Richtlijn bestandsformaten

Inleiding

Dit document is een uitwerking van de kennisniveaus zoals beschreven in het preservingsplan. Het is de voorwaarden om uiteindelijk tot functionele preservering[1] te komen. In dit document worden de drie kennisniveaus (opgeslagen, geïdentificeerd, gekend) uit het preservingsplan verder uitgewerkt.

Dit document is voor de KB om een onderdeel van het preservingsplan uit te werken. Daarnaast is de uitwerking van de richtlijn ook belangrijk om te laten zien aan de designated community welke bestandsformaten op welke manier bewaard worden.

Kennisniveaus

Het Digitaal Magazijn bevat heel wat bestandsformaten[2], maar de controle en identificatie van deze bestandsformaten is beperkt. Wanneer voor een project een groot aantal Excel bestanden geanalyseerd werd, bleek dat heel wat van de Excel bestanden in praktijk andere bestandsformaten zijn of zelfs systeemfouten bevatten. Daarom is het belangrijk om beter de bestanden in het Digitaal Magazijn te controleren en te identificeren[3].

Om dit te kunnen doen, is het belangrijk om bestandsformaten beter te kunnen identificeren, controleren, valideren, extraheren van technische metadata,.... Hiervoor is meer kennis over bestandsformaten nodig in de praktijk. Door het grote aantal bestandsformaten, is het niet mogelijk om dit onmiddellijk voor alle bestandsformaten te doen. Daarom wordt er met kennisniveaus gewerkt. Zo kan duidelijk gecommuniceerd worden over de kennis van bestandsformaten en welke formaten opgeslagen worden en welke bewaard worden aan de designated community. Het is mogelijk dat sommige bestandsformaten nooit tot het niveau van gekend bestandsformaat komen, want niet alle bestandsformaten hoeven tot op het hoogste kennisniveau gebracht te worden.

Het moet ook duidelijk zijn dat deze kennisniveaus losstaan van de controles die tijdens ingest gebeuren. Een bestand van het kennisniveau ‘opgeslagen bestandsformaat’ moet nog altijd op 0 byte gecontroleerd worden en de aangeleverde checksum moet nog kloppen. Bit preservation is het minimumniveau voor alle bestanden. Het niveau van gekend bestandsformaat heb je

---

[1] Zie preserveringsplan 2019-2022 voor meer uitleg over de term functional preservation
[2] Er zijn ongeveer 1130 bestandsextensies
[3] De bestandsformaten worden nu vooral geïdentificeerd op basis van bestandsextensie.dit is een onbetrouwbare manier. Een betere manier is om op basis van ‘signature’ bestandsformaten te herkennen. Een signature is data in het bestand dat aangeeft welk formaat een bestand is. Hiervoor zijn tools om te zoeken naar de signatures in bestanden.

nodig om functional preservation uit te voeren. Concreet betekent dit dat er uitgebreidere controles worden uitgevoerd op sommige bestandsformaten. Bij een gekend bestandsformaat wordt er bijvoorbeeld validatiesoftware gebruikt worden om probleembestanden te identificeren in het Digitaal Magazijn. Dat betekent niet dat we geen bestanden met problemen opslaan, maar wel dat de problemen met een bepaald bestand gekend zijn. Er kan dan door Preservation Watch en risicoanalyses onderzocht worden welke acties uitgevoerd kunnen worden om de duurzaamheid te garanderen.

Het bereiken van een volgend kennisniveau zal stapsgewijs gebeuren. Sommige bestandsformaten kunnen al wel geïdentificeerd worden, maar nog niet gevalideerd. Als aan alle voorwaarden voor een kennisniveau voldaan is, zal het kennisniveau stijgen. Deze status zal bijgehouden worden per bestandsformaat en er wordt ook bijgehouden welke stappen nog gezet moeten worden om naar het volgende kennisniveau te stijgen.

1. Kennisniveau ‘opgeslagen bestandsformaat’

Bij opgeslagen bestandsformaten worden enkel checksumcontroles uitgevoerd om te kijken of er geen bit rot optreedt. Er is weinig bekend over het formaat aangezien het bestandsformaat niet geïdentificeerd kan worden[4]. Er is ook geen ondersteuning is in de identificatiesoftware, zoals Droid. Er is dus geen PRONOM ID vastgelegd voor het bestandsformaat. Er is mogelijk wel een mimetype gekend voor het bestandsformaat, maar dit is onvoldoende.

2. Kennisniveau ‘geïdentificeerd bestandsformaat’

Een geïdentificeerd bestandsformaat heeft een PRONOM ID en er is dus ook een tool die bestanden kan herkennen en een PRONOM ID kan toevoegen.
De identificatie met PRONOM wordt gezien als best practice in de digitale duurzaamheidsgemeenschap. Een PRONOM ID geeft ook specifieker[5] aan wat het bestandsformaat is, dan mimetype. Deze specifieke informatie is belangrijk voor de duurzame bewaring en wordt daarom als minimum gezien.

Het is de basisstap voor het duurzaam bewaren. Het is een eerste stap richting functional preservation, maar is nog onvoldoende om van functionele preservation te spreken. Maar is meer dan enkel bit preservation, bit preservation+.

3. Kennisniveau ‘gekend bestandsformaat’

Bij een gekend bestandsformaat is goede bewaring mogelijk aangezien de KB de resultaten van de identificatie, validatie en technische metadata extractie kan interpreteren en richtlijnen heeft opgesteld voor elk formaat.

---

[4] Bestandsextensie wordt niet gezien als een betrouwbare indicator voor het bestandsformaat.
[5] Bijvoorbeeld of een bestand een pdf is versus of een bestand een pdf versie 1.4 is

De stappen van geïdentificeerd naar gekend zullen geleidelijk genomen worden. Zo kan het zijn dat bij sommige bestandsformaten al technische metadata extractie mogelijk is, maar geen validatie. Dan blijft het formaat nog altijd op het kennisniveau van geïdentificeerd bestandsformaat steken.

De voorwaarden om een gekend bestandsformaat te zijn, zijn:

A. Alle voorwaarden van geïdentificeerd bestandsformaat
B. Technische metadata extractie
C. Validatie van bestandsformaten
    a. Classificatie van output van validatiesoftware
D. Documentatie van het bestandsformaat, met o.a.
    a. Ondersteunde applicaties
    b. De leesbaarheid van het bestandsformaat in de leeszaal of voor publiek toegankelijke bestandsformaten ook de leesbaarheid in de browser
    c. Toegevoegd aan Technology watch
    d. Minimale bestandsgrootte is gekend
    e. Er wordt onderzoek gedaan naar risico's voor langetermijnbewaring

Met deze informatie kan het bestand als duurzaam bewaard beschouwd worden en wordt functional preservation uitgevoerd.

Gecomprimeerde archiefformaten

Naast het kennisniveau van bestandsformaten, zijn gecomprimeerde archiefformaten[6], zoals zip bestanden, een speciale categorie. Deze bestanden worden gebruikt om de omvang van opslag van andere bestanden te verkleinen. Op dit moment worden deze bestanden niet uitgepakt, maar opgeslagen zoals aangeleverd. Aangezien het moeilijker is om de bestanden in deze gecomprimeerde archiefformaten op een goede manier te bewaren, worden deze bestanden in de toekomst uitgepakt opgeslagen en komen in principe de gecomprimeerde archiefformaten niet meer voor.

Bijlage 1: lijst opgeslagen, geïdentificeerde en gekende bestandsformaten

Inleiding

Deze bijlage bevat een overzicht van de huidige bestandsformaten ingedeeld volgens kennisniveau.

[6] zie bijlage 2 voor een lijst met de huidige gecomprimeerde archiefformaten in Digitaal Magazijn

Gekende bestandsformaten

Er zijn op dit moment geen gekende bestandsformaten

Geïdentificeerde bestandsformaten

Er zijn op dit moment geen geïdentificeerde bestandsformaten, omdat het toekennen van een PRONOM identifier nog niet mogelijk is.

Er zijn al wel formaten waar we heel wat ervaring mee hebben:

Bestandsformaat	**Voorwaarden**			
	Identificatie	**Technische metadata extractie**	**Validatie**	**Documentatie**
pdf	Ja, Versie	Nee	Nee	Nee
ePub	Ja, Versie	Ja, epubcheck	Ja	Draft richtlijnen over de validatiefoutmeldingen
jp2	Nee	Ja, jplyzer	Ja, jplyzer	Heel wat kennis bij Onderzoek

Opgeslagen bestandsformaten

Opgelet onderstaande lijst zijn de extensie zoals ze in het Digitaal Magazijn opgeslagen zijn. Bestandsextensie is een onbetrouwbare vorm van identificatie, vandaar dat deze lijst fouten bevat. Na de migratie in het nieuwe Digitaal Magazijn kan deze lijst geüpdatet worden en betrouwbaar gemaakt worden.
Verder onderzoek is enkele jaren geleden door Johan van der Knijff uitgevoerd op deze informatie[7].

gif	xml	jpg	csv	tif	doc	arc
sml	raw	oa3	htm	wav	mp3	docx
txt	bmp	swf	xls	lmx	zip	class
epq	js	mp4	xlsx	rtf	suppl	mov
png	abg	avi	ppt	dat	mpg	aif
cab	page	exe	pptx	dll	ini	dib

[7] Johan van der Knijff, 2015 en 2016, Quickscan bestandsformaten Digitaal Magazijn

x32	db	fig	inf	drv	phd	asc
swa	onx	kmz	ani	dxr	css	eps
vxd	eks	format	tar	lcp	epv	icns
flo	kml	bmr	wmv	hed	bak	bin
jar	pge	pm	api	flp	template	pic
ico	mht	hdr	cat	lst	fba	gz
fb0	wri	cxt	mno	px	al	nls
val	pfb	pfm	tst	grf	xg0	yg0
cct	abt	ex_	hlp	log	lib	ins
grx	xg1	yg1	mol	ps	utx	xtu
pdb	rar	ddd	did	cer	nld	dig
wma	mdb	pl	pdd	ttf	nwk	dcr
dbf	new	wld	ccj	fbd	avx	fas
inx	h	dir	psd	r	boot	cdx
pkg	stp	pod	cnt	fpt	dl_	epm
wassessment	wquestion	ocx	bundle	svg	config	id
vus	dxf	idx	mb	cst	pdx	plist
gid	tag	tre	lng	vbx	htc	fbf
lid	a4r	fbe	sgm	url	xg2	yg2
lic	std	fasta	rsd	rst	bat	lsr
syd	syx	cmp	msf	nex	tgz	m
bs	dwg	exp	x	msi	ndx	tpl
m1v	fla	x16	bed	trn	z	tar
plc	scm	bz2	000	cfg	dbr	bik
thd	cif	grp	tlb	hw4	tab	ddb
fbc	properties	hwn	otf	xg3	yg3	fbg
stc	fast	tbk	tsv	1	pse	apl
reg	iss	ent	jbf	mmm	chm	fmt
xg4	yg4	dot	asm	m4v	wmz	cnf
ods	fbp	lb0	ldb	c	fbs	src
xg5	yg5	aicon6	dbd	dic	qtc	mst
enc	lix	lsf	cs	wpl	mpp	phy
pps	qdat	vmp	xg6	yg6	fa	mso
aln	cys	f6p	rcm	fbj	fst	lnk

py	xrf	cdt	fdt	ifp	l01	l02
nib	n01	n02	pft	conf	dah	dal
cyto	f6s	oc_	xg7	xg8	yg7	yg8
dtd	dxx	init	its	prj	and	ctf
cur	lba	wcm	wpd	cac	hts	prc
a05	a09	sif	bag	dbk	ibk	old
ssh	xg9	xpi	yg9	a07	fbx	gtr
mml	pir	psp	rm	wrd	xsl	a01
a10	enl	fot	gpr	inc	nav	olb
sas	wig	fbv	map	msg	01	a08
di_	dil	fbu	mdl	nb	sfk	cgi
daw	fbm	hqx	md8	rsrc	six	tan
tbl	u32	xib	acv	cpr	csp	fbh
gpk	ipj	ix	mnu	mop	out	rcs
sps	wmf	xga	xgb	xgc	xgd	xge
yga	ygb	ygc	ygd	yge	7z	atr
ex2	rt_	asf	bngl	btr	cyp	dmg
dsp	fbn	f60	hsc	lck	mat	nxs
qt	zdt	386	art	as	dict	dis
emf	ind	mac	par	res	sdf	tmb
trt	typ	xmo	xsd	arv	bst	ckb
ckw	cty	do	epb	fam	fo_	fts
hmm	idl	lex	md_	mdx	ovr	sid
sld	tit	tt_	tv	vb_	wip	afa
apm	a03	cp2	dct	dlx	fby	fit
fon	fpr	frm	gff	gtf	hpf	ihe
irt	isu	oud	sta	the	vst	v12
wdz	xyz	9	bdef	btl	en1	en3
fbb	fbi	fdb	fna	gpd	mcd	ohe
ort	owl	ply	ptt	rdata	r32	sbk
shs	ssi	sto	tex	tfw	toc	uc
w32	xgf	ygf	118	abs	au	bas
cel	chr	cmd	cpp	cwt	en5	fbq
fbr	flt	fsa	f6q	gml	hsh	hwt

img	lsp	mca	mid	mom	msk	net
noi	opx	osx	ppz	shlb	shlib	smi
tl_	txtn	vbs	vdb	2	3	4
warc	aa	acm	ac1	ac3	addi	akc
ande	bkg	cdi	cfm	cga	cpl	cps
cp1	dab	dbe	dc1	dc3	dft	d2i
d32	ega	eml	fbt	fbw	fbz	f6r
gbf	gp	he0	he2	he4	his	hr2
hw2	iaa	ic1	ic2	ic3	idb	idh
i2n	java	ld	lo	lo1	lo3	mega
mod	mrg	mswm	mw1	mw2	mx	_n_
ngm	null	nw1	obj	oc1	oc3	odp
pag	pcx	pem	prm	pub	pzl	rdat
read	ref	rsm	sbml	sfl	sh	spf
srg	thb	thh	thw	tnt	tree	ttx
twl	tw1	tw2	ufl	uni	vga	wa_
wp5	xtr	0	ali	bbk	blb	bsc
cdr	daf	dbz	ddf	dls	dwt	end
esp	fld	frx	gb	ghe	gms	gmx
gps	grt	hl_	hli	hwa	in_	isk
kix	ls	lvl	lxt	meg	mmp	mpd
mpl	mws	m8cs	nec	nfo	obd	ol_
plm	pp1	re_	rsr	seq	spk	sql
start	stockholm	sys	x86	eps	nwk	3gp
38_	adm	ahe	alb	all	ana	asp
gz	bhe	bloc	bmf	bmk	bnm	bok
brt	bsp	b6p	cdb	ckt	classic	de_
dendro	diff	drl	dsw	dta	edm	eng
expression	faa	fastq	fbk	fda	fdi	fpi
ftp	gct	ged	gex	gi_	gpml	graphml
hhc	hhk	hhs	hta	htv	hwi	hwr
hwy	idp	jp_	jrs	jsp	jtv	krl
krs	layout	lp	lwf	mad	mail	man
matr	~mc	meta	mhe	mm	mpc	mrt

mta	mts	mw	mzid	m12	nexml	obo
odc	ode	odi	odt	off	opa	_pa
paml	path	ped	pgt	ph	phylip	pif
plt	pml	pot	ppsx	qda	rel	rxc
sav	scop	slb	slw	smm	sq1	stf
stx	sty	tdt	td1	toe	trl	trx
types	vcf	vgw	vmo	web	wrl	w02
w03	xa	xfasta	xlsm	01	64	aaa
aab	accdb	ace	ade	ado	af2	ai
amu	apf	apr	apr2	aps	apt	arb
asu	asx	aw	bash	bbb	bio	blib
brd	brh	brl	cath	ccd	cf_	ch_
chip	clg	cli	cls	clus	cm_	cmdx
cmmcd	cmt	cod	col	com	core	cp
cpd	cr	ctm	cvb	c21	c32	daa
dac	dad	dae	dai	daj	data	dau
dav	dax	day	daz	dba	dbb	dbc
dbg	dbh	dbi	dbj	dbl	dbm	dbn
dbo	dbp	dbq	dbs	dbt	dbu	dbv
dbw	dbx	dby	db1	db4	db5	dca
dcb	dcc	dcd	dcl	def	del	_dh
dia	_dj	_dl	drw	dr1	dr4	dr5
ds	dsj	dsl	dtx	dum	dump	_dv
egt	embl	emz	en_	en2	en7	epg
eug	ext	fat	fbo	fdx	flg	flx
fnd	fnt	fpkm-tracking	functions	f4v	gal	gbk
gdb	gdl	genbank	gens	gff3	gls	gmt
gnum	gobpm	gosfunction	gpc	gvw	hdb	hi_
hst	hs6	htmla	hyp	ia	ias	ijm
ion	is	jac	jjj	key	kgg	kin
kit	kls	lbb	ld_	lhs	lisp	list
local	lsrh	lx1	lx2	lzh	mak	manifest
mase	ma4	mbo	mbp	mbq	mbr	mbs

mbt	mbu	mbv	mbw	mbx	mby	mbz
mcb	mcc	mch	mct	me	mfw	mmd
mng	mnt	mor	mpg4	ms_	msb	msh
mus	mxp	nab	name	nbp	newi	nex
nexu	nexus	nh	nhx	nlog	nlogo	no_
not	npr	nt_	ntt	nvw	nw	n3
obi	odb	ol2	opf	pak	pau	pct
perl	pgn	phb	phyloxml	phyt	pict	pid
pk	pop	pos	ppj	prf	prt	ptn
pux	p01	qst	qtif	qtx	raw1	rda
rlx	rnw	rpt	scb	scc	sdq	seq58
sit	slv	ss	stl	str	_sv	sxw
sym	syt	s1	tage	tbr	tcl	td_
tde	tdf	td2	tnc	tok	top	tsk
tt	twf	tx_	_u_f_l	uml	usn	utd
vb	vdp	voi	vxr	wll	wpr	wtr
xgmml	xmfa	zon	ztz	FEXT	003	004
mol2	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	5	6	7	8	pse
onx	kmz	kml				

Bijlage 2: gecomprimeerde archiefformaten

zip
rar
7z
tar
gz
lzh
bz2
tgz
sit