CODA-WEBB

SLUTRAPPORT

Lena Lindbäck
LDB-centrum

2009-05-05

Abstract

This final report from the CODA-WEBB project consists of three parts:

- Strategy for archiving a website
- A literature review of tools for web crawling
- Review of ”Open repositories” to publish, store and give access to digital documents

The first part of the report is intended as a handbook for authorities and other organizations working in the first phase of web archiving. The report concludes that a number of areas need to be investigated, for example: aim and requirements, selection, users and access. Other areas discussed in part 1 are frequency and date for collecting the web, methods, tools and technical questions.

In part 2 three web crawlers were evaluated and compared: HTTrack Website Copier, Heritrix and PageNest. The aim of the survey was to choose an appropriate tool for the Test platform project at the LDP Centre. The tool selected for this purpose was Heritrix.

Part 3 consists of an evaluation of Open repository tools, with the aim to choose the best tool to use in the Test platform. Five tools were evaluated by a literature review: EPrints, DSpace, Fedora, Greenstone and DAITSS. The study suggested Fedora as most suitable for the Test platform.

The full report is written in Swedish.

Sammanfattning

Denna slutrapport vid namn CODA-WEBB består av tre delar:

- Strategi för webbarkivering
- Undersökning av webbcrawler-verktyg för insamling av webbsidor samt
- Utvärdering av ”Open repositories” för att publicera, lagra och tillgängliggöra digitala dokument

Den första delen av rapporten är tänkt att fungera som en handbok för myndigheter och andra organisationer som ska göra en förstudie inför ett webbarkiveringsprojekt. Slutsatsen av rapporten är att ett antal områden behöver utredas, som till exempel: syfte och krav på bevarandet, urval, användare och tillgängliggörande. Frekvens och tidpunkt för insamling samt metoder, verktyg och tekniska frågor är andra områden som behandlas i del 1.

I del 2 utvärderas och jämförs tre verktyg för webbcrawling: HTTrack Website Copier, Heritrix och PageNest. Syftet med denna undersökning var att utse lämpligt verktyg för LDB-centrums interna projekt Testplattformen. Heritrix utsågs till bästa val för detta ändamål.

Del 3 består av en utvärdering av Open repository-verktyg, även detta i syfte att välja lämpligt verktyg för användande i Testplattformen. Fem verktyg utvärderades med hjälp av en litteraturstudie, dessa var: EPrints, DSpace, Fedora, Greenstone och DAITSS. Som mest lämpat för användande i Testplattformen utsågs Fedora.

Innehåll

1. Bakgrund

LDB-centrum[1] (Centrum för långsiktigt digitalt bevarande) är ett kompetenscentrum som arbetar med modeller, metoder och verktyg för arkivering och återskapande av digitalt material. De partner som ingår i LDB-centrum är Luleå tekniska universitet[2] (LTU), Riksarkivet[3] (RA), Bodens kommun[4], Kungl. biblioteket[5] (KB) samt Statens ljud- och bildarkiv (SLBA). Den första januari 2009 upphörde SLBA att vara en egen myndighet och blev istället Avdelningen för audiovisuella medier inom KB.

1.1 CODA

Samarbetet mellan parterna i LDB-centrum drivs sedan år 2007 under namnet CODA (Curation of Digital Assets). Inom CODA-gruppen kommer olika delprojekt att drivas, vilka områden gruppen ska arbeta med och vilka resurser som ska tilldelas projekten beslutas på gemensamma möten. Arbetet regleras av det samarbetsavtal som skrevs inför det första årets samarbetsprojekt. Varje delprojekt avrapporteras när det är färdigt. Rapporterna skrivs på svenska med ett engelskt abstract och kan vid behov översättas i sin helhet.

1.2 Tidigare projekt i CODA

Under 2007 drevs två projekt mellan LDB-centrums parter, CODA-FORM och CODA-POP. Projektet CODA-FORM fokuserade på frågor om filformat medan CODA-POP framför allt arbetade med organisatoriska frågor samt framtagande av kriterier för bevarandeprocessen.

Under 2008 startade två nya projekt, det första fick namnet CODA-META och hade som syfte att utvärdera och jämföra fyra olika verktyg för att extrahera metadata. Det andra projektet är det som redovisas i denna slutrapport, CODA-WEBB.

CODA-META avslutades under september 2008 medan arbetet med CODA-WEBB fortsatte även under första delen av år 2009. Under våren 2009 blev delar av CODA-FORM samt slutrapporten för CODA-META översatta till engelska.

Mer information om CODA-samarbetet samt rapporterna för CODA-FORM, CODA-POP och CODA-META finns publicerade på LDB-centrums webbplats: http://www.ldb-centrum.se under Projekt/CODA. Där kommer också information om CODA-WEBB att finnas tillsammans med denna slutrapport.

[1] Webbplats: http://www.ldb-centrum.se
[2] Webbplats: http://www.ltu.se
[3] Webbplats: http://www.statensarkiv.se
[4] Webbplats: http://www.boden.se
[5] Webbplats: http://www.kb.se

2. CODA-WEBB och Testplattformen

På ett möte med CODA-gruppen i januari 2008 diskuterades vilka områden som skulle behandlas inom projektet under det kommande året. Det beslutades att två områden skulle prioriteras: verktyg för att extrahera metadata (se kapitel 1.2) samt webbarkivering.

Eftersom Luleå tekniska universitet (LTU) just var i startgroparna för att inleda ett webbarkiveringsprojekt beslöts att LTU:s projekt skulle utgöra ett pilotprojekt inom CODA. Utöver det förslag till en teknisk lösning som universitetet önskade skulle projektet även arbeta med att beskriva förarbetet när en myndighet eller annan organisation beslutar sig för att börja arkivera sin webbplats. Det ansågs viktigt att utreda vilka områden som behöver analyseras och vilka frågor som behöver besvaras i projektets inledningsskede för att insamling och bevarande ska kunna utföras på bästa sätt.

I början av år 2008 startade ett internt projekt vid namn Testplattformen på LDB-centrum. Projektet har som syfte att ge resultat till nya tekniska framsteg, produkter eller tjänster. Det första delspåret blev att testa, utvärdera och utveckla metoder och verktyg för webbarkivering.

För att bygga upp testmiljön behövde vi bland utvärdera vilka verktyg som kunde vara lämpliga att använda för insamling av webben samt om användande av ett s.k. Open Repository kunde vara lämpligt till bevarandet av de digitala filerna. Resultaten av dessa utvärderingar ingår som Del 2 respektive Del 3 i denna rapport.

Parallellt med denna rapport skrevs även en slutrapport för pilotprojektet inom LTU som utöver de strategiska frågorna kommer med ett förslag till teknisk lösning. Detta i sin tur bygger på erfarenheter ur både CODA-WEBB och från Testplattformen.

Pilotprojektet med LTU samt skrivandet av slutrapporten för CODA-WEBB var planerat att avslutas i och med december månads utgång 2008. Under senhösten 2008 utökades dock Testplattformen med ett uppdrag från Riksarkivet. Och eftersom frågorna i Testplattformen till största del även berör de områden som avhandlas i CODA-WEBB beslöts att förskjuta arbetet med denna slutrapport till första halvåret 2009.

2.1 Syfte

Denna rapport har som syfte att:

- Svara på vilka områden som behöver utredas i inledningsfasen för webbarkivering
- Göra en litteraturstudie om verktyg för att samla in webbplatser
- Genom en litteraturstudie bedöma ett antal Open Repositories

2.2 Målgrupp

Rapporten riktar sig i första hand till myndigheter i Sverige vilka har att följa Tryckfrihetsförordningens regler om allmänna handlingar, men förhoppningsvis kan den även användas av andra organisationer som vill göra en bra förstudie inför ett webbarkiveringsprojekt.

Flera organisationer arbetar redan med att samla in kopior av webbplatser, som Internet Archive (IA) i USA och Kungl. biblioteket (KB) i Sverige. Detta innebär dock inte att frågan anses löst, utan varje svensk myndighet är *själv* ansvarig för att bevara den egna webbplatsen. Denna rapport riktar sig till en enskild organisation som har som inritning att bevara den egna

domänen. Rapporten är tänkt att fungera som hjälp för att ta beslut på frågor som: varför webbplatsen ska bevaras, vad som ska bevaras, när det ska göras och hur arbetet ska bedrivas.

Den kan också vara till hjälp när en organisation ska skapa en ny webbplats och/eller införskaffa ett nytt verktyg att göra sin webbplats i, för att i ett tidigt skede underlätta framtida bevarande.

Del 1 är tänkt att tillföra kunskap av allmän karaktär till alla som ska arbeta med arkivering av webbplatser medan Del 2 och Del 3 mest riktar in sig till dem som ska avgöra vilken teknik som ska användas vid insamling och lagring/arkivering av webbplatserna.

Open repositories kan även användas till lagring av annat digitalt material än webbsidor. Del 3 kan därför även vara intressant i andra sammanhang. Nämnas kan att dessa verktyg mestadels används som publiceringsverktyg inom universitetsvärlden.

2.3 Disposition

Rapporten är uppdelad i tre olika delar vilkas disposition ser ut som följande:

Del 1 – Strategi för webbarkivering, inleds med allmän information om arkivering av webbplatser. Kapitel 3 handlar framför allt om vikten av en dokumenterad policy, begreppsanalys och beskrivning av organisation och webbplats.

Kapitel 4 behandlar syftet med att bevara, kapitel 5 vilka externa och interna krav som finns medan kapitel 6 handlar om att bestämma urval och omfattning av insamlingen. I kapitel 7 analyseras användaren medan kapitel 8 diskuterar när och hur ofta insamling bör göras.

Kapitel 9 handlar om olika tekniska lösningar för insamling och lagring, kapitel 10 beskriver modeller för insamling och arkivering medan kapitel 11 ger ett förslag till en handlingsplan. Kapitel 12 är en slutsats på hela Del 1.

Del 2 – Webbcrawlers. Denna del innehåller information om hur verktyg för att samla in webbplatser fungerar, följt av en studie som beskriver tre stycken olika verktyg: HTTrack Website Copier, Heritrix samt PageNest. En kort slutsats avslutar Del 2.

Del 3 – Open Repositories beskriver först vad dessa verktyg har som syfte och hur de kan användas samt även något om fördelar med att använda ”open source-verktyg”. Efter detta kommer resultaten av en litteraturstudie över fem olika verktyg: Eprints, DSpace, Fedora, Greenstone samt Daitss. Resultaten sammanfattas och en slutsats redovisas.

Sist i rapporten finns en källförteckning över vilka källor som har använts i de olika delarna.

DEL 1: STRATEGI FÖR WEBBARKIVERING

3. Webbarkivering

Arkivering av en organisations webbsidor innebär att man lagrar de tidigare versionerna av webbplatsen för att kunna se på, leta i, visa upp och bevisa vad man förut har publicerat. För myndigheter i Sverige gäller att myndighetens webbsidor är allmänna handlingar och därför skall bevaras. Begreppet ”allmän handling” definieras i Tryckfrihetsförordningen, TF kap 2 § 3[6] på detta sätt:

> ”Med handling förstås framställning i skrift eller bild samt upptagning som kan läsas, avlyssnas eller på annat sätt uppfattas endast med tekniskt hjälpmedel. Handling är allmän, om den förvaras hos myndighet och enligt 6 eller 7 § är att anse som inkommen till eller upprättad hos myndighet.”

Ett omfattande regelverk styr hanteringen av allmänna handlingar hos svenska myndigheter. För att kunna gallra bort handlingar krävs gallringsbeslut, i annat fall ska de bevaras. Förutom att följa de lagar och förordningar som styr en myndighet är syftet med att bevara handlingar i arkiv även att de ska finnas tillgängliga för framtida forskare.

Men arkivering av webbplatser kan även spela viktig roll för andra än statliga myndigheter och framtida forskare: webbsidorna kan ha ett stort juridiskt värde genom att organisationen vid en eventuell juridisk process kan bevisa vad som har och inte har varit publicerat vid en viss tidpunkt. Bevarat material kan också bygga upp kunskap och därför ha ett värde i att återanvändas.

3.1 Strategi för webbarkivering

När beslut om webbarkivering har tagits finns ett antal strategiska frågor att utreda:

- Varför ska webbplatsen bevaras, i vilket syfte?
- Vilka krav ställs på bevarandet?
- Vad ska bevaras?
- Vad ska inte bevaras, vilka avgränsningar ska gälla?
- Vem ska använda det bevarade materialet?
- Hur ska materialet användas?
- Hur ofta ska insamling göras?
- Vid vilka tidpunkter är det lämpligast att samla in webbsidorna?
- Hur skall insamling göras, med vilken metod?
- Hur ska filerna insamlas, i vilka format och med vilka programvaror?
- På vilka lagringsmedium ska filerna lagras?
- Ska arbetet göras internt, av extern organisation eller i samverkan med andra?
- Hur säkerställs bevarande på lång tid?

Var och en av punkterna ovan måste brytas ner, analyseras och utredas inom den egna organisationen. Varje val som görs och beslut som tas ska bygga på ett medvetet agerande. Detta innebär att andra alternativ väljs bort vilket också måste bedömas och motiveras. Mer om punkterna ovan kommer i de följande kapitlen, 4-12 i denna rapport.

[6] Tryckfrihetsförordningen: http://www.notisum.se/rnp/sls/lag/19490105.htm

3.2 Policydokument och handlingsplan

De beslut som tas på frågorna i punkten ovan ska leda till ett skrivet policydokument med en handlingsplan. Implementering av en webbarkiveringspolicy är en viktig del av en organisations rättsliga och reglerade ansvar för material som publicerats digitalt.

Policydokumentet ska beskriva arbetet med att utreda och ta beslut om frågorna ovan. Men det är inte enbart besluten som bör bevaras utan även en summering av hur diskussionerna drevs under förstudien. På så sätt får man till exempel en bättre förståelse för varför valet av ett visst antal insamlingar per år gjordes av olika delar av webbplatsen, något som kan vara viktigt att förstå när beslut ska omprövas i framtiden.

Handlingsplanen ska beskriva hur arbetet med webbarkivering ska bedrivas i framtiden, från och med när förstudien är färdig och så länge webbarkivering ska pågå. Det gäller frågor som vilken metod som ska användas, vilken mjukvara som är lämplig och med vilka inställningar.

Förutom att beskriva val, metoder och verktyg är det även viktigt att besluta om och dokumentera vem som är ansvarig för de olika uppgifterna, vilka roller som ska arbeta med de olika aktiviteterna, när arbetet ska göras och hur långt fram i tiden en eventuell översyn ska göras och vem som då har till ansvar att initiera den.

3.3 Definiera begrepp

För att underlätta förståelsen för webbarkiveringspolicyn och även undvika missförstånd emellan personerna som ska arbeta med detta rekommenderar vi att göra en begreppsanalys i projektets inledningsskede. Speciellt viktigt är det att ha ett gemensamt språk och därför ha klara definitioner av begrepp som exempelvis: domän, webbplats eller sajt, hemsida, webb, startsida, insamling, webbcrawl, att arkivera, att bevara. Bifoga även denna begreppsanalys till policy och handlingsplan.

3.4 Beskriv organisationen och webbplatsen

Beskriv också organisationens syfte och syftet med webbplatsen. Vilka uppgifter ansvarar organisationen för? Hur är man organiserad? Hur har organisationen utvecklats/förändrats över tiden?

Vad används webbplatsen till? Vilken funktionalitet finns? Vilka typer av objekt? (filmer, ljud, formulär, databasgenererade sidor osv). Om det finns formulär, hur används de och vad hände med de data som skrevs in i de olika fälten? Om sidor skapades ur innehåll som fanns i databaser, vilken information var det och var finns den arkiverad?

Denna information ska även finnas med som beskrivande metadata till den insamlade webbplatsen och uppdateras vid behov, som till exempel vid förändringar inom organisationen eller när större förändringar har gjorts på webbplatsen.

4. Syfte

Syfte: Varför webbplatsen ska bevaras och vad det arkiverade materialet ska användas till är själva grunden för bevarandestrategin. Dessa frågor måste utredas och besvaras först, innan tankarna på hur det tekniskt ska göras kommer in i diskussionen. Insamling med hjälp av en speciell metod kan i värsta fall innebära att materialet inte kan användas som det var tänkt.

4.1 Varför bevara webbplatsen?

Det kan finnas flera olika syften för att bevara en webbplats och dessa syften kan ibland ställa olika krav på vilken metod som är lämpligast att använda.

Som exempel på olika syften kan nämnas:

- Uppfylla krav på myndigheten
- Juridiskt bevisvärde
- Historiskt värde
- Ökad trovärdighet
- Återanvändning

Läs mer om dessa syften i följande stycken:

Uppfylla krav på myndigheten
Det kan finnas många anledningar till att starta arbetet med att bevara den egna webbplatsen. I en del fall kan det röra sig om en myndighet som anser sig tvungen att börja insamling av webben eftersom webbplatsen räknas som en allmän handling och därför måste bevaras. (Läs mer om detta i avsnitt 5.1.)

Juridiskt bevisvärde
Att juridiskt kunna bevisa vad som har varit respektive inte varit publicerat på en webbplats under en viss tidpunkt kan vara viktigt för vissa organisationer. Detta kan kräva full tillgång till all skriven text inklusive tidstämpel för under vilken tidsperiod det varit publicerat. Det kan också krävas andra bevis på autenticitet, som en loggfil som visar vem som har publicerat texten eller att det kan styrkas att informationen inte är förändrad efter publicering. Å andra sidan kan det innebära att bevarande av sidans utseende är av underordnad betydelse.

Historiskt värde
Vi kan tänka oss att det bevarade materialet ska finnas kvar för att uppfylla forskarnas önskan om att kunna studera en organisation om hundra år eller mer, och då ser kraven annorlunda ut. Vad framtida forskare har för syfte med att undersöka webbplatser från i dag kan vi spekulera om. Men vi kan inte säkert veta om det i första hand är webbplatsens utseende och dess funktion och möjligheter som är målet för undersökningen eller om det viktigaste är den text som förmedlas för att få en bild av vår tid just nu.

Ökad trovärdighet
Genom att studera historiska webbsidor kan man få en kontinuerlig överblick över vad som hänt inom organisationen vid olika tidsperioder. Att bevara gamla webbsidor för att kunna visa upp vad man tidigare har publicerat visar att organisationen ser förmedlande och bevarande av information som en viktig uppgift. Det bör samtidigt öka trovärdigheten för organisationen som informationsspridare i nutid.

Återanvändning, Knowledge management
När webbmaterialet är välstrukturerat och kan hittas kan det också ses som ett värdefullt informationslager som kan återanvändas och på så sätt effektivisera verksamheten. Mycket av informationen finns kanske inte lagrat någon annanstans och skulle kunna användas på nytt till exempel i form av informationskampanjer.

4.2 För hur lång tid?

Frågan om hur lång tid materialet ska bevaras är nära förknippat med syftet för att bevara. Webbsidor som insamlas enbart för sitt juridiska värde kan ofta ha en relativt kort livslängd medan material som har historiskt värde ska finnas till för framtida forskning bör ha minst ett hundraårigt perspektiv. Om olika delar av webbplatsen bevaras med olika frekvens för olika syften kan de också behöva olika livslängd. Definiera i så fall vilka avgränsningar som gäller och hur länge de olika delsamlingarna ska bevaras.

4.3 Syftet som grund för teknisk lösning

När frågan om syftet med att bevara webbplatsen är analyserad och besvarad ska detta vara grund för hur insamling, bevarande och tillgängliggörande tekniskt ska göras.

Det är inte ovanligt att syftet med webbinsamling inte är ett enda utan att det handlar om en mix av olika syften. Organisationen kan till exempel dels ha krav på sig att bevara webbplatsen för framtida forskning men också ha behov av att kunna bevisa vad som har publicerats. Önskan om att uppfylla ett bevisvärde kan dessutom ibland bara gälla för delar av webbplatsen.

Om syftet innebär att informationen är det enda som behöver bevaras (text, figurer, osv.), ställer det helt andra krav på metoden än om webbplatsens utseende, uppträdande och funktion också måste bevaras (stillmallar, menyer, länkar osv.).

Om det finns två eller flera olika syften till att bevara bör bedömning göras om det räcker att använda sig av en enda metod eller om det är bättre att göra separata insamlingar och på olika sätt för att därefter gallra det som är bevarat för kortare period efter en viss tidpunkt.

5. Kravanalys

Syfte: Att få en helhetsbild av de krav som ställs på organisationen när det gäller bevarande av webbplatsen. Vilka externa krav vilar på organisationen och vilka interna krav finns?

5.1 Externa krav

Vilka skyldigheter som gäller för en organisation kan skilja sig beroende på till exempel om det är en statlig myndighet eller ej, som nämndes i början av kapitel 3. Därför behöver det utredas i ett tidigt skede vilka lagar och förordningar som kan innebära krav för webbplatsens insamling och bevarande.

Som exempel på lagar och förordningar kan nämnas: Tryckfrihetsförordningen, Arkivlagen, Arkivförordningen, Riksarkivets författningssamling samt myndighetsspecifika föreskrifter, Personuppgiftslagen, Sekretesslagen, Förvaltningslagen och Offentlighetsprincipen. Förutom dessa kan även frågor om copyright och licenser behöva utvärderas.

Tryckfrihetsförordningen (TF) säger att svenska myndigheters webbsidor är upprättade allmänna handlingar och Arkivlagen, som anknyter till TF, slår fast att ett arkiv är en del av det nationella kulturarvet.

Enligt detta skall arkivet bevaras, hållas ordnat och vårdas så att det tillgodoser:

1. Rätten att ta del av allmänna handlingar
2. Behovet av information för rättskipningen och förvaltningen
3. Forskningens behov

Riksarkivet säger att det vid leverans av webbsidor för arkivering ska medfölja en kompletterande beskrivning om vilken funktionalitet som har funnits på webbplatsen när den var tillgänglig för användarna. Navigeringen på en webbplats är kritisk och man ska även spara layout och klickbara länkar. Har det funnits en databas ska den levereras separat tillsammans med en beskrivning över hur användaren från webbsidan har kunnat söka i databasen.

Riksarkivet ger också ut föreskrifter (RA-FS) om bland annat tekniska krav på ADB-upptagningar. Läs mer på Riksarkivets webbplats http://www.statensarkiv.se

5.2 Interna krav

Förutom de externa krav som vilar på organisationen finns även ett antal interna krav och önskemål att uppfylla när insamling av webbplatser ska starta. Det kan vara viktigt för vissa organisationer att bevara tidigare publicerade webbsidor för att kunna bevisa vad som har publicerats och vad som inte har varit publicerat vid en speciell tidpunkt i händelse av en juridisk process.

Andra organisationer kanske inte ser bevarandet av gamla webbplatser i första hand som bevisvärde utan vill kunna återanvända material eller nå ökad trovärdighet. Avsnitt 4.1 beskriver detta mer i detalj.

Förutom de interna kraven ovan behöver organisationens tekniska krav utredas i ett tidigt skede. Vilken teknisk lösning som ska väljas kan komma att bero på organisationens nuvarande tekniska plattform och/eller på vilken slags kompetens som redan finns inom organisationen.

5.3 Kravbild

Beskriv vilka externa respektive interna krav som gäller för just den aktuella organisationen och vad dessa krav betyder för bevarandet av webbplatsen. Finns det krav som inte går att kombinera måste det utredas vilka krav som är överordnade.

Den sammanställda kravbilden ska, tillsammans med det formulerade syftet för insamlingen (kapitel 4.3) utgöra grund för val av teknisk lösning. Vid val av teknisk lösning ska denna utvärderas och kontrolleras så att den är förenlig med både syfte och kravbild
.

6. Urval

Syfte: Definiera webbplatsens avgränsning och avgöra vad som ska bevaras och vad som inte ska bevaras inom ramarna för webbarkiveringen.

6.1 Vad ska bevaras?

I detta projekt utgår vi från att bevarandet gäller en organisations egna publika webbplats och att denna ligger publicerad på en specifik huvuddomän, som t.ex. www.ldb-centrum.se[7]. Men det är sällan så enkelt att allt som är publicerat fysiskt ligger på den domänen eller att allt som finns publicerat också ska samlas in för att bevaras. Därför behöver det utredas vad som ska och inte ska insamlas.

En webbplats består ofta av information från flera olika källor, som till exempel från ett publiceringsverktyg, från databaser och/eller från andra informationssystem. Beskriv dessa delar och vilken information som kommer från (och skickas till) vilka system. För bättre översikt bör beskrivningen i text även kompletteras med en bild som visar dessa samband. Det finns också många webbplatser som har en egen domän men som länkar in information som fysiskt ligger på andra domäner. Kontrollera om sådana finns och besluta om de i så fall ska tas med i insamlingen eller inte.

MoSCoW-metoden[8] kan vara till hjälp för att klassificera webbmaterialet. Enligt denna hör allt material till av en av dessa fyra grupper:

- M: Things your institution **must** preserve *(måste bevaras)*
- S: Things you **should** preserve, if at all possible *(borde bevaras)*
- C: Things you **could** preserve, if it does not affect anyting else *(skulle kunna bevaras)*
- W: Things you **won´t** preserve *(ska inte bevaras)*

Vilka sorters material finns på webbplatsen? En del typer är svårare att samla in som t.ex. information som hämtas från databaser, HTTPS- eller lösenordsskyddade sidor, diskussionsforum, intranät, kartor, spel eller videofilmer. Utred vilka typer av material som kan tänkas ställa till problem vid a) insamlingen eller b) bevarandet på lång sikt. (Läs mer i CODA Slutrapport 2007[9].)

Besluta om insamlingens omfattning. Ska hela domänen bevaras eller bara ett urval av den? Jämför för- och nackdelarna med selektivt urval respektive insamling av en hel domän (6.2 resp. 6.3 nedan). När denna fråga är besvarad, utred i grova drag hur mycket data det handlar om och hur mängden förväntas växa med tiden. Denna information kan påverka tekniska val, som till exempel frågan om vilka lagringsmedia som är lämpligast.

Urval kan göras efter olika kriterier: allt som publicerats av en speciell avdelning, för vissa typer av material, inom vissa ämnesområden eller kring specifika händelser, material som riktar sig till vissa användare, material som inte bevaras i annan form och så vidare.

Om ett mindre urval ska bevaras – beskriv urvalet och motivera det. Vem ska välja/värdera informationen och för hur lång period gäller urvalet? När ska urvalet ses över igen och av vem? Hur ska man besluta om ny information som tillkommer ska bevaras? Skriv en

[7] I exemplet är ldb-centrum huvuddomän medan ändelsen se (för Sverige) är toppdomän.
[8] JISC-PoWR handbook: http://jiscpowr.jiscinvolve.org/files/2008/11/powrhandbookv1.pdf
[9] Kan laddas ner på LDB-centrums webbplats: http://www.ldb-centrum.se

urvalspolicy som besvarar dessa frågor. Denna policy måste ha en ansvarig person och den måste uppdateras med lämpliga intervaller och ändras vid behov.

6.2 Selektivt urval

Att göra ett selektivt urval från en webbplats och enbart samla in vissa delar kan ge flera fördelar:

- Det ger bättre kontroll över kvalitet och funktionalitet
- Materialet kan katalogiseras och bli enklare att få tillgång till
- Det blir mindre mängd att lagra fysiskt
- Inget ”skräp” lagras, vilket bör göra det enklare att hitta värdefull information

Men samtidigt finns det nackdelar med selektivt urval:

- Dyrare metod för insamling på grund av att det är mer tidskrävande
- Svårt att bestämma urval, vi vet inte vad som kan komma att bli intressant i framtiden
- Intressant/viktigt enligt vem? Urvalet görs efter subjektiva bedömningar
- Man tappar sammanhang, vad som har varit länkat till vad och vilka objekt som har legat under vilka kategorier, man ser aldrig helheten

6.3 Insamling av hel domän

Fördelarna med att samla in en hel domän är:

- Man fångar in allt på ett automatiskt sätt och med regelbundna intervaller
- Att crawla in en stor domän kan ta lång tid men kräver inte mycket mänskligt arbete
- Informationen på den insamlade webbplatsen kan visas upp med samma utseende som det hade när en person besökte sidorna
- Länkar inom domänen fungerar som tidigare, det vill säga att man kan klicka sig fram mellan sidor som man gjorde när webbplatsen var publicerad på webben
- Nya sidor eller subdomäner som skapas under tidens gång kommer automatiskt med i insamlingen, under förutsättning att de är publicerade under samma huvuddomän

De nackdelar som är finns är till exempel dessa:

- Man får in mycket material vilket innebär att det kan bli svårare att hitta i det
- Sämre kontroll över materialet eftersom insamling sker med vissa intervaller, information som funnits på webben mellan dessa intervaller kommer aldrig med

6.4 Vad ska inte bevaras?

En avgränsning över vad som *inte* ska bevaras är nästa område att besluta om. När en hel domän ska samlas in är det vanligt att avgränsningen dras så att allt material som ligger under domänen samlas in medan exempelvis länkar som går till andra domäner bryts.

Om det på någon sida finns RSS-flöden[10] för att visa information från andra webbplatser ska man bestämma om och i så fall hur sidans ursprungliga funktionalitet ska dokumenteras.

[10] RSS används för att visa sammanfattande eller fullständigt innehåll av text från exempelvis webben, tillsammans med en (permanent) länk till ursprungsplatsen. (Wikipedia 2009-03-25)

Webbsidor som hämtar information från en databas och publicerar denna på sidan får olika utseende och innehåll beroende på vilken data som visas. Ibland är det användaren som väljer vilken data som ska visas och det innebär att en och samma webbsida kan visas upp i ett mycket stort antal olika varianter.

Om databasens innehåll kommer att bevaras i sin fullständiga form är det kanske lämpligt att webbinsamlingen bara visar ett exempel på hur sidan har sett ut, kompletterat med dokumentation över var innehållet togs ifrån och var framtida användare kan hitta materialet. En handbok från JISC[11] gör denna avgränsning för viss information på webben som kommer från externa system och menar att dessa ses som hermetiska system med egen administration och egna resurser och att de därför inte riskerar att försvinna som annat material som kanske enbart publiceras på webbplatsen.

Vissa typer av innehåll är kända för att kunna bli problematiska vid insamling, som till exempel kalendrar. Om man väljer att samla in sidorna genom användande av verktyg för webbcrawling riskerar sidorna att bli så kallade ”crawler traps” där mjukvaran kan fastna i en loop utan slut. Mer om detta finns att läsa i del 2 som handlar om crawlingsverktyg.

En annan fråga att ta hänsyn till är de ingående filernas kvalitet. Finns det material som kan bevaras i annan kvalitet än det som publiceras på webben? Exempel på detta kan vara om en webbplats innehåller många och stora fotografier som kanske kan lagras i en lägre kvalitet än den ursprungliga för att ta mindre plats. Som alltid gäller det att göra en avvägning, hur viktig är den sparade mängden gentemot de förluster i kvalitet som förminskningen innebär?

Ska alla filformat som används på webbplatsen samlas in eller bör filtrering av vissa format göras för att underlätta för bevarande på lång tid? Eller ska vissa filformat undvikas och istället konverteras till andra? Ska det i så fall göras innan eller vid insamling? Rent generellt kan det vara klokt att överföra dessa beslut till den aktiva webbplatsen, alltså införa regler på vilka filformat som ska få användas på webbplatsen proaktivt för att på så sätt underlätta det kommande bevarandet.

De beslut om avgränsning som tas ska dokumenteras men också motiveras. Beskriv metoder och rutiner för gallring: vilka delar som ska gallras bort, varför de anses mindre viktiga, när, hur och av vem arbetet ska göras.

När det handlar om att avgränsa information som ligger inom egna domänen är det viktigt att återigen komma ihåg att för en svensk myndighet är webbplatsen en allmän handling. Det innebär att de regler som gäller för gallring av allmänna handlingar även gäller för det material som är publicerat på myndighetens webbsidor och därför kan en gallringsföreskrift från Riksarkivet krävas.

[11] http://jiscpowr.jiscinvolve.org/files/2008/11/powrhandbookv1.pdf (2009-04-21)

7. Användare

Syfte: Definiera vem eller vilka som ska använda sig av de arkiverade webbsidorna och på vilket sätt dessa personer ska få tillgång till materialet.

7.1 Vem ska använda materialet?

Ska det överhuvudtaget finnas möjligheter till åtkomst förutom för arkivpersonal? Om så är fallet: gruppera de framtida användarna och beskriv vad dessa har för syfte med att använda materialet. Beskriv även vilken behörighet de ska ha till olika delar av materialet och vilka rättigheter varje grupp ska ha när det gäller att titta på, flytta och göra förändringar. Kan rättigheterna behöva ändras och i så fall vem får besluta om detta? Förväntas det tillkomma nya grupper av användare under tidens lopp och vem ska besluta om vilken behörighet dessa ska få till arkivmaterialet?

Planera också för hur man kan säkerställa att obehöriga hindras från att få åtkomst till materialet. Om vissa delar av materialet är under sekretess eller av annan anledning måste behandlas extra varsamt ska det utvärderas hur detta praktiskt ska göras, som om det ska finnas inloggning med användarnamn och lösenord till systemet och om aktiviteterna ska loggas även i de fall som en användare bara har sökt bland materialet utan att göra förändringar. Kommer restriktionerna att gälla för en viss tidsperiod eller för all framtid?

Även om ändringar inte bör göras på arkivmaterial kommer det att uppstå situationer när detta blir nödvändigt för att bevara de digitala filerna för lång tid. Det kan handla om överföring till andra datalagringsmedium och därmed nya sökvägar till filerna eller konvertering från ett filformat till ett annat som är mer stabilt och säkert. Information om dessa förändringar, vem som har gjort ändringar, vilka filer som berörts och tidpunkt för ändringarna, bör lagras tillsammans med materialet. För ökad säkerhet bör det göras i form av loggfiler som genererats av systemet.

7.2 Hur ska materialet användas?

På vilka sätt ska man komma åt de bevarade sidorna? Hur ska man kunna söka? Efter tid för när sidan samlades in, i en enda insamling eller i flera samtidigt? Krävs det fungerande fulltextsökning på alla filer? Dessa åtkomstkrav ställer direkta krav på vad som måste finnas med vi inleverans och därmed krav på insamlingsverktyget.

Även här bör de olika användargruppernas behov analyseras var för sig eftersom deras behov av material och syfte för att komma åt det kan skilja sig. I slutänden kan det innebära att det behövs flera olika vägar till åtkomst, som att olika gränssnitt skapas för de olika grupperna.

Ska materialet bara kunna ses i vissa datorer inom organisationen eller är webbaccess lämpligt för delar av eller för alla insamlade webbsidor? Detta kan skilja sig mycket mellan olika organisationer men framför allt beroende på materialets karaktär och känslighet och av de lagar som gäller, som till exempel Personuppgiftslagen (PUL).

7.3 Beskriv användningsfall

För att kunna förutse framtida användande av lagrat material rekommenderas att beskriva ett antal olika användningsfall för de olika grupperna.

Framtida användare kan med fördel grupperas enligt de syften som analyserades fram i kapitel 4.1. Beskriv till exempel följande användningsfall:

- Person som ska kontrollera de insamlade filerna (se kap 9.7)
- Arkivarie som vill kontrollera att organisationen lever upp till de krav som ställs, som att kunna visa upp webbplatsen med det utseende som det hade
- Personal inom organisationen som letar bevis till en juridisk process
- En forskare som om etthundra år studerar organisationen
- Personal som internt vill återanvända lagrat material samt
- IT-personal som utför bevarandeåtgärder som migrering eller konvertering

Användningsfallen kan belysa speciella krav som i sin tur påverkar hur insamling av webbsidorna tekniskt måste göras. Exempel på detta kan vara krav på tidstämpel för all publicerad text för att duga som bevis i en juridisk process eller om användning kräver att fulltextsökning ska kunna göras på all text som har publicerats.

8. Frekvens och tid

Syfte: Att bestämma hur ofta och vid vilka tidpunkter webbplatsen bör samlas in.

8.1 Hur ofta ska webbplatsen bevaras?

Även här är det syftet med att bevara (som diskuterades i kapitel 4) tillsammans med kravbilden (från kapitel 5) som utgör grund för avgörandet om hur ofta och vid vilka tider webbplatsen bör samlas in. Om syftet är att ha kvar exakt all information som bevis kan någon form av ständig övervakning krävas, så att samtliga förändringar bevaras.

Om syftet i stället är att bevara webbplatsen för dess historiska värde och det inte finns någon lag som säger annat kan organisationen själv avgöra vad man anser vara tillräckligt ofta. Viktigt är att de val som görs är väl underbyggda så att man är medveten också om vad som inte kommer med när insamling endast görs ett visst antal gånger per år.

En annan möjlighet är att webbplatsen enbart ska samlas in vid ett enda tillfället. Detta kan vara fallet exempelvis om organisationen som äger webbplatsen kommer att läggas ner eller slås ihop med en annan organisation eller om själva webbplatsen läggs ner, eventuellt för att återuppstå i en helt annan form.

Förutom syfte och krav finns ännu ett par parametrar som påverkar beslutet om hur ofta insamling ska göras, nämligen de tekniska och ekonomiska begränsningar som råder inom organisationen. En grov kalkyl kan därför behöva göras i syfte att besluta om hur ofta insamling ska utföras.

Kostnadsberäkningen bör utreda frågor som: hur stor är webbplatsen och hur mycket kommer att behöva lagras under ett år? Hur mycket kan detta material förväntas växa under de kommande åren? I hur många kopior ska det insamlade materialet lagras? Hur mycket tid förväntas insamling, lagring och hantering kräva?

En beräkning som denna är inte i första hand tänkt för att räkna ut den totala kostnaden för insamling och lagring utan för att ha som grund för avgörandet om hur ofta webbplatsen ska bevaras. I slutänden handlar det om att väga fördelar mot nackdelar.

8.2 När ska sidorna samlas in?

Om webbplatsen ska bevaras ett visst antal gånger per år ska det beslutas om vilka tidpunkter som är mest lämpliga för insamling. För framtida forskning bör webbplatsens insamlande så långt som möjligt spegla verksamheten, även om det alltid bara blir en ögonblicksbild över hur webbplatsen såg ut vid en specifik tidpunkt. För exempelvis en utbildningsorganisation, som i pilotprojektet vid LTU, följer verksamheten skolårets indelningar i terminer. I ett fall som detta bör insamlingen ha som mål att visa viktiga skeden under året istället för att göras mitt i sommaruppehållet eller under årsskiftet då aktivitetsnivån normalt är som lägst.

Finns det delar av webbplatsen som behöver bevaras mer frekvent för att leva upp till de tidigare ställda kraven? Bestäm i så fall även vilka sidor det rör sig om, hur ofta de ska samlas in och vid vilka tidpunkter.

Beskriv även inom detta område de val som gjordes. Varför ansågs det lämpligast att utföra insamling n gånger per år och varför valdes de speciella tidpunkterna? Vad kommer vi att gå miste om i och med detta och hur viktigt är det? För att besvara denna fråga på bästa sätt rekommenderas att analysera vad organisationen har för verksamhet under ett år och utifrån detta välja lämpliga tidpunkter för insamling av webben.

En parameter som kan komma att påverka val av tidpunkt är när den personal som ska utföra arbetet med insamling, kvalitetskontroll och lagring har bäst tid för detta. En annan är när en insamling bör göras för att i så låg grad som möjligt störa den övriga verksamhetens behov av datakraft.

Om beslutet blir att använda ett webbcrawlingsverktyg kan dessa ofta schemaläggas i förväg, så att en insamling automatiskt startar vid en tidigare vald tidpunkt, exempelvis nattetid eller under en helg om aktiviteterna i verksamheten är lägre då. Det kan också vara möjligt att vid schemaläggning bestämma hur mycket kraft insamlingen får använda, något som då i stället medför att den tar längre tid.

Viktigt är att direkt efter varje insamling göra kvalitetskontroll av det insamlade materialet för att så snabbt som möjligt upptäcka eventuella problem. Upptäckten av dem innebär ofta att en ny insamling behöver köras igång, och det bästa är om det kan göras så snart som möjligt. (Läs mer om detta i kapitel 9.7.)

9. Teknisk lösning

Syfte: Att besluta om vilken teknisk lösning som bäst lever upp till de syften och den kravbild samt övriga beslut som analyserats fram i föregående kapitel.

9.1 Vilken metod bör användas?

I skrivande stund (hösten 2008) finns tre vanliga metoder för att samla in webbsidor:

- Webbcrawlingsverktyg
- Export ur publiceringsverktyg
- Kopiering av befintlig mappstruktur

Viktigt är att jämföra metoderna ovan med om och hur väl de uppfyller kraven i kravanalysen och om det är nog för att svara mot syftet med insamlingen. Räcker det med en metod eller krävs en kombination av olika metoder? Ska båda metoderna i så fall användas på hela webbplatsen eller på vilken del passar vilken metod?

Uttag med ett **webbcrawlingsverktyg** ger en kopia av webben som den såg ut för en användare i just det ögonblick som insamlingen gjordes. De sidor som man kunde klicka sig fram till på den publicerade webbplatsen kan man också klicka sig fram till när sidan är nedladdad. De sidor som ligger på webbplatsen utan att vara tillgängliga för en användare utifrån kommer däremot inte med om man väljer att crawla ner innehållet. De olika verktygen kan lagra filerna i olika form, vissa gör en kopia av mappstrukturen medan man i andra som exempelvis Heritrix väljer om filerna ska lagras en och en eller i form av ett paket. Insamling med hjälp av webbcrawlingsverktyg är den vanligaste metoden att använda. (Läs mer om webbcrawling och olika verktyg i Del 2 i denna rapport.)

Vissa crawlingsverktyg ger användaren möjlighet att utföra en inkrementell insamling av webben. Detta innebär att hela webbplatsen samlas in den första gången verktyget körs medan senare insamlingar bara tar med sidor som har förändrats från föregående crawling. Detta kan kännas som en klok strategi som gör att insamlingen sker snabbare och för att det blir mindre material att lagra. Men samtidigt innebär det vissa risker på lång tid eftersom det skapar beroenden mellan de olika versionerna av arkiverade webbplatser. Överföring till annan databärare, där sökväg till filer förändras eller gallring av en version kan göra att allt material, förutom den första och fullständiga crawlingen, blir oanvändbart. Däremot kan inkrementell insamling vara ett lämpligt tillvägagångssätt om mindre delar av en webbplats behöver samlas in ofta men däremot inte bevaras på lång sikt.

Numera är många och framför allt de stora webbplatserna skapade med hjälp av ett **publiceringsverktyg** av något slag, som till exempel ett CMS, (Content Management System). I ett sådant lagras innehållet på en sida ofta som en post i en databas medan sidans utseende bestäms av en separat stilmall. Ofta har dessa CMS någon typ av exportfunktion som kan användas för att ta ut den text som ligger på webbsidorna. I vilken form texten kan exporteras varierar, men XML verkar vara ett vanligt exportformat. Om export ur publiceringsverktyget kan vara en lämplig metod att använda bör man kontrollera i vilket filformat de exporterade filerna tas ut, om det går att följa länkar inom domänen och om utseendet på sidorna bevaras, i de fall där detta är ett krav. Att bevara utseendet kan ofta vara helt omöjligt, vilket gör detta till en metod som kan fungera för att bevara *innehållet* från en webbplats men inte webbplatsen själv.

Kopiering av befintlig mappstruktur ger en exakt kopia av webbplatsen, inklusive den struktur som den låg lagrad i. Alla filer kommer med, oavsett om de vid det valda tillfället var publicerade eller inte. Det är också enkelt att kontrollera att allt verkligen har kommit med genom att jämföra ursprungsmapparna med de kopierade mapparna. Metoden är dock relativt resurskrävande att använda och inte automatiserad som användande av crawlingsverktyg.

För att kunna följa den lagrade webbplatsens länkar är det enklast om relativa länkar har använts, i annat fall behöver alla interna länkar ändras från absoluta till relativa.

Absoluta länkar är sådana där koden består av en fullständig sökväg till filen där den ligger publicerad på webben, som t.ex: Hem

Relativa länkar däremot visar var filen ligger, relaterat till den fil som länken finns i, till exempel: Hem

Bevarande av absoluta länkar kommer att innebära att när man klickar på en sådan länk i arkivet kommer webbläsaren att försöka komma ut på Internet för att söka efter filen enligt den fullständiga sökvägen.

Om webbplatsen är dynamiskt uppbyggd med hjälp av en eller flera databaser eller på andra sätt direkt beroende av någon typ av mjukvara är kopiering av befintlig mappstruktur inte heller någon bra metod. Det kräver i så fall att även den tekniska lösningen bevaras och hålls körbar för all framtid. Det i sin tur kräver också att kompetens för att använda tekniken och dessutom för att anpassa den för framtida hård- och mjukvara hela tiden finns kvar inom organisationen.

Metod två och tre, export ur publiceringsverktyg respektive kopiering av befintlig mappstruktur kräver både tillstånd från och ett aktivt samarbete med webbplatsens ägare. Däremot är det oftast enkelt att samla in en webbplats även utan ägarens hjälp genom att använda en webbcrawler. Insamling med webbcrawlers görs i stora drag på samma sätt oavsett hur webbplatsen ser ut och är uppbyggd medan de andra metoderna kräver individuell anpassning för varje domän som ska samlas in.

Om filerna ska kunna användas som bevis måste en noggrann undersökning göras av vilka krav som detta ställer på materialet. Måste det kunna bevisas under vilken tidsperiod som informationen varit publicerad på webbplatsen? Hur ska det kunna bevisas att ingenting har ändrats eller tagits bort efter skapandet/lagringen? Hur kan man bevisa att obehöriga inte har kunnat komma åt materialet? Hur loggas förändringar, som de processer som är nödvändiga att utföra vid till exempel konvertering till annat filformat? Krävs användande av tidstämplar och/eller checksummor?

9.2 Hur ska filerna samlas in?

Om insamling med hjälp av webbcrawlingsverktyg anses vara den bästa metoden är nästa steg att besluta om vilket verktyg som ska användas. Det viktigaste i detta val är att kontrollera att man i och med användande av det specifika verktyget verkligen kommer att uppnå sitt syfte och svara på de krav som ställts. (Läs mer i Del 2.)

Innan verktyget ska användas i drift behöver många undersökningar göras och beslut tas, som till exempel: hur ska verktyget konfigureras? Ska alla filformat tillåtas? Vilken eller vilka

startadresser (URL:er) ska användas? Kan insamlingen schemaläggas i förtid? Vems ansvar är det att schemalägga? Och om det inte finns funktion för att schemalägga insamling i förväg: vem är ansvarig för att starta insamlingen i rätt tid? Vilken tid på dygnet är lämpligast att göra insamling? Hur mycket kraft (bandbredd) tar processen i anspråk och kommer det att påverka den övriga verksamheten?

Om filerna i stället ska exporteras ur publiceringsverktyg eller om hela mappstrukturen ska kopieras behöver liknande frågor utredas, besvaras och dokumenteras.

9.3 I vilka filformat ska sidorna lagras?

Förutom val av verktyg ska beslut även tas om i vilket eller vilka filformat filerna ska bevaras. Ska filerna lagras hoppackade till ett paket per insamling? Enligt OAIS-modellen[12] är ett grundkoncept det arkivinformationspaket (AIS) som innehåller det digitala objektet självt plus all den information som behövs för att förstå, presentera och hantera det.

Val av en viss metod eller ett specifikt verktyg kan innebära att organisationen själv inte kan välja vilka filformat som ska användas pga. begränsningar i det valda verktyget eller metoden. Kopiering av en hel webbplats, utan användande av något webbcrawlingsverktyg ger en kopia som ser ut och är strukturerad i samma form som den ursprungliga webbplatsen. Användande av ett verktyg som Heritrix däremot ger fler valmöjligheter, dels kan filerna bevaras i ursprunglig form, dels som paket av filtypen ARC eller efterföljaren WARC vilka fungerar som omslutande format, där de ursprungliga filerna ligger inuti paketen.

Strategi för att välja lämpliga format, se CODA-FORM (sid 11 i CODA Slutrapport 2007[13]).

Det kan också vara klokt att före insamlingens start göra en inventering över vilka filformat som används på organisationens webbplats och analysera om dessa format är lämpliga eller om de kan förväntas innebära problem vid insamlingen eller senare vid bevarandet. Om man vid inventeringen upptäcker att det finns filformat som inte är bra bör man undersöka om de ska konverteras till andra format redan innan insamlingen startar.

9.4 På vilka lagringsmedium ska filerna lagras?

Svaret på denna fråga beror i första hand på vilken lagringsmiljö som redan finns inom organisationen men även på hur materialet ska användas. Krävs det snabb åtkomst till lagrade webbsidor? Kommer access att ske ofta eller sällan? Vad använder organisationen för lagringsmedium i dag? Är det mest effektivt att använda samma sorts hårdvara även till filer som ska bevaras på lång sikt?

Om insamling görs av två eller flera olika syften, ska filerna lagras olika länge? Hur ska detta markeras? Hur och när utsorteras den ena sortens filer bort och av vem?

Fler frågor av teknisk natur måste också besvaras, som: var ska filerna ligga rent fysiskt? På vilken server/vilka band och i vilken struktur? Finns det möjlighet till automatisk överföring av filerna till valt lagringsmedium och vem ansvarar för att detta initieras? Om automatisk

[12] http://public.ccsds.org/publications/archive/650x0b1.pdf
[13] Slutrapport CODA 2007, finns på LDB-centrums webbplats: http://www.ldb-centrum.se

överföring inte är möjlig – vem ansvarar för att överföringen blir gjort och hur och när ska det göras?

De arkiverade webbfilerna ska dessutom omfattas av samma säkerhetsrutiner som gäller för övrigt digitalt material inom organisationen. Hur ska backup göras och när? Ska filerna finnas i ett antal kopior? Ska olika typer av databärare användas, till exempel en kopia på disk och en på band? Lagring på optiska skivor (CD och DVD) rekommenderas inte på grund av deras relativt korta livslängd.

Om samma typ av databärare används till de olika kopiorna, ska man i så fall välja olika tillverkare eller generation för att öka säkerheten? Ska någon av kopiorna fysiskt bevaras på en annan plats, utanför organisationen? Ska ett exemplar betraktas som ”master” och endast användas för framställning av nya brukskopior? (Läs mer på sid 35 i CODA Slutrapport 2007[14].)

9.5 Lagringsstruktur och namngivning

Hur det insamlade materialet lagras på databäraren är till stor del beroende av vilken metod och vilket verktyg som har valts för insamlingen. Som exempel kan nämnas att olika webbcrawlingsverktyg lagrar i olika, förutbestämda strukturer där förändring kan innebära att sammanhanget mellan filerna går förlorat.

Kopiering av en hel webbplats kan verka vara en enkel metod och i vissa fall kan det vara det också, om webbplatsen är relativt enkelt uppbyggd till exempel av enbart statiska HTML-filer. Men problem kan uppstå som till exempel att navigeringen mellan sidor inte fungerar på grund av att länkarna består av absoluta sökvägar som pekar på filerna där de ligger publicerade. De flesta problem uppstår dock när dynamiska sidor ska kopieras, där innehåll kommer från databaser, mallar eller skapas via script. Det kommer troligtvis att bli en utmaning att få till samma miljö i arkivet som för den aktiva webbplatsen.

Utred i förväg hur materialet ska lagras på hårdvaran men också vilka regler som ska användas för konsekvent namngivning av mappar och filer. Tidsaspekten är viktig och det enklaste sättet att skilja de olika insamlade versionerna av webbplatsen från varandra. Datum bör därför alltid finnas med i namnet, inte bara för mapparna där webbsidorna lagras utan även för de dokument som ska medfölja och som finns med för att öka förståelsen för materialet.

9.6 Metadata

För att människor ska kunna använda och förstå det lagrade materialet även efter att en lång tid har passerat behövs en hel del dokumentation. Metadata brukar beskrivas som ”data om data” och att den finns och är läsbar är en förutsättning för framtida användning. Beskriv därför hur insamlingen är gjord, med vilken metod och med hjälp av vilka verktyg eller programvaror och vilka versioner.

Om webbcrawler har använts ska även inställningar, valda seeds[15] och andra gjorda val beskrivas och medfölja de lagrade digitala filerna. Verktygen brukar också spara information

[14] Kan laddas ner på LDB-centrums webbplats: http://www.ldb-centrum.se
[15] Start-URL:er, de webbadresser som väljs ut som startpunkter för insamlingen. Ofta en domänadress.

om insamlingen, när den gjordes, hur lång tid det tog och hur resultatet blev. Denna information bör lagras med filerna men dessutom kompletteras med annan metadata som behöver skrivas in manuellt. Lämpligt är att lagra filerna efter år och datum för när de samlades in, för att underlätta framtida användning.

Förutom tenisk metadata över insamlingen i stort behövs även metadata om de enskilda filerna och om de lagras i paket även information om själva paketen.

Administrativ och beskrivande metadata krävs också för att framtida forskare ska få förståelse för materialet. Tänk därför på att även beskriva organisationen och dess verksamhet samt den miljö som den verkade i. Större händelser eller förändringar som berör organisationen och/eller webbplatsen och som har inträffat mellan de olika insamlingarna bör också dokumenteras kontinuerligt för att öka förståelsen.

Enligt vilket metadataschema ska informationen lagras? Vilka metadata är minimikrav och vilka är extra/utökad metadata. Hur ska arbetet ovan utföras och vem är ansvarig?

Även i denna fråga är det en balansgång mellan vad som kan tänkas behövas för framtiden och hur det kan göras så automatiskt som möjligt. Att föra in metadata om insamlingarna manuellt är resurskrävande men kan ändå vara nödvändigt.

Om webbplatsen analyseras med något verktyg som rapporterar om hur många gånger sidorna har visats och hur länge besökare stannat på webbplatsen och dylikt bör även denna information lagras tillsammans med webbfilerna, företrädesvis i PDF/A eller XML-format. Dessa format ses i nuläget (våren 2009) som de mest lämpade för allt långsiktigt bevarande av information som ska kunna läsas direkt av människan, naturligtvis under förutsättning att det kan visas på en datorskärm eller dras ut på papper.

9.7 Kontroll

Vilken metod för insamling som än används kan det uppstå fel, det kan till exempel bero på att verktyget inte fungerar som det ska eller på en felaktigt gjord inställning av verktyget. För att upptäcka sådant så fort som möjligt krävs det en rutin för att kontrollera varje nedladdning så snart den är gjord så att man vid behov kan göra en ny insamling när fel har upptäckts.

Två typer av kontroll behöver göras: dels om allt material har kommit med vid insamlingen och dels att sidorna går att se och att navigeringen mellan sidorna fungerar. Om länkar finns av olika slag, som textlänkar, drop-down menyer och knappar, bör alla typer kontrolleras. Undersök också att det fungerar att följa navigeringen nedåt i de olika nivåerna samt att länkarna verkligen leder till de insamlade filerna och inte ut till material som ligger publicerat på nätet.

Kontrollera olika filtyper: att text bevaras i rätt form så att även tecken som svenskans å,ä och ö är läsbara, om olika bildformat finns att alla kan ses och att video och/eller ljudfiler är möjliga att spela upp. Om webbplatsen länkar till dokument i format som till exempel Microsoft Word eller PDF, kontrollera att de har fångats in och kan öppnas och läsas. Om webbplatsen ska arkiveras enligt Riksarkivets regler är format som Word ej godkända utan ska konverteras till lämpligare format före arkivering.

Om webbplatsen innehåller information om datum och tid ska denna information frysa (bli statisk) och inte visa aktuell tid utan den tid som insamlingen gjordes. Om webbplatsen är uppbyggd med hjälp av ramar (frames) ska det kontrolleras att sådana sidor visas korrekt. Innehåller webbplatsen någon form av söktjänst bör det kontrolleras om den fungerar även efter insamling. Om insamling gjorts med hjälp av webbcrawler fungerar sökning normalt inte efter insamling, något som man bör vara medveten om.

Om det är en stor webbplats som har samlats in kan det vara omöjligt att kontrollera allt utan man kan behöva avgränsa sig till att enbart utföra ett antal stickprov. När den första typen av kontroll ska göras är det en fördel att veta hur stor webbplatsen är, räknat dels i byte dels i antal filer.

Frågor att besvara för detta område är till exempel: Vilken kontroll ska göras på nedladdade filer för att se så att de fungerar och kan visas upp? Ska stickprov göras slumpmässigt eller bör speciella testfall skrivas som ska följas vid varje kontrollsituation? Vem är ansvarig för detta? Vad ska hända om fel upptäcks?

9.8 Arkivförteckning

En arkivförteckning är ett systematiskt uppställt register över ett arkivs innehåll, inbundna volymer, arkivboxar med lösa dokument, pärmar, fotografier etc. Ofta finns också text som närmare beskriver innehållet i arkivets olika volymer. Allt som arkiveras inom svenska myndigheter måste också förtecknas och beskrivas[16], och det är genom arkivförteckningens register som åtkomst görs till det arkiverade materialet.

Även det digitala webbmaterialet som arkiveras måste förtecknas. Det kan göras antingen i traditionellt pappersbaserad form eller med hjälp av arkivförteckningsprogram på dator. I dags dato finns två stycken välkända arkivförteckningsprogram för svenska användare: Visual Arkiv[17] från Arkivnämnden i Göteborg samt Klara[18] som tidigare hörde till företaget Imagon Systems, numera Know IT.

9.9 Att komma igång

När en förstudie har gjorts ska framför allt syfte och krav utgöra grund för vilken teknisk lösning som ska användas för insamling och arkivering av organisationens webbplats. Om ett nytt verktyg ska användas ska det inhandlas eller laddas ner, installeras, konfigureras och eventuellt även anpassas så att det går att använda i den specifika organisationen på ett effektivt sätt. Personalen som ska använda det behöver också få tid att lära sig verktyget. Om verktyget som ska användas är mer avancerat, som till exempel Heritrix kräver det en hel del tekniskt kunnande. Andra verktyg är enklare och kan köras med standardinställningar så att man i princip bara behöver skriva in vilken domän som ska insamlas.

Det kan därför i vissa fall krävas att organisationens IT-personal ger utbildning åt den som ska göra insamlingen men även till den eller de som ska ha behörighet till att leta i materialet.

[16] Statens arkiv: Ordna och förteckna arkivet: http://www.statensarkiv.se/default.aspx?id=1314&ptid=0
[17] Visual Arkivs webbplats: http://produktinfo.visualarkiv.se/index.php?page=3
[18] Klara, Know IT: http://www.knowit.se/KIT_templates/PageWide____1952.aspx

10. Modeller för webbarkivering

Syfte: Att diskutera olika modeller för webbarkivering, om arbetet ska göras internt, externt eller genom samarbete med andra.

Varje organisation som ska börja samla in sin webbplats med syfte att bevara den för lång tid framåt måste ta egna beslut om frågorna i föregående kapitel. Det innebär däremot inte att organisationen själv också måste utföra det praktiska arbetet.

Vi ser tre möjliga upplägg:

- Allt arbete görs internt inom organisationen
- Samarbete mellan flera organisationer
- Arbetet utförs av ett externt företag

Var och en av dem har för- och nackdelar, vilka diskuteras här:

10.1 Allt arbete görs internt inom organisationen

Fördelar med denna modell är att organisationen har total kontroll över aktiviteten och kan vara flexibel när helst förändringar behöver göras. Samtidigt har man det totala ansvaret för att allt ska fungera, vilket kräver till exempel ett stort kunnande inom organisationen såväl som en teknisk miljö som kan hantera materialet. Organisationen måste själv bekosta allt arbete, vilket inte minst på lång sikt kommer att innebära stora summor.

10.2 Samarbete mellan flera organisationer

Om flera organisationer med liknande behov samarbetar krävs det mindre resurser från var och en om lösningarna kan samordnas när det gäller hårdvara, utveckling, drift m.m. Ansvaret delas på alla parterna och även om varje organisation själv måste bestämma om hur den egna insamlingen ska göras kan parterna dra nytta och lärdom från varandras överväganden och beslut.

Men det finns även nackdelar med gemensamma lösningar, det tar tid att skapa former för organisationen såväl som för tekniken som ska användas. Varje samarbete bygger på såväl strategiska som operativa överenskommelser och varje organisation måste vara beredd på att förhandla och jämka och därmed inte alltid nå den optimala lösningen för egen del. Kontrollen och flexibiliteten blir också lägre än om allt arbete görs internt.

10.3 Arbetet utförs av ett externt företag

En organisation som varken har resurserna eller vill investera i att bygga upp eget kunnande eller teknisk lösning kan istället anlita ett annat företag till att göra det praktiska arbetet. Fördelarna är uppenbara, inga stora egna investeringar krävs och förhoppningsvis kan arbetet starta snabbt. En annan fördel är att det externa företaget ansvar för att ständigt ha rätt kompetens och de mest lämpade verktygen för att arbeta effektivt.

Nackdelen är att det inte finns så många företag som har kompetensen att arbeta med just detta område än. Det kräver också att kontrakt skrivs där det preciseras vad företaget har för ansvar och vad det leder till om något inte fungerar. Stora organisationer har enklare för att

förhandla fram en specialanpassad lösning medan små företag kan ha svårt att hitta någon som är intresserad av att ta på sig ett relativt litet uppdrag som inte följer en standardlösning. En annan nackdel är beroendet som uppstår gentemot ett externt företag. Problem kan exempelvis uppstå om företaget läggs ner, byter inriktning på sina tjänster eller köps upp av annat företag. Om bara en eller ett fåtal personer arbetar med denna fråga inom företaget innebär det också en risk.

Om ett externt företag ska användas för insamling och/eller bevarande av en organisations webbplats ska man kontrollera att företaget inte gör det med metoder som försvårar bevarande över tid. Exempel på detta är om företaget använder egna metoder, olämpliga verktyg eller teknikberoende sådana, inte följer standarder eller sparar materialet i osäkra filformat.

Det är också möjligt att kombinera dessa tre så att visst arbete utförs av organisationens egen personal medan andra delar utförs av externa företag eller i samarbete med andra.

Hur det än görs måste beslut om vad som ska göras, hur det ska göras och vilka tidpunkter som är mest lämpliga, bero på de val som gjorts i förstudien och inte överlåtas till annan person eller organisation. Om ett externt företag anlitas för insamling och/eller bevarande ska förundersökningen och handlingsplanen utgöra grunden för en kravanalys.

11. Handlingsplan

Syfte: Att klargöra för vad som ska göras, hur det ska göras, av vem och när.

11.1 Vem är ansvarig för arbetet?

I de olika kapitlen i denna rapport finns ett stort antal frågor som varje organisation behöver besvara inför arbetet med att börja samla in och bevara sin webbplats. De beslut som tas ska dokumenteras och diskussionerna inför besluten förklaras. Vilka valmöjligheter som fanns och varför man tog vissa beslut kan förklara för framtida beslutsfattare när handlingsplanen i framtiden ska uppdateras.

Speciellt viktigt är att samtliga uppgifter som ska göras när det gäller webbplatsens aktiviteter och övervakning har en ansvarig person och att denne person även vet hur han eller hon ska agera när någonting inte går enligt planerna. Ett exempel på detta är diskussionen i avsnitt 9.7 om hur de insamlade filerna ska kontrolleras och hur ansvarig ska agera om någonting inte ser ut att ha fungerat som det borde. Ytterst är det myndighetens arkivarie som bär ansvaret för det digitala material som enligt lag ska bevaras för eftervärlden.

Exempel på arbetsflöden som kan behöva beskrivas:

- Analysera (och utföra) hur materialet ska vara sökbart och hur det ska visas upp
- Välja parametrar för insamling (domäner, djup, länkar osv.)
- Starta och övervaka insamling
- Kvalitetskontrollera filerna (och hantera ev. problem)
- Lägga in metadata om filerna/insamlingen
- Lagra i arkivsystem
- Dokumentera processerna och ansvaret för dessa
- Ge behörighet för användare av arkivet (ev. flera olika typer av användare)

Handlingsplanen bör dels finnas i digital form tillsammans med de insamlade filerna och dels i pappersformat där det även framgår var de digitala filerna finns lagrade och hur man hittar bland dem. Den digitala versionen bör även den lagras i lämpligt filformat (t.ex. PDF/A) för att försäkra sig om att den går att läsa även i framtiden.

11.2 Bevarande över lång tid

Digitalt bevarande inbegriper alla de aktiviteter som krävs för att försäkra sig om att digitala filer kan tillgängliggöras så länge det finns behov av det. Det innebär aktiviteter som planering, insamling, dokumentation, bevarande, tillgängliggörande, resursallokering, val och användande av metoder och teknik för att försäkra sig om att digitalt material kan användas under lång tid. Med andra ord handlar det om en mix av både tekniska och organisatoriska frågor om vilka man kan läsa mer om i artiklar och avhandlingar av Runardotter samt Runardotter et al (2009, fortcoming).

Organisatoriskt behövs en stark förankring för projektet så att det tilldelas resurser, inte bara under den första förstudiefasen eller i uppstartsskedet utan på lång sikt. Ledningen måste känna till de krav som ställs på webbplatsens bevarande men även vilken nytta man kan uppnå. Som vid all digital lagring ska man vara medveten om att det finns kostnader som är svåra att beräkna, allmänt kan man dock räkna med att kostnaden för lagringsmedia bara motsvarar ca 20 % av den totala kostnaden, resten är administrativa kostnader. Att underskatta

kostnaderna är ingen god idé utan kommer med stor sannolikhet att leda till att ge projektet dåligt rykte, att man tvingas kompromissa med kvaliteten eller att projektet avslutas i förtid[19].

Digitala filer är enkla att förändra, förstöra och radera. För att öka trovärdigheten för att en fil verkligen är äkta kan man arbeta med att skapa kontrollerade processer, loggfiler och övervakade och dokumenterade händelser när någonting har gjorts med en fil (se kap 4.1). För att kvalitetssäkra sitt digitala arkiv rekommenderas att utvärdera och förbättra arkivet enligt konceptet ”Trusted digital repository[20]”. Enligt detta ska digitala arkiv leva upp till ett antal krav inom sju olika områden för att betraktas som säkra. De områden som utvärderas är: överensstämmelse med OAIS[21], administrativt ansvar, organisatorisk livskraft, ekonomisk hållbarhet, säkra teknologier och procedurer, systemsäkerhet samt ansvarsfullhet (egen översättning).

För att kunna ta hand om en digital samling på lång sikt krävs det att man har en bra kontroll över vad man har för material. Detta gäller även för det som bevaras från webbinsamlingarna och en fråga att besvara är just hur materialet och själva insamlingen av den ska dokumenteras för att kunna hållas under kontrollerade former.

Utred därför frågor som: Vilken statistik behövs för bevarandet? Exempel på statistik: Hur stor är samlingen i sin helhet? Vad finns i arkivet, hur många filer och av vilka filformat? Om filerna är samlade i paket eller mappar, hur kan vi ha kontroll över vad som finns inuti varje paket eller mapp? Vilka beroenden finns mellan filformat som lagras och den mjukvara och/eller hårdvara som krävs för att visa upp dem?

Som det nämns i avsnitt 9.3 är det klokt att arbeta proaktivt när det gäller bland annat vilka filformat man tillåter på webbplatsen. Lagring av filformat som misstänks kunna vara problematiska innebär annars att konvertering kan bli nödvändigt ganska snart. Det kan därför vara ett klokt drag att byta till lämpligare format redan för det material som är publicerat på webbplatsen, innan insamling startar.

I handlingsplanen bör beskrivas vem som ska vara ansvarig för att kontrollera när filformat som finns i arkivet behöver konverteras, hur ofta kontrollerna ska göras och hur det ska gå till. Var kan man hitta information om vilka format som är riskformat, eller i annat fall: vem är bäst lämpad att ta beslut om vilka format som är riskfyllda och vad man bör konvertera till?

Omvärldsbevakning krävs även för andra områden än filformat, man måste följa hur tekniken i övrigt utvecklas (lagringsmedia, mjukvara, operativsystem osv.) men också andra händelser som kan påverka tillgången till dessa som om företag går omkull, förändrade licensmodeller eller nya lagar som stiftas.

På en övergripande nivå behöver man inom organisationen avgöra också hur ofta webbarkiveringsstrategin och handlingsplanen behöver ses över och vem som ska vara ansvarig för att aktiviteten blir gjord.

Tekniskt finns ännu fler frågor som kan påverka livslängden på det bevarade materialet, som till exempel på vilken hårdvara det har lagrats på och vilka säkerhetsrutiner som krävs, oavsett

[19] Slutrapport CODA 2007: http://www.ldb-centrum.se
[20] Rapport om TDR: http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf
[21] Open Archival Information System, http://public.ccsds.org/publications/archive/650x0b1.pdf

typ av digitala samlingar. Dessa frågor ligger dock utanför detta projekt eftersom de redan i dag måste hanteras för den dagliga IT-verksamheten (se kap 9.4).

11.3 Lärdomar

Analysera syftet med bevarandet av webbplatsen och vilka krav som ställs från externt och internt håll innan beslut tas om hur insamlingen ska göras och vilka verktyg som kan vara lämpliga.

Redan när en webbplats planeras och skapas bör det göras med tanke på framtida bevarande. Framför allt bör man tänka på hur den är strukturerad och vilka filformat som ska tillåtas. Följ riktlinjerna i Vägledningen 24-timmarswebben[22], utgiven av Verva[23] år 2006.

Om inte hela webbplatsen ska bevaras utan bara delar av den bör det som ska bevaras märkas på något sätt redan vid skapandet eller publiceringen.

Förståelsen för att webbsidorna ska bevaras för framtiden är också viktig att ha med sig när en helt ny webbplats ska skapas och vid upphandling av nya publiceringsverktyg.

Följ standarder inom området och undersök även vad Riksarkivet förespråkar, som till exempel när det gäller vilka format som är lämpliga för bevarande.

Undvika så långt som möjligt användande av PowerPoint- ,Word- och andra fristående dokumentformat på webbplatsen. All text bör så långt möjligt skrivas direkt i HTML-format.

Skriv en bevarandeplan för organisationens digitala material, i det kommer även de insamlade filerna från webbarkiveringen att ingå. En del i bevarandeplanen är en handbok för vem som är ansvarig för de olika områdena, när och hur aktiviteterna ska bedrivas och vad som ska göras och av vem när problem eller avvikelser inträffar.

När webbplatserna har samlats in och bevarats under en tid bör en uppföljning göras. Syftet med det är att analysera frågor som exempelvis: för vilka syften har det arkiverade materialet använts/behövts/gjort nytta? Skulle det kunnat göras om så det blir ännu mer användbart? Finns det fall när man trodde sig kunna få nytta av den arkiverade webbplatsen men där det av någon anledning inte lyckades? Hur skulle det kunna fungera bättre? Vilka brister har upptäckts?

[22] Vägledningen 24-timmarswebben (2006) kan laddas ner på http://www.ldb-centrum.se
[23] VERVA, Verket för förvaltningsutveckling, lades ner den 31 dec 2008

12. Slutsats – Strategi för webbarkivering

Många frågor behöver utredas innan en organisation startar arbetet med att bevara sin webbplats. Det är frågor som vilket syfte som finns med bevarandet, vilka krav som åligger organisationen, vad som ska bevaras, vem som ska använda materialet samt hur ofta och när insamling ska göras.

Dessutom måste beslut tas om hur arbetet ska utföras rent tekniskt, vilken metod som ska användas, vilka verktyg som passar bäst för den specifika organisationen, vilka filformat som är lämpligast osv.

Det viktigaste enligt oss är att börja arbeta i rätt ände, det vill säga med att utreda syfte, krav, urval, användare och andra strategiska beslut före val av verktyg.

Komprimerat till en mening:

”VAD före HUR, STRATEGI före TEKNIK”

Från arbetet med att bevara den egna webbplatsen kommer organisationen troligen också fram till att man genom att förändra sättet att publicera på webben i dag kan underlätta för framtida bevarande.

Det kan gälla exempelvis beslut om hur webbplatsen ska struktureras eller vilka filformat som är lämpliga respektive olämpliga att använda, till att granska och ställa krav på framtida publiceringsverktyg ur ett bevarandeperspektiv.

Det är svårt, eller kanske till och med omöjligt, att i dag ge råd för hur insamlat webbmaterial kan *säkerställas* för all framtid, så långt har vi inte kommit. Vår inställning är att bevarandet av alla typer av digitalt material aldrig blir färdigt utan att det är ett ständigt pågående arbete. Det vi däremot kan och måste göra redan i dag är att planera för hur hanteringen av digitalt material bäst kan göras samt att bygga upp ett skydd så att det inte skadas eller försvinner, vare sig avsiktligt eller av misstag.

Till sist en uppmaning till dig som ska starta ett projekt för att bevara din organisations webbplats: projektet behöver inte resultera i den ”perfekta lösningen”. Det är bättre att starta insamling nu fastän de långsiktiga planerna inte är lösta, för det som *inte* har samlats in kan aldrig bevaras!

DEL 2: WEBBCRAWLERS

VERKTYG FÖR ATT SAMLA IN WEBBSIDOR

13. Om studien

Inför starten av projekt Testplattformen gjorde LDB-centrum en litteraturstudie om olika metoder för att samla in webbplatser för arkivering.

Studien visade att det finns tre vanliga metoder:

- Uttag med webbcrawlingsverktyg
- Export ur publiceringsverktyg
- Kopiering av befintlig mappstruktur¨

Den första metoden samlar in material direkt från vilken publicerad webbplats som helst medan de andra två kräver aktivt medverkande av domänens ägare. Den första metoden kan också användas oavsett hur webbplatsen ser ut, om den är statiskt uppbyggd av enkla HTML-filer eller dynamiskt byggd på en databas eller dylikt. Av dessa anledningar valde vi inom projekt Testplattformen att välja metoden ”uttag med webbcrawlingsverktyg”. Läs mer om de olika metoderna i kapitel 9.1 i denna rapport.

Efter att ha valt metod övergick vi till att välja lämpligt verktyg att använda i Testplattformen, och det är denna undersökning som redovisas här. Vi beskriver hur verktyg för insamling av webbsidor fungerar. Dessutom jämförs tre olika verktyg genom litteraturstudie och vissa praktiska tester. De valda verktygen är HTTrack Website Copier, Heritrix och PageNest. Resultat och jämförelser ses i nästföljande kapitel.

På Riksarkivets webb finns följande information om programvaror som används för att bevara webbplatser:

> ”Det finns idag ett flertal programvaror som kan användas för att spara ned webbsidor till den egna datorn. Dessa skiljer sig bland annat åt i komplexitet och pris. Exempel på fria programvaror är HTTrack Website Copier och Heritrix. Det finns en mängd program att botanisera bland, dock är resultatet bland annat beroende av hur programmen används samt hur webbsidan som ska sparas är konstruerad, och enbart användandet av en specifik programvara garanterar inte ett godtagbart resultat.”
>
> http://www.statensarkiv.se/default.aspx?id=4018&refid=4020

14. Webbcrawlers

”Web crawler”, ”web spider”, ”web robot” eller ”web scutter” är ett datorprogram eller en sökmotor som samlar in innehållet från webbsidor automatiskt genom att följa sidornas länkstruktur. Webbcrawlers används framför allt av söktjänster som ”Google” eller ”Yahoo” och för webbarkivering.

Något riktigt bra svenskt ord för ”web crawler” och ”crawling” har vi inte hittat, därför kommer vi att använda orden ”webbcrawler” (-crawlers i pluralis) för verktyget samt verben ”webbcrawling” respektive ”att crawla” i texten.

Kända användare av webbcrawlers för insamling av webbplatser är bland annat Internet Archive[24] (IA) som samlar in kopior av webbplatser från hela jorden och Kungl. biblioteket[25] (KB) som sedan 1997 har samlat ihop svenska webbsidor 2 gånger per år. KB:s arkivering av webbplatser sker dock inte enligt Riksarkivets arkiveringsregler. Mer information om KB:s insamling, med namnet ”Kulturarw[3]”, finns på adressen: http://www.kb.se/soka/internet/sv-webbsidor/om/

Att dessa organisationer redan samlar in alla svenska webbplatser regelbundet innebär dock inte att frågan anses löst. Varje svensk myndighet är själv ansvarig för att bevara den egna webbplatsen. Till skillnad mot IA och KB vars mål är att samla in hela webben respektive hela den svenska delen av webben fokuserar denna rapport på en enskild organisation som har som inritning att bevara den egna domänen.

Webbcrawling innebär att man skapar en kopia av besökta webbsidor, sidor som lagras på den egna hårdvaran och indexeras så att det går snabbare att hitta bland dem. Bland de insamlade sidorna kan man sedan surfa runt genom att klicka på länkar, precis på samma sätt som om sidorna hade legat ute på Internet, med samma utseende och funktion som de hade när en besökare såg dem.

Endast ”aktiva” sidor samlas in, det vill säga sidor som har länkar till sig från andra sidor. Sidor som har gjorts inaktiva, till exempel gamla sidor som det inte länkas till längre, kan inte hittas när man klickar sig fram på en webbplats och samlas heller inte in av webbcrawlern.

Vanligen startar man crawlingen med att skriva in vilka URL-adresser som ska besökas, dessa kallas för seeds (fröer). När programmet besöker en webbsida identifierar den hyperlänkar till andra sidor och lägger in dem på listan över sidor som ska besökas. Så länge det finns nya sidor som inte har samlats in kvar på listan fortsätter crawlingen. Processen är den samma, oavsett om man begränsar crawlern till att bara gå igenom en liten domän eller hela världens webbplatser.

14.1 Strategier

Olika crawler-verktyg jobbar på olika sätt, som t.ex. enligt vissa algoritmer. Man kan som användare välja att begränsa insamlandet till ett antal filtyper, exempelvis att bara kopiera bara HTML-sidor och inte ta in sidor som slutar på .asp, .php, ? (query string) eller dylikt.

[24] Webbplats: http://www.archive.org/web/web.php
[25] Webbplats: http://www.kb.se

Man brukar också kunna välja djup på insamlingen, det vill säga hur många nivåer nedåt som ska samlas in.
Vid så kallad ”path-adcending crawling” – försöker programmet att spara ner så många sidor som möjligt från en specifik domän. Dessa program kallas ibland för ”harvester software” eftersom de ”skördar” allt innehåll från en speciell sida eller domän.

Vid insamlandet behövs också en strategi för att kolla vilka sidor som redan har samlats in. Genom val av lämplig strategi slipper man dubbellagring bland annat genom att man då inte samlar in två exemplar av samma sida, som till exempel adresser både med och utan ”/” efter.

14.2 Problem

Vissa webbplatser kan vara uppbyggda på sådant sätt att de försvårar crawling. Detta gäller till exempel platser där läsaren måste uppge lösenord för att kunna se sidorna.

Dynamiska sidor som förändras utifrån val som användaren gör är svårare att samla in än statiska sidor. Det är till exempel problematiskt att fånga upp information från sidor som hämtar upp information från exempelvis server-side scripts eller där navigeringen bygger på .asp-sidor.

Webbplatser där sidans arkitektur bygger på element (t.ex. textstycken) som ligger lagrade i databaser ställer också till problem eftersom webbcrawlers ofta kan inte komma åt posterna i dessa databaser.

”Robots Exclusion Protocol” är en konvention för att förhindra webbcrawlers att komma åt delar av eller hela webbsidor. Protokollet fungerar som en begäran att ignorera specifika sidor vid sökningen. I de flesta verktyg kan man välja om man vill följa denna begäran eller om man ska ignorera den och ändå samla in sidorna. Vissa webbplatser returnerar andra sidor för insamlande program än vad som visas vid normalt surfande och andra innehåller så kallade ”Crawler traps” där webbcrawlern ”fastnar” så att insamlandet stannar av eller i värsta fall kraschar. Detta kan förhindras genom att konfigurera crawlern till att begränsa antalet dynamiska sidor som den ska gå igenom.

Om det är möjligt att välja tid då crawling ska göras rekommenderas att välja en tidpunkt då så lite aktivitet som möjligt utförs på sidan. I annat fall kan det uppstå problem om förändringar görs samtidigt som webbplatsen förändras, så att länkar till nya sidor exempelvis inte kommer med.

14.3 Olika webbcrawlers

I detta projekt har vi valt att utvärdera tre olika webbcrawler-verktyg, nämligen de två open-source verktyg som Riksarkivet nämner: ”HTTrack Website Copier” samt ”Heritrix”. Förutom dessa har även en kommersiell produkt utvärderats, ”PageNest” (tidigare känd under namnet ”Webstripper”). Programvarorna har även laddats ner, installerats och provkörts mot en liten privat webbplats för att utvärdera hur enkla de är att använda.

14.3.1 HTTrack Website Copier är ett av de program som Riksarkivet nämner. Det är ett gratis open-source verktyg för att kopiera webbsidor från Internet till en lokal dator. HTTrack samlar in html-sidor, bilder och andra filer från servern till din dator. Med programmet kan man också se och navigera bland sidorna i offline-läge, på samma sätt som

när sidorna var publicerade på Internet. HTTrack arrangerar de nedladdade sidorna enligt domänens länkstruktur. Den enda mjukvara som krävs för att använda de lagrade sidorna är en webbläsare.

HTTrack är skriven i C av Xavier Roche. Det kan konfigureras enligt val och olika filter och har en integrerad hjälpfunktion. HTTrack finns i olika versioner, med kommandotolk respektive två olika versioner med GUI. Programmet klarar också <av att fortsätta ladda ner en webbplats där tidigare nedladdning har blivit avbruten.

HTTrack kan följa länkar som genereras av JavaScript eller länkar som finns inne i Applets eller Flash, men inte komplexa länkar som genereras av funktioner.

HTTracks finns att ladda mer på webbadressen http://www.httrack.com/ På sidan finns också ett forum som är väldigt aktivt och har ett flertal inlägg varje dag. Programmet är väl dokumenterat, det finns flera manualer för hur man ska och inte ska använda HTTrack och på webbsidan finns en lång FAQ-sida.

Den aktuella versionen av HTTrack heter 3.42-2 och släpptes i mars 2008.

HTTrack är mycket enkel att ladda ner, installera och köra. Inga som helst tekniska kunskaper krävs för att komma igång och ingen konfigurering är nödvändig. Det enda som krävs är att skriva in vilken domän som ska hämtas.

Programmet kan köras på Windows 95/98/NT/2K/XP, Linux/Unix/BSD, OSX m.fl.

14.3.2 Heritrix är det andra program som nämns i Riksarkivets information om att bevara webbsidor. Heritrix skapades speciellt för webbinsamling och arkivering och används av Internet Archive (www.archive.org). Heritrix är skriven i Java för en Linuxmiljö.

Heritrix samlar ihop ”allt”: ljudfiler, bilder, stylesheets etcetera och kan ladda ner upp flera filer samtidigt.

Heritrix har utvecklats av Internet Archive (IA) i samarbete med de nordiska nationalbiblioteken (bland annat KB). Den första versionen släpptes i januari 2004 och den senaste versionen, 1.14.0, kom ut i april 2008.

Ända sedan 1996 har Internet Archive lagrat filerna som samlats in i formatet ARC file format, men Heritrix kan även konfigureras för att lagra filerna i andra format. Det speciella med ARC är att ett många webbsidor lagras i en enda stor fil för att undvika att behöva hantera många små filer. En efterträdare är på gång till ARC, med namnet WARC (Web ARchive file format). Utvecklingen sker inom IIPC (International Internet Preservation Consortium) och IA. En ansökan om ISO-standard har gjorts men är ännu inte färdigbehandlad.

Precis som HTTrack kan Heritrix accessas både med kommandotolk och med webbläsare. Utöver Heritrix utvecklas olika program att använda för tillgängliggörandet av de lagrade webbsidorna. WERA är ett verktyg för att söka med en tidslinje och NutchWax hanterar fritextindexering, även den med en tidslinje. Ett ”Curator tool” har också tagits fram för att ge möjlighet till icke-tekniker att sköta och kvalitetskontrollera insamlingar. (Även fler verktyg finns.)

På webbadressen http://crawler.archive.org finns mer information samt programmet för nedladdning. Där finns också manualer, systemkrav, mailinglista, wiki och FAQ. Heritrix används av en relativt stor grupp användare.

Att ladda ner, installera och använda Heritrix är inte lika enkelt som HTTrack. Det krävs ett ganska stort tekniskt kunnande och JRE (Java Runtime Environment) installerat för att verktyget ska gå att använda. I manualen står också att färdigheter i Linux krävs. Heritrix lämpar sig därför bäst till medelstora och stora organisationer samt till mindre organisationer med höga krav på säkerhet och autenticitet för de lagrade webbfilerna.

14.3.3 PageNest (tidigare namn Webstripper) används bland annat av Högskolan i Borås. Detta är en kommersiell programvara med webbadress http://pagenest.com/index.html

En licens kostar 29.95 EUR (2008-06-11). PageNest har även en gratisversion av sitt program vid namn PageNest Free Edition som får användas för icke-kommersiellt bruk.

PageNest/Webstripper har funnits sedan 1999 och har använts av över en miljon användare. Det påstås vara en av de mest pålitliga webbcrawlers och samtidigt en av de enklaste att användas.
Programmet kan ladda ner upp till fyrtio filer samtidigt. Det konverterar alla länkar till relativa länkar och sidorna kan öppnas med PageNest självt eller med annan webbläsare. En nedladdning kan pausas och senare återupptas.

PageNest kan ladda ner vilka filtyper som helst: html, shtml, php, asp och andra. Programmet kräver Windows 98/ME/NT4/2000/XP/Vista samt Internet Explorer 4 eller senare.

Att ladda hem och installera gratisversionen av PageNest för privat bruk är väldigt enkelt. Inga specifika tekniska kunskaper krävs för att komma igång med att ladda ner en domän.

15. Slutsats - webbcrawlers

Webbcrawler-verktyg är oftast mycket enkla att både ladda ner och använda. Detta gör också att man enkelt själv kan installera och utvärdera ett antal verktyg för att avgöra vilket som passar bäst beroende på vilka behov man har och på vem som ska använda verktyget.

Av de tre verktyg som här har undersökts (HTTrack, Heritrix och PageNest) är det bara Heritrix som kräver ett stort tekniskt kunnande. Samtidigt är Heritrix det enda av programmen som har utvecklats speciellt med tanke på långtidsarkivering.

En annan viktig styrka hos Heritrix är att det används av de stora aktörerna i världen, som Internet Archive till exempel. Det innebär att programmet knappast kommer att försvinna under överskådlig tid utan att utvecklingen lär fortsätta.

Heritrix samlar in alla filer som finns på en webbplats och bevarar dem en i ARC- eller WARC-fil. ARC/WARC är ett omslutande format och inuti det finns filerna i de format som de hade från början tillsammans med extra metadata. WARC bygger på ARC men har dessutom utökad funktionalitet. Båda formaten har skapats av arbetsgruppen vid Internet Archive/IIPC[26] och WARC är just nu (vintern 2008) inlämnad som draft för att bli ISO-standard.

Konvertering av ingående filer kommer att behövas förr eller senare, men vi menar att det är en trygghet i att använda samma filformat som används av Internet Archive och andra stora institutioner på området.

[26] IIPC: http://www.netpreserve.org/about/index.php

DEL 3: OPEN REPOSITORIES

16. Om studien

I detta dokument har fem olika verktyg för lagring, publicering och tillgängliggörande av digitalt material undersökts och jämförts genom en litteraturstudie.

Arbetet gjordes för att välja det mest lämpliga verktyget att ha som lagringsplattform i projekt Testplattformen, med syfte att testa, utvärdera och utveckla metoder och verktyg för insamling och långtidsbevarande av webbplatser.

Det finns ett antal verktyg av detta slag som är skapade av kommersiella företag, men denna studie kommer enbart att undersöka verktyg som bygger på ”Open source”, det vill säga principen om att all källkod ska vara fri och att användaren alltid ska ha möjligheten att använda, läsa, modifiera och vidaredistribuera den. Distributionen ska också vara fri och källkoden ska följa med programmet. Principen om öppen källkod förvaltas av Open Source Initiative (OSI)[27].

[27] Webbplats: http://www.opensource.org/

17. Open repositories

Open repository-verktygen utvecklades i första hand med syfte att inom universitetsvärlden lagra och hantera sina digitala publikationer så att exempelvis forskare själva kunde publicera sina forskningsdokument och intresserade läsare enkelt skulle kunna få åtkomst till desamma. Numera används verktygen även i andra institutioner än i universiteten för att lagra, bevara och erbjuda access till digitalt material, ett användande som med all sannolikhet kommer att bli allt vanligare med tiden.

De olika verktygen har utvecklats med olika syften, ett exempel på detta är att vissa verktyg redan från början fokuserat mer på bevarandet av det digitala materialet medan andra mer strävat efter att skapa verktyg där enkelheten att publicera och komma åt filerna står i centrum.

Det är viktigt att organisationens eget syfte och mål med att bevara digitala dokument står som grund för valet av verktyg. Inom LDB-centrum har vi som syfte att testa långtidsbevarande av webbsidor för eftervärlden, med detta i fokus finns fyra krav på ett Open repository:

1. Data i arkivet måste kunna hanteras utan att skadas, försvinna eller av misstag raderas
2. Data måste kunna hittas och extraheras från arkivet och tillhandahållas en användare
3. Data måste kunna visas upp och förstås av en användare
4. Krav 1, 2 och 3 måste uppfyllas även på mycket lång sikt

17.1 Begrepp

Många begrepp används för dessa verktyg: ”Open Access Repositories”, ”Institutional Repositories”, ”Repository platforms” och ”Öppna arkiv” är några av dem. Samma sak gäller för mjukvaran som ligger till grund för att skapa ett ”repository”, här kan nämnas ”Digital asset management system” (DAMS), “Institutional repository software” eller på svenska “Publiceringsverktyg”.

Här kommer ”Open Repositories” att användas, med förkortningen OR. Repository betyder översatt till svenskan förråd eller magasin, en plats för lagring. Begreppet ”Öppna arkiv” menar många är fel, att repository inte är samma sak som ett arkiv. Trots det kommer vi här att använda begreppen arkiv och arkivering, just eftersom vi inte hittat något bra svenskt ord för ”repository”.

17.2 OR-mjukvara

En viktig komponent för att framgångsrikt kunna lagra stora mängder digitalt material är mjukvara som kan hantera bevarandeprocessen. Till skillnad från priset på hårdvara har priset på mjukvara inte sjunkit och detta har lett till att många projekt har dragits igång för att gemensamt skapa Open source-verktyg som kan ersätta de dyra kommersiella mjukvaror som finns.

Mjukvaran har som syfte att erbjuda ett antal funktioner, som:

- Hantera intag av data
- Hantera digital arkivlagring

- Generera och hantera metadata
- Hantera access (åtkomst) till lagrad data
- Kontrollera bevarandeåtgärder
- Migrera datafiler till andra format
- Exporterar filer till auktoriserade konsumenter

Enligt OpenDOAR[28] (The Directory of Open Access Repositories) så fanns det i början av år 2007 ungefär 800 öppna arkiv runt om i världen, varav ca 30 stycken i Sverige.

17.3 Fördelar med ”Open source”

Varför ska man då använda sig av Open source-programvaror? Följande punkter är de stora fördelarna med att välja öppna system hellre än kommersiella verktyg:

- Ger organisationen större kontroll över sina program
- Bygger på öppna standarder vilket även förenklar framtida arkivering
- Eliminerar beroendet till en leverantör
- Kan enklare modifieras för att passa den egna verksamheten
- För populära mjukvaror finns aktiva communityn som granskar källkoden, täpper till säkerhetsluckor och gör förbättringar
- Utvecklingen går snabbt vilket är en förutsättning på områden där förändring är en del av vardagen
- Bidrar till samarbete genom communityn
- Organisationen slipper utveckla egna mjukvaror från grunden utan kan bygga vidare på kända och beprövade verktyg
- Kostnadsbesparingar, dels sådana som hänger ihop med punkterna ovan, dels slipper man licenskostnader
- Förespråkas av stora aktörer som exempelvis VERVA

17.4 Programvaror

Det finns ett stort antal fria programvaror för e-publicering: aDore, Archimede, ARNO, CDS Invenio, DSpace, DPubS, ECL, Eprints, Fedora, Greenstone, Harvestroad, Hyperjournal, IntraLibrary, Intrallect, i-Tor, LOKSS, MyCoRe, Open Journal Systems, Opus, Topaz med flera (våren 2008).

Det finns också ett antal kommersiella produkter, som DigitalCommons, CONTENTdm och Digitool samt flera egenutvecklade system som exempelvis DiVA, Digitala vetenskapliga arkivet utvecklat på Uppsala universitetsbibliotek.

Parallellt med utvecklingen av olika typer av programvaror för att bygga upp digitala arkiv, så har Open Archives Initiative[29] tagit fram standarden OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting) för att möjliggöra interoperabilitet mellan olika arkiv.

[28] Webbplats: http://www.opendoar.org
[29] Webbplats: http://www.openarchives.org/

17.5 Litteraturstudie

Fem av Open source-verktygen har här utvärderats genom litteraturstudier. Dessa är:

- EPrints
- DSpace
- Fedora
- Greenstone
- DAITSS

Anledningen till att just dessa fem valdes var att **EPrints** och **DSpace** har funnits på marknaden länge och är de två mest använda inom området Open repositories.
Fedora är inte lika stort, men har utvecklats senare och bygger på erfarenhet från de tidiga verktygen. Fedora är också ett mycket populärt verktyg idag, det finns med i många jämförande undersökningar och det sker mycket utveckling på Fedoramoduler.

Greenstone bygger som Fedora på separata moduler och har som mål att vara enkelt att använda och ha lågra krav på hårdvara för att även fungera i länder med lägre materiell standard. Utvecklingsarbetet sker i samverkan med UNESCO.

DAITTS slutligen är utvecklat med ett helt annat syfte än de andra verktygen, nämligen för att vara ett slutet arkiv utan kopplingar utåt. Programvaran är trots det öppen för alla att ta del av och modifiera efter eget behov.

18. Jämförelse av OR-programvaror

Här följer resultaten av den litteraturstudie som har gjorts i syfte att utvärdera och jämföra olika programvaror av typen Open Repository.

18.1 EPrints

www.eprints.org

Det första publiceringsverktyget, lanserades år 2000.
Föregångare till Dspace och övriga verktyg.
Version 3.0 (totalt omgjord version) släpptes i januari 2007.
EPrints utvecklats och administreras av University of Southampton.

Användare: c:a 200 arkiv (mars 07), med tonvikt på Storbritannien.
Tillsammans med DSpace den mest spridda plattformen.
Sverige: SLU Uppsala (kallas där Epsilon). Handelshögskolan i Göteborg, Lunds universitet (LU Research), Högskolan i Kristianstad, SICS (Swedish Institute of Computer Science) Kista.
Internationellt: California Institute of Techonlogy, Institut Jean Nicord – Paris, University of Bath, Glasgow ePrints Service m.fl.

EPrints var från början avsett för forskare att kunna sprida elektroniska artiklar genom att själva deponera dem för lagring i ett sökbart arkiv.
Syfte: vem som helst skulle kunna sätta upp ett öppet arkiv.

OAI-kompatibelt.
Programvaran är Open Source, alltså fri att använda och utveckla som man vill. Det finns dock tilläggstjänster, som hjälp med installation etc. som man får betala för.

Materialet för arkivering kan utgöras av en mängd olika typer, allt från textfiler till multimedia och ljudfiler. Sökgränssnitt för att söka i arkivet och dess fulltexter.

Det är en fördel att EPrints är väletablerat, det innebär att programvaran uppdateras och utvecklas fortlöpande. Det stora antalet användare gör också att det finns ett forum för utbyte erfarenheter.

Bygger på OAIS.
Utvecklats GNU/Linux, även Solaris och MacOS.
Version 3 Apache på Windows.
PHP, MySQL.
Programmeringsspråk: Perl.
Kommandotolk för installation och administration.
Webbinterface för all administration.

Fördelar: Lättast och snabbast att sätta upp.
Kräver ett minimum av teknisk expertis.
Koden är väldokumenterad.
Flexibelt för egna anpassningar.
Klarar av en mängd olika filer, text, multimedia, ljudfiler osv.
Vilka metadata-scheman som helst kan användas.
Fulltext-sökning på arkivets innehåll.
Flerspråkigt, en svensk översättning är på gång.
Väletablerat. Stort antal användare.
Forum för utbyte och erfarenheter.
E-postlistor, Wiki och FAQ.
Programvaran uppdateras och utvecklas fortlöpande.

Nackdelar: Skalbarhet ett problem.
I huvudsak avsett för små arkiv.
Kräver att ePrints installeras på en egen server.
Konfigurering kräver stora tekniska kunskaper.
Långsiktigt bevarande var aldrig en kärnfråga vid skapandet av EPrints.
Saknas funktioner för statistik, stöd för import och export mellan olika databaser.
Objekt kan endast presenteras i det format som det inkom till lagringen.
Sökfunktionen skulle behöva förbättras, enligt SLU som använder EPrints (”Epsilon”).
Communityn har få medlemmar och arbetar slutet.

18.2. DSpace

www.dspace.org

Utvecklad av MIT Libraries (Massachusetts Institute of Technology) och Hewlett-Packard.
Första utgåva 2002.
Version 1.5.0 kom i mars 2008.

Användare: drygt 200 arkiv, flest i USA (mars 07)
Tillsammans med Eprints den mest spridda plattformen.
Sverige: Göteborgs universitet, Malmö, Borås och Hamlstads högskolor.
Internationellt: National Library of Finland, Cambridge University, University of Edinburgh, University of Montreal, Hong Kong University of Science & Technology Library m.fl.

Open source-programvara.
OAI-kompatibelt.

Designad för att låta slutanvändare deponera digitala objekt via webbgränssnitt.
Lagrar i nivåer: Community, collections, subcollections.

Kan lagra olika typer av material så som text, bilder, data, ljudfiler etc.
Register över format.
Fulltext-sökning + söka i metadata med pekare till objektet.
Användare – roller med olika behörighet.
Webbgränssnitt för lagring och sökning.

Identifierare = ”handles”, unika avgiftsbelagda.

Satsvis (batch) input.
Workflows: kan skräddarsydda för olika grupper av användare.

Optimalt i en Unix/Linux-miljö men ska fungera i alla miljöer.
Valfri webbserver, Java, PostgreSQL/Oracle.
BSD Open source-licens.

METS för SIP.
Qualified Dublin Core.
Metadata lagras i databaser.

Fördelar:
Relativt enkel att installera och att använda.
Enkelt för producenter att leverera material in i arkivet.
Access till visst material kan ges åt endast auktoriserade användare.
Olika supportnivåer för olika typer av objekt (Known, Supported & Unsupported).
Kan lagra olika typer av material så som text, bilder, data, ljudfiler etc.
Vid utvecklingen var långtidsbevarandet ett viktigt syfte.
Väletablerat med många användare, vanligast bland alla verktyg.
Väldigt stort internationellt.
Bra statistikfunktioner.
Stort forum för utbyte av erfarenheter.
Öppen utvecklingsgrupp.
Väldigt aktiva ”listservs”.
Intensiva utvecklingsinsatser förväntas under de närmaste tre åren.

Nackdelar:
Via webbgränssnittet kan endast enskilda objekt köras in.
Dåligt med verktyg för batch import.
Klarar inte versionshantering av objekt.
Dålig skalbarhet, problem med större objekt.
Monolitiskt byggd, stor kod.
”Koden skulle behöva göras om från grunden”, finns dock inga sådana planer.
Koden är dåligt dokumenterad samt komplex och i med detta svår att modifiera.
Kräver en erfaren systemadministratör för installation och konfiguration.
För att skräddarsy krävs en Javaprogrammerare.
Svårt att leta upp objekt.
Klarar inte av relationer mellan objekt.
Låst vid Dublin Core som metadatastandard.
Utvecklingsplan (roadmap) saknas, framtiden känns osäker.
Mer skapat som ett publiceringsverktyg än som ett ”repository”.

Jämföranden:
DSpace är anpassat till större multidisciplinära institutioner än EPrints.
Vid utvecklingen fokuserades något mer på långtidsbevarande jämfört med EPrints.

18.3. Fedora

http://www.fedora.info/

Fedora är en förkortning av "Flexible Extensible Digital Object and Repository Architecture", (ej att förväxla med Linux distributionen Fedora).

Utvecklat av The University of Virginia Library och Cornell University i USA.
Forskningsprojekt 1997, mjukvaran lanserades 2003.
Version 2.2 kom i februari 2007.

Användare: 127 kända (registrerade) användare (juni 2008). Tusentals nedladdningar av programvaran.
Sverige: KB sedan årsskiftet 2007/2008.
Internationellt: Det Kongelige Bibliotek i Danmark, Arrow-projektet i Australien, University of South Australia (BORSA, dark archive), Library of Congress, USA, Bibliotheque national de France, New York University.
Förväntas växa. Finns även 7 kända kommersiella företag (juni 2008) som utvecklar lagringslösningar som bygger på Fedora.

OAI-kompatibelt.
Erbjuder Open source programvara för att skapa digitalt arkiv bestående av en mängd olika typer av digitala objekt.

Systemet är gjort mer för de som arbetare professionellt med att hantera det, än för enskilda personer som vill lagra några få dokument på enkelt sätt.

Open source programvara.
Inte så mycket ett verktyg utan mer en metod.
Modulär.
En samling verktyg för att skapa webbaserade lagringsmöjligheter.
Flexibel objektorienterad datamodell (både positivt och negativt).

Identifierare: "Fedora PID" (persistant identifier) på enskilda objekt/dataströmmar.
Checksummor på dataströmmar.

Satsvis (batch) körning för både import och export.

Optimal prestanda i Unix/Linux-miljö men klarar även Windows.
Apache, Java, MySQL/Oracle 8i.

Lagrar objekt + metadata som XML-filer.
METS för att skapa SIP och producera DIP
Dublin Core

Fördelar:
Klarar av en mängd olika typer av digitala objekt.
Mycket god skalbarhet, kan hantera miljontals objekt.
Robust arkitektur.
Flexibel.
Access till visst material kan ges åt endast auktoriserade användare.
God integrering med andra system med hjälp av flera olika metoder.
Hanterar relationer mellan objekt, även aggregationer (relationer av typen ”är medlem av” och ”har medlemmar”).
Även relationerna är indexerade och kan sökas på.
Versionshantering (varje förändring av ett objekt skapar en ny version av objektet)
Flera olika metadatascheman kan användas.
Kan användas både för enkla som väldigt komplexa arkiv.
Stark utvecklingsgrupp.
Långtidsplan (roadmap) finns för Fedoras fortsatta utveckling.
Intensiva utvecklingsinsatser förväntas under de närmaste tre åren.
Koden är mycket väldokumenterad.
”Fedora Preservation Services Working Group” arbetar specifikt med bevarande.
Det finns mycket skrivet om Fedora: tester, jämföranden osv. Väl utvärderat.
Mest sofistikerad av samtliga OR-mjukvaror.

Nackdelar:
Långtidsbevarande var inte i fokus vid utvecklingen, men har blivit alltmer viktigt i nyare versioner av verktyget.
Stor kodbas.
Svårare att lära sig använda systemet än Dspace och EPrints.
Något svårare att specialanpassa än DSpace och EPrints, kräver större tekniskt kunnande.
Relativt låg aktivitet i användargrupperna, mycket av arbetet sker i den interna gruppen.
Utvecklingen går långsamt.
Många slags olika mjukvara behöver sättas ihop (även positivt!)

Projekt från Yale/Tuft: “*Does Fedora have the ability to serve as the basis of a trustworthy electronic records preservation system?”* (svaret är Ja)

Jämföranden:
Jämfört med Dspace erbjuder Fedora fler möjligheter för att hantera data, även data som lagras utanför den egna lagringsytan, som olika sätt att presentera objektet på.
Inte lika enkelt som i Dspace att som användare leverera objekt till lagret.

18.4. Greenstone

http://www.greenstone.org

Utvecklat av University at Waikato, Nya Zeeland, i samarbete med UNESCO och Human Info NGO i Belgien.
Tidigt verktyg, började utvecklas 1996.
Version 2.80 för att skapa ett OR, Version 3.0 (alfaversion, utkom i feb 07) för utvecklingsarbete.

Användare: Okänt antal, men ett ca 3000 nedladdningar av programvaran/månad (mars 07). 750 anmälda till e-postlistan.
Sverige: (okänt)
Internationellt: New Zealand Digital Library Project, iArchives, Oxford Digital Library, Peking University Digital Library.

Open source-programvara föra att skapa digitalt bibliotek för olika typer av material.
OAI-kompatibelt.

Designad för att vara enkel att installera och använda för icke-specialister.
Två gränssnitt, Reader interface (webb) respektive Librarian interface (Java grafiskt gränssnitt)

GNU-licens
Unix/Linux/Windows/Mac OS/X
Version 2 skriven i Perl, version 3 i Java
MySQL
Relationsdatabas för metadata (version 3)

METS
Dublin Core (både Qualified och Unqualified) samt flera andra.
Export från Greenstone till CD-ROM

Fördelar:
Enkel att installera och använda.
Klarar många typer av objekt.
Flerspråkigt, 35 språk i nuläget.
Fulltext-sökning.
Flexibel.
Samarbete vid UNESCO lång tid framåt.
Låga krav på hårdvaran, kan installeras på t.ex. en enskild laptop.
Installering på Windowsdator kräver ingen som helst konfiguration.
Enkelt att skräddarsy hur sökresultat ska visas upp.
Anpassat bruk av metadata, kan använda vilket schema som helst.
Stort internationellt, används i många olika länder.
Aktiva utvecklingsgrupper.
Mycket aktiv mailinglista (flera inlägg varje dag).
Roadmaps över pågående utvecklingsprojekt.
Wiki, FAQ, blogg, manualer, övningsexempel, tutorials mm på webbsidan.
Kan distribueras (och installeras) via CD.

Nackdelar:
Inte orienterad mot bevarande.
Inte utvecklat mot större organisationers behov.

18.5. DAITSS

http://daitss.fcla.edu/

Förkortning för "Dark Archive in the Sunshine State"

Utvecklat av FCLA, Florida Center for Library Automation.
Arbetet började år 2000. Första utgåva 2005.
Open source sedan 2006.
Version 1.2.6 (juli 07) Arbete pågår med en omgjord version 2.0.

Användare: Florida Digital Archive. Ett femtiotal publika- och universitetsbibliotek i Florida

Syfte: Designad för långtidsbevarande, genom att implementera funktionsmodellen definierad i OAIS-modellen.

Byggdes som ett internt arkiv men har nyligen släppts som Open source.
Bygger på OAIS, och är enkel att mappa mot funktionerna i OAIS Reference Model även om alla funktioner inte finns utvecklade. (information om vad som saknas jfr mot OAIS finns).

Skriven för Linux.
Java
My SQL
GPL licens.
Använder METS för att definiera SIP, AIP och DIP enligt OAIS.
PREMIS (inte fullständigt)

Egen identifierare, DFID på objektnivå och IEID på paketnivå.

Dark archive, bevarandet i fokus.
Inte tänkt för publik åtkomst, men kan kopplas ihop med andra system för access.
Två nivåer av bevarande "Bit level" där objektet lagras i flera kopior i den form den inkom till arkivet, respektive "Full" där även normalisering och formatmigrering inkluderas.

Fördelar:
Utvecklat med bevarandet i fokus.
Aktiva åtgärder för bevarande, som normalisering redan vid intag.
Klarar av ett stort antal olika slags objekt.
Kan hantera relationer mellan format.
Fleranvändarstöd.
Mailinglista,
Fixity checks görs på alla lagrade filer, två separata algoritmer används på varje fil och felaktiga filer ersätts automatiskt av en fungerande kopia.

Nackdelar:
Hela koden går ej att ladda ner.
Felaktigheter i kod och i tutorials.
Dåligt fungerande access.
Kräver en hel del programmering mot befintliga lagringssystem.
Ingen automatisk back-up i DAITSS, måste göras utanför systemet.
Mycket få som använder DAITSS.
Mycket låg aktivitet i mailinglistan (3 inlägg de senaste 9 månaderna).

DAITSS kan användas tillsammans med andra OR-verktyg. DAITSS används då för bevarandet medan det andra verktyget används till inleverans och uttag ur arkivet.

19. Sammanfattning

Att bygga system för digitalt bevarande är en komplex process bestående av kravanalys, noggrann planering och val som bygger på relevant information. Det är mycket mer komplicerat än att bara installera någon mjukvara och börja ladda upp innehåll. Men oavsett vilket system du väljer - redan att få in sitt digitala material i en organiserad lagringsmiljö är ett bra första steg på vägen mot långtidsbevarande.

Det är en utmaning att jämföra och utvärdera för- och nackdelar med olika verktyg och att välja det som bäst når upp till ens egna behov. I vissa fall kan det vara en bra att implementera mer än ett verktyg eller att integrera komplementerande system för att få den funktionalitet man eftersträvar. En lovande kombination är att kombinera Fedora och Dspace. För inleverans och lagring används Dspace där enkelheten är en stor fördel medan Fedora används till access.

I en undersökning från 2006 undersöktes sex stycken olika OR-verktyg för att välja bästa system inför skapandet av ett nationellt Open repository på Nya Zeeland. Följande kriterier betygsattes: *skalbarhet, enkelhet att anpassa koden, säkerhet, möjligt att integrera med andra system, stöd för fleranvändare och administration, flerspråksstöd, open source-licens, verktyg för konfigurering och workflows samt hur stark community som finns för varje system*.

Från sex stycken verktyg gallrades hälften bort i en första omgång (Arno, CDSware och i-Tor). Kvar till grundligare undersökning i en andra omgång blev Fedora, DSpace och EPrints.

Man kom fram till följande rekommendationer:

- Som kärna i en centraliserad arkitektur är Fedora mest lämpad. Fedora erbjuder en bra infrastruktur med god skalbarhet och möjlighet att integrera med andra system. Att anpassa Fedora efter interna behov är svårare än för de andra systemen men det förväntas inte krävas speciellt mycket anpassning.
- EPrints är bästa kandidat för mindre institutioner som vill skapa och driva egna repositories. EPrints har problem med skalbarhet men är enklast att komma igång med och att använda.

20. Slutsats – Open repositories

Efter att ha gjort denna litteraturstudie valde vi på LDB-centrum att välja Fedora som Open Repostitory-verktyg. De stora fördelarna med Fedora är:

- Robust arkitektur
- God skalbarhet
- Flexibelt
- Modulbaserat
- Klarar många typer av objekt
- Hanterar relationer
- Hanterar versioner av objekt
- Olika metadataschema kan användas
- Stark utvecklingsgrupp
- Långtidsplan för fortsatt utveckling finns
- Relativt stor aktivitet i användargrupperna
- Koden är mycket väldokumenterad
- Mest sofistikerad av alla verktyg

De negativa aspekterna av Fedora är framför allt att det krävs stort tekniskt kunnande för att installera, anpassa och utveckla Fedora som lagringsmiljö. För LDB-centrum anser vi det viktigt att omedelbart börja bygga upp denna kunskap för att vara kapabla att utvärdera och utveckla metoder för webbarkivering på ett trovärdigt sätt.

Efter detta dokument hade skrivits färdigt meddelas att diskussioner om samarbete/samverkan har inletts mellan Fedora och DSpace. Det är dock oklart vad detta kommer att leda till.

För mer läsning rekommenderas dokumentet ”Creating an Institutional Reposirtory: LEADIRS Workbook” skriven av Mary R. Barton & Margret M Waters för MIT Libraries. (LEADIRS står för LEarning About Digital Institutional Repositories.)

Källförteckning

Del 1. Webbarkivering
Brown, Adrian (2006). *Archiving websites.* Facet Publishing, London. ISBN: 978-1-85604-553-7.
Hanzo Archives - http://www.hanzoarchives.com
JISC, UKOLN and ULCC (2008). *PoWR – The Preservation of Web Resources Handbook.*
LDB-centrum. *CODA Slutrapport 2007–* http://www.ldb-centrum.se
Masanés, Julien (Ed.) (2006), *Web Archiving.* Springer Verlag Berlin Heidelberg. ISBN: 978-3-540-23338-1
National Library of Australia - http://www.nla.gov.au
The National Archives of the United Kingdom - http://www.nationalarchives.gov.uk
Riksarkivet - http://www.statensarkiv.se
Runardotter, Mari; Mirijamdotter, Anita and Mörtberg, Christina (submitted). *Long-term Digital Preservation – Whose Responsibility?* Scandinavian Journal of Information Systems.
Runardotter, Mari (submitted). Organizational Cooperation for the Cultural Heritage – a Viable System Approach. Systemic Research and Behavioral Science.
Ulfsparre (red), (1995). *Arkivvetenskap.* Studentlitteratur, Lund. ISBN: 91-44-60731-8.

Möten:
Partnermöte 1 i LDB-centrum, 2008-01-24
Partnermöte 2 i LDB-centrum, 2008-06-16

Del 2. Webbcrawlers
Masanés, Julien (Ed.) (2006), *Web Archiving,* Springer Verlag Berlin Heidelberg

HTTrack Website Copier: http://www.httrack.com/
Heritrix: http://crawler.archive.org
PageNest: http://pagenest.com/index.html

Del 3. Open repositories
”A Guide to Institutional Repository Software”, Open Society Institute (2004)
”A Survey and Evaluation of Open-Source Electronic Publishing Systems”, Cyzyk & Choudhury, (2008)
”A Technology Analysis of Repositories and Services: A Proposal Submitted to the Mellon Foundation”
”CONTENTdm vs DSpace vs Fedora”, Ralph LeVan, OCLC Research
”Creating an Institutional Repository: LEADIRS Workbook”, Barton & Waters (2005)
”DAITSS 1.x Overview”, The Florida Center for Library Automation, (2006)
”Digital Repositories Review”, Rachel Heery, UKOLN & Sheila Anderson, AHDS (2005)
DSpaceFeatures: https://wiki.library.jhu.edu/display/RepoAnalysis/DSpaceFeatures
”DSpace vs Fedora”, Ralph Levan, OCLC Research
”E-Archiving: An Overwiew of Some Repository Management Software Tools”, Prudlo (2005)
”Factsheet Greenstone” http://www.greenstone.ofg/factsheet

Fedora Service Framework
First Monday: http://www.firstmonday.org/issues/issue9_5/dilauro/
”Foundations of Excellence – DSpace vs Fedora: Or ehat I do on my summer vacation”
Greenstone:http://aabc.bc.ca/aabc/electronic_records_workshop/presentation_slides.ppt#474, 165,Greenstone con't
”Institutional Repositories in the context of Digital Preservation”, Wheatley (2004)
Konferens: ”Infrastructure for Future Research”, Kungl. Bliblioteket 2008-06-24
Lifefeed: http://www.lifefeed.se/lifefeed/index.jsp?blogId=23713 (2008-06-25)
”Memory of the world. Towards an Open Source Repository and Preservation System”, Kevin Bradley, Junran Lei & Chris Blackall (2007)
Mötesanteckningar Uppsala: http://www.tec.hkr.se/moodle/mod/resource/view.php?id=1964
Open Access Information: http://www.searchguide.se/oa/?cat=18
Open Access News: http://www.earlham.edu/~peters/fos/2006/09/evaluating-dspace-eprints-and-fedora.html
Solidaritetshuset: http://www.solidaritetshuset.nu/usrd/rvo195.pdf
”StoneD: A bridge between Greenstone and DSpace”, Witten mfl (2005)
”Techincal Evaluation of selected Open Source Repository Solutions”, Open Access Repositories in New Zealand (2006)
The ARROW project: http://openrepositories.org/2007/program/files/1/treloar.pdf
”The Florida Digital Archive and DAITSS: A Working Preservation Repository Based on Format Migration”, Priscilla Caplan (2007)

LDB-centrum
Centrum för långsiktigt digitalt bevarande

Skapa Företagsby
Teknikvägen 3-13
961 50 BODEN

Telefon: 0921 573 00
E-post: kontakt@ldb-centrum.se
Webb: www.ldb-centrum.se