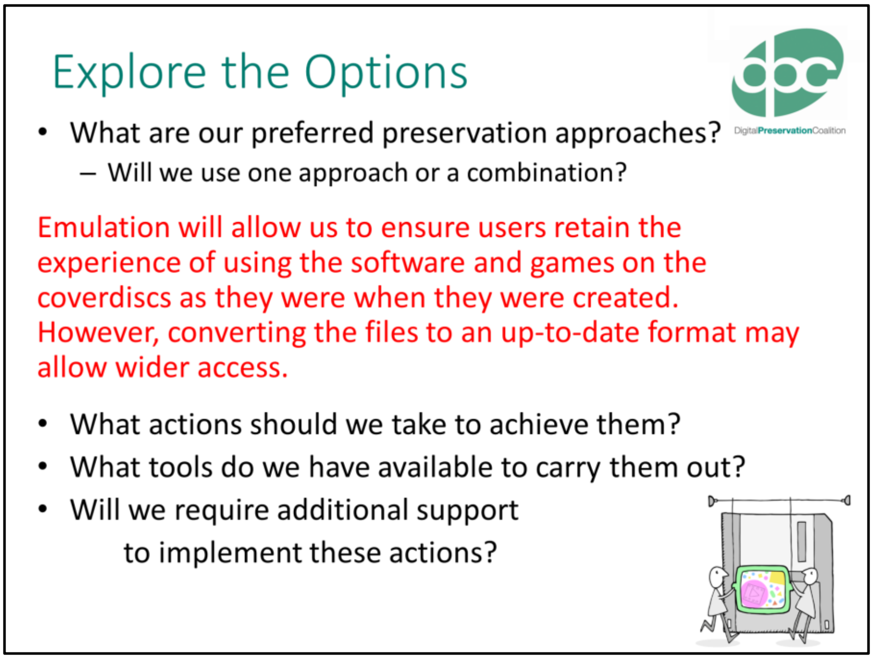
The coverdiscs run on a number of different computers and operating systems that were popular throughout the life of the magazine.
The coverdiscs are still wrapped up in plastic and have never been used.



Understand the Data, Context & Risks

- What is the collection?
  - How does it break down?
- Do we want to retain everything?
- Are there any data protection or other legal issues?
- What risks do the different parts of the collection face?
- What are the highest priorities for action?

The oldest discs may be more difficult to recover and should be addressed first.

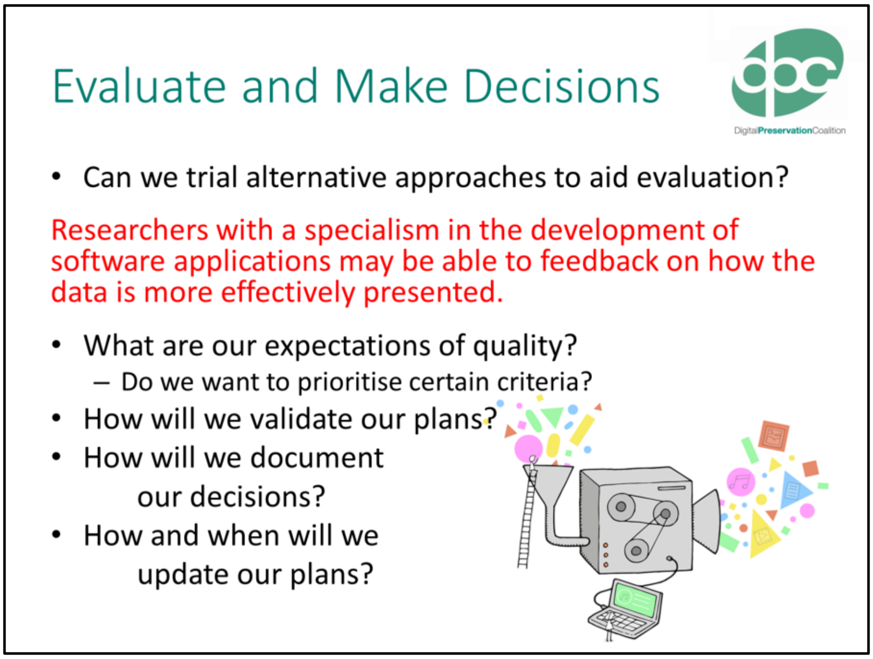


Explore the Options

- What are our preferred preservation approaches?
  - Will we use one approach or a combination?

Emulation will allow us to ensure users retain the experience of using the software and games on the coverdiscs as they were when they were created. However, converting the files to an up-to-date format may allow wider access.

- What actions should we take to achieve them?
- What tools do we have available to carry them out?
- Will we require additional support to implement these actions?

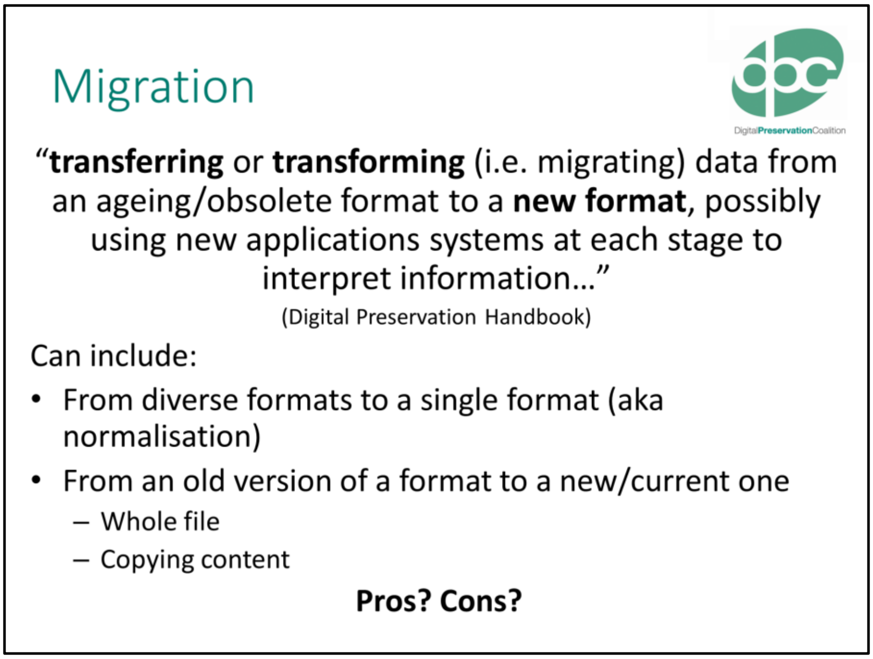


Evaluate and Make Decisions

- Can we trial alternative approaches to aid evaluation?

Researchers with a specialism in the development of software applications may be able to feedback on how the data is more effectively presented.

- What are our expectations of quality?
  - Do we want to prioritise certain criteria?
- How will we validate our plans?
- How will we document our decisions?
- How and when will we update our plans?



Migration

"**transferring** or **transforming** (i.e. migrating) data from an ageing/obsolete format to a **new format**, possibly using new applications systems at each stage to interpret information…"

(Digital Preservation Handbook)

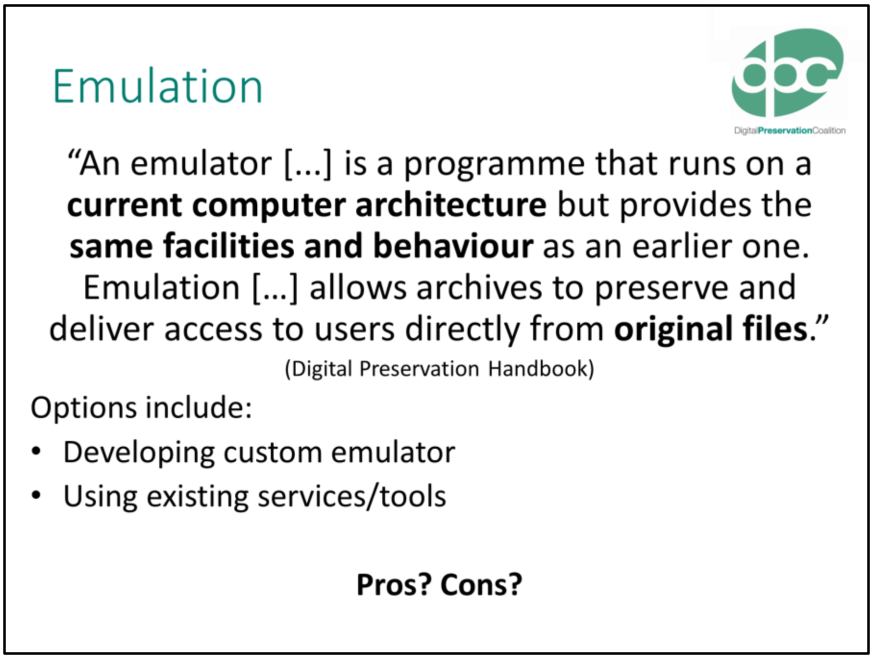
Can include:

- From diverse formats to a single format (aka normalisation)
- From an old version of a format to a new/current one
  - Whole file
  - Copying content

Pros? Cons?

Group Discussion

Pros
Create homogenous collections
Access through current software
Potential for automation of process
Cons
Possible loss of functionality
Possible loss of content
Create errors



Emulation

"An emulator [...] is a programme that runs on a **current computer architecture** but provides the **same facilities and behaviour** as an earlier one. Emulation […] allows archives to preserve and deliver access to users directly from **original files**."

(Digital Preservation Handbook)

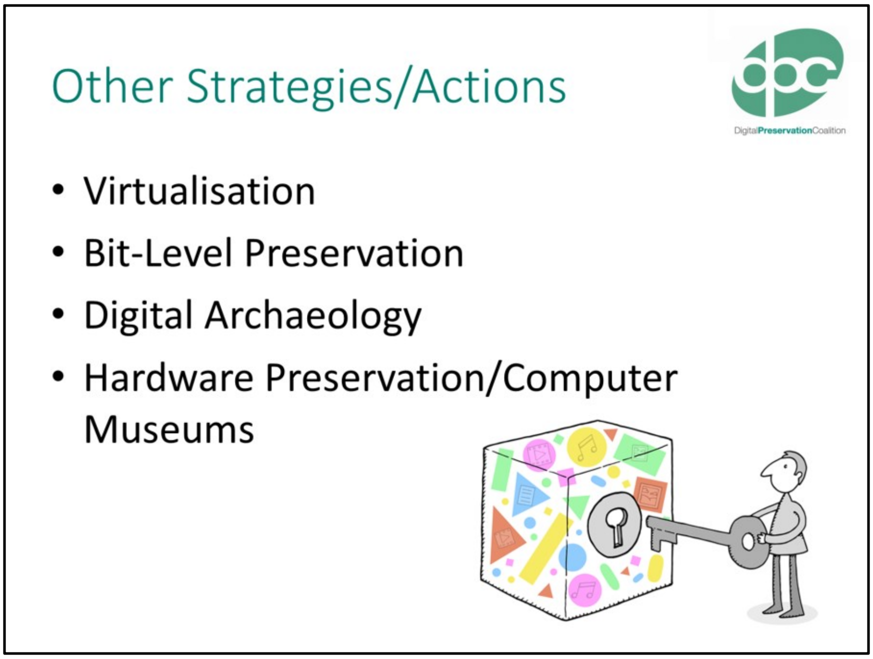
Options include:

- Developing custom emulator
- Using existing services/tools

Pros? Cons?

Group Discussion

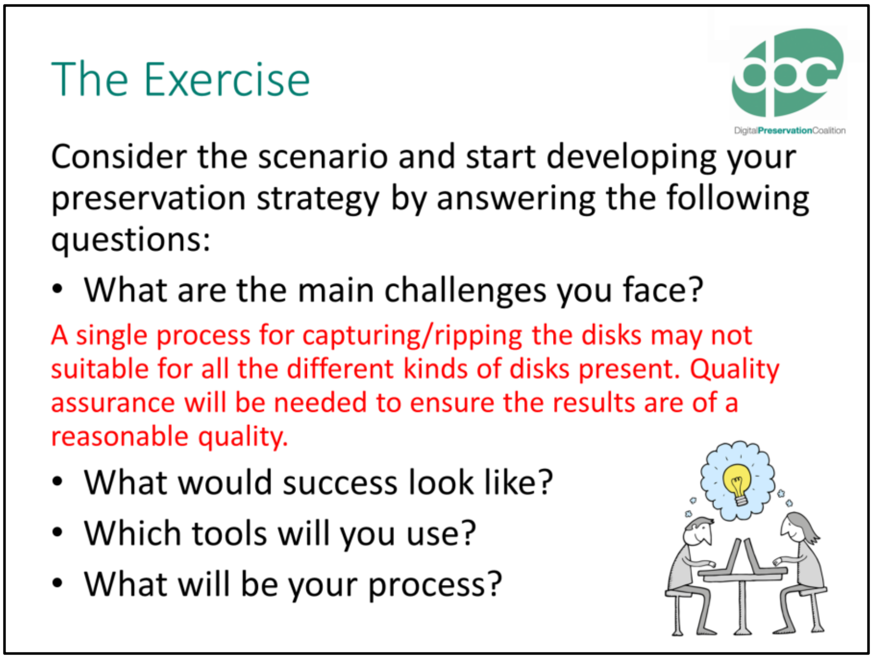
Pros
File remains unchanged
File accessed in original environment
Authentic experience for user
Cons
Significant programming input required and/or
Reliance on 3rd party tools
Need skills to work original software



Other Strategies/Actions

DigitalPreservationCoalition

- Virtualisation
- Bit-Level Preservation
- Digital Archaeology
- Hardware Preservation/Computer Museums



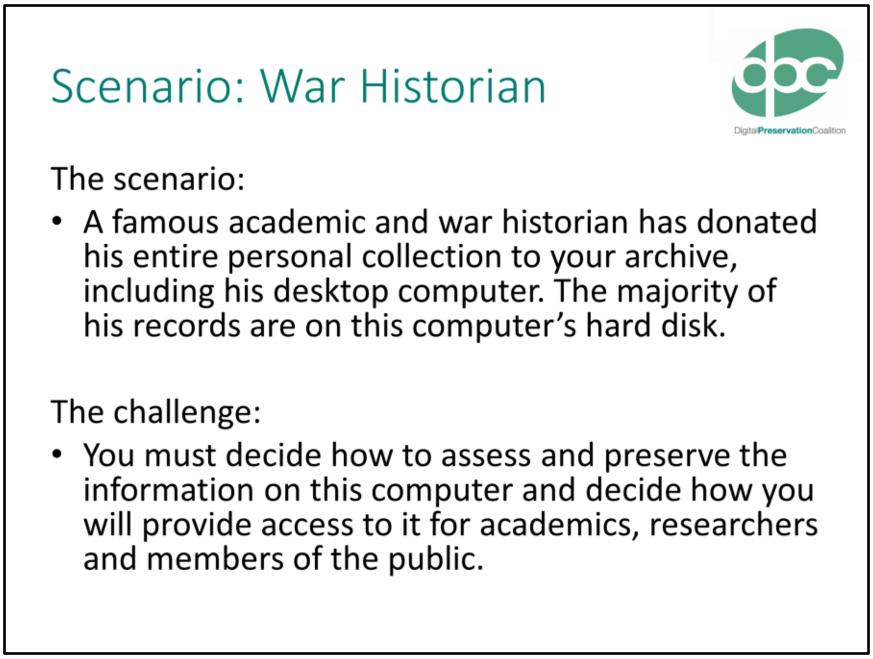
The Exercise

Consider the scenario and start developing your preservation strategy by answering the following questions:

- What are the main challenges you face?

A single process for capturing/ripping the disks may not suitable for all the different kinds of disks present. Quality assurance will be needed to ensure the results are of a reasonable quality.

- What would success look like?
- Which tools will you use?
- What will be your process?



Scenario: War Historian

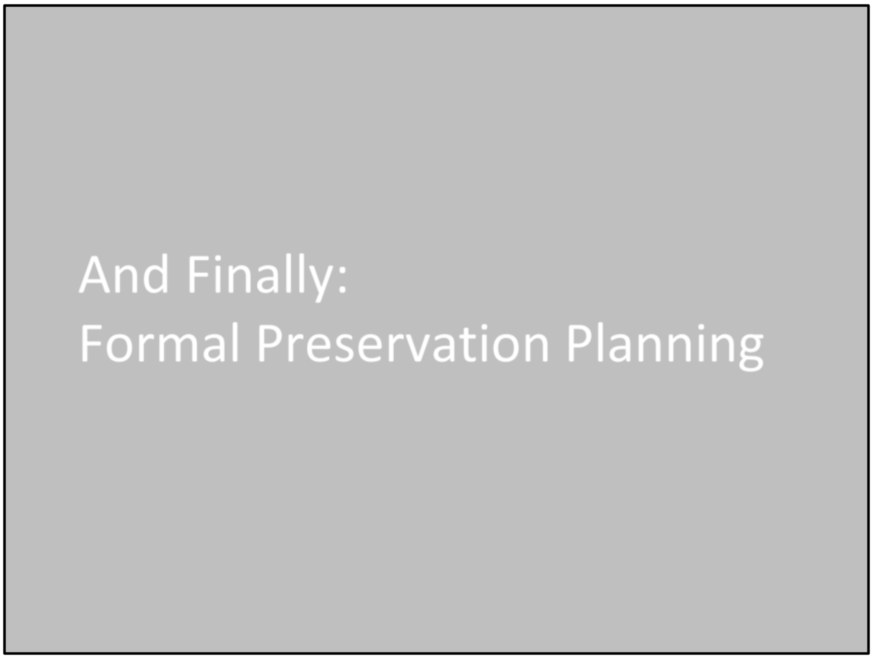
DigitalPreservationCoalition

The scenario:

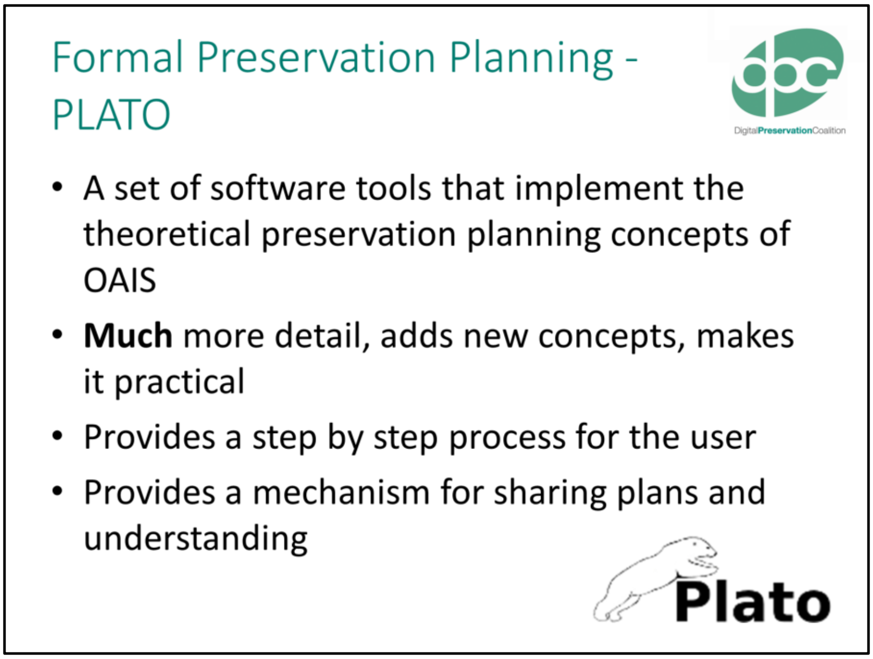
- A famous academic and war historian has donated his entire personal collection to your archive, including his desktop computer. The majority of his records are on this computer’s hard disk.

The challenge:

- You must decide how to assess and preserve the information on this computer and decide how you will provide access to it for academics, researchers and members of the public.



And Finally:
Formal Preservation Planning

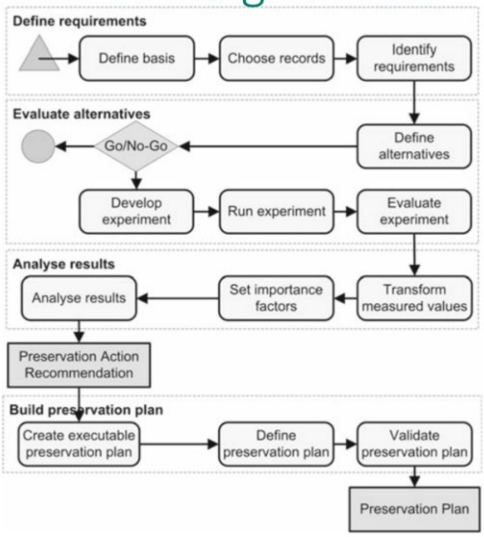


Formal Preservation Planning - PLATO

- A set of software tools that implement the theoretical preservation planning concepts of OAIS
- **Much** more detail, adds new concepts, makes it practical
- Provides a step by step process for the user
- Provides a mechanism for sharing plans and understanding

dpc
DigitalPreservationCoalition
Plato

PLATO Planning Process



Define requirements

Define basis

Choose records

Identify requirements

Evaluate alternatives

Go/No-Go

Define alternatives

Develop experiment

Run experiment

Evaluate experiment

Analyse results

Analyse results

Set importance factors

Transform measured values

Preservation Action Recommendation

Build preservation plan

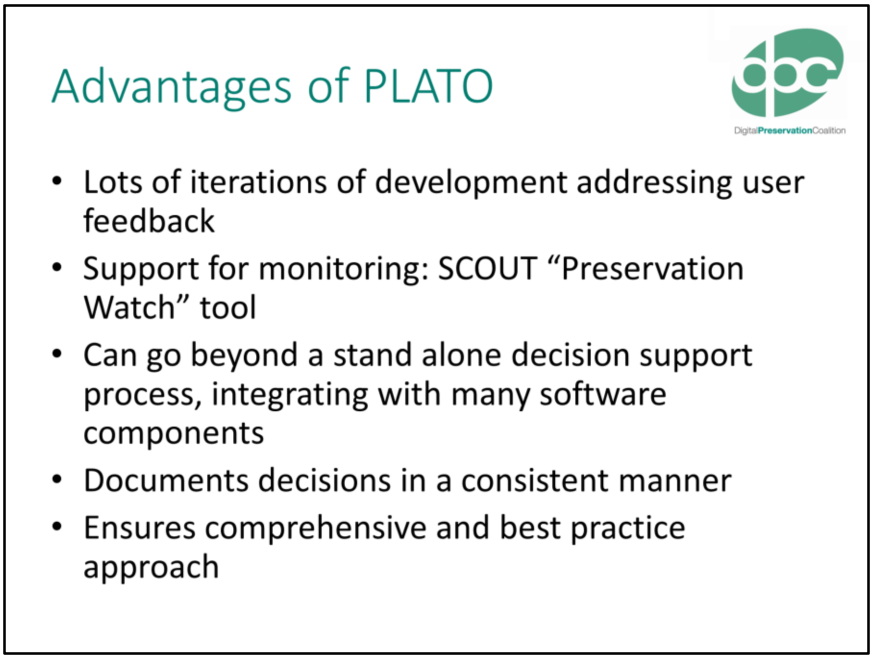
Create executable preservation plan

Define preservation plan

Validate preservation plan

Preservation Plan

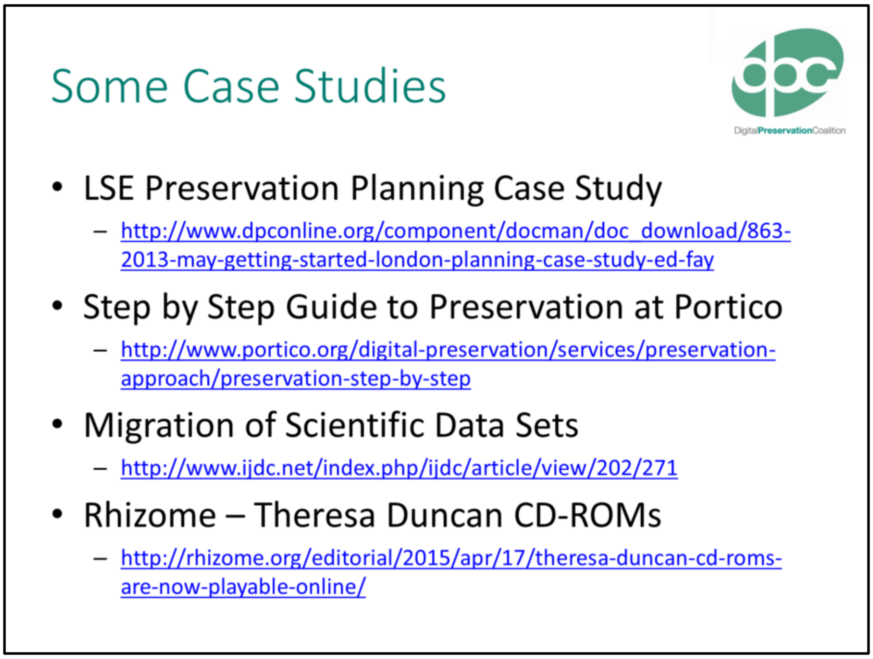
Diagram taken from: "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans“



Advantages of PLATO

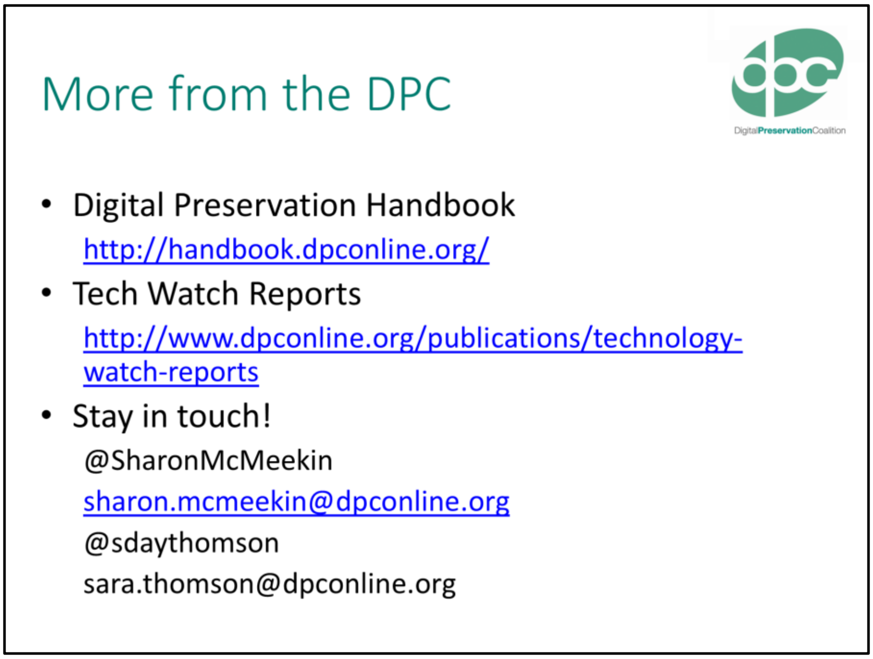
- Lots of iterations of development addressing user feedback
- Support for monitoring: SCOUT “Preservation Watch” tool
- Can go beyond a stand alone decision support process, integrating with many software components
- Documents decisions in a consistent manner
- Ensures comprehensive and best practice approach

Some negatives:
Appears intimidating, but the bar to entry is actually quite low
Some of the newer automated components of the toolset are unproven



Some Case Studies

- LSE Preservation Planning Case Study
  - http://www.dpconline.org/component/docman/doc_download/863-2013-may-getting-started-london-planning-case-study-ed-fay
- Step by Step Guide to Preservation at Portico
  - http://www.portico.org/digital-preservation/services/preservation-approach/preservation-step-by-step
- Migration of Scientific Data Sets
  - http://www.ijdc.net/index.php/ijdc/article/view/202/271
- Rhizome – Theresa Duncan CD-ROMs
  - http://rhizome.org/editorial/2015/apr/17/theresa-duncan-cd-roms-are-now-playable-online/



More from the DPC

- Digital Preservation Handbook
  http://handbook.dpconline.org/
- Tech Watch Reports
  http://www.dpconline.org/publications/technology-watch-reports
- Stay in touch!
  @SharonMcMeekin
  sharon.mcmeekin@dpconline.org
  @sdaythomson
  sara.thomson@dpconline.org