

by René Mittå



With focus on TIFF, PDF & PDF/A



ACM	BAT	CHK	CRT	DRW	FPX	HT	LBM	MDZ	OPT	PM6	REG	SVX	WAD
ACP	BFC	CHP	CSC	DSG	FRM	HTM	LDB	MED	ORG	P65	RES	SYS	WAV
ACT	BG	CIL	CSP	DSM	FRM	HTML	LEG	MER	OR2	PNG	RFT	S3M	WB1
ACV	BI	CIM	CSV	DSP	FRM	HTX	LHA	MET	OR3	PNG	RLE	TAR	WB2
AD	BIF	CIN	CT	DSQ	FRM	ICA	LIB	MI	P10	PNT	RM	TAZ	WBK
ADB	BIN	CK1	CTL	DSW	FRM	ICB	LIS	MIC	PAB	POT	RMI	TEX	WBL
ADD	BK	CK2	CUE	DWG	FRX	ICM	LOG	MID	PAK	PP4	ROV	TGA	WCM
ADM	BK\$	CK3	CUR	DXF	GAL	ICO	LOG	MMF	PAL	PPA	RPT	TGZ	WDB
ADP	BKS	CK4	CUT	EMF	GCP	IDD	LPD	MMM	PAT	PPM	RTF	THEME	WEB
ADR	BMK	CK5	CV	ENC	GED	IDQ	LRC	MOD	PBK	PPS	RTM	THN	WGP
AFM	BMP	CK6	CV	EPS	GEM	IFF	LST	MOD	PBM	PPT	SAM	TIF	WID
AF2	BM1	CLASS	CWK	ER1	GEN	IGF	LWO	VOM	PCD	PRE	SAV	TIFF	WIL
AF3	BRX	CLP	CWS	ERX	GFC	IIF	LWP	MPE	PCL	PRF	SCC	TIG	WIZ
AI	BSP	CLS	CXX	EVY	GFI	IMA	LZH	MPEG	PCS	PRN	SCD	TLB	WK1
AIF	BS1	CMD	DAT	EWL	GFX	IMG	LZS	MPG	PCT	PRS	SCD	TMP	WK3
ALB	BTM	CMD	DAT	EXE	GID	INC	M3D	MPP	PCX	PRZ	SCH	TRM	WK4
ALL	B4	CMD	DBF	F	GIF	INF	MAC	MPP	PDF	PS	SCN	TRN	WKS
AMS	C	CMF	DBX	F77	GIM	INI	MAD	MP3	PDF	PSD	SCP	TTF	WKS
ANC	CAB	CMP	DCR	F90	GIX	INP	MAF	MSG	PDQ	PST	SCR	TWF	WLF
ANI	CAL	CMP	DCS	FAR	GNA	INS	MAK	MSN	PFA	PTM	SCR	TWW	WLL
ANS	CAL	CMV	DCX	FAV	GNX	INS	MAM	MSP	PFB	PUB	SCT	TX8	WMF
API	CAS	CMX	DDF	FAX	GRA	ISO	MAP	MTM	PFC	PUB	SC2	TXT	WOW
APR	CAT	CNF	DEF	FH3	GRD	ISP	MAP	MUS	PFM	PWD	SDL	T2T	WP
APS	CB	CNM	DEF	FIF	GRP	ISU	MAQ	MVB	PGL	PWZ	SDR	UDF	WPW
ARC	CCB	CNQ	DER	FITS	GTK	IT	MAR	MWP	PGM	PXL	SDT	ULT	WP4
ARJ	CCF	CNT	DIB	FLC	GT2	IW	MAS	NAP	PIC	QAD	SEA	URL	WP5
ART	CCH	COB	DIC	FLI	GWX	JAR	MAT	NCB	PIC	QBW	SEP	USE	WP6
ART	CCM	COD	DIF	FLT	GWZ	JAVA	MAX	NSF	PIC	QDT	SHB	VBP	WPD
ASA	CDA	COM	DIR	FLT	GZ	JBF	MAZ	NST	PIF	QLB	SHB	VBW	WPD
ASC	CDF	CPD	DIZ	FMB	H	JFF	MB1	NTF	PIF	QRY	SHG	VBX	WPG
ASD	CDI	CPE	DLG	FMT	HED	JIF	MCC	OBD	PIG	QT	SHS	VCF	WPS
ASM	CDR	CPI	DLL	FMT	HEL	JMP	MCS	OBD	PIN	QTM	SHW	VDA	WPT
ASP	CDT	CPL	DMF	FMX	HEX	JN1	MCW	OBJ	PIN	QXD	SIT	VI	WQ1
ASP	CDX	CPL	DOC	FOG	HGL	JPEG	MDA	OBZ	PIX	R	SLK	VLB	WQ2
AST	CEL	CPP	DOC	FON	HLP	JPG	MDB	OCX	PJ	RA	SND	VOC	WRI
ATT	CFB	CPR	DOC	FOR	HOG	JS	MDE	OFN	PKG	RAM	SND	VP	WRL
AVI	CFG	CPT	DOC	FOT	HPJ	JTF	MDL	OFT	PL	RAS	SQC	VSD	WS1
AWD	CGI	CPX	DOT	FP	HPP	KDC	MDL	OKT	PLT	RAW	SQR	VST	WS2



I made a small survey of file types used in 122 submissions.

Among 50 million documents, I found more than 200 different file formats had been used before conversion:

- 98% Office documents, mails, pictures and other text based formats
- 1% Spreadsheet
- 0,5% GML (geodata)
- 0,2% CAD
- 0,2% Video
- 0,01% Audio



Problem

- Impossible to handle so many different file types inside the archive, because of limited resources.
- Password protection
- SW licenses

Solution (until now)

Files have to be converted before delivery to the archive by document owner



File Formats - How to choose?

Our guidelines for choosing file formats

- support in the IT-sector
- being a standard or well described
- being platform independent
- free / open
- stable
- simple (can it be validated)





If a format is closed or unnecessary complex it will be difficult to validate and migrate!

The amount of potential errors will increase!

We decided to go for:

- MP3 (audio)
- MPEG2 & 4 (video)
- TIFF & JPEG2000 (office formats and pictures)
- GML (Geodata)



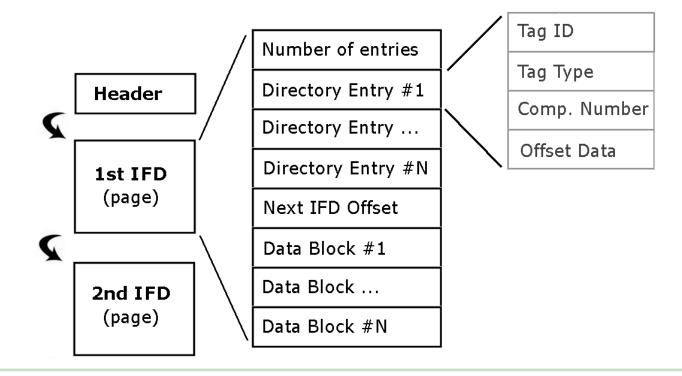
TIFF

- First TIFF specification released in 1986 by Aldus Corporation as a standard method of storing black & white images created by scanners, fax and desktop publishing applications.
- First public release of TIFF was Revision 3.0
- Version 4.0 added support for uncompressed RGB color images
- Version 5.0 added support for storing palette color images and support for the LZW compression algorithm.
- Version 6.0 was released in 1992 and added support for additional color profiles and JPEG compression.
- Today, TIFF is a standard file format found in most paint, imaging, and desktop publishing programs.



TIFF Structure

- Container with one or more pages (Image File Directories or IFD)
- Each page contains its own data.
- Metadata such as size, compression and page offset are defined in sections with Tags
- Page offset information makes it fast to access and manipulate pages





TIFF

- Simple and solid (100 million documents with less than a handful un-correctable failures)
- Page oriented (Single and multipage)
- Flexible (mix of bit depths and compressions I same file)
- Supporting metadata of different types (ex EXIF)
- Commonly used and widely supported
- Well described and license free
- 14 different bit depths and several compression methods
- Used for: fax, scan, photos, printing, archives, libraries museums and private
- Offering effective lossless compression using LZW and group 4 (especially true when talking about monochrome text documents)



TIFF limitations

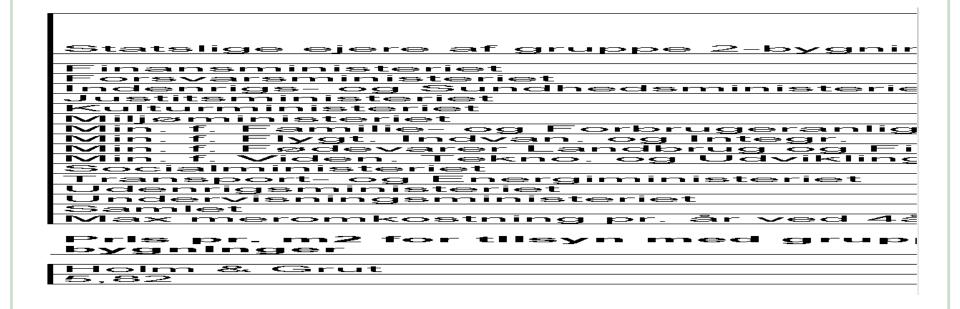
- The format uses 32bit offsets, and is limited to 4 gigabytes (this limitation is shared by the majority of other file formats)
- Be aware of viewers missing support for different DPI scale and odd bit depths (examples)



TIFF - DPI scaling

It's a feature not a bug!

In this case a 600x96 DPI image shown without DPI scaling





TIFF – DPI scaling

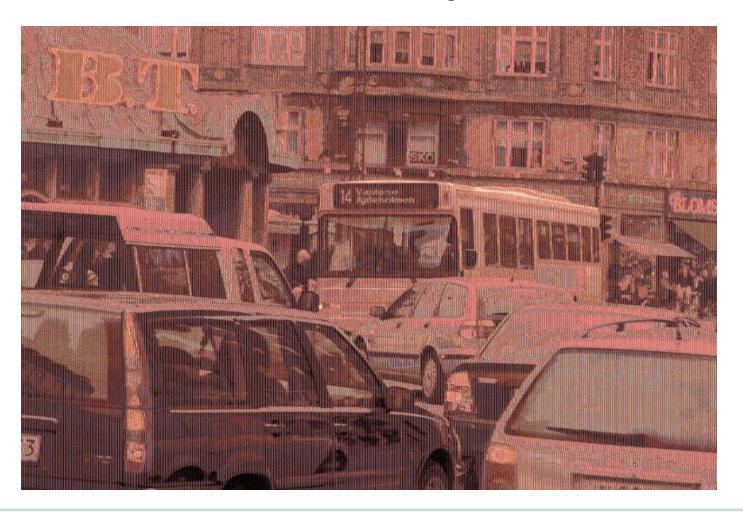
- In a viewer that use the DPI information for prober scaling ... everything looks fine.
- Important to use this information when doing a convertion!

Statslige ejere af gruppe 2-bygnir Finansministeriet Forsvarsministeriet
Finansministeriet Forsvarsministeriet
Forsvarsministeriet
Indenrigs- og Sundhedsministerie
Justitsministeriet
Kulturministeriet
Miljøministeriet
Min. f. Familie- og Forbrugeranlig
Min. f. Flygt. Indvan. og Integr.
Min. f. Fødevarer Landbrug og Fi
Min. f. Viden. Tekno. og Udvikling
Socialministeriet
Transport- og Energiministeriet
Udenrigsministeriet
Undervisningsministeriet
Samlet
Max meromkostning pr. år ved 4έ
Prls pr. m2 for tllsyn med grup
bygninger
Holm & Grut
5,82



TIFF - odd bit depths

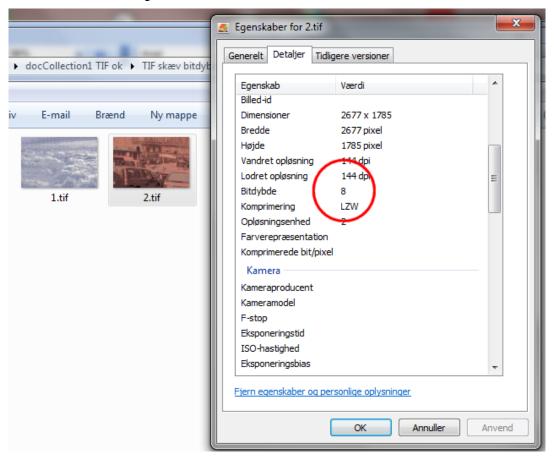
Yet another feature ... or is it a bug?!





TIFF - odd bit depths

It is a feature ... but in this case Windows claims it to be a 8 bit file when it actually isn't. Advice is not to use odd bits!





TIFF - odd bit depths

- In a viewer that supports odd bit depths ... everything looks fine.
- Important to use this information when doing a convertion!





TIFF – creation and conversion

TIFF printer driver

- Many on the marked.
- Most with access to an API (Application Programming Interface) enabling automated creation of TIFF files

Conversion

- SW (Image Magick and Ghostscript)
- Open source class libraries (LibTIFF, Java TIFF library and FreeImage)
- Native support in Visual Studio .NET



TIFF → PDF

TIFF as we know it today is based on version
 6.0 from 1992

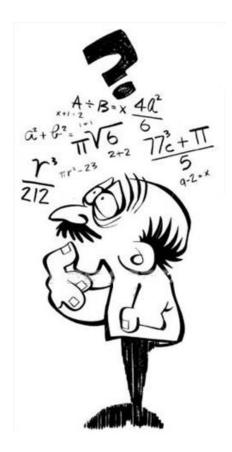
 PDF was invented as a replacement for TIFF but have expanded over the years

 Latere PDF/A was defined because the industry realized that PDF wasn't suited as format for long time preservation



PDF/A – why not?

The short answer is ... it's very complex!





PDF/A

PDF/A is a standard that comprises three parts:

- PDF/A-1
 - ISO 19005-1:2005 Part 1: based on PDF 1.4
- PDF/A-2
 - ISO 19005-2:2011 Part 2: based on ISO 32000-1 (PDF 1.7)
- PDF/A-3
 - ISO 19005-3:2012 Part 3: based on ISO 32000-1 with support for embedded files

Note: PDF/A-2 and PDF/A-3 are based on the ISO standard for PDF 1.7,

PDF/A-1 is based on PDF 1.4 which is not an ISO standard.



PDF/A

3 versions and 8 PDF/A flavours:

- PDF/A-1b PDF 1.4, visual reproduction
- PDF/A-1a PDF 1.4, level b + structure
- PDF/A-2b PDF 1.7, visual reproduction
- PDF/A-2a PDF 1.7, level b + structure
- PDF/A-2u PDF 1.7, level b + unicode map
- PDF/A-3b PDF 1.7, visual reproduction + attachments
- PDF/A-3a PDF 1.7, level b + structure + attachments
- PDF/A-3u PDF 1.7, level b + unicode + attachments



PDF & PDF/A-1

What is PDF/A ISO 19005-1?

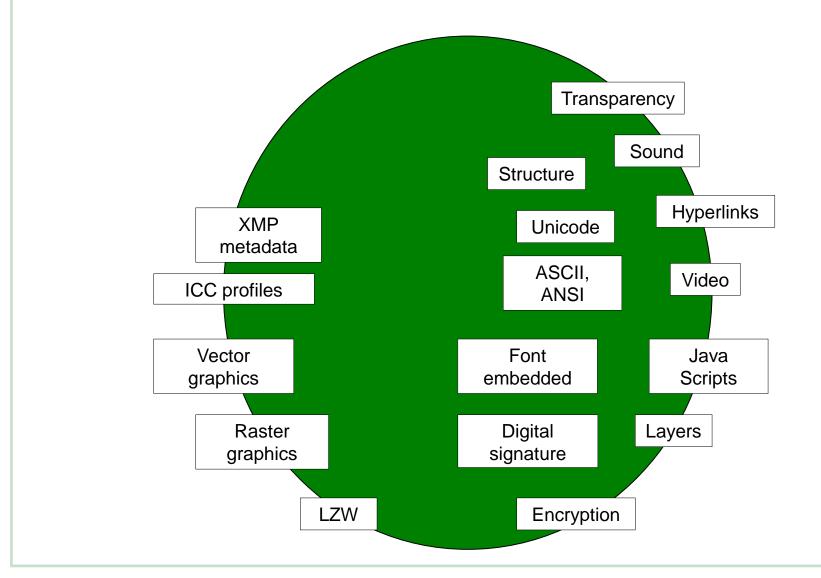
- It's a ISO standard counting <u>29 pages</u>
- PDF/A-1 is based on PDF version 1.4
- PDF/A-1 ISO 19005-1 only defines some limitations in the way you use PDF version 1.4

What is PDF 1.4?

- It's not a ISO standard!
- PDF version 1.4 is based on many standards and specifications!
- Each standard and specification are complex and rich (Unicode 750 pages, Open Font Format 420 pages, SVG 719 pages)
- PDF/A specifications is estimated to reach > <u>5000 pages</u>, when we put it all together!

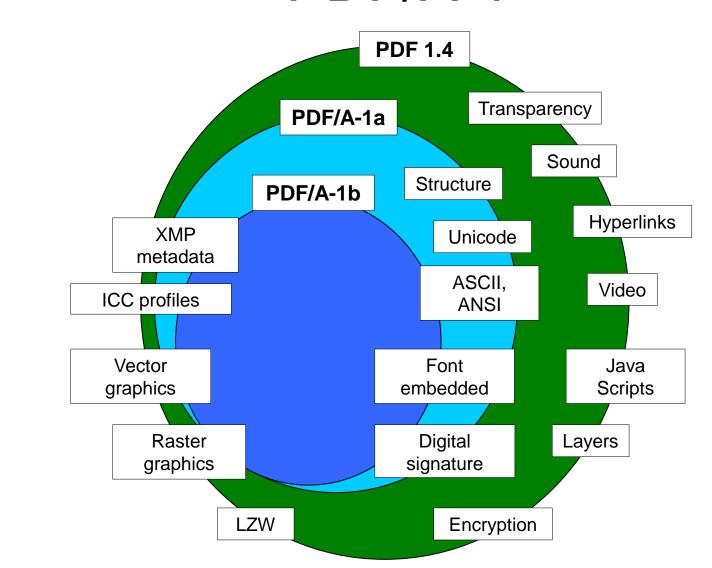


PDF 1.4





PDF/A-1





PDF & PDF/A

- PDF files are backward compatible and that is normally good ... but it also adds a lot of complexity to the format!
 - LOTS of functionality !!!
 - Many and complex font types
 - Missing support of transparency in PDF/A-1
 - Impossible to validate 100%
 - PDF/A is still under development
 - Vulnerable against bad SW



PDF Association investigations in March 2013

PDF Association made a series of investigations in March 2013.

At "PDF Technical Conference Cologne 2013", Duff Johnson made a presentation named "Inability of full validation of PDF" covering two studies ..

Market Research

Estimating the extent of defective PDFs

Technology and support survey

Faulty PDF's impact on daily business and operations.



2. What proportion of your electronic documents are PDF files?

Don't know / rather not say	0	0%	
We don't use PDF	0	0%	
We almost never use PDF	1	2%	
< 5 in 100 documents are in PDF	3	5%	
5 - 10 in 100 documents are in PDF	2	3%	
10 - 25 in 100 documents are in PDF	6	9%	
25 - 50 in 100 documents are in PDF	8	13%	
50 - 75 in 100 documents are in PDF	13	20%	
> 75 in 100 documents are in PDF	22	34%	
PDF is all we use	10	15%	





3. How common are "bad" PDF files?

Don't know / rather not say	3	5%
Never happens	6	9%
Very occasionally	11	17%
Rare, but certainly not unheard-of	24	37%
I get these files a few times a year	7	11%
1-2 in 100 files	4	6%
2-3 in 100 files	3	5%
3-5 in 100 files	5	8%
5-10 in 100 files	1	2%
>10 in 100 files	1	2%

34% affirm that they personally encounter bad files, or think they are very common.





4. Do Bad PDF files cause business problems?

Don't know / rather not say	7	11%	
No	12	18%	
A little	26	40%	
Significant	10	15%	
A lot	5	8%	
Bad PDF is a very serious problem	5	8%	

31% say bad files cause "significant" or worse business problems!





5. Do differences in the way various PDF software behave cause problems?

			1
Don't know / rather not say	8	12%	
No	7	11%	
A little	16	25%	
Not after standardizing our software	9	14%	
Significant	11	17%	
A lot	12	18%	
Other (significant)	2	3%	

38% say variations in s/w behavior cause "significant" or worse problems.

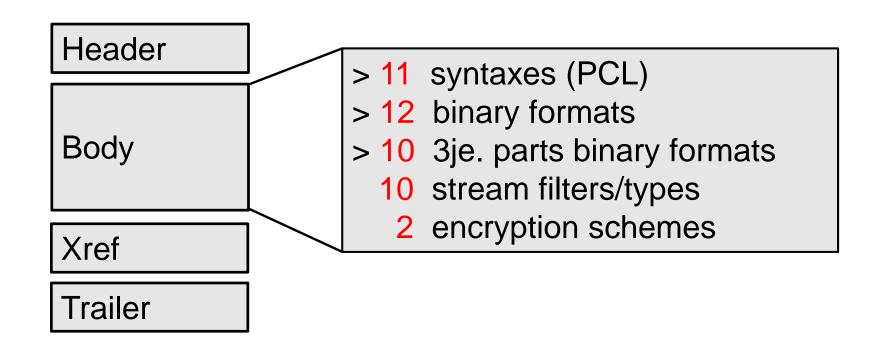
Other: "PDF/A prevents use of form-fields"

"We receive the blame for the PDFs produced by inferior software"





PDF Complexity How much can be validated?





PDF and PDF/A

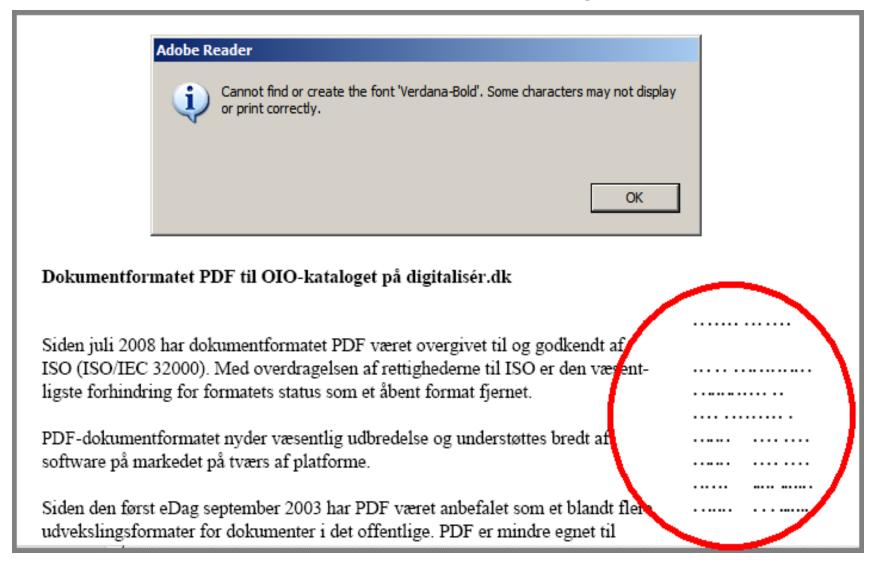
Complexity and hidden risk - examples

- Missing fonts
- File size
- Glyphs but no text
- Layers
- Transparency





Example PDF "missing font"





Example PDF/A "embedded font"

Dokumentformatet PDF til OIO-kataloget på digitalisér.dk

Siden juli 2008 har dokumentformatet PDF været overgivet til og godkendt af ISO (ISO/IEC 32000). Med overdragelsen af rettighederne til ISO er den væsentligste forhindring for formatets status som et åbent format fjernet.

PDF-dokumentformatet nyder væsentlig udbredelse og understøttes bredt af software på markedet på tværs af platforme.

Siden den først eDag september 2003 har PDF været anbefalet som et blandt flere udvekslingsformater for dokumenter i det offentlige. PDF er mindre egnet til publicering på nettet, hvor brug af HTML anbefales.

PDF-dokumenter er som udgangspunkt ikke tilgængelige for handicappede, men ved hensigtsmæssig strukturering af dokumentet kan det blive tilgængeligt. IT- og Telestyrelsens vejledning herom findes på:

http://www.itst.dk/kommunikation-og-

tilgaengelighed/tilgengelighed/verktojer/tilgengelige-pdf-filer/vejledningen-ihtml-udgave/vejledning-om-tilgengelige-pdf-filer. 3. november 2008

IT- og Telestyrelsen

Holsteinsgade 63

2100 København Ø

Telefon 3545 0000

Telefax 3545 0010

E-post itst@itst.dk

Netsted www.itst.dk

Sagsnr.

Dok nr.

Side 1/1



Example file size

Dette er en meget lille fil © Dette er en meget lille fil ©

Dette er en meget lille fil 😊

Dette er en meget lille fil 🕲

Dette er en meget lille fil © Dette er en meget lille fil ©

•	Word	21 kb
•	TIFF Group 4	15 kb
•	PDF plain v 1.7	20 kb
•	PDF embedded font subsets	78 kb
•	PDF/A-1 fully embedded fonts	489 kb











This PDF can't be converted!

DIES & DAS

BADEN-BADEN

- ▼ Adventskalender, mit Geschichten und Liedern, für Gross und Klein, mal lustig, mal besinnlich, Theater am Goetheplatz, 17 Uhr.
- ▼ "Finstere Mächte, dunkle Machenschaften": "In der Hölle: Ewige Verdammnis oder grauer Alltag – Zwischen grässlicher Fiktion und gängiger Seelenpein", Vortrags- und Diskussionsreihe, Stadtbibliothek, Gartenhaus, 19 Uhr.

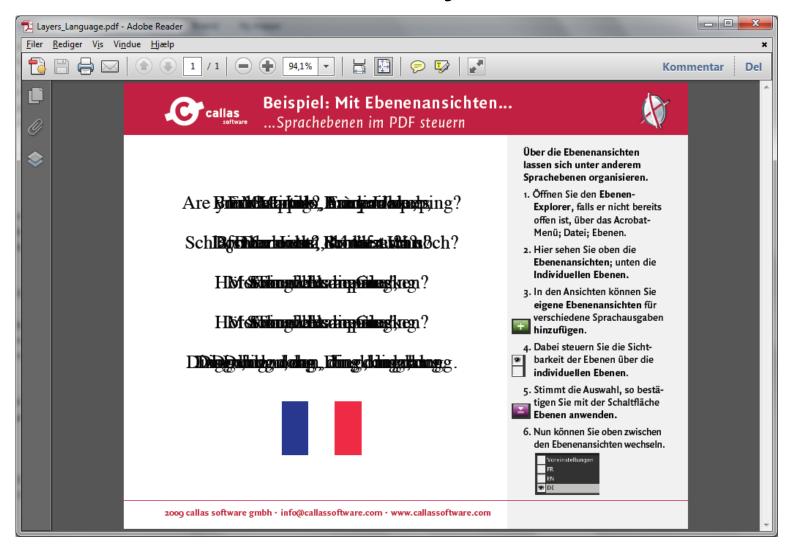
```
)* =
- 8 =
;
= ),$
```





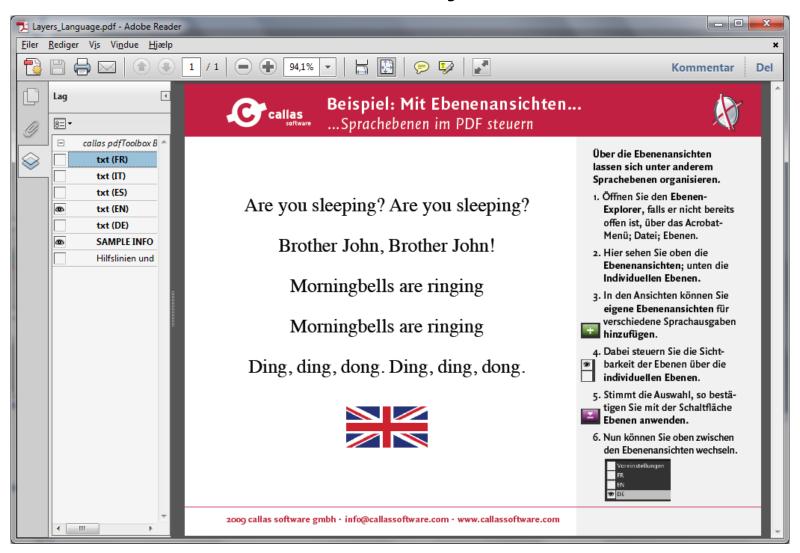


PDF Layers





PDF Layers





PDF transparency

Velkommen til en sjov eftermiddag med 4a.





Så nærmer sommeren sig og vores "små" børn skal snart op i 5.kl. Vi, i klasserådet, synes det er vigtigt at vi som forældre mødes af og til og hygger os sammen.

Derfor inviterer vi jer alle til en eftermiddag på Vigen. Vi håber, at alle vil prioritere dette,

Hvor? Vi mødes på Vigen, ved fodboldbanen/ gyngestativ.

Hvornår? D.15. juni kl.14.00-?

Hvem? Hele familien

Hvad? Planen er, at vi drikker kaffe sammen og spiller/ leger så længe vi har lyst.

Hvad skal I medbringe?

- Medbring kaffe og kage til eget forbrug.
- Evt. penge til en is i kjosken.
- Husk tæpper, badetøj, udelegetøj, bolde osv.

Vi sørger for at arrangere lidt konkurrencer og lege.

Meld tilbage om I kommer, til en af os fra klasserådet. Gerne ugen inden.

Meld også tilbage, hvis I har plads i bilen, så vi sikrer at alle børn kommer med, selvom forældrene ikke kan.

Vi glæder os til at være sammen med jer.

Hilsen klasserådet

Jørn (Malou), Sune (Sille), Betina (Jonathan), Anne (Signe), Kirsten (Nora), Ewa (Frederik), Jesper (Nikolaj), Lene (Mikkel), Signe (August) og Kamilla (Marie).

Velkommen til en sjov eftermiddag med 4a.



Så nærmer sommeren sig og vores "små" børn skal snart op i 5.kl. Vi, i klasserå er vigtigt at vi som forældre mødes af og til og hygger os sammen.

Derfor inviterer vi jer alle til en eftermiddag på Vigen. Vi håber, at alle vil midt i en travl tid.

Hvor? Vi mødes på Vigen, ved fodboldhanen / gyngestativ

tvornar? D. 15. juni kl.

Hvem? Hele famil

Hvad? Planen er, at vi drikker kaffe sammen og spiller/leger så længe vi har lyst

Ivad skal I medbringe?

Medhring kaffe og kage til eget forbrug

Evt. penge til en is i kiosken

Husk tæpper, badetøj, udelegetøj, bolde osv

/i sørger for at arrangere lidt konkurrencer og lege

Meld tilbage om I kommer, til en af os fra klasserådet. Gerne ugen inden.

Meld også tilbage, hvis I har plads i bilen, så vi sikrer at alle børn kommer med, selvom forældrene ikke kan.

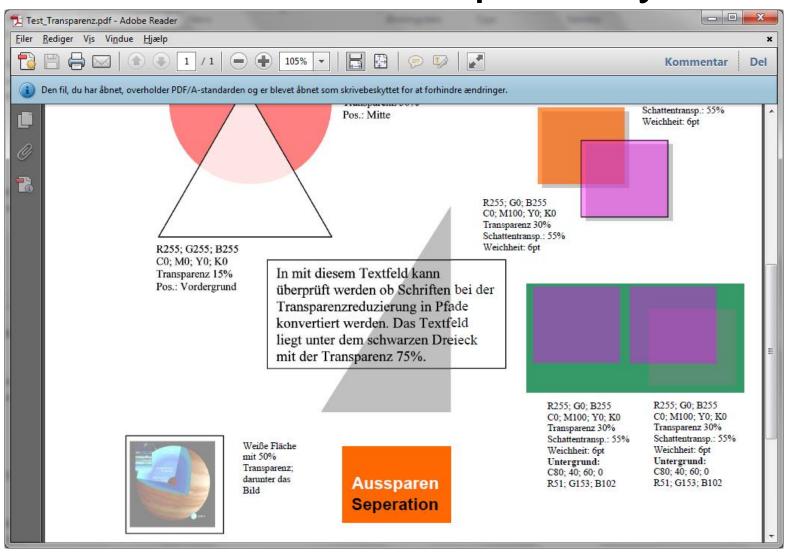
Vi glæder os til at være sammen med jer.

Hilsen klasserådet

Jørn (Malou), Sune (Sille), Betina (Jonathan), Anne (Signe), Kirsten (Nora), Ewa (Frederik), Jesper (Nikolaj), Lene (Mikkel), Signe (August) og Kamilla (Marie).

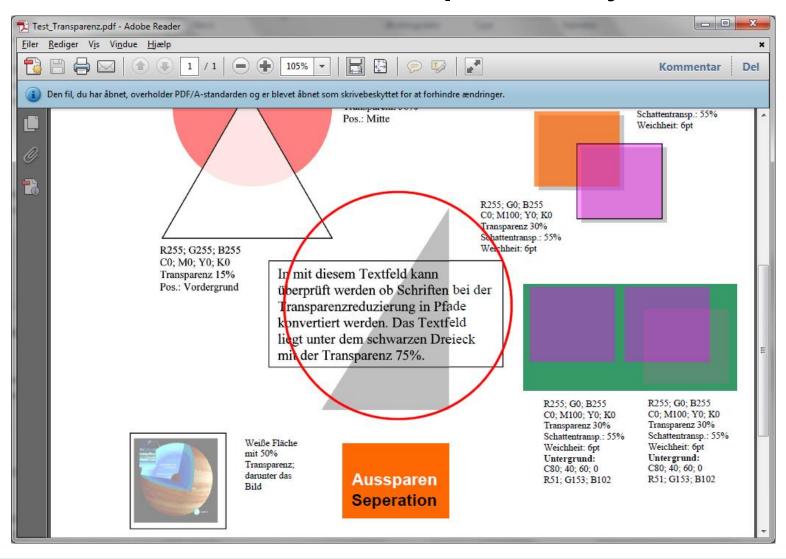


PDF/A-1 Transparency





PDF/A Transparency





PDF/A Transparency





