PM Digitalt bevarande

Karl Pettersson

Inledning

Avsikten med detta PM är att diskutera vilka frågeställningar som behöver utredas och vilka åtgärder som är nödvändiga för att möjliggöra bättre dokument- och arkivhantering i den i specifikationerna för PM-uppgiften beskrivna myndighet *A*. Frågeställningarna kan delas in i följande huvudkategorier:

1. Vilka krav kan ställas när det gäller bevarande, gallring och tillgängliggörande av dokument hos myndigheten?
2. Vilka hinder som finns för att möta dessa krav hos myndigheten?
3. Hur skall dessa hinder hanteras?

Var och en av dessa huvudfrågor skulle kunna diskuteras ur många olika aspekter. I detta PM har jag valt att fokusera på problemet att se till att hålla elektroniska dokument läsbara över tid. Jag diskuterar vad som behöver utredas närmare men ger också en del antydningar om vilka lösningar vi kan förväntas komma fram till, givet beskrivningen av myndighet *A*. I den avslutande diskussionen om möjliga lösningar lägger jag speciell tonvikt speciell tonvikt på problemet med val av filformat för bevarande.

En referensmodell som ofta används för elektroniska arkiv är OAIS (*Open Archival Information System*). Grundläggande i denna modell är åtskillnaden mellan de tre olika typerna av informationspaket: paket för leverans till arkiv (SIP, *Submission Information Package*), paket för arkivering (AIP, *Archival Information Package*) och paket för distribution (DIP, *Dissemination Information Package*). Sett ur det perspektivet kan detta PM sägas fokusera på i första hand skapande av SIP, då det handlar om vilka format som skall väljas när information levereras till arkivet (Toller 2009). Även information som finns i arkivet kommer emellertid att behöva konverteras och överföras till nya databärare om den skall hållas läsbar över lång tid. Elektroniska handlingar kan aldrig betraktas som ”färdiga”, vilket är något man tagit fasta på i *records continuum*-modellen, som karakteriseras av att man inte skiljer mellan olika tidsmässigt åtskilda faser i livscykeln hos

handlingar (”records”), utan tänker sig att de befinner sig i olika dimensioner: ”create”, ”capture”, ”organize” och ”pluralize” (McKemmish 2001).

Krav

Eftersom *A* är en myndighet gäller att de dokument som hanteras där i många fall kan vara allmänna handlingar, som omfattas av offentlighetsprincipen. Deras arkiv bildas av de allmänna handlingarna från deras verksamhet, och de måste fastställa vilka handlingar som skall vara arkivhandlingar. Enligt arkivförordningen skall t.ex. allmänna handlingar i ett ärende arkiveras när ärendet slutbehandlats, och anteckningar i olika typer av fortlöpande register skall betraktas som arkiverade så fort de registrerats (Jarborn och Gäfvert 2013a, s. 8). Det är nödvändigt att myndighet *A* upprättar en gallrings- och bevarandeplan, där det mer i detalj fastställs vilka typer av handlingar som skall bevaras och hur länge. Hos myndighet *A* finns följande system för hantering av elektroniska handlingar, och en gallrings- och bevarandeplan måste innehålla föreskrifter för de uppgifter som finns i de olika systemen:

- Forskningsstödssystemet, som innehåller uppgifter ur ansökningshandlingar från bidragssökanden.
- E-postsystem för registrering av elektroniska ansökningar.
- En rad andra IT-system för sådant som personaladministration, ekonomi, ärendehantering, forskningsdatabaser och statistik.

Huvudprincipen i arkivlagen är att allmänna handlingar skall bevaras, även om gallring får ske i enlighet med Riksarkivets föreskrifter.

Myndigheter måste ha ett ”proaktivt” förhållningssätt, så att de redan när dokument färdigställs, eller åtminstone så snart det är möjligt ur ett verksamhetsperspektiv, ser till att de har format som möjliggör bevarande (Jarborn och Gäfvert 2013a, s. 8–9; jämför även Borglund 2008). Samtidigt är införandet av ett proaktivt förhållningssättet vid framställning av nya dokument inte nog: det finns också en stor mängd redan existerande dokument hos myndighet *A* vars framtida läsbarhet behöver säkras.

Riksarkivets föreskrifter (RA-FS 2009:2) innehåller detaljerade tekniska krav för elektroniska handlingar, bl.a. vilka filformat som kan användas för bevarande. Alla de handlingstyper som tas upp där behöver dock inte vara relevanta för myndighet *A*, och frågor kring vilka delar av föreskrifterna som är relevanta kommer att diskuteras i de följande avsnitten.

National Archives (2009) diskuterar en rad allmänna, i viss mån överlappande, adekvansvillkor när det gäller filformat för långtidsbevarande utan att föreskriva några specifika filformat:

Ubiquity Format med utbredd användning är att föredra.

Support Det skall finnas gott om mjukvara som stöder formaten.

Disclosure Teknisk dokumentation av formaten skall vara tillgänglig, och formaten skall helst vara öppna, så att specifikationerna är fria att implementera.

Documentation quality Dokumentationen skall hålla god kvalitet.

Stability Formaten skall inte genomgå stora förändringar över tid.

Ease of identification and validation Det skall vara lätt att validera om en fil överensstämmer med ett visst format.

Intellectual property rights Format som inte skyddas av patent är att föredra.

Metadata support Format med stöd för metadata är att föredra.

Complexity Format skall vara så komplexa som behövs för att stödja de funktioner som är viktiga att bevara. Detta måste dock vägas mot att komplexa format tenderar att bli mer kostsamma att bevara och därmed lätt kommer i konflikt med de övriga villkoren.

Interoperability Format som möjliggör utbyte av information mellan olika IT-system är att föredra.

Viability Format som innehåller algoritmer för att upptäcka korruption vid filöverföring är att föredra.

Reusability Format som gör det möjligt att processa informationen på nytt är att föredra, om sådana möjligheter behövs. Om vi t.ex. konverterar en kalkylbladsfil till PDF försvinner möjligheten att skapa nya beräkningar, diagram och liknande utifrån kalkylbladet (om vi inte gör potentiellt kostsam och osäker extrahering av information ur PDF-filen).

Toller (2009), s. 110 ger också en lista med generella villkor för arkivformat, som har mycket gemensamt med National Archives (2009) men inte är lika detaljerad: punkter i hennes lista handlar om att formaten skall kunna läsas ”utan alltför speciell och svåråtkomlig programvara”, att de skall vara ”*de jure*-standard” eller åtminstone ”*de facto*-standard”, att specifikationen skall vara öppen och att de skall passa den aktuella typen av arkivinformation.

Det kan uppkomma situationer där en överföring av information till de specifika bevarandeformat som anges i (RA-FS 2009:2) skulle innebära oacceptabelt

stora uppoffringar när det gäller adekvansvillkor av den typ som anges i National Archives (2009). I sådana fall kan det vara en lösning att parallellt bevara informationen i ett format som inte är officiellt sanktionerat av Riksarkivet men uppfyller adekvansvillkoren i större utsträckning. Exempel på detta kommer att diskuteras i avsnittet om lösningar.

Det finns inga detaljerade föreskrifter om fysiska databärare i (RA-FS 2009:2). Däremot ger Riksarkivet *funktionskrav*: det sägs att myndigheter vid överföring av elektroniska handlingar till bevarande skall välja databärare ”med hänsyn till livslängd, vårdbehov, klimatkrav, bevarandetid samt krav på handlingarnas tillgänglighet” (RA-FS 2009:1, 4 kap. 16§). Det föreskrivs dessutom att myndigheter regelbundet skall framställa säkerhetskopior av allmänna handlingar, och att dessa skall förvaras geografiskt åtskilt från de kopierade handlingarna (RA-FS 2009:1, 6 kap. 5§).

Hinder

Vi vet en del om vilka hinder det finns när det gäller att hålla dokument hos myndighet *A* läsbara. Det rör sig om filer som inte är läsbara, filer som är utspridda på personliga och delade hårddiskar, bristfällig kompetens i organisationen när det gäller bevarande och bristfälliga uppgifter om vilken information som finns i de olika IT-systemen. Alla dessa områden behöver utredas närmare.

Om det finns datafiler som inte är läsbara, kan det vara ett problem på hård- eller mjukvarunivå. De fysiska databärarna (hårddiskar, disketter, magnetband etc.) kan ha blivit skadade, eller det kan saknas teknisk utrustning för att läsa dem. Problemet kan också ligga på mjukvarunivå: filernas innehåll kan ha blivit korrupt, eller det kan vara så att tillgänglig programvara inte kan läsa de format i vilka filerna sparats. Det är nödvändigt att utreda vilka av dessa problem som föreligger hos myndighet *A* och i vilken mån de berör de dokument som behöver bevaras.

Det behöver utredas i vilken mån problemet med utspridda filer berör information som skall bevaras och i vilken mån de olika IT-systemen innehåller sådan information. När det gäller de IT-system som används för att skapa ny information behöver det utredas i vilken mån dessa möjliggör lagring av information i format lämpliga för bevarande, och om överföring till sådana format är automatiserad i systemen.

Likaså bör det undersökas vilken utbildning personalen har när det gäller frågor relaterade till digitalt bevarande.

Lösningar

På hårdvarunivån gäller det att se till att information som skall bevaras lagras på databärare som uppfyller relevanta krav och att det sker regelbunden säkerhetskopiering. Information som skall långtidsbevaras kommer att behöva överföras till nya databärare med jämna mellanrum, dels därför att databärarnas livslängd är begränsad, dels därför att hårdvara som kan läsa databärarna inte kan förväntas finnas tillgänglig för evigt. Om existerande information finns lagrad på databärare som inte är läsbara med utrustning tillgänglig hos myndighet *A* kan det finnas det olika sätt att få tillgång till informationen. Rör det sig om standarddisketter finns diskettenheter som passar moderna datorer tillgängliga för ringa kostnad. I andra fall kan det behövas mer specialiserade lösningar, och då måste det göras en kostnadsavvägning hur viktigt det är att få tillgång till den aktuella information. Ett standardval när det gäller lagringsmedia för långtidsbevaring är bandkassetter, och det kan vara lämpligt att kopiera över deras innehåll till nya databärare vart femte år (Toller 2009, s. 108).

När det gäller val av filformat för bevarande varierar det naturligtvis beroende på vilken typ av information det är som skall bevaras. (RA-FS 2009:2, 3 kap.) anger krav för följande dokumenttyper:

1. Databaser och register.
2. Strukturerade dokument baserade på märkspråk.
3. Kontorsdokument.
4. Elektroniskt underskrivna handlingar.
5. E-postmeddelanden.
6. Digitala bilder och skannade bilder.
7. Digitala kartor och ritningar.
8. Webbsidor.

Vilka av dessa typer som finns representerade när det gäller den information hos myndighet *A* som skall bevaras är något som behöver utredas. Vi kan dock vara ganska säkra på att det finns en lång rad IT-system som innehåller databasinformation, och att forskningsstödssystemet och det elektroniska ansökningssystem innehåller e-postmeddelanden och kontorsdokument och kanske även sådant som strukturerade dokument och digitala bilder. Avgränsningen mellan vissa av kategorierna, som ”kontorsdokument” och ”strukturerade dokument baserade på märkspråk” är heller inte alltid helt klar, vilket kommer att diskuteras i det följande.

Databaser skall enligt Riksarkivet sparas som textfiler med fast fält- och postlängd eller teckenseparerade fält (t.ex. kommaseparerade fält, CSV) eller

XML (RA-FS 2009:2, 3 kap. 1§). Om de sparas som textfiler skall dessa vara kodade enligt ISO/IEC 8859-1 eller ISO/IEC 10646 (Universal Character Set, vilket är ekvivalent med Unicode), men då med en begränsning av teckenmängden som i stort sett innefattar ASCII-tecken (i huvudsak engelska bokstäver, siffror och några skiljetecken), svenska tecken och andra latinska bokstäver med accent. När det gäller myndighet *A* kan restriktionerna av teckenmängd för textfiler eventuellt innebära problem, därför att databaser med information relaterad till vetenskapligt arbete kan innehålla Unicodetecken utöver detta, t.ex. tecken i icke-latinska alfabet eller matematiska symboler.

Toller (2009), s. 115–117 tar upp de nämnda bevarandeformaten för databaser och hävdar att XML ofta är det enklaste att använda, under förutsättning att databassystemet är så modernt att det finns funktioner för XML-export och att databasen inte är alltför stor. I annat fall är text med fast bredd eller separatortecken att föredra.

Jarborn och Gäfvert (2013b) påtalar att dessa bevarandeformat inte är de format som normalt används internt i databashanterarna, och att de inte ger den ”funktionalitet och effektivitet som kommer med SQL-server eller Oracle och deras gränssnitt”, något som innebär att det krävs konvertering, vilket nästan alltid medför ”en större eller mindre katastrof ur bevarandesynpunkt”, till följd av sådana effekter som ”omstrukturering av uppgifter, bristande kvalitet” och ”förlust av samband” (Jarborn och Gäfvert 2013b, s. 7). Med andra ord råkar bevarandeformaten i konflikt med t.ex. komplexitetsvillkoret från National Archives (2009), att formaten för bevarande skall stödja den funktionalitet som behöver bevaras.

Vilka databashanterare som används, och i framtiden skall användas, för de olika IT-systemen vid myndighet *A* är något som behöver utredas. Behovet av konvertering för bevarande kommer dock sannolikt att kvarstå oavsett detta. De flesta databashanterare använder sig av den s.k. relationsmodellen, och både proprietära program, som de av Jarborn och Gäfvert (2013b) nämnda Oracle och Microsoft SQL Server, och fria program av denna typ, som MySQL/MariaDB (MariaDB Corporation Ab 2014b) och PostgreSQL (PostgreSQL Global Development Group 2014), använder sig normalt av binära, icke standardiserade format[1]. Dessa format tenderar att komma i konflikt med sådana kriterier som utbrett mjukvarustöd, öppenhet (framför allt i fallet med proprietära program) och interoperabilitet.

Ett sätt att hantera denna problematik kan vara att parallellt med de i (RA-FS 2009:2) officiellt sanktionerade bevarandeformaten lagra databasinformationen

[1] Dock kan det finnas möjlighet att direkt använda sig av bevarandeformat som CSV och XML för relationsdatabaser, t.ex. via databasmotorn CONNECT i MariaDB (MariaDB Corporation Ab 2014a).

i format som ger en bättre avvägning mellan olika relevanta adekvansvillkor. Det är t.ex. ofta möjligt att exportera tabeller som SQL-frågor, som kan lagras i textfiler. På så sätt bibehålls databasstrukturen, samtidigt som innehållet i fälten kan läsas i vilket program som helst som kan hantera oformaterad text. SQL är också ett öppet, ISO-standardiserat format. Emellertid avviker många databashanterare från standarden på olika sätt, vilket kan innebära problem när det gäller interoperabilitet mellan olika databashanterare.

Det finns även databashanterare som inte följer relationsmodellen, exempelvis s.k. dokumentorienterade databaser, som är centrerade kring ”dokument” med taggar organiserade i olika nivåer, i stället för den fixa uppsättningen fält i relationsdatabaser. Det kan tänkas att sådana databassystem har använts, eller kommer att användas, för vissa av uppgifterna i en organisation som myndighet *A*, där det är nödvändigt att hantera dokument relaterade till forskning. Dokumentorienterade databaser kan använda sig av XML, eller format som ligger nära detta, som standardformat. En av de mest använda dokumentorienterade databashanterarna är MongoDB, som använder sig av BSON, en binär kodning av JSON (*JavaScript Object Notation*) (MongoDB, Inc. 2014). JSON är ett öppet XML-liknande format, även om det inte finns upptagen bland Riksarkivets bevarandeformat (RA-FS 2009:2).

Kontorsdokument skall enligt Riksarkivet sparas som antingen text kodad enligt ISO 8859-1 eller som PDF/A-1 (RA-FS 2009:2, 3 kap. 4§). Sparas de som text går t.ex. formatering förlorad, och 8859-1 är kanske inte tillräcklig för att representera alla tecken som används i forskningssammanhang, som diskuterades ovan i samband med databaser. PDF/A betecknar en uppsättning ISO-standardiserade versioner av PDF (det finns A-1, A-2 och A-3), som bl.a. innebär att information nödvändig för att visa dokumentet på ett konsistent sätt (bilder, teckensnitt etc.) skall bäddas in i PDF-filen och att vissa typer av innehåll som kan befaras medföra kompatibilitetsproblem, t.ex. ljud, video och JavaScript, inte får förekomma.

En nackdel med att överföra kontorsdokument till PDF/A (eller PDF generellt) är, som nämnts, förlust av återanvändbarhet, t.ex. om man vill använda data i ett kalkylblad för nya beräkningar. Originalformat för kontorsdokument är ofta proprietära och olämpliga för långtidsbevarande. När det gäller nya versioner av kontorspaket som Microsoft Office och Libre Office använder sig dessa emellertid av öppna, XML-baserade format. Detta innebär att sådana dokument också kan betraktas som en form av dokument baserade på märkspråk, och för sådana dokument anger Riksarkivet SGML, (X)HTML och XML som tillåtna bevarandeformat (RA-FS 2009:2, 3 kap. 3§).

Det kan också förekomma andra typer av strukturerade dokument baserade på öppna format som inte finns upptagna i (RA-FS 2009:2). Ett exempel är LaTeX, som är vanligt i vetenskapliga sammanhang, och därför kan tänkas

förekomma bland de dokument som hanteras som myndighet *A*. I sådana fall kan en lösning vara att spara källfilen i märkspråket parallellt med en PDF/A-fil, för att möjliggöra återanvändbarhet. Ett intressant alternativ i sammanhanget kan vara s.k. lättviktiga märkspråk, som markdown. Dessa gör det möjligt att enkelt skapa kod, innehållande bl.a. olika typer av formateringsmarkeringar och metadata, som är lätt att läsa i textredigerare och som kan konverteras till och från andra märkspråk (inklusive XML-baserade format), PDF etc. med hjälp av program som Pandoc (MacFarlane 2013).

När det gäller e-post och bilder finns inte så mycket specifika problem relaterade till valet av format för bevarande, utöver vad som tagits upp ovan. E-postmeddelanden kan sparas som databaser, kontorsdokument eller strukturerade dokument (RA-FS 2009:2, 3 kap. 6§). Digitala eller skannade bilder kan sparas som JPEG, TIFF eller PNG (RA-FS 2009:2, 3 kap. 7§).

Om det finns existerande dokument som skall bevaras och som inte är läsbara till följd av att de sparats i filformat som inte kan hanteras av tillgänglig programvara, bör det undersökas i fall det går att få tillgång till program som kan konvertera dem till nyare format. En lösning kan eventuellt vara emulering av äldre systemmiljöer (Toller 2009, s. 103–104).

Förkortningar

RA-FS 2009:1 Riksarkivet (2009a). *Riksarkivets föreskrifter och allmänna råd om elektroniska handlingar (upptagningar för automatiserad behandling)*. Riksarkivet.

RA-FS 2009:2 Riksarkivet (2009b). *Riksarkivets föreskrifter och allmänna råd om tekniska krav för elektroniska handlingar*. Riksarkivet.

Referenser

Kompendium i e-arkiv

Jarborn, Elisabeth och Thomas Gäfvert (2013a). ”Anteckningar rörande RA-FS 2009:1”.

— (2013b). ”Ursprungligt skick och bevarande av uppgifter i databaser”.

Övrig kurslitteratur

Borglund, Erik (2008). *Design for recordkeeping : areas of improvement*. Diss. Sundsvall : Mittuniversitetet, 2008. Sundsvall: Department of Natural Sciences, Mid Sweden University. ISBN: 978-91-85317-95-0.

McKemmish, Sue (2001). ”Placing Records Continuum Theory and Practice”. I: *Archival Science* 1, s. 333–359.

National Archives (2009). *Selecting File Formats for Long-Term Preservation*. Tekn. rapport. National Archives. URL: https://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.

Riksarkivet (2009a). *Riksarkivets föreskrifter och allmänna råd om elektroniska handlingar (upptagningar för automatiserad behandling)*. Riksarkivet.

— (2009b). *Riksarkivets föreskrifter och allmänna råd om tekniska krav för elektroniska handlingar*. Riksarkivet.

Toller, Eva (2009). ”Planering och genomförande av leverans till e-arkiv”. I: *E-arkivera rätt : sju perspektiv på hantering av digital information med hjälp av OAIS*. Utg. av Katrin Askergren och Maria Dahlgren. Stockholm: Näringslivets arkivråd, s. 99–130. ISBN: 9789197386333.

Specifikationer online

MacFarlane, John (2013). *Pandoc User's Guide*. URL: http://johnmacfarlane.net/pandoc/README.html (hämtad 2014-12-28).

MariaDB Corporation Ab (2014a). *CONNECT Table Types - Data Files*. URL: https://mariadb.com/kb/en/mariadb/documentation/storage-engines/connect/connect-table-types/connect-table-types-data-files/ (hämtad 2014-12-26).

— (2014b). *Storage Engines*. URL: https://mariadb.com/kb/en/mariadb/documentation/storage-engines/ (hämtad 2014-12-26).

MongoDB, Inc. (2014). *BSON*. URL: http://docs.mongodb.org/meta-driver/latest/legacy/bson/ (hämtad 2014-12-26).

PostgreSQL Global Development Group (2014). *Database Physical Storage*. URL: http://www.postgresql.org/docs/9.4/static/storage.html (hämtad 2014-12-26).