PDF and Long-Term Archival (1)

UNIVERSITY OF SKÖVDE

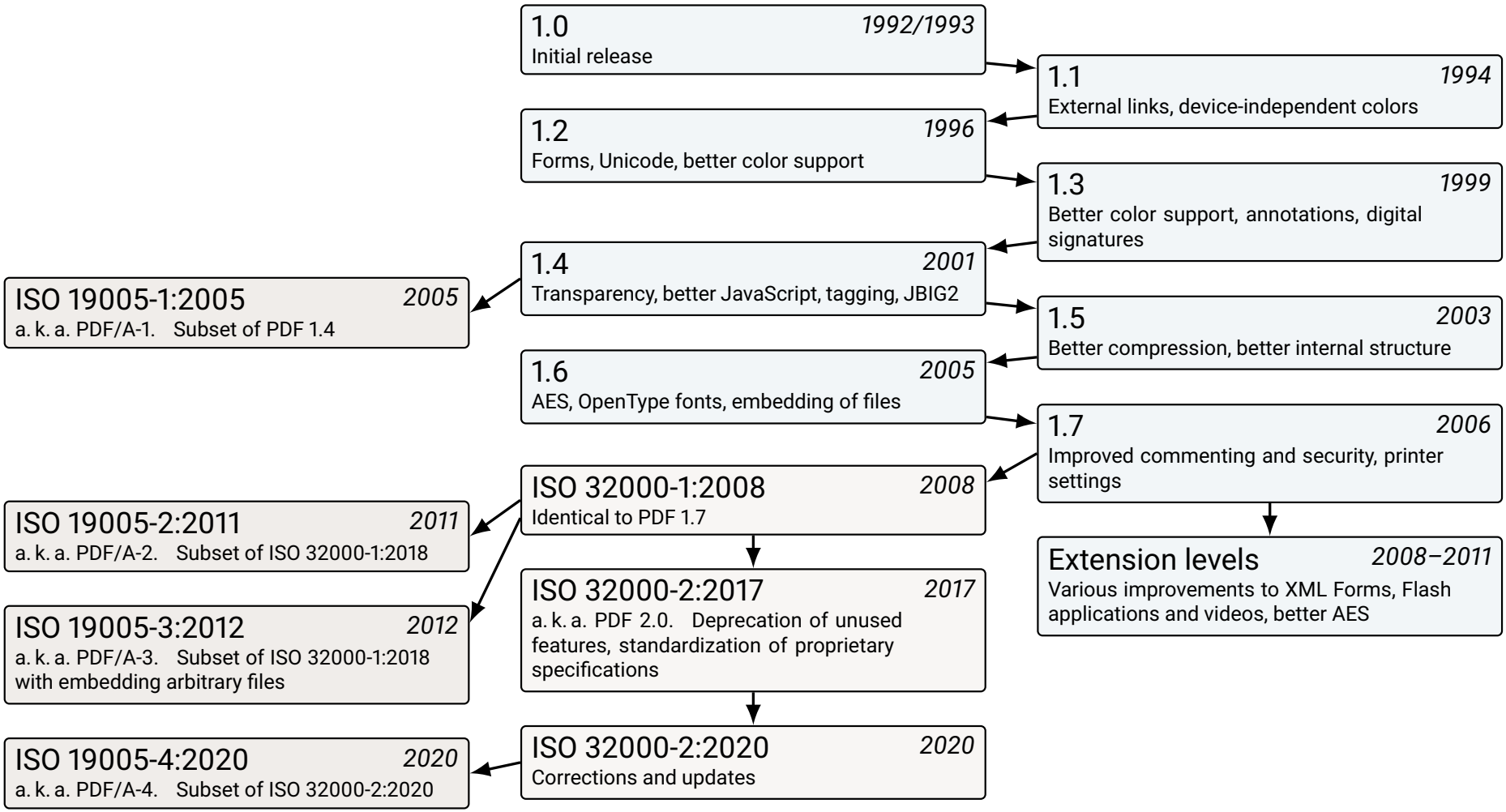
Portable Document Format (PDF)

- Developed by **Adobe** since **1992** as successor to **PostScript**
- Represents **pages**, meant to be **read-only** and for **printing**
- ⊕ Tools for **viewing and generating PDF documents** available on virtually all computing platforms, provided by **multiple vendors** including **open source implementations**
- ⊕ **Works in most cases as expected**, i. e. document receiver sees document as intended by sender
- ⊖ Format is **complex**, relies on **external references**, known to be **ambiguous**
- ℹ Despite existing ‘official’ documentation, historically how **Adobe’s tools interpret PDF is the authoritative way**

PDF and Long-Term Archival (2)

UNIVERSITY OF SKÖVDE

PDF and PDF/A Versions



PDF and Fonts (1)

UNIVERSITY OF SKÖVDE

- For a PDF viewer to show text, it must know the **shape** (‘glyph’) of each letter
- Glyph data can be stored/retrieved from different sources
  - (a) As **part of a PDF file**
    1. An existing font data file is embedded into the PDF document
    2. A **subset** of an existing font data file is embedded (only those glyphs that are necessary)
  - (b) From **outside of the PDF file**
    3. 14 ‘standard’ fonts every PDF viewer has to have available
    4. Referring the font by name, viewer has to find an alternative locally
    5. Storing the glyphs’ metrics, viewer has to find an alternative locally

ℹ Only alternatives in group (a) **guarantee visual fidelity**

❗ **PDF/A** only allow alternative from group (a)

PDF and Fonts (2)

UNIVERSITY OF SKÖVDE

- Most jurisdictions allow to **copyright the font program** that generates glyphs
- You need **approval from the copyright holder** for use, modification, and **re-distribution**

(a) **Proprietary font vendors** like Adobe or Microsoft

- Today's font licenses **allow embedding of font data**
- . . . but only if original **font data can not be extracted** (technically realized by subsetting and removal of hints)
- If font was **part of a software**, it may only be used with software
- Creator must have valid license
- Editing PDF documents may require to have the same license

(b) **Fonts under open licenses** in analogy to open source

- **Re-distribution unrestricted** for (unmodified) font data

Research on PDF and Fonts (1)

UNIVERSITY OF SKÖVDE

1. **Identification and retrieval** of **PDF files**
   Focus on Swedish public sector organizations (PSOs)
   1. **Doctoral dissertations** published between 2018 and 2021

      9341 files retrieved
   2. **Svensk författningssamling** (SFS, Swedish Code of Statues) published since 2018 on a dedicated webpage

      5931 files retrieved
   3. **Government's investigation on secure and cost-effective IT operations** (SOU 2021:1, SOU 2021:97) containing submissions by public-sector organizations (national, regional, municipal)

      155 files retrieved

Research on PDF and Fonts (2)

UNIVERSITY OF SKÖVDE

2. **Analysis of PDF files**
   1. Conformance to **PDF/A**
   2. Interesting **metadata** (e. g. PDF version, used tools)

> *To what extent do PDF files from public sector organizations (PSOs) conform to the PDF/A standard and what characterizes those files?*

RQ 1

3. **Which fonts are used** and under **which license** are those

> *How are different fonts used in the collected PDF files?*

RQ 2

PDF/A Conformance

UNIVERSITY OF SKÖVDE

- How to assess a PDF file's conformance to PDF/A?
- Two tools: **veraPDF** and **3-Heights PDF Validator Online**

  Both tools must agree to count a file as conforming
- **Doctoral dissertations**
  - 0.4% of all files **claim** conformance
    0.2% of all files achieve **conformance**
  - **PDF/A-1a** is the single most popular standard part/level
    (18 of 39 and 11 of 22 files, respectively)
- **SOU**
  - 40% of all files **claim** conformance
    29% of all files achieve **conformance**
  - **PDF/A-1b** and **PDF/A-3a** are the most popular parts/levels
- **SFS** No file claims conformance

PDF Creators and Producers (1)

UNIVERSITY OF SKÖVDE

- Which **tools** have been used to generate PDF documents?
- Two metadata fields in PDF with unstructured text string set by involved tools

  Creator Tool used to generate the original document

  Producer Tool that generated the PDF document

- 'Unstructured text string' requires **heuristic guessing**

Examples

`Acrobat Distiller 11.0 (Windows)`	Vendor: Adobe, Product: Acrobat Distiller, Version: 11.0, Operating System: Windows
`Antenna House PDF Output Library 6.6.1437 (Linux64); modified using iText 2.1.7 by 1T3XT`	$Vendor_1$: Antenna House, $Product_1$: PDF Output Library, $Version_1$: 6.6.1437, $Operating\ System_1$: Linux $Vendor_2$: iText a.k.a. 1T3XT , $Product_2$: iText, $Version_2$: 2.1.7

PDF Creators and Producers (2)

UNIVERSITY OF SKÖVDE

Creators			Producers		
Doctoral dissertations, 9341 documents					
Microsoft Word	2859	30.6%	Adobe PDF Library	2199	23.5%
– non-O365	2520	27.0%	Adobe Distiller	1533	16.4%
– O365	339	3.6%	Apple Quartz	1241	13.3%
Adobe InDesign	1458	15.6%	Microsoft Word	1232	13.2%
TEX tool chain	1028	11.0%	– non-O365	911	9.8%
Adobe Acrobat Pro	715	7.7%	– O365	321	3.4%
Microsoft PScript5.dll	524	5.6%	TEX tool chain	848	9.1%
SOU submissions, 155 documents					
Microsoft Word	97	62.6%	Microsoft Word	75	48.4%
– non-O365	61	39.4%	– non-O365	40	25.8%
– O365	36	23.2%	– O365	35	22.6%
Acrobat PDFMaker	16	10.3%	Adobe Acrobat Pro	29	18.7%
OmniPage	3	1.9%	Adobe PDF Library	15	9.7%
RICOH MP C4504ex	2	1.3%	Apose.PDF for .Net	11	7.1%
SFS, 5931 documents					
Microsoft Word	5930	>99.9%	iText	5931	100.0%
– non-O365	5927	>99.9%	PixEdit PixToolsLib	4729	79.7%
– O365	3	<0.1%	PixEdit Converter Server Core	1197	20.2%
Microsoft PScript5.dll	1	<0.1%	Microsoft Word	3	<0.1%
			– non-O365	3	<0.1%

PDF Creators and Producers (3)

UNIVERSITY OF SKÖVDE

Tool	Part/level	# Files	Claims	Conform.	# Files	Claims	Conform.
		Doctoral dissertations, 9341 documents			SOU submissions, 155 documents		
Microsoft Word		2861			97		
– non-O365	A-1b	2522	2	1	61	12	10
	A-1a		18	11		6	5
	A-3b		1	0			
	A-3a		2	1		5	5
– O365	A-3b	339	3	3	36		
	A-3a		3	3		19	19
Adobe Acrobat Pro	A-1b	739			29	5	2
	A-1a					1	0
	A-2b		2		1	1	0
Adobe PDF Library	A-2b	2199	2	0	15	1	1
Adobe Distiller	A-1b	1533	1	1			
	A-2b		1	0			
Adobe InDesign	A-2b	1458	2	0			
TEX tool chain	A-2b	1063	1	1			
	A-2u		1	0			
Acrobat PDFMaker	A-2b				16	1	1
Apose.PDF for .Net	A-1b				11	9	0
PixEdit PixToolsLib	A-1b				10	10	8
iText	A-1b	5	2	0			
	A-3a		1	1			
OmniPage	A-1a	3	3	3			
Apose.Words for .Net	A-1a	3	2	2			

Fonts in PDF Files (1)

UNIVERSITY OF SKÖVDE

Open The font is available under an open license, i.e. a font license which does not restrict the modification or redistribution of the font program

- **Example license** SIL OFL
- **Example font** Liberation

Proprietary A commercial license must be acquired or the font's use and/or distribution is otherwise restricted such as 'for personal use only'

- **Example fonts** 'Arial' and 'Times New Roman'

Ambiguous The font's name is so generic that there exist multiple independent fonts both under open and proprietary licenses

- **Example fonts** 'Garamond' and 'Symbol'

Unknown The font's name as identified in the PDF file is inconclusive to determine the font's original name

- **Example font** 'AdvOT40514f85'

Fonts in PDF Files (2)

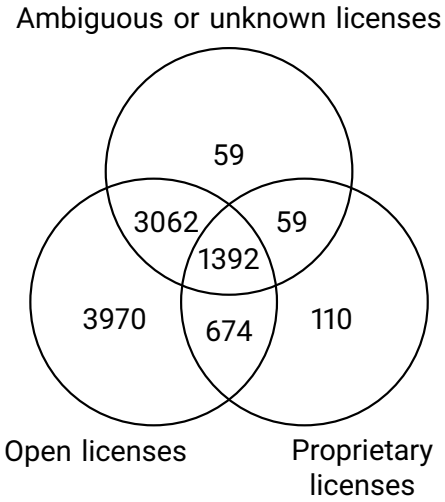
UNIVERSITY OF SKÖVDE

Font name	License cat.	D. diss.		SOU		SFS	
Times New Roman	proprietary	8062	86.3%	103	66.5%	5931	100.0%
Arial	proprietary	7789	83.4%	121	78.1%	13	0.2%
Calibri	proprietary	5754	61.6%	52	33.5%	207	3.5%
Cambria	proprietary	4091	43.8%	13	8.4%	7	0.1%
Symbol MT	proprietary	3474	37.2%	13	8.4%	2	<0.1%
Helvetica	proprietary	2850	30.5%	9	5.8%		
Symbol	ambiguous	1615	17.3%				
Times	ambiguous	1608	17.2%	11	7.1%		
Georgia	proprietary	1531	16.4%	16	10.3%		
Computer Modern	open	1501	16.1%				
Verdana	proprietary	992	10.6%	13	8.4%	1	<0.1%
Garamond	ambiguous	860	9.2%	21	13.5%	52	0.9%
Minion Pro	proprietary	650	7.0%	12	7.7%		
Segoe UI	proprietary	434	4.6%	4	2.6%	3	<0.1%
Microsoft Sans Serif	proprietary	55	0.6%	20	12.9%		
Baskerville Old Face	proprietary	47	0.5%			2	<0.1%
Liberation Serif	open	42	0.4%			1	<0.1%
No fonts		15	0.2%				

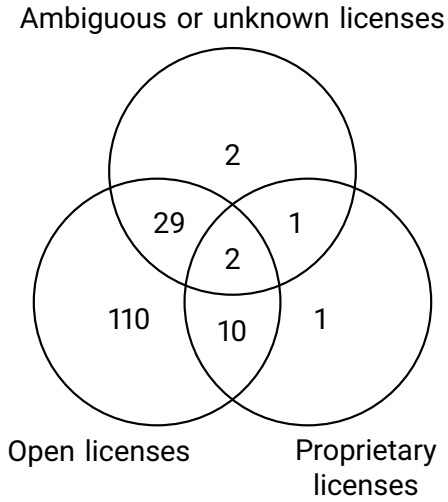
Fonts in PDF Files (3)

UNIVERSITY OF SKÖVDE

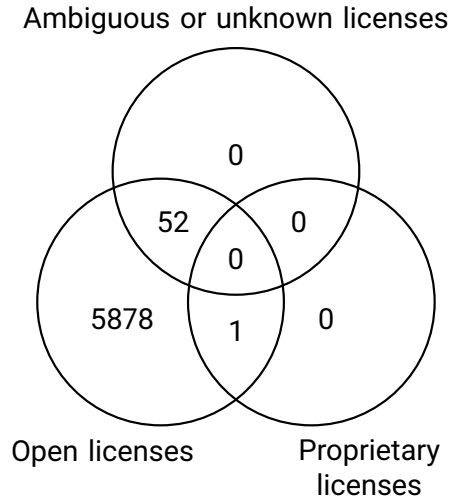
Doctoral dissertations
9341 documents



SOU submissions
155 documents



SFS
5931 documents



Fonts in PDF Files (4)

UNIVERSITY OF SKÖVDE

PDF/A part	Doctoral dissertations		SOU		SFS	
			No fonts used			
PDF/A-1	0		0		0	
PDF/A-2 or -3	0		0		0	
Not PDF/A	15	0.2%	0		0	
			Only open licenses			
PDF/A-1	0		0		0	
PDF/A-2 or -3	0		0		0	
Not PDF/A	110	1.2%	1	0.6%	0	
			Open licenses in combination with other licenses			
PDF/A-1	1	<0.1%	1	0.6%	0	
PDF/A-2 or -3	3	<0.1%	5	3.2%	0	
Not PDF/A	2121	22.7%	7	4.5%	1	<0.1%
			Combinations of non-open licenses			
PDF/A-1	11	0.1%	19	12.3%	0	
PDF/A-2 or -3	7	<0.1%	20	12.9%	0	
Not PDF/A	7073	75.7%	102	65.8%	5930	>99.9%

Fonts in PDF Files (5)

UNIVERSITY OF SKÖVDE

Font	# Files	Font reference count per embedded status not	subset	fully	⋆
Doctoral dissertations, 9341 documents					
Times New Roman	8062	6609	58483	1677	1400
Arial	7789	5159	33657	896	699
Calibri	5754	12	19226	239	235
Cambria	4091	4	16225	178	167
Computer Modern	1501	2	25147	257	245
SOU submissions, 155 documents					
Arial	121	68	143	93	92
Times New Roman	103	65	85	207	207
Calibri	52	0	73	34	34
Garamond	21	4	35	0	0
Microsoft Sans Serif	20	0	0	41	41
SFS, 5931 documents					
Times New Roman	5931	10	30634	608	33
Calibri	207	0	213	0	0
Garamond	52	0	52	0	0
Arial	13	0	13	0	0
Liberation Serif	1	0	1	0	0

Discussion

UNIVERSITY OF SKÖVDE

- Great variation in extent of achieving PDF/A conformance across PDF sets
  1. **SFS** Not at all
  2. **Doctoral dissertations** Minor fraction
  3. **SOU submissions** About one third
- Non-cloud **Microsoft Word** is dominating in PDF generation
  Most prominent **open source alternative** are TEX tool chains
  LibreOffice and alternatives virtually not existing
- Most used fonts are under **proprietary licenses**
  Doctoral dissertations are an exception due to TEX tool chains
- **Subset embedding** most popular but many documents do not embed font data for fonts ‘everyone’ has available

Conclusions

UNIVERSITY OF SKÖVDE

RQ 1: PDF/A Conformance and Files' Characteristics

- Surprising differences among PSOs regarding achieving **PDF/A conformance** given that similar limitations apply
  ➔ **Challenge** is not only technical, but also **administrative**
- Swedish National Archives' requirement for **PDF/A-1** (but not A-2 or later) is not respected

RQ 2: Font Usage in PDF Files

- **Proprietary fonts dominate**
  . . . but unclear whether author had valid license
- Many **documents lack font data** for 'standard' fonts like 'Arial'

➔ PDF document **maintenance** (editing) **hindered**

✅ **Adopt open fonts** and adjust document templates

Do You Recognize Those Fonts?

UNIVERSITY OF SKÖVDE

ABC abc 123 åäö

ABC abc 123 åäö

ABC abc 123 åäö

ABC abc 123 åäö

ABC abc 123 åäö

ABC abc 123 åäö

Top till bottom: Georgia (proprietary), Gelasio (open), Arial (proprietary), TeX Gyre Heros (open), Times New Roman (proprietary), TeX Gyre Termes (open)