



EPrints Preservation



What will you know after this tutorial?

- Understand the challenges in digital Preservation
- Understand why we need to plan preservation activities
- Be able to choose a simplistic preservation plan strategy.
- Be able to deploy this plan on your repository and control the outputs.





(I) Introduction

(2) File Classification & Risk Analysis

(3) Simple Preservation Planning

(4) Preservation Action & Provenance







EPrints Preservation

Introduction





() file:///media/disk/070704 dp tuwien.ppt - KHexEdit

<u>File Edit View Documents Bookmarks Tools Settings Help</u> 12 17 16 🗐 🕲 🕹 🥱 ĉ 🖰 67 67 19 🔍 🗭 🏓 🔵 🖽 ÐĪ.ài±.á....>...þÿ.w...e2. 00.008 00:00 -2. . 2. . /2. . 62. . 12. . 22. . 32. . 42. . 52. . 62. . 72. . 82. . 92. . : : 2. . : 2. : : 2. . : 2. . : 2. . : 2. . : 2. . : 2. : : 2. . : 2. . 00.010 00:0140 000:0180 M2. N2. 02. P2. 02. R2. S2. T2. U2. V2. W2. X2. Y2. Z2. [2. \2 000:01C0]2..^2. 2..2.a2.b2.c2.d2.g2..2..2..3.v4.′4.µ4.ú5. 000:0200 F.ô.q..Û*¢. 'M⁻ÛX3.£;ÿØÿà.JFIF...``.ÿÛ.C.... 888-8246ÿÀ....¶ 999-928 00:020 999-939 900:0340 Qa.! 1ARS...006." 2Bg±N.#3Cbi³áñ.Tcrs.\$.ÅÅ4.¢2...£.%V.&5EFUeâD.ÅÖdÿÜ.......?.0. 00.000 999-9369 999-9449 00.048 ðÓĪÔ¿¦¦ð»î.¢¢ð®É.¢¦ª«©dÉà.ô¢Ç¹..39V..L@ÌÌF..Ìþ÷5É|zMé.ÓèåÉ-Ñyß· -Us]tU.¹⁻¹/J.z^{**}xc. ú§-pm0Xf.2...m@...Éi...Ao[A800..do'q.|«kô0T ÝDÆ:..-Y.ÿ.ht.ÄWSOE0{*.0êf..Y ē{B°Å#sìN²å.k.òä2.ÿ.sn..CHr10.ÞBÿ. BB-85c 00.004 00:0686 00.070 00:0740 00.028 00:07c0 00.0840 00.000 00-08c 000-000 00:0940 00-0080 00-00-0 D/Popundka a -Gouo. fattu polm. jd ∈ Jryv. pol. (10.6 zerov krijito @bēc - filmoju j. 0. GYAD (16.6 x) 13. a · fey ... 1. 0. 1. 0. 1. A sll @bēc A filmoju j. 0. GYAD (16.6 x) 13. a · fey ... 1. 0. 1. 0. 1. A sll @bēc A filmoju j. 0. GYAD (16.7 x) 14. a · fey a · fey a filmoju j. Jvali @bec A filmoju ... 7 x Ytlēt (17.6 x) 4. a · fey a 00-0-40 00:0a80 .DID0._B..e.au2TêK81«51]ªê6p.100ā0K_EUkéo§;Û´÷*'~.ت.é«f·Y}·kūů adýšeji×i=P1.@eiŶÜY×...°þå=50ÖhhE...×Y[b]D.0=uþý.10ÿ.cZ^e.4£0. .8s.ci.%t\ÅÄ.3...0.±|.üD¶Ü.r#yÜ.«élT.,.£Åtæ′|<R+..I'.³4à#2^) 9929-996 99 - 944 8h6 : 66 \$øg.þ.ÞÞÖ.3.áÇ.ED.Ä2" sF"?§zq.×ÄÇ.3Ç6Úÿ.|É¿µTþ7w.§%dBÚº..»Ä×;İ.? Hex 🔻 Find Backwards X Signed 8 bit: Signed 32 bit: 646593717 Hexadecimal: -75 **B5** Unsigned 8 bit: 181 Unsigned 32 bit: 646593717 Octal: 265 Signed 16 bit: 15541 32 bit float: 9.592128E-16 Binary 10110101 Unsigned 16 bit: 15541 64 bit float: 6.898045E+258 Text: Show little endian decoding Show unsigned as becadecimal Stream length: Fixed 8 Bit -Encoding: Default OVR Size: 7768064 Offset: 0000:0725-7 Txt RW

leiprints

If the file:///n	nedia/disk	/070704_0	lp_tuwien	ppt - K	HexEdit					X
<u>F</u> ile <u>E</u> dit	<u>V</u> iew [<u>D</u> ocuments	<u>B</u> ookma	rks <u>T</u> o	ols <u>S</u> etting	s <u>H</u> elp				
	6	1	5) B	6 6	0.		۲		0
0000:0570	11111111	00000000	01110011	011011	10 10011110	10011111	01000011	01001000	ÿ.snCH	
0000:0578	01110010	00110001	00110000	100101	88 11011110 81 01001101	11011111	111111111	00000000	r10.ÞBÿ. ÀÒ.ÑM}	
0000:0588	00101100	10111110	11101111	010001	01 10111010	11110101	10001111	10001111	,≒īEºö.	
0000:0590	01111001	01000000	11111001	010110	00 00001110 00 00010000	01000011	00100010	0 11110010	y@uX.C=0	
0000:05a0	01101101	10101001	10100110	100001	01 01101000	10100001	00110111	00101111	m©¦.hi7/	
0000:05a8	01100011	01110001	00110111	010100	01 01101001 01 10110100	11001101	11111011	01101001	cq7QiIûi Āe1ñ //1E	
0000:05b8	00011101	01001110	00110101	011101	10 00010110	00110011	10111110	01000010	.N5v.34E	
0000:05c0	11011101	00101100	10111001	101111	01 10101101	10101110	01110011	01101110	Ý,1½,-0sr	
0000:05d0	01011100	11000110	01110000	110001	10 10010010	00110001	01101000	01000011	\ÆpÆ.1hC	
0000:05d8	01001100	00101010	11101010	011001	10 00100110	01101010	00111001	10001011	L*êf&j9.	
0000:05e0 0000:05e8	00000000	11110101	11000111	011000	11 1000101011	10110101	10111001	01011011	.01c.μ ¹ [5
0000:05f0	10100011	01011111	00111000	100011	00 01010001	11110011	00001011	00011100	£_8.Qó.	
0000:0518	11101010	11100101	00110100	011010	11 11111001	11011111	10100011	111111111	eå5kùߣv	
0000:0608	00000000	11011010	10100111	101000	11 11011110	10010010	01111010	01000110	.Ú§£Þ.zF	
0000:0610	01011110	10001011	01010010	110100	11 11011111 11 01001101	01110111 01000101	01101101	011110111	∩.ROBwm+ ö.úKME}z	
0000:0620	00111111	01111010	01011101	010010	11 10101111	10111000	11010101	01011001	?z]K⁻,Ôĭ	r
0000:0628	01001001	01000000	11101011	110001	10 11101110 10 10101110	10101011	10101100	01001000	I@ēÆi≪⊸H ÓWÓ:®9	1
0000:0638	00101010	01101010	10010110	010011	00 01001101	01010010	10001010	00100000	*j.LMR.	
0000:0640	10000100	10001000	10000111	1111110	10 00110111 10 10000000	10100111	00011111	11011000	ú7§.@	1
0000:0650	01101001	01011110	01010101	101101	11 10100010	00010111	11010101	000000000	i^W∙¢.Ō.	
0000:0658	11101111	01111111	11111110	010111	11 10101101	10101000	11110100	01001011	ī.þ - ôk	5
0000:0668	1010010	10101011	101010111	000010	10 00001011	11001000	011011110	10101001	§©2.En2	
0000:0670	10000111	01101110	01001010	100101	10 00100110	10000110	10110000	10000101	.nJ.&.°.	
0000:0678	10000100	00011001	01011010	111000	00 11001101 10 01110011	00010110	10101000	01100111	.3Zâs{90	
0000:0688	01010010	00010001	10011010	001100	10 10010001	10001100	10011101	10000110	R. 2	
0000:0690	01010111	01000110	11110011	100000	11 01110011	10001001	10111011		WFó.s.».	
0000:06a0	11100001	11011111	10110011	010001	01 11100011	01111101	00101110	10111010	áß³Eā}.9	
0000:06a8	11000101	00110100	01000001	110001	10 10001001	11011110	11010001	11011101	Å4AÆ.ÞŇÝ	<u></u>
0000:06b8	001101001	10100100	10000110	001101	00 01111100	11101010	010000011	10000110	5¤.< êC.	
0000:06c0	10011001	10010100	10010101	100011	10 01101110	11101001	01001011	01010011	néKS	
0000:06C8	11010111	0100101	00101100	001000	10 11101011	110101111	0100010101	001010001	×M, nē×E)	
0000:06d8	01110101	10110110	00101011	011101	10 10111110	11100101	00100111	00111110	u¶+v¾å'>	
0000:06e0 0000:06e8	00111101	11010111	10101101	101111	11 01011011 80 10010111	100110011	01011010	0 11011101	1Km2[.21	
0000:06f0	10110010	11101011	00110100	010111	10 10110111	11100001	01100000	00110010	²ē4^•á`2	2
0000:06f8	11100000	11111010	00101010	001010	10 00001100 10 01101001	01101001	01101010	01101001	àú**.iji ¤áÅŌii	•
0000:0708	00010010	00101001	01101110	100110	10 01011101	00101100	00110001	00110111	.)n.],17	'
0000:0710	10101010	01010110	00110011	010110	01 10110011 10 110010011	10101011	10101111	10111010	°V3Y3«°°	
0000:0720	11010111	01011101	01110000	010111	10 00111010	10110101	00111100	10001010	×]p^:μ<.	-
Hex 💌						Find	B <u>a</u> ckwa	rds 📃 Igr	ore case	\mathbf{X}
Signed	8 bit:	-77	Signed 3	2 bit:	-1483133005	Hexa	decimal:		В3	
Unsigned	8 bit:	179	Unsigned 3	2 bit:	2811834291		Octal:		263	
Signed 1	16 bit:	12211	32 bit f	loat:	4.251775E-15		Binary:	1	0110011	
Unsigned 1	16 bit:	12211	64 bit f	loat: 1	.775997E-255		Text:		3	
X Show I	ittle endian	decoding	Show <u>u</u>	insigned	as hexadecim	al Strear	n length:	Fixed 8 Bit	-	
		Encoding: D)efault		OVR Siz	e: 7768064	Offse	et: 0000:066	5-7 Bin	RW

(i) file:///	media/disl	c/070704	_dp_tuwien.p	pt - KHexEdit		×
<u>F</u> ile <u>E</u> dit	View	<u>D</u> ocumen	its <u>B</u> ookmark	s <u>T</u> ools <u>S</u> ettings	<u>H</u> elp	
۵ 🛛		1	\$ 6	800	🧟 🗢 Þ 🔵 🕻	I 🖉
0000:0660 0000:0670	d2 eb 9 87 6e 4	7 d2 6f a <mark>96</mark> 26	b3 2f 99 a7 a 86 b0 85 6c 1	9 aa 0a 0b c8 6e a 9 cb 34 cd 16 c0 1	a Òē.Òo³/.§©ºÈnº 4 .nJ.&.°.l.Ë4İ.À.	200
0000:0680	84 33 5	a <mark>e2</mark> 73	7b a9 67 52 1	1 9a 32 91 8c 9d 8	6 .3Zâs{©gR2	111
0000:0690 0000:06a0	el df b	3 45 e3	89 DD 87 22 8 7d 2e ba c5 3	4 /3 at Da d3 44 b 4 41 c6 89 de d1 d	4 WF0.S.≫.⁻.S ≌UUd d áß³Eā}.ºÅ4AÆ.ÞÑÝ	
0000:06b0	a9 dc a	b <mark>37</mark> 0e	11 d3 45 35 a	4 86 3c 7c ea 43 8	6 ©Ū«7ÓE5¤.< êC.	
0000:06C0 0000:06d0	99 94 9 d7 4d 2	o 8e 6e c 6e eb 1	e9 40 53 66 e d7 45 29 75 b	6 2b 76 be e5 27 3	<pre>4neKSta.#ut e xM.nēxE)u¶+v3å'></pre>	
0000:06e0	ef <mark>4b</mark> 6	d <mark>bf</mark> 5b	13 5a dd 3d d	7 ae ac 97 9e 06 2	1 īKmż[.ZÝ=×®¬!	
0000:06f0 0000:0700	b2 eb 34	4 5e b7 2 d6 69	el 60 32 e0 f 6a 9a 8a 12 2	a 2a 2a 0c 69 6a 6 9 6e 9a 5d 2c 31 3	9 2ē4^.á`2àú**.iji 7 ¤áÅŌii)n.l.17	
0000:0710	aa 56 3	3 59 b3	ab af ba 3d 6	b 90 7e c9 46 ea b	7 ºV3Y ³ « [°] °=k.~EFê·	
0000:0720	d7 5d 7	9 <mark>5e</mark> 3a 9 44 26	b5 3c 8a 26 8 65 97 2b 17 3	c b6 ac f5 6a ee d 5 1c b5 ec a5 63 9	5 ×]p^:µ<.&.¶∽õjî0 7 1ă D+e + 5 uì¥c	
0000:0730	fe c7 4	d 75 d5	07 49 2b 65 b	7 fa 34 dl 44 fd c	a þÇMuÖ.I+e·ú4NDýÉ	
0000:0750	96 fd 8	b b4 d6	bf a8 a5 94 e	f 09 12 08 5b 46 3	1 .ý. Ő¿ ¥.ī[F1	
0000:0700	3f c8 9	B 2f 51	c4 c1 e7 5c c	e f6 31 19 a3 7e 3	1 ?É./QĀÁç\Íō1.£~1	
0000:0780	89 9d e	B c7 6c	46 b9 c2 26 d	9 d3 d5 bb eb 0c e	fèÇlF¹Á&ÙÓŐ»ë.ī	
0000:0790 0000:07a0	09 d8 b	r f1 9e 2 75 a3	08 88 C9 5/ 3 17 8e 85 40 8	3 ac dd 48 8e ad 9 8 4a 12 6c f7 56 4	4 0A.n., EW3¬YH d .زu£@.J.l+VM	
0000:07b0	24 b4 6	5 <mark>bd</mark> d0	00 d2 3e a9 d	6 bl f2 0c le e8 d	e \$´f½Ð.Ò>©Ō±òèÞ	
0000:07c0	ff 00 4	2 da b4 2 3a e6	7a cd c6 96 7 ee d4 f7 6f 6	a a7 28 dd 7f ec 1 2 9c ab 16 ea fb e	a ÿ.BU`zIÆ.z§(Y.i. d e\$â:æî0+ob.«.êûî	
0000:07e0	d7 5b d	5 b9 bd	a9 a8 d5 75 3	5 d2 df 47 ac 75 1	e ×[01½0⁻0u50ßG¬u.	
0000:07f0	bb 37 2	a fd 8d . A 4a 31	34 dl 45 af d	3 4f 4e 9a 69 ae b	a »7*ý.4NE ⁻ OON.i® ù11Vťaì-v1.4%	
0000:0800	c9 f3 5	0 d9 14	49 d8 8c a6 2	8 45 92 31 57 e2 6	c ÉóPÚ.IØ.¦(E.1Wâl	
0000:0820	5a 28 fa	a 32 41	2c 89 93 f1 f	c b6 b3 6d 27 86 2	4 Z(ú2A,ñū¶³m'.\$	
0000:0830	e8 a8 f	f 00 dc	4e 1f 46 99 6 7a 2f e8 79 7	a 47 75 7a 43 5b 4	9 è ⁻ ÿ.Üz/èyzGuz <mark>t</mark> [I	
0000:0850	45 5b 5	f <mark>59</mark> 73	95 cf 09 a5 a	5 ad 45 12 c8 6f 2	7 E[_Ys.I.¥¥-E.Éo'	
0000:0860	d5 29 c a2 73 6	/ 50 fa 3 11 18	b6 a6 50 41 2	4 T9 97 03 21 33 3 e 3c ab 2c d1 33 0	9 0)ÇPu4u!39 b csc¶!PA.<«.Ň3.	
0000:0880	23 29 2	B 11 91	81 10 <mark>89</mark> 99 c	d 38 c7 b3 3b f3 1	c #)(İ8dz;ó.	
0000:0890	18 c6 19	9 bd d6 1 5 b8 99	66 06 8e c9 b 1f 08 2e 07 2	0 38 6a 68 c8 e8 6 4 10 9d 95 87 6a d	7 .Æ.½0fE°8jhEèg	
0000:0860	65 ea b	2 cd 1b	93 81 44 d9 7	4 67 a8 1d 9d d1 2	7 eê²ÍDÙtg ⁻ N'	
0000:08c0	f7 29 6	5 47 0b	c5 3d 68 dd 5	5 5d al 7a la 65 5 6 7a e8 a2 1f 76 b	f ÷) kG. Ä=h YU] iz.e	
0000:08e0	5b ca e	5 68 db	ee 4a 83 17 9	a 1b 6e 9e 3f 9f c	7 [ÊâhÛîJn.?.Ç	
0000:08f0	f0 e8 9	9 58 20	a9 4d 96 8f e	3 ed c8 97 4d 38 6	4 õè.X ©MäíÉ.M8d	
0000:0900	fa 56 0	d 65 64	a2 1b b6 8c 8	5 78 f7 bd 52 18 0	c e<.1:№.K.5N;g. 9 úV.ed¢.¶x÷§R	
0000:0920	47 50 1	3 <mark>8d</mark> 99 -	a7 a3 97 6e 2	2 24 90 6a df 4b a	d GP§£.n"\$.jBK-	
0000:0930	40 55 d	c aa c9 5 aa a9	64 d⊎ 47 45 5 a7 67 ba f7 5	5 D9 24 93 d5 55 b 7 a9 65 ba 6b 76 9	a u?v@©&g@+W©e@kv.	
0000:0950	7a 73 e	8 <mark>85</mark> 37	a2 75 97 05 1	6 8e a9 15 17 97 a	3 zsè.7¢u0£	
0000:0960	7/ 75 f	1 5c 8a 1 2a 4a '	ba d4 d6 4d 3 9c 53 4e a3 8	5 50 60 40 72 80 4 d e6 4c c9 11 4b 8	a wun\.º00M5jmMr.J 8 }-=*J.SN£.æLĖ.K.	
0000:0980	4a 05 7	8 d0 d6	32 b5 46 d1 8	2 11 17 b5 43 04 b	9 J.xĐÔ2μFÑμC. ¹	
0000:0990	02 91 59 de 8c 6	9 c0 89 2 df 99	10 9c 09 0c 9 ch ch 19 5d 6	8 ce 6c 0a 06 70 9 5 ae ch 65 1a dc a	8YAIlp. 7 b bR FF le®Fe II6	
0000:0960	d9 b3 4	7 0e 21	12 d6 ed 3d 3	e 8f 4e 9e ee 26 d	f Ú ³ G.!.Ōí=>.N.1&B	
0000:09c0	ec 7d 3	f 63 dd	6b f6 3d 3f 6	3 d3 e9 fb 15 fa 1	3 ì}?cÝkō=?cÓéů.ú.	-
Hex 🔻				Find	B <u>a</u> ckwards	nore case 🗙
Signed	i 8 bit:	67	Signed 32 b	oit: 1162435395	Hexadecimal:	43
Unsigned	i 8 bit:	67	Unsigned 32 b	oit: 1162435395	Octal:	103
Signed	16 bit:	23363	32 bit flo	at: 3.221704E+03	Binary:	01000011
Unsigned	16 bit:	23363	64 bit flo	at: 4.435060E+247	Text:	С
X Show	little endiar	n decoding	Show uns	signed as hexadecimal	Stream length: Fixed 8	3 Bit 💌
	Enc	oding: De	fault	OVR Size: 776	8064 Offset: 0000:084	d-7 Hex RW



- Digital Objects require specific environment to be accessible :
 - Files need specific programs
 - Programs need specific operating systems (-versions)
 - Operating systems need specific hardware components
- SW/HW environment is not stable:
 - Files cannot be opened anymore
 - Embedded objects are no longer accessible/linked
 - Programs won't run
 - Information in digital form is lost (usually total loss, no degradation)
- Digital Preservation aims at maintaining digital objects authentically usable and accessible for long time periods.





- Essential for all digital objects
 - ▶ Office documents, accounting, emails, ...
 - Scientific datasets, sensor data, metadata, ...
 - Applications, simulations,...
- All application domains
 - Cultural heritage data
 - eGovernment, public administration
 - Science / Research
 - Industry
 - Health, pharmaceutical industry
 - Aviation, control systems, construction, ...
 - Nerivate data
 - **...**





Migration

- Transformation into different format, continuous or on-demand (Viewer)
 - + Wide-spread adoption
 - + Possibility to compare to un-migrated object
 - + Immediately accessible
 - Unintended changes, specifically over sequence of migrations
 - Cannot be used for all objects
 - Requires continuous action to migrate





Emulation

Emulation of hardware or software (OS, applications)

- + Concept of emulation widely used
- + Numerous emulators are available
- + Potentially complete preservation of functionality
- + Object is rendered identically
- Object is rendered identically
- Requires detailed documentation of system
- Requires knowledge on how to operate current systems in the future
- Complex technology
- Emulators must be emulated or migrated themselves
- Emulators potentially erroneous/incomplete





Open Archival Information System (OAIS) reference model











Digital Preservation

- Is a complex task
- Requires a concise understanding of the objects, their intellectual characteristics, the way they were created and used and how they will most likely be used in the future
- Requires a continuous commitment to preserve objects to avoid the "digital dark hole"
- Requires a solid, trusted infrastructure and workflows to ensure digital objects are not lost
- ls essential to maintain electronic publications & data accessible
- Will become more complex as digital objects become more complex
- Needs to be defined in a preservation plan

eprints





EPrints Preservation File Classification & Risk Analysis



The Preservation Process

Preservation - Check

• Bit checking & checksum calculation

Preservation - Analyse

- What is the type of file, is the file valid?
- Is the file at risk of not having an editor/reader?
- Is there a better format available? Lossless or Lossy?

Preservation - Action

- File migration to avert risks found by analysis.
- Movement of file to new storage.





File Format Analysis

Preservation - Analyse

Preserv 2

EPrints File Classification



Search

Home About Browse by Year Browse by Subject

Logged in as Mr David C Tarrant | Manage deposits | Profile | Saved searches | Review | Admin | Logout

Formats/Risks



Risks analysis functionality is currently not available. This feature is due to be made available by The National Archives (UK) in the near future. This page will automatically pick up the data when this feature becomes available.







Preservation - Analyse

- What is the type of file, is the file valid?
 - Droid is a good classification tool for this.



- Is the file at risk of not having an editor/reader?
- Is there a better format available? Lossless or Lossy?

Risk Information obtained from factual data Objective risk information is local





Risk Analysis In EPrints

Preservation - Analyse

2

EPrints File Classification + Risk Analysis

Training Repository - 10

eprints

Search



Formats/Risks

This EPrints install may be referecing a trial version of the risk analysis service. If you feel this is incorrect please contact the system administrator.

High Risk	Objects
Graphics Interchange Format (Version 1987a) 🖶	9
Low Risk	Objects
Acrobat PDF 1.4 - Portable Document Format (Version 1.4)	20
Graphics Interchange Format (Version 1989a)	3
Preservation Plans	





Risk Analysis In EPrints - Detailed View

Preservation - Analyse

EPrints File Classification + Risk Analysis







Collection Gathering

	Preservation Actions		
Downle	oad File Seclection		
No. of Files: 5 Download			
Upload	Preservation Plan		
	Browse		
	Upload		

- If more than I file requested:
 Provide Newest and Oldest
- If morn than 3 files requested:
 - Also provide Largest and Smallest
- Then
 - Provide a random selection





Exercise Time





Preservation - Check

• Handled by our storage manager.

Preservation - Analyse

- Parallels can be drawn with storage, in that we are integrating with and utilising currently available services to perform our analysis.
- Processing of the results leads to a powerful interface which tells us many things about the repository ecosystem and it's future.

Preservation - Action

• Next part of workshop...







EPrints Preservation

Preservation Planning

Preservation workflow

Check



 Format identification, versioning
 File validation
 Virus check
 Bit checking and checksum calculation Preservation planning Characterisation:

Significant properties and technical characteristics, provenance, format, risk factors

Risk analysis

Tools e.g. DROID JHOVE FITS

Tools

Plato (Planets) PRONOM (TNA) P2 risk registry (Keeplt) INFORM (U Illinois)

KB

eprints

Migration

Action

- Emulation
- Storage selection



What file formats do you accept? Do you convert any to a different format?

- ALL: Accept any format.
- Two: Convert everything to PDF, but store the source files in the background for preservation reasons.
- Four: Mention specifically converting Word to PDF: one seeks permission from the author to do this, and uploads as Word if permission is not granted.
- ► One: Mentions converting ZIP files to PDF.

Sue Ashby, University of Portsmouth Library, Summary of responses to IR questionnaire, JISC-REPOSITORIES, 18 February 2010





Format risks

Ubiquity: degree of adoption of the format

- Support: number of tools available which can access the format
- **1002** Disclosure: extent to which the format documentation is publicly disclosed
- **1003** Document Quality: completeness of the available documentation
- Stability: speed and backwards-compatibility of version change
- Ease of Identification: ease with which the format can be identified
- Ease of validation: ease with which the format can be validated
- Lossiness: does the format use lossy compression
- Intellectual Property Rights: whether or not the format in encumbered by IPR
- Complexity: degree of content or behavioural complexity supported

From PRONOM documentation (The National Archives), July 2008





Format risks

Ubiquity: degree of adoption of the format

- Support: number of tools available which can access the format
- **1002** Disclosure: extent to which the format documentation is publicly disclosed
- **1003** Document Quality: completeness of the available documentation
- Stability: speed and backwards-compatibility of version change
- Ease of Identification: ease with which the format can be identified
- Ease of validation: ease with which the format can be validated
- Lossiness: does the format use lossy compression
- Intellectual Property Rights: whether or not the format in encumbered by IPR
- Complexity: degree of content or behavioural complexity supported

From PRONOM documentation (The National Archives), July 2008





A group task on format risks

- I. Choose two formats to compare (e.g. Word vs PDF, Word vs ODF, PDF vs XML, TIFF vs JPEG)
- 2. By working through the (surviving) list of format risks **select a winner** (or a draw) between your chosen formats **for each risk category** (I point for win)
- 3. Total the scores to find an overall winning format
- 4. Suggest one reason why the winning format using this method may not be the one you would choose for your repository





Exercise Time





Some thoughts about formats

Free vs open source vs open standard:

- **MS Office** XML open standard
- **Open Office** free XML open standard
- **PDF** page representation
- XML generic Web format, computational





Rosenthal: Why we can relax about preservation

"

Historically, the open source community has developed rendering software for almost all proprietary formats that achieve wide use

Even the formats which pose the greatest problems for preservation, those protected by DRM technology, typically have open source renderers"

Format Obsolescence: Scenarios (April 29, 2007) http://blog.dshr.org/2007/04/format-obsolescence-scenarios.html





Work with, not against, your authors and contributors

- "Preservation begins with the author"
- ► U. Rochester (USA) has written its own repository software IR+ to give its authors a Web-based authoring workspace
- But which applications are widely used and popular among your authors? Digital content authoring tools are typically chosen on the basis of purpose, utility, familiarity (what is provided, supported by Information Systems?) Rarely are they chosen for format or preservation.
- Authors will craft their output in the chosen application, but will often throw away that craft if asked to convert to another format
- One approach that builds on popular formats is ICE: Integrated Content Environment, which converts formats from popular content authoring tools





- Studies and user reports claim JPEG 2000 to be or at least will become – the next archiving format for digital images
- The format offers new possibilities, such as streaming, and reduces storage consumption through lossless and lossy compression. Another often claimed advantage of JPEG 2000 is that the master image can possibly serve as the access copy as well, and thus replace derived compressed, low resolution access copies.

Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, *D-Lib Magazine*, Vol15 No. 11/12, Nov/Dec 2009, http://www.dlib.org/dlib/november09/kulovits/11kulovits.html





TIFF vs JPEG 2000?

Who's for JPEG? The major players line up

- The National Library of the Netherlands evaluated JPEG 2000 against uncompressed TIFF (currently used) for storage capacity, image quality, long-term sustainability, functionality. JPEG 2000 is recommended as future archive format.
- 2. The British Library recently moved forward to migrate their 80terabyte newspaper collection from TIFF to JPEG 2000
- 3. The Wellcome Library announced they will use JPEG 2000 for their upcoming digitization projects

Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, *D-Lib Magazine*, Vol15 No. 11/12, Nov/Dec 2009, http://www.dlib.org/dlib/november09/kulovits/11kulovits.html





What does Plato say?

"At this point in time **not migrating** the TIFF v6 images is the best alternative."

"However, in one year we'll look at this plan again to see if there are more tools available and whether or not the ones we considered in this year's evaluation have been improved."

> Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings, *D-Lib Magazine*, Vol15 No. 11/12, Nov/Dec 2009, http://www.dlib.org/dlib/november09/kulovits/11kulovits.html







EPrints PreservationPreservation Action & Provenance



The Preservation Process

Preservation - Action

- Uploading a Preservation Plan in EPrints
- Viewing resultant actions
- Managing your plans
- Re-enacting the Plan
- Viewing Provenance Information





Uploading a Plan

	Preservation Actions			
Downle	oad File Seclection			
No. of Files: 5 Download				
Upload	Preservation Plan			
	Browse			
	Upload			

- Each set of "at risk" classified files can have a single related preservation plan.
- Once uploaded, any defined actions will be performed on **all** files of that classification.

Preservation Plan Upload Successful

Actions have been queued to be executed shortly and changes will be reflected below once completed. In order to view these changes please revisit or refresh this page later.



S



No plan can cause files to be deleted.

- A plan controls any files it has created.
- While these files exist, the plan cannot be deleted.







Viewing the Result

Previously high risk objects are still represented by a red bar, but are now in the low risk category.







Preservation Actions Panel

Preservation Actions
Download File Seclection
No. of Files: 5 Download
Download Preservation Plan
Download Enact Plan

Download plan for reviewing in planning software.

Re-enact plan





Viewing the Result

Before



After



Image (PNG) (Migrated (Preservation) from Document ID: 41 (image/gif)) Download (76Kb) | Preview

• Image (GIF) (Original Version)





Provenance Information

- Open Provenance Model (OPM) compliant
- Stored in RDF triple form using the EPrints relation manager added in 3.2







Exercise Time













Many Thanks







