SND Svensk Nationell Datatjänst

Dokument och Digital Text: En guide till god hantering

2017-03-13

Översatt, omarbetat och kompletterat av Ulf Jakobsson

PDF/A-1 (ISO 19005-1), skapat i Microsoft Office 2010 från formatet “docx”

Dokument och digital text: En guide till god hantering

Översatt från Archaeology Data Service's "Documents and Digital Texts: A Guide to Good Practice" (http://guides.archaeologydataservice.ac.uk/g2gp/Main), varefter det har omarbetats och kompletterats för att även passa andra datamaterial med annat ursprung än arkeologi.

Innehåll

1. Introduktion till dokument och text

De vanligaste filtyperna som skapas under ett forskningsprojekt är olika typer av dokument och textfiler. Oavsett typ av forskningsprojekt så kommer, om inte annat, en slutrapport i form av ett textdokument att skrivas. Bortsett från olika rapporter skapas normalt sett dokument som beskriver olika arbetsprocesser, enkäter, metadatadokument osv.

Denna guide är tänkt att ge en överblick över olika typer av binära[1] och rena textfiler[2]. Förutom att gå igenom vanliga filtyper och filtyper lämpade för långtidsbevaring, kommer denna guide ta upp de olika element/objekt som ett dokument består av och hur olika sätt att skapa element påverkar dessa. Vi kommer också gå igenom vilken strategi för långtidsbevaring man kan använda sig av för att säkerställa att dessa bibehålls efter förberedelser för långtidsbevaring.

1.1 Vad är dokument och text?

Förenklat kan majoriteten av textdokument variera i storlek och komplexitet från enkla rapporter och korta uppsatser till betydligt större dokument som avhandlingar eller böcker. Textfilerna består huvudsakligen av strukturerad text (meningar, stycken, sidor, kapitel) men inkluderar ofta andra element som bilder, figurer och tabulerad data.

Textdokument kan skapas på flera olika sätt men de flesta skapas via olika ordbehandlingsprogram som t.ex. Microsoft Word eller OpenOffice-baserade program (LibreOffice, Apache OpenOffice, NeoOffice etc.). Tittar man på filformaten så har dokument skapade med ordbehandlingsprogram tidigare huvudsakligen sparats i proprietära[3] binära format. Dock har de senaste programpaketen för ordbehandling som Microsoft Word 2007 och OpenOffice visat på en förändring mot xml-baserade format och standarder ämnade för mänsklig läsning som .docx (Office Open XML[4] format) och .odt (OpenDocument[5] format). Förutom dessa format sparas många slutversioner av dokument (av användaren) i systemoberoende format, vanligen Adobe's Portable Document Format[6] (PDF), vilket möjliggör att formatet och strukturen i dokumentet bibehålls oavsett plattform men också förhindrar möjligheten att redigera dokumentet.

Förutom att dokument skapas via ordbehandlingsprogram, skapas en stor del dokument som ett resultat av en digitalisering. Digitalisering av facktidskrifter i syfte att bevara eller tillgängliggöra samlingar skapade före den digitala åldern är oftast den största källan till digitala texter bortsett från de som skapas med ett ordbehandlingsprogram. Denna process börjar normalt sett med en digitaliserad bild av pappersdokumentet som sedan behandlas med hjälp av ett program för

---

[1] **Binärfil**, en fil som innehåller data i ett format avsett att läsas av specifika datorprogram, med en väsentlig del av informationen kodad som annat än text. Binärfiler är i allmänhet inte möjliga att tolka utan kännedom om filformatet, annat än möjligen till vissa delar.

[2] En **textfil** är en fil som enbart innehåller text. Texten kan vara sparad i olika teckenkodningar (som UTF-8 eller ISO 8859) och har ofta filändelsen .txt. Det som skiljer en textfil från övriga filformat är att den är avsedd att kunna läsas och vara begriplig som sådan (förutsatt att man använder rätt teckenkodning), utan att man använder något specifikt program. Filen bör inte innehålla kontrolltecken annat än sådana med vedertagen betydelse, såsom radbyten.

[3] **Proprietär** programvara är programvara som har restriktioner (vanligtvis satta av ägaren) vad gäller att använda, modifiera eller kopiera den.

[4] http://www.ecma-international.org/publications/standards/Ecma-376.htm

[5] http://xml.openoffice.org/

[6] http://www.adobe.com/products/acrobat/adobepdf.html

maskinläsning/optisk teckenläsning (OCR) för att på det sättet förvandla bilden till ”riktig” (redigerbar/sökbar osv.) text. Den slutliga texten, som kan innehålla bilder och tabeller m.m., lagras normalt i PDF-format även om användandet av ett XML-baserat format blir mer och mer vanligt, särskilt när texten publiceras online.

Inom mjukvaruutvecklingen har man riktat in sig mot XML-baserade öppna format, som docx och odt, eftersom man vill standardisera formaten så att olika mjukvarusystem kan använda sig av varandras format. Liknande problem fanns för PDF-formatet (inkompabilitet), varför den öppna standarden med PDF/A[7] togs fram för att lösa dessa problem och möjliggöra långtidslagring (se sid 10 om pdf/a-1). Bortsett från de vanligaste ordbehandlingsformaten och PDF, kan text existera i ett stort antal varianter av ren text eller uppmärkta format som t.ex. SGML, HTML och XML.

[7] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50655

2. Att tänka på när man skapar texter och dokument

2.1 Allmänna överväganden

Olika textfiler och dokument är normalt sett en del av mängden filer som skapas i ett projekt och används ofta även i dokumentationen av projektplaneringen samt i olika steg av projektet. Även om det är förhållandevis lite planering som behövs för att kunna hantera denna typ av filer finns det några saker man bör tänka på:

- Det är viktigt att infogat material i textdokument, som bilder och kalkylblad, lagras separat tillsammans med dokumentet. Bortsett från att detta säkerställer att filerna lagras i sitt ursprungsformat alternativt i ett arkivvänligt format, möjliggör det dessutom att de bibehåller sin ursprungliga upplösning och storlek.
- Undvik att ta med länkar och dynamiskt innehåll[8], eftersom länkar mm kan sluta gälla med tiden.
- Om tanken är att slutversionen av dokumentet ska sparas i PDF, se till att det ursprungliga formatet (t.ex. Word .docx, OpenOffice .odt, etc.) också sparas separat.
- Om dokument sparas som PDF, säkerställ att det sparas i PDF/A.[9] Här gäller också att filen inte ska användarskyddas, inte inkludera javascript/video/ljud och liknande samt att fonter och bilder är korrekt infogade. Se också till att filen är korrekt uppmärkt[10].
- Säkerställ att icke-proprietära fonter används och då UTF-8 kodning. Detta för att dokumentets utseende kan ändras om det öppnas av någon som ej har tillgång till den specifika fonten. Anledningen till att man bör använda UTF-8 kodning är att den innehåller majoriteten av tecken (bokstäver och siffror mm).

Generellt sett; Det är viktigt att se till att all information i dokumentet är fullständig och förståelig. Källor ska vara citerade osv.

Undvik infogat material

En av de vanligaste typerna av infogat material i textdokument är bilder. För vissa typer av dokument (t.ex. Microsoft Word och PDF) kan mer komplexa material (t.ex. kalkylblad och videos) infogas och då i format som bör bevaras separat tillsammans med textdokumentet. Rekommendationen är därför att förutom att infoga materialet i textdokumenten så ska dessa filer lagras och bevaras separat (i enlighet med respektive filformat) för att på det sättet säkerställa att filerna bibehåller ursprungskvalitén (t.ex. bildupplösning).

---

[8] Att skapa ett **dynamiskt innehåll** kan t.ex. vara att länka innehållet från en Excel-fil till en tabell i ett Word-dokument. När man ändrar innehållet i Excel-filen ändras innehållet i tabellen. Till skillnad från ett infogat objekt, där man kopierar in värdena till tabellen, lagras informationen för ett objekt med dynamiskt innehåll i en separat fil. Problemet med detta är att om man flyttar filerna bryts länken mellan Excel-filen och tabellen i Word-dokumentet.

[9] I vissa program går det enbart att spara i PDF varvid man senare får spara om filen som PDF/A genom att välja ”spara som” och sedan välja formatet PDF/A.
För att spara i PDF/A-1 från Microsoft Word 2010 välj Spara som (eng. Save as type), välj PDF i listan, klicka på knappen Alternativ (eng. Options…) och för PDF alternativ (eng. PDF options) välj ISO 19005-1 compliant (PDF/A).

[10] **Uppmärkning** (eng: markup) innebär att man förser text eller andra filer med instruktioner eller extra information för att underlätta automatisk hantering och informationssökning. Instruktioner gäller oftast textens grafiska utseende (stilsättning) och uppställning. Extra information brukar vara nyckelord (metadata) som läggs in för att underlätta automatiska sökningar.

3. Arkivering av texter och dokument

3.1 Vilka filer ska arkiveras

Som nämnts ovan så sker filformatsförändringar för dokument inte i någon högre grad under framtagandeprocessen. Detta med undantag av pdf-filer vilka ofta skapas i slutet av denna process i syfte att användas för tillgängliggörande/spridning av materialet. Det är dock att rekommendera att ursprungsfilen sparas parallellt med pdf-filen. Infogat material som bilder bör också lagras separat tillsammans med textdokumentet för att säkerställa att de bevaras på bästa sätt i enlighet med filtyp.

Var noga med att det som ska bevaras verkligen är den slutliga versionen av dokumentet. För att underlätta det så är det bra att hålla ordning på alla olika arbetsversioner, men också att man för slutversionen tar bort anteckningar/kommentarer och så vidare.

3.2 Hur ska det arkiveras

När man ska besluta vilket filformat som ska användas för långtidsbevaring av dokument så är det bra att välja ett format som både bevarar viktiga egenskaper i dokumentet samtidigt som formatet bör vara vanligt förekommande och, om det är nödvändigt, kunna migreras[11] av olika applikationer.

Viktiga egenskaper

Viktiga egenskaper, dvs. de grundläggande element i texter och dokument som ska bevaras och underhållas, beskrivs nedan:

- Ord och ordföljd i dokumentet
- Den hierarkiska strukturen i dokumentet (t.ex. olika rubriknivåer)
- Formateringen inom dokumentet (t.ex. fetstil, kursiv stil)
- Sidnumreringen av dokumentet. Detta är viktigt om dokumentet är en rapport eller en avhandling, publicerat eller ej. Om en användare vill citera och referera till dokumentet så måste sidangivelsen vara korrekt. Det gäller att vara extra observant om dokumentet migreras ett flertal gånger
- Infogat material, som bilder och datatabeller. Säkerställ att de bevaras separat

Det finns även egenskaper som inte alltid ses som viktiga att bevara. Däribland font-typ och fontstorlek (förutsatt att det inte påverkar formatering och sidbrytning) samt funktionen för att Spåra ändringar.

Viktiga egenskaper i ett dokument kan dock förändras beroende på dokumentet som ska bevaras. Oavsett så bör man vid genomgång av ett dokument som ska långtidslagras bedöma vilka av ovanstående element som måste bevaras.

Filformat för långtidslagring

Tittar man på filformat för långtidslagring och arkivering finns det idag en generell rekommendation att använda sig av standardiserade XML-format som Microsofts OOXML (docx) och OpenOffice ODF (odt). En teknisk rapport från JISC, 'XML-based Office Document Standards' (Ditch 2007)[12], går igenom och jämför de olika specifikationerna. Den största fördelen med båda formaten är att de är

[11] **Migrering** kan innebära överföring mellan olika medier men också överföring mellan olika filformat.

[12] http://www.webarchive.org.uk/wayback/archive/20140615220449/http://www.jisc.ac.uk/media/documents/techwatch/tsw0702pdf.pdf

internationellt erkända öppna standarder och textbaserade (till skillnad från binära filer, vilka är enbart maskinläsbara) och därigenom även avsedda för mänsklig läsning. Båda formaten är ömsesidigt accepterade liksom även accepterade av ett antal tredjepartslösningar som Google Docs. Formaten liknar varandra i det att de använder sig av ett zippat arkivformat där de olika delarna av dokumenten lagras separat för att tillsammans bilda en fil.

ODF använder sig på ett bättre sätt av öppna och existerande standarder som t.ex. SVG (Scalable Vector Graphics). Den tillgängliga dokumentationen av ODF är betydligt kortare, och möjligen också mer komplett, än OOXML vilket kan innebära att tredjepartsstödet för formatet snabbare kommer att spridas. Microsofts OOXML har däremot ett bättre stöd för tidigare versioner av MS Word, då bakåtkompabiliteten var ett av syftena med framtagandet av standarden. Vid konvertering från MS Word till ODF blir formatet inte helt korrekt. Bland annat så fungerar inte konverteringar av grafiska element fullt ut på grund av inkompabilitet mellan formaten.

Som komplement till dessa XML-baserade format så kan PDF/A vara ett potentiellt format för långtidslagring men då i huvudsak för dokument som annars bara existerar i PDF-format. Även om PDF/A är ett binärt format så är det en öppen standard där programvara för att kunna läsa filerna är gratis och lätt att hitta, bland annat genom ett ökat tredjepartsstöd. Eftersom uttag ur eller migrering från PDF-dokument till andra format är problematisk så erbjuder PDF/A ett bra och noggrant sätt att bevara befintliga PDF-material i ett känt öppet standardformat, även om det är binärt.

Det finns även ett par andra och mer generella problemområden som gäller för textdokument och dess långtidsbevaring. Det första, vilket också gäller många andra filformat, är den kontinuerliga förändring som sker med de filformat som används i ordbehandlingsprogram. Det andra problemet gäller filformat där mjukvaran inte längre finns. Ytterligare ett uppstår när utveckling och förbättring av filformat, som används av existerande ordbehandlingsprogram, resulterar i inkompabilitet mellan äldre filversioner och nuvarande version av mjukvaran.

3.3 Metadata och dokumentation

Fastän många dokument och texter är självförklarande så bör man för sökbarhetens och historikens skull dokumentera viss metadata för varje enskilt dokument eller grupp av dokument. Det finns ett stort antal olika metadatastandarder som har utvecklats i olika ämnesområden, t.ex. MARC[13] inom biblioteksområdet. Den typ av metadata som behövs för att kunna göra sökningar och ursprungskontroll, till skillnad från teknisk metadata om filen som dokumenteras med format som textMD[14], beskrivs nedan och upptar bara ett minimum av metadataposter vilka också existerar i de flesta metadatastandarderna. Bland dessa ingår bl.a. de 15 element som finns i Dublin Core Metadata Element Set[15]. För dokument är det vanligt att viss metadata överlappar med metadata för projektet (t.ex. för rapporter som beskriver projektet).

[13] http://www.loc.gov/marc/
[14] http://www.loc.gov/standards/textMD/
[15] http://dublincore.org/documents/dces/

Tabellen nedan tar upp information som bör finnas för varje dokument/grupp av dokument:

Element	Beskrivning
Titel	Dokumentets titel.
Abstract/ Sammanfattning	Kort beskrivning/sammanfattning av dokumentet.
Publikationsdatum	Publikationsår.
Publicerad	Fullständig referens för publikation. Serie/tidskrift, utgåva, upplaga, start och slutsida eller antal sidor mm bör noteras.
Förlag eller motsvarande	Uppgift om förlaget eller motsvarande (namn, plats etc).
ISBN	ISBN[16], (där sådan finns).
DOI	Digital Object Identifier (DOI)[17] (där sådan finns).
URL	URL (där sådan finns).
Relaterat material	Information om relaterat material såsom filer, databaser och annat material.
Språk	Ange det språk dokumentet är skrivet på.
Författare	Namn på författare.
Medarbetare	Namn på medarbetare.
E-post	E-post till författare/kontaktperson.

Bortsett från de element som angivits ovan finns det ytterligare ett antal element (se nedan) som bör registreras på dokumentnivå. Några av dessa element kan även gälla för en grupp dokument vilka skapats på projektnivå. Metadata kan därför finnas redan på den nivån, men det är ändå rekommenderat att dessa element även registreras för varje dokument:

[16] http://www.isbn-international.org/
[17] http://www.doi.org/

Element	Beskrivning
Projektnamn	Namn på tillhörande projekt, samt ev delprojekt.
Ämne	Ange ämne för projektet. Mappa mot DC subject eller annan kontrollerad vokabulär.
Typ av undersökning	Ange hur undersökningen genomförts t.ex. via enkätundersökning, observation, experiment, fältundersökning osv.
Geografisk täckning	Församling, kommun, län, socken, landskap, land osv.
Tidsperiod för undersökningen	Nyckelord för tidsperioder, start/slutdatum för undersökning, datering av material (t.ex. C14).

4 Filformat

Tabellen nedan syftar till att ta upp några vanliga filformat och tillhörande applikationer samt lite information om dem och deras potential för att användas för långtidsbevaring av dokument.

Adobe PDF	
Filformat/-ändelse	PDF/.pdf
Format	PDF (Portable Document Format), skapat av Adobe, är huvudsakligen en öppen standard för överföring mellan olika system. Formatet har utvecklats flera gånger sedan det skapades första gången och det finns ett antal varianter däribland PDF/A.
Beskrivning	PDF är ett format ämnat att användas över olika plattformar. Även om det är ett proprietärt och binärt format så fungerar det mycket bra för att tillgängliggöra material då det är designat för att bibehålla formatet på ursprungsdokumentet. Formatet kan förutom att visa upp vanlig text även innehålla ett stort antal olika infogade filformat eller länkad media inklusive raster och vektorgrafik, JavaScript och även 3D-data. PDF-filer kan skyddas så att det inte går att redigera alternativt att skriva ut dokumentet.
Rekommendationer	Även om pdf är en öppen standard, så är formatet proprietärt och binärt vilket gör att det inte är lämpligt att använda för deponering eller långtidslagring.[18] I de flesta fall skapas pdf-filer utifrån andra filformat (t.ex. docx). Det är bättre att deponera och långtidslagra ursprungsformatet, men där det inte är möjligt så föredras PDF/A. Man måste då kontrollera att allt infogat material kommer med på rätt sätt. Om formatet ska användas så bör man säkerställa att funktioner som textsökning, inbäddade fonter, icke-förstörande komprimering (lossless compression), högupplösta bilder, standardiserad information om färger[19] samt taggning av innehållet inkluderas i dokumentet.[20]
Filformat/-ändelse	PDF/A / .pdf/a-1
Format	Formatet är baserat på PDF-formatet och skapat av Adobe. PDF/A togs fram som en öppen standard för att kunna användas för långtidslagring.

[18] För ytterligare läsning om problem vid användandet av pdf för långtidslagring läs rapporten 'Preserving the Data Explosion: Using PDF' (Fanning 2008), skriven av DigitalPreservationCoalition (DPC). Se även http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml samt https://riksarkivet.se/Media/pdf-filer/pdf-a-07-08-10.pdf s2.

[19] Eng: device-independent specifications of colorspace, d.v.s. standardtermer för att beskriva färger som ljushet (brightness), kulörton (hue), mättnad (saturation) och intensitet (intensity).

[20] http://www.loc.gov/preservation/resources/rfs/textmus.html

Beskrivning	PDF/A är baserat på PDF version 1.4 och syftar till att fungera som ett tillförlitligt öppet standardformat för långtidsbevaring. Formatet är en ISO standard (ISO 19005-1:2005[21]). Det finns två olika nivåer för PDF/A-dokument. Grundnivån (PDF/A1-b) där nivån ska säkerställa att det visuellt ska kunna reproduceras över tid, medan PDF/A1-a dessutom ska inkludera uppmärkning samt sökbar dokumentstruktur. För ytterligare information om de olika PDF/A-formaten så har en bra överblick skrivits av Bentley Historical Library[22] vid University of Michigan.
Rekommendationer	PDF/A har blivit accepterat som ett fungerande format för långtidsbevaring (bl.a. av Library of Congress[23] samt Riksarkivet[24]) och har blivit genomgånget och bedömt av DigitalPreservationCoalition (DPC)[25]. Det är rekommenderat att filer som har annat ursprung (t.ex. .doc eller .odt) behålls och sparas parallellt med PDF/A-filen. För att kunna ses som ett säkert format för långtidsbevaring krävs det att vissa element (fonter och färger) är specificerade eller infogade i filen, medan andra element (javascript, ljud, video, krypterat material) inte får finnas.

Microsoft Word

Filformat/-ändelse	DOC/.doc
Format	Ett proprietärt binärt format för Microsoft Word.
Beskrivning	Ett populärt filformat och standardformatet för alla versioner av MS Word från 1.0–6.0, 95 och 97-2003. Filerna kan även läsas av OpenOffice samt konverteras till .pdf. Även om bakåtkompabilitet har funnits mellan olika versioner av Word, så har man med service pack 3 tagit bort stödet för version 2.0 och tidigare. Efter 2008 har specifikationerna för ett antal av Microsofts binära filformat tillgängliggjorts på Microsofts websida samt British Library[26].
Rekommendationer	Även om filformatet inte är lämpligt som arkivformat eller som format för tillgängliggörande av material så är det så pass vanligt att det inte föreligger hinder att använda.

[21] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920
[22] http://bentley.umich.edu/uarphome/bestprac/pdfafaqs.php
[23] http://www.digitalpreservation.gov/formats/fdd/fdd000125.shtml
[24] https://riksarkivet.se/Media/pdf-filer/pdf-a-07-08-10.pdf, för ytterligare information från Riksarkivet gällande regler om PDF/A: https://riksarkivet.se/pdfa
[25] 'Preserving the Data Explosion: Using PDF' (Fanning 2008), skriven av DigitalPreservationCoalition (DPC)
[26] http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/formats/index.html

Filformat/-ändelse	DOCX/.docx
Format	Del av Office Open XML (OOXML) formatet som Microsoft skapat. En ECMA (ECMA-376[27]) och ISO (ISO/IEC 29500-1:2008[28]) standard.
Beskrivning	Filformatet introducerades med Office 2007. Microsoft valde att utveckla en egen standard (OOXML) istället för att använda sig av standarden ODF (ISO/IEC 26300:2006, se ODT nedan) för att därigenom skapa bättre förutsättningar för bakåtkompabilitet med tidigare versioner av MS Word-filformat. Formatet består av läsbara XML-filer vilka är packade tillsammans med andra filer (bilder mm) i ett zippat arkiv.[29]
Rekommendationer	Lämpligt för deponering, tillgängliggörande och långtidsbevaring, men infogat material bör lagras separat. Då filformatet är byggt som ett zippat arkiv bör det lagras i okomprimerat format.[30]

Oformaterad text

Filformat/-ändelse	TXT/.txt och oformaterad text
Format	Enkel oformaterad textfil. Skiljer sig från formaterad text där formatmallar inkluderas, samt binära filer där del av informationen är kodad. Oformaterade textfiler är grunden för uppmärkta texter (se nedan).
Beskrivning	Enkla oformaterade textfiler är det enklaste formatet för textinformation och är kompatibelt över ett stort antal plattformar och mjukvaror. Eftersom formatet knappt stödjer någon form av formatering bör det endast användas till den enklaste formen av dokument. Man bör för alla varianter ange någon form av kodning (ASCII eller UNICODE)[31].
Rekommendationer	Fungerar bra för långtidslagring och tillgängliggörande av textfiler men endast om det är enkla filer.

[27] http://www.ecma-international.org/publications/standards/Ecma-376.htm
[28] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=51463
[29] För att packa upp filen: Byt filsuffix från .doc till .zip. Packa upp zip-filen. Resultatet blir en xml-fil samt ett antal mappar innehållandes fler mappar samt xml-filer. Själva texten finns i mappen ”word”, i dokumentet ”document.xml”. http://www.digitalpreservation.gov/formats/fdd/fdd000397.shtml
[30] MS Office 2007/2010 stödjer ej ODF 1.2-formatet vilket gör att man kan få felmeddelande om man försöker öppna ett odf-dokument i t.ex. ms word. Konvertera ODF till 1.0/1.1 så ska det fungera. Office 2013 stödjer ODT 1.2-format.
[31] För ytterligare information om oformaterad text se AHDS “Preservation Manual on Plain Text”: http://ota.ahds.ac.uk/documents/index.xml

OpenDocument Text	
Filformat/-ändelse	ODT/.odt
Format	Open Document Text[32] är ett av formaten som tagits fram som en del av OpenDocument Formatet, en ISO standard (ISO/IEC 26300:2006[33]) för XML-baserade dokumentformat. Formatet stöds och används av flera olika kontorsapplikationer.
Beskrivning	Liksom .docx består .odt i huvudsak av XML-filer, ämnade för mänsklig läsning, vilka är packade tillsammans med andra filer (bilder mm) i en ZIP-fil.
Rekommendationer	Eftersom .odt är ett öppet XML-format fungerar det bra både för deponering och långtidslagring. Vid långtidslagring bör filerna vara okomprimerade. Om filerna innehåller bilder eller annat infogat material så bör det materialet lagras separat i lämpligt format för långtidslagring.

OpenOffice/StarOffice	
Filformat/-ändelse	Sxw/.sxw
Format	XML-format som använts av bland annat OpenOffice/StarOffice från version 1.0 till 2.0. Senare ersatt av OpenDocument Format (.odt).
Beskrivning	Även om formatet ersatts av .odt så är det strukturellt likartat (zippade xml-filer) och kan läsas av OpenOffice.org 2.0.
Rekommendationer	Fungerar för långtidslagring men .odt bör användas där det går.

Rich Text Format	
Filformat/-ändelse	RTF/.rtf
Format	RTF (Rich Text Format) är ett uppmärkt textformat utvecklat av Microsoft.
Beskrivning	Även om formatet huvudsakligen är läsbar ren text, och därför lämpligt för både långtidslagring och användning, så finns en del kompabilitetsproblem gällande formatering (t.ex. textrutor och tabeller) när man öppnar filerna i

[32] http://www.digitalpreservation.gov/formats/fdd/fdd000427.shtml
http://www.digitalpreservation.gov/formats/fdd/fdd000428.shtml
[33] http://www.iso.org/iso/catalogue_detail.htm?csnumber=43485
http://www.digitalpreservation.gov/formats/fdd/fdd000247.shtml

	vissa ordbehandlingsprogram. Filstorleken på en .rtf är normalt sett större än motsvarande .doc-, .pdf- eller .odt-fil.
Rekommendationer	Även om formatet fungerar för deponering och lagring finns det bättre och nyare format som .docx och .odt vilka är mindre i formatet samt har bättre kompabilitet.

WordPerfect	
Filformat/-ändelse	WPD/.wpd
Format	Ett binärt och proprietärt filformat framtaget för WordPerfect.
Beskrivning	Användandet av det en gång populära WordPerfect har minskat sedan dess introduktion på tidigt 1980-tal, mycket på grund av Microsoft Word. Även om .wpd (för tidigare versioner .wp och .wp5) är standardformatet så stödjer senare versioner av programmet import och export av olika filformat. Filformatet kan läsas av Microsoft Office Word samt OpenOffice.
Rekommendationer	Ej rekommenderat format för tillgängliggörande av material eller som format för lagring/långtidsbevaring. Senaste versionen av WordPerfect kan visserligen importera och exportera ODF och OOXML-filer, men det är bättre att använda de senare XML-baserade öppna alternativen som filformat.

Format för uppmärkning av text

Visserligen används vanligtvis inte märkspråk för att skapa rapporter och liknande dokument (HTML används vanligen för websidor och XML för utbyte av data) men nedan finns lite information om några av formaten.

HTML/XHTML	
Filformat/-ändelse	HTML/.htm/.html, XHTML/.xhtml/.xht
Format	Hypertext Markup Language (HTML) är ett uppmärkningsspråk/märkspråk skrivet som oformaterad text och utvecklat som en undergrupp av SGML.
Beskrivning	HTML är ett märkspråk som huvudsakligen används för hemsidor. Bortsett från HTML-filens oformaterade textinnehåll (inklusive antingen en formatmall alternativt länk till formatmall) består websidor ofta av olika typer av länkad media (bilder, video, ljud, dokument etc).

Rekommendationer	Formatet fungerar för långtidsbevaring samt tillgängliggörande men teckenkodning måste anges. För versioner tidigare än HTML5 måste dokumentet dessutom följa och ange ett giltigt DTD (dokumenttypsdefinition). Om CSS (stilmallar) används ska de antingen specificeras i dokumentet eller bifogas separat. Bilder och annan media ska behandlas som enskilda objekt enligt filtyp.

SGML	
Filformat/-ändelse	SGML/.sgml
Format	Standardised Generalised Markup Language (SGML). En certifierad ISO standard (ISO 8879:1986 SGML[34]) för märkspråk/uppmärkningsspråk.
Beskrivning	SGML är ett metaspråk som används för att definiera andra märkspråk som HTML och XML.
Rekommendationer	Fungerande format för långtidsbevaring och tillgängliggörande men dokumenten måste följa ett fastställt schema enligt standarden ISO 8879.

XML	
Filformat/-ändelse	XML/.xml
Format	Extensible Markup Language (XML), är en öppen standard baserad på oformaterad text och utvecklad av W3C (World Wide Web Consortium).
Beskrivning	XML utvecklades som en undergrupp till SGML[35] och används huvudsakligen för web och vid utbyte av data mellan olika system (t.ex. databaser).
Rekommendationer	Fungerar för långtidsbevaring och tillgängliggörande men dokumenten måste följa, och ange ett giltigt XSD/DTD och teckenkodning.

De flesta typer av textfiler förblir i det format de skapades i. Ett undantag från detta är pdf-formatet där väldigt få dokument skapas i pdf. Majoriteten av dessa dokument har sitt ursprung i någon form av ordbehandlingsprogram (Word eller OpenOffice) som sedan sparas som pdf för att underlätta tillgängliggörandet av dokumentet när det är klart.

[34] http://www.iso.org/iso/catalogue_detail.htm?csnumber=16387
[35] http://www.digitalpreservation.gov/formats/fdd/fdd000075.shtml

Dokument och digital text: Bibliografi

Ditch, W. (2007) *XML-based Office Document Standards*. JISC: Bristol. http://www.jisc.ac.uk/techwatch

Fanning, Betsy A. (2008) *Preserving the Data Explosion: Using PDF*. DPC Technology Watch Series Report 08. http://www.dpconline.org/docs/reports/dpctw08-02.pdf

Morrison, A., Popham, M. & Wikander, K. (2001) *Creating and Documenting Electronic Texts: A Guide to Good Practice*. AHDS. http://ota.ahds.ac.uk/documents/creating/cdet/

Morrison, A. & Wynne, M. (2005) *AHDS Preservation Handbook: Marked-up Textual Data*. AHDS. http://ota.ahds.ac.uk/documents/preservation/preservation_markup.pdf