



# **The National Archives**

[www.nationalarchives.gov.uk](http://www.nationalarchives.gov.uk)



# File formats for digital preservation: criteria & thinking

Malcolm Todd, Archives Sector Development  
Collaboration with: [www.dpconline.org](http://www.dpconline.org)



- Forthcoming report commissioned by Digital Preservation Coalition
- Major theoretical & practical issue
- OAIS issue “content information”
- Research community, including an archival science viewpoint
- Conclusions



- 
- Review literature explicitly “about” file formats
  - Look at leading edge research and its consequences for the issue, special need to look at linkage to other DP issues
  - Evaluate whether anything like consensus exists
  - Make recommendations for repository managers and the DP community



- 
- Published papers on file format selection, including one survey
  - Internal TNA reports, including consultancy
  - Recent archival science literature
  - Relevant resources on linked information, especially Representation Information, e.g. PREMIS, PRONOM, GDFR



# Major file format literature

---

- One resource [McLellan] is a synthesis of 20[+] repository deposit guidelines and research / national and other major libraries' and archives' recommendations
- Harvard [Abrams: DCC Manual, GDFR papers]
- National Library of the Netherlands [Rog and van Wijk]
- Library of Congress [Arms and Fleischhauer]
- OCLC [Stanescu]
- Danish National Library [Netarkivet]
- Adrian Brown, TNA [PRONOM]
- French 'PIN' group [Huc et al.; web archiving focus]
- OAIS Reference Model



# Broad consensus on key criteria

The National Archives

	<i>Adoption</i>	<i>Platform independence</i>		<i>Core criteria Disclosure</i>		<i>Transparency</i>	<i>Metadata support</i>	
		<i>Support</i>	<i>Interoperability</i>	<i>Disclosure</i>	<i>Documentation quality</i>			
<b>Brown TNA, UK (2008 No.1)</b>	<i>Ubiquity</i>	<i>Support</i>	<i>Interoperability</i>	<i>Disclosure</i>	<i>Documentation quality</i>	<i>Ease of identification and validation</i>	<i>Metadata support</i>	
<b>Arms &amp; Fleischhauer LoC, USA (2005)</b>	<i>Adoption</i>	<i>External dependencies</i>		<i>Disclosure</i>	<i>Impact of patents</i>	<i>Transparency, incl.human readability;lack of encryption; natural reading order of textual files' content; standardisation of source code</i>	<i>Self documentation</i>	
<b>Rog &amp; van Wijk KB, NL (2008)</b>	<i>Adoption</i>	<i>Dependencies</i>		<i>Openness</i>		<i>Complexity</i>	<i>Self-documentation</i>	
<b>McLellan InterPARES2, CAN (2007)</b>	<i>Widespread use</i>	<i>Platform independence</i>		<i>Non-proprietary origin</i>	<i>Availability of documentation</i>	<i>Compression</i>	<i>-</i>	
<b>Christensen Nerarchivet, DK (2004)</b>	<i>-</i>	<i>Dependencies</i>		<i>-</i>		<i>-</i>	<i>Metadata support</i>	<i>Support for authenticity information</i>
<b>Huc et al PIN group .v.5 [FR]</b>	<i>-</i>	<i>-</i>		<i>Public standardisation</i>		<i>Inspectability</i>	<i>Extractability of metadata</i>	
<b>*Stanescu OCLC (2005)</b>	<i>Adoption</i>	<i>-</i>		<i>Disclosure</i>	<i>Documentation quality</i>	<i>-</i>	<i>Metadata support</i>	



# Even some wider agreement

	<b>Wider criteria</b>				
	<b>IP / DRM</b>	<b>Stability /backward compatibility</b>	<b>Robustness /Complexity / Viability</b>		<b>Re-usability</b>
<b>Brown TNA, UK (2008)</b>	<i>IPR</i>	<i>Stability / backward compatibility</i>	<i>Complexity</i>	<i>Viability</i>	<i>Re-usability</i>
<b>Arms &amp; Fleischhauer LoC, USA (2005)</b>	-	-	-	-	-
<b>Rog &amp; van Wijk KB, NL (2008)</b>	<i>Technical protection mechanism</i>	<i>Robustness</i>			
<b>McLellan InterPARES2, CAN (2007)</b>	-	-	-	-	-
<b>Christensen Nerarchivet, DK (2004)</b>	-	-	<i>Robustness</i>		
<b>Huc et al PIN group .v.5 [FR]</b>	-	-	<i>Simplicity</i>		<i>Manipulability</i>
<b>*Stanescu OCLC (2005)</b>	<i>DRM, signature, encryption facilities</i>	<i>Stability / backward compatibility</i>	-		<i>(as regards metadata interoperability)</i>



- Oddities: DK 'authenticity' criterion
- Much of literature is digital library focused, some is explicitly aimed at web harvesting
- Broad mappings conceal many differences of detail and emphasis
- .....even some major contradictions



# Irreconcilable examples

---

- Is 'complexity' a good thing in a file format? [richness of content vs. inherent level of dependencies]. Is it measurable?
- Digital Rights Management
- Compression [Verbosity]
- Standard formats [What?]



- Scoring methods
- Groupings, scales & weightings
- Complex and essentially arbitrary arithmetic
- Reorients discussions of familiar issues, e.g. standardisation



*“....as the weighing of these criteria is connected to an institution’s policy, the KB wonders whether agreement on the relative importance of the criteria can be reached at all .....  
**the examples in [their] paper are the weights as assigned by the KB based on its local policy, general digital preservation literature and common sense.....”***

<b>Robustness</b>	
Format should be robust against single point of failure (2)	
	2 Not vulnerable
	1 Vulnerable
	0 Highly vulnerable
Support for file corruption detection (2)	
	2 Available
	0 Not available
File format stability (2)	
	2 Rare release of new versions
	1 Limited release of new versions
	0 Frequent release of new versions
Backward compatibility (2)	
	2 Large support
	1 Medium support
	0 No support
Forward compatibility (2)	
	2 Large support
	1 Medium support
	0 No support

# A Standardisation [different levels]

The National Archives

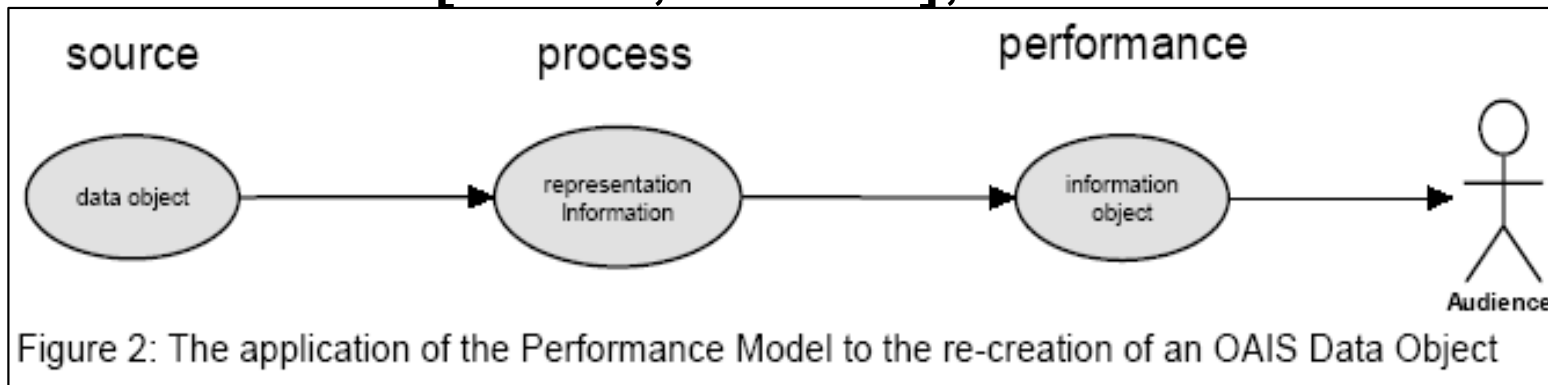
Type / level of 'standardisation'	Arbitrated / handled by...	Displayed by	Fall back	Issues [Criterion]
<ul style="list-style-type: none"> <li>Textual encoding</li> </ul>	Varies	Viewers.....	-	Rarely proprietary, but exceptions.....
<ul style="list-style-type: none"> <li>Mark-up languages [HTML, XML, XHTML...]</li> </ul>	Web standards, browser	Style sheets / DTD	None [human reading?]	Community has a choice: treat these as "formats" and remember to highlight additional levels of requirements to maintain OAIS representation information or exclude from definition
<ul style="list-style-type: none"> <li>Format specifications</li> </ul>	Operating system	<i>Compatible</i> application software	<i>Compatible</i> viewers	Platform independence
<ul style="list-style-type: none"> <li>....</li> </ul>				



# Standardisation [reoriented]

<b>Criterion</b>	<b>Associated with...</b>	<b>Business model</b>	<b>Comments</b>
<b>Adoption</b>	Mass-market proprietary software	Proprietary software; IPR in code	See <i>Disclosure</i> ; Wide adoption <i>ought</i> to mean support of tools owing to market forces
<b>Platform independence</b>	Interoperability	<i>Sometimes</i> a relationship to <i>disclosure</i> [through de jure standardisation]	-
<b>Disclosure</b>	“Open standards”	de jure standardisation activity	Quality of documentation may be an issue
<b>Transparency</b>	Lack of compression, encryption, reading order of textual files	?DRM requirements	-
<b>Metadata support</b>	Extraction and registry tools	-	[Most sources focus on representation metadata]

- ISO15489
- Significant properties [InSPECT Project]
  - ‘Performance’ of record as it applies to relatively simple documents [image, text, email]
  - Australian [Xena, VERS], InterPARES1





- InterPARES2 case study environments: dynamic, interactive and experiential
- InterPARES2 conceptual analyses of concepts of authenticity in arts, sciences and government
  - Irrespective of “recordness” of objects
  - Problems of fixed content based on point in time and business activity / transaction
  - Protocols / interactions, appearance [*“look and feel”*]
  - ... if a static grab not a viable means.....



- Archival / library usage, e.g. authenticity, provenance
- Library of Congress does not impose a finite specification as an essential characteristic of a file format
- A lot of indistinct discussion of markup languages
- OAIS terms and concepts
- Meaning / desirability of file format criteria
- A mess!



- Glossary has 20 terms where the meaning is emerging, domain-sensitive or contested
- OAIS concepts are sound
- OAIS terms are not completely consistent, nor always helpful



# OAIS local view

*“Information object  
.....digital object”*

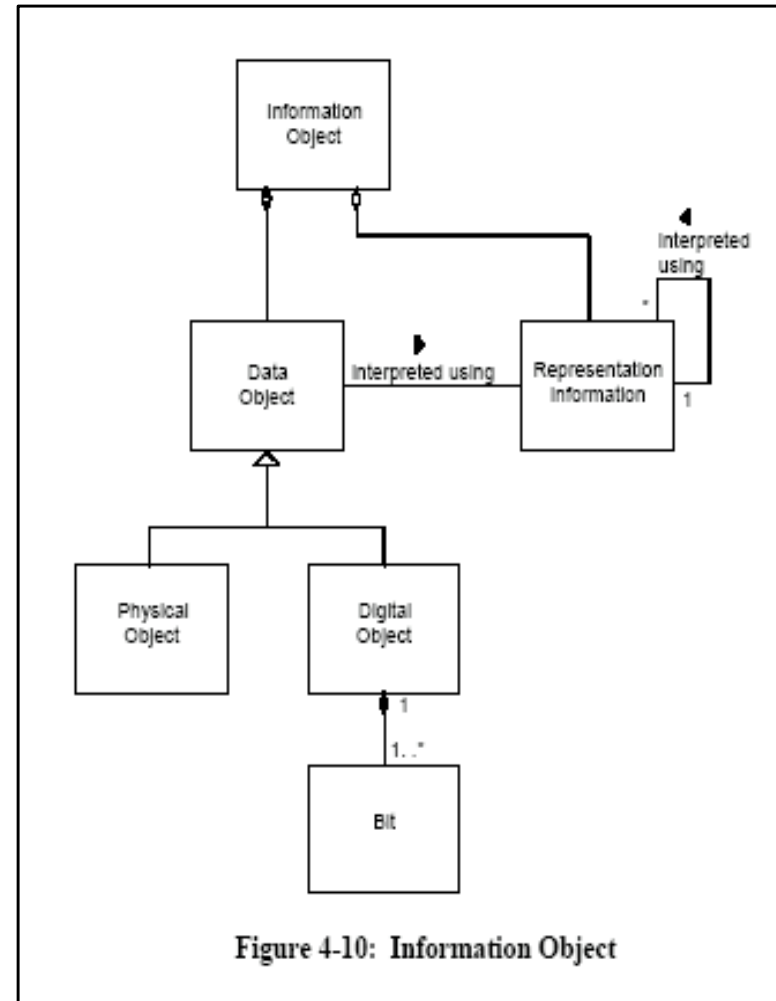
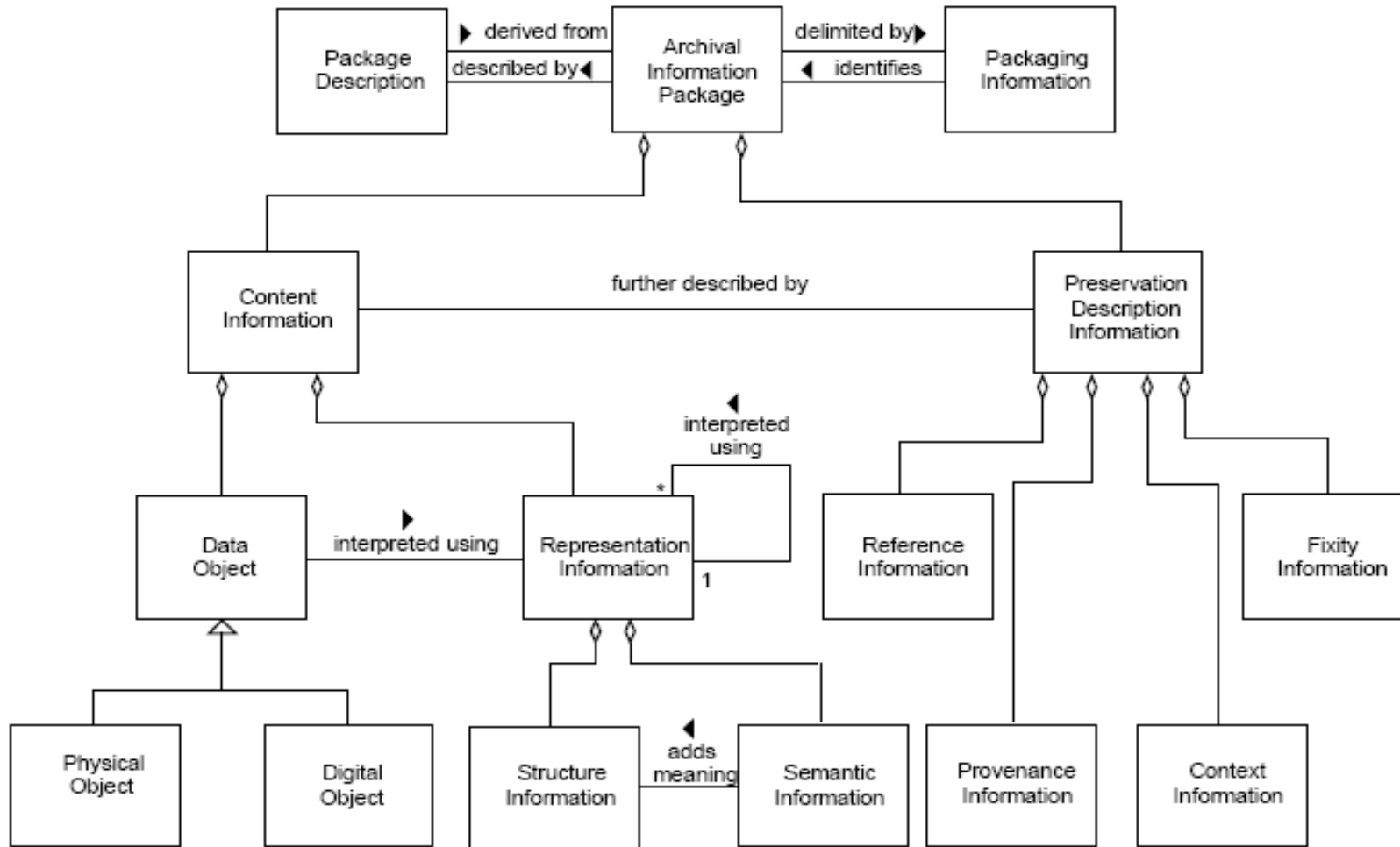


Figure 4-10: Information Object



# OAIS Archival Information Package view



**Figure 4-18: Archival Information Package (Detailed View)**



- There is a broad consensus on the main issues but differences of detail
- Agree with Rog & van Wijk that a clear preservation strategy is only way for a preserving institution to resolve its selection of file formats based on what is important to its mission and its collection
- Risk-based method, regular review
- A distinct archival viewpoint seems to be emerging but has a lot of ground to make up
- A distinct archival viewpoint ought to have mappings to concerns of other professionals, but a lot of explanation and translation is required
- The terminology is a mess



- E. McLellan [www.interpares.org](http://www.interpares.org)
- S. Abrams [DCC Manual [www.dcc.ac.uk](http://www.dcc.ac.uk), GDFR papers]
- Rog and van Wijk [Netherlands National Library] [www.kb.nl](http://www.kb.nl)
- Arms and Fleischhauer [Library of Congress] <http://memory.loc.gov>
- A. Stanescu [OCLC] [www.dlib.org](http://www.dlib.org)
- S. Christensen [Danish National Library, Netarkivet] [www.netarchive.dk](http://www.netarchive.dk)
- A. Brown, TNA [PRONOM] [www.nationalarchives.gov.uk](http://www.nationalarchives.gov.uk); [www.planets-project.eu](http://www.planets-project.eu)
- Huc et al. [French 'PIN' group; web archiving focus] [www.ssd.rl.ac.uk](http://www.ssd.rl.ac.uk)
- InSPECT Project [www.significantproperties.org.uk](http://www.significantproperties.org.uk)
- Duranti & Thibodeau, Archival Science 6 [1], 2006
- OAIS Reference Model



# **The National Archives**

[www.nationalarchives.gov.uk](http://www.nationalarchives.gov.uk)