Projektrapport: Förvärv av digitala personarkiv och digital forensis

2014-2015

Göteborgs Universitetsbibliotek

Umeå Universitetsbibliotek

Sammanfattning

I jämförelse med den klassiska hanteringen av arkivmaterial är handhavandet av digitalt material en relativt ny företeelse inom ABM-sektorn. Det har därför i stor utsträckning saknats riktlinjer för hur man ska gå till väga med digitalt fött material.

Projektet har velat medverka till att ta fram rutiner, metoder och avtal för hur bibliotek, arkiv och andra förvärvande institutioner bör hantera digitala delar av ett personarkiv. Projektet har undersökt de arkivtekniska, datatekniska , juridiska och etiska aspekterna av förvärv av digitalt material, särskilt material som inkommer till arkiven på fysiska bärare.

Projektet har tagit fram en guide för hanteringen av materialet, från förvärv till bevarande och testat programvaror och metoder för analys av diskinnehåll. Dessutom har det inom projektets ram utvecklats en programvara för visning av diskinnehåll och långtidslagring.

På senare år har flera internationella projekt fört utvecklingen framåt på detta område och projektet kan konstatera att tröskeln idag inte är oöverstiglig för att att på ett ändamålsenligt sätt behandla digitalt material.

Innehållsförteckning

Inledning

Bakgrund

Personarkiv som samlas in, ordnas, vårdas och tillgängliggörs på Sveriges forskningsbibliotek är en viktig del av det svenska kulturarvet. Ett personarkiv kan innehålla brev, manuskript, anteckningar, dagböcker, biografiskt material, klippsamlingar och fotografier men det kan också förekomma andra materialkategorier. Personarkivet är ofta komplext ur ett fysiskt bevarande-perspektiv. Den ordning som arkivbildaren skapat, medvetet eller omedvetet, är ofta svåra att hantera genom att olika materialkategorier är blandade. Det kan också finnas post-it lappar, plastfickor, gem, tejp och pärmar eller omslag av dålig kvalitet och som riskerar att påskynda en nedbrytning av innehållet.

Utöver detta ställs nu bibliotek och andra förvärvande institutioner inför en ännu större utmaning. Institutionerna får i högre utsträckning in digitala dokument på olika typer av datalagringsmedier och ibland hela datorer som en del av en donation av ett fysiskt arkiv från t.ex. en författare, artist, forskare etc. Den nya stora utmaningen i förvärvet av personarkiv kommer att bli omhändertagandet av det digitala materialet. Det finns inget utarbetat system för att hantera arkiv där bäraren kan vara både papper, disketter, CD-skivor, USB-minnen och hårddiskar. Nästan allt kan finnas digitalt i ett personarkiv; e-post, utkast, arbetsdokument, databaser, bilder, videor, chatt, bloggar etc. Smarta telefoner, surfplattor, elektroniska utrustningar som kan spara data kommer med tiden också att bli aktuella. För forskare kan det handla om e-post arkiv och databaser.

Risken är stor att vi av okunskap förstör och förvanskar den digitala informationen. Det är inte bara en fråga om att kopiera digitalt lagrad information från olika digitala medier utan också säkerställa att informationen inte förändras.

Förutom de digitala dokumenten finns det också spår på lagringsmedierna som kan ge information om hur dokumenten har använts och om dokument som tagits bort. I de fall som man vill bevara dessa spår krävs det att man går metodiskt och försiktigt tillväga när man överför datat från ett lagringsmedia till ett annat. Digital forensis är verktyg och metoder för hur man hanterar digitalt långtidsbevarande särskilt med hänsyn till att kunna skydda digitala spår så att dessa kan beforskas i framtiden. En annan viktig aspekt är att man vid hanteringen av de digitala dokumenten bibehåller och säkrar ett dokuments integritet genom bevarandeprocesser.

Digital Forensis kommer ursprungligen från polisens och underrättelsetjänstens arbete med att söka och säkra digital information. För arkiven innebär tekniken nya kraftfulla verktyg och metoder för hantering av de digitala delarna av arkiv, men samtidigt väcker det frågor av etisk, juridisk och hermeneutisk natur om var gränsen mellan det publika och privata går, samt vilket ansvar som donatorn, arkivarien och forskaren har gentemot arkivet. Dessa frågor måste klargöras och hanteras i samband med ett förvärv och institutionen som har kunskapen måste ta initiativet till att dessa behandlas på ett riktigt sätt.

Projektinformation

Projekttid

2014-01-01 – 2015-03-31

Finansiering

Projektet finansierades av Kungliga biblioteket.

Deltagare

Göteborgs universitetsbibliotek (huvudsökande):
Lennart Stark (projektledare), Erik Siira, Jimmy Carlsson, Anders Larsson

Umeå Universitetsbibliotek:
Mats Danielsson, Anders Lindberg, Alexander Näslund

Begrepp och Definitioner

BitCurator
BitCurator var ett projekt som finansierades av the Andrew W.Mellon Foundation och leddes av School of Information and Library Science at the University of North Carolina, Chapel Hill (SILS) och Maryland Institute for Technology in the Humanities (MITH). Projektet löpte 2011-2014 med syfte att ge arkivarier bättre möjligheter att hantera digitalt fött material. Projektet har nu omvandlats till ett konsortium för att förvalta ramverket med samma namn.

Digital forensis (eng. Digital Forensics)
Ett begrepp som har sitt ursprung inom kriminaltekniken och som innebär undersökning av alla typer av hårdvaror som kan innehålla digital data.

Dold information
Information som donatorn inte är medveten om. Ex. maskinellt skapad information, raderad men återskapad information i skivavbildningsprocessen.

Donator
Den som donerar något, givare. Någon som överlåter något till någon annans ägo utan att ta betalt eller kräva en ersättning.

Filsystem
Hur ett lagringsutrymmes, t. ex. en hårddisk, data är ordnad för att ge möjlighet att läsa från och skriva till lagringsutrymmet. Det finns en hel mängd både nya och gamla filsystemstyper, te x FAT, NTFS, HFS.

Gallring
En aktiv handling i syfte att avlägsna och förstöra handlingar/uppgifter efter en bedömning av informations- eller bevisvärdet.

Informationsbärare
Ett fysiskt objekt som agerar lagringsutrymme åt digital data. Ex. Diskett, CD, Hårddisk, USB-minne.

Hybridarkiv
Ett fysiskt arkiv som innehåller både traditionella arkivobjekt (papper, brev, fotografier) och bärare av digital information (hårddiskar, CD-skivor, USB-minnen).

Kontrollsumma (eng. checksum)
En, med hjälp av en algoritm, uträknad summa för att unikt beskriva t ex en datafil. Kontrollsumman fungerar som ett fingeravtryck och gör att man i efterhand kan avgöra om kopian stämmer överens med originalet.

Metadata
”Data om data”, t ex bibliografisk metadata (exempelvis titel, författare) om en bok eller teknisk metadata (storlek, filtyp) om en datafil.

Proveniens
Härkomst, ursprung. Uppgifter om ursprung, upphovsman, tidigare ägare och annan historia.

Skivavbildning (eng. disk image)
En exakt kopia av det digitala innehållet i en fysisk informationsbärare helt oberoende av filsystemstyp.

Skrivblockerare (eng. Write Blocker)
Hård- eller mjukvara som förhindrar oavsiktlig skrivning och därmed förändring av en skivavbildning under kopieringsprocessen.

Tidigare studier

Under den senaste 10-årsperioden har flera rapporter kommit som undersökt verktyg och tekniker för digital forensics i ABM-sektorn (arkiv, museer och bibliotek). En rad institutioner har dessutom inkorporerat digital forensics i sina arbetsflöden. De tidigaste stora aktörerna var Bodleian Library (Oxford) the British Library, Emory University, King´s College London, The National Library of Australia, the New York Public Library, Stanford University och Yale University.

The Paradigm project (2005-2007) var ett projekt som undersökte hela processen för digitala personarkiv speciellt avseende frågor om långtidsbevarande. Medverkande var bibliotek vid universiteten i Oxford och Manchester.
http://www.paradigm.ac.uk/

The AIMS project 2009-2011. (Inter-Institutional Model for Stewardship) syftade till att identifiera metoder och ramverk för att förvalta digitalt fött arkivmaterial. Medverkande var Stanford University, the University of Hull och Yale University.
http://www.digitalcurationservices.org/aims/project-team-and-overview/

Library of Congress erbjuder generella anvisningar hur man bevarar eget digitalt fött material under rubriken ”Personal Archiving. Preserving Your Digital Memories”
http://digitalpreservation.gov/personalarchiving/

Projektet Bitcurator (2011-2014) har satt samman en svit av programvaror för att underlätta skapande och analys av diskavbildningar.
http://www.bitcurator.net

Det finns ingen information om att svenska ABM-institutioner använt sig av metoder och tekniker för digital forensis, i alla fall inte som rutin i den dagliga verksamheten.

Metod

Hanteringen av material i ett personarkiv eller motsvarande kan förenklat delas upp i stegen

- Förvärv
- Förtecknande
- Bevarande

Utifrån dessa tre steg har ett antal problemställningar identifierats i relation till digitalt fött material:

- Hur hanteras ett mottagande av material i juridisk mening?
- Hur lagras och beskrivs bärare lämpligen i ett befintligt fysiskt arkiv?
- Hur extraheras data ur bäraren till en skivavbildning?
- Hur identifieras relevant information utifrån skivavbildningen?
- Hur arkiveras skivavbildningen på ett långsiktigt och beständigt vis?

Juridiska och etiska frågeställningar

För digitala leveranser av personarkiv uppstår flera juridiska och etiska frågeställningar. Till exempel kan kopior av innehållet i cd/dvd-skivor, hårddiskar eller ett helt digitalt personarkiv göras relativt enkelt. Det innebär att det kan vara svårt att garantera att det levererade materialet är unikt och det kan vara svårt att veta om donatorn är upphovsman. Dold information för donatorn kan komma fram i ett senare skede.

För donatorer

För donatorn kan det finnas information som kan kränka den personliga integriteten. Information som donatorn inte är medveten om som webbhistorik, medicinsk information, finansiell information, känslig information i e-post, annan konfidentiell information kan finnas och bli tillgängligt med de verktyg som finns idag för att spara och analysera bärare av digital information.
Där kan också finnas information som donatorn har raderat men som ändå fortfarande finns kvar vilket donatorn måste vara medveten om.

För tredje man

På bärare av digital information kan det finnas information om tredje man. Dessutom kan det finnas material som tillhör någon annan upphovsman.
Mottagande arkiv kan skada donator och tredje man genom intrång i den personliga integriteten, immaterialrätter och ärekränkning. Det hela kan kompliceras ytterligare om tredje man finns utomlands.

Etiska frågor

Arkivleveranser av digitala personarkiv medför dessutom etiska ställningstaganden. Möjligheten att återskapa raderad information eller automatiskt sparad information. Hur ställer sig donatorn till detta? Vad ville en donator som är avliden?
I bästa fall kan man diskutera frågor med donatorn i förväg och klargöra dessa men arkiven kommer även att behöva hantera frågor på egen hand.

Övriga frågor

För arkiven är det angeläget att inte skrämma bort eventuella donatorer genom att lägga för stort ansvar på donatorn. Det är inte alls säkert att donatorn kan redogöra för hela innehållet i sin leverans vilket innebär att arkiven måste ta ett större ansvar i dessa frågor.
Ny rättslig utveckling kan få stor inverkan så som rätten att bli glömd.
Arkiven måste hantera förhållandet till enskilda arkivariers och bibliotekariers tillgång till känslig information som kan finnas i digitala personarkiv.
Dessutom kan det finnas straffrättsligt sanktionerat innehåll som till exempel barnpornografi.

Att idag hantera och arkivera digitala personarkiv är svårt eftersom rättsläget är så oklart. Samtidigt prioriteras digitala satsningar av bibliotek och arkiv av EU inom ramen för ”i2010: digitala bibliotek”.

Arkivtekniska frågeställningar

Vid arkivleveranser av digitalt material behöver man hantera och dokumentera både hårdvaran och det digitala innehållet. Dokumentationen bör innehålla leveransens historik och vad som sker när det kommit in till arkivinstitutionen. Slutligen bör det finnas någon form av överenskommelse gällande ägarförhållanden, bevarande och tillgängliggörande.

Digital information

Den digitala informationen som levereras till ett arkiv ligger vanligtvis på någon typ av bärare, eller hårdvara om man så vill. Hårdvaran blir ganska snart omodern och till följd av detta blir informationen svår eller rentav omöjlig att nå även om det inte är något fel på hårdvaran i sig. Exempelvis behöver man en diskettstation för att läsa disketter och olika typer av disketter kräver förstås olika typer av diskettstationer. Lyckligtvis är det förstås informationen i sig som är intressant att bevara och strategin för detta är migrering, dvs informationen flyttas till nya databärare.

Till detta kommer förstås mjukvara eller datorprogram som behövs för att läsa informationen på filerna. Problemen man ställs inför är att det eftersökta programmet inte finns ute på marknaden och svårt eller omöjligt att få tag i. Nya versioner av programmen utvecklas vilket kan göra att äldre filer inte kan läsas av aktuella program.

Den fysiska miljön för att bevara hårdvaran på bästa sätt skiljer sig inte mycket från pappersarkiv (18 °C 35 % luftfuktighet) vilket även gäller för optiska databärare. För magnetiska databärare rekommenderas 15 °C 30 %. (Se RA-FS 2013:4) Rekommendationen när man tar emot digital information till ett arkiv bör vara att så snart som möjligt att säkra den, dels med tanke på ovan nämnda dels för att livslängden på digitala media är mycket kort jämfört med till exempel papper.

Leveranser av digital information

Vid leverans av digitala informationsbärare bör en allmän teknisk beskrivning av media göras. Vilken proveniens har mediet? Ta reda på så mycket som möjligt om dess historia, ägare, användning, nuvarande skick. Vilket fabrikat, ålder, serienummer, total diskstorlek, storlek på nuvarande informationsinnehåll samt eventuellt andra uppgifter.

Är det hårddiskar, cd, dvd, disketter av olika slag, minneskort, USB-minne osv. Är det PC eller Mac?
Gå igenom om det finns skriven information direkt på cd:n till exempel. Står det något på disketternas etiketter osv. Det behöver naturligtvis inte motsvara det faktiska innehållet men antyder att det kan ha funnits annat innehåll tidigare som helt eller delvis kan återskapas. Använd strategier för att gå igenom materialet utan att ändra innehållet med hjälp av skrivblockerare.

Notera övrig teknisk information. Vilka filformat finns och hur stor datamängd utgör materialet?
Finns det till exempel ett stort antal disketter kan det vara idé att ge dem löpnummer och upprätta ett enklare register. Om möjligt kan en kronologisk sortering vara till stor hjälp. Digital information kan förstås också levereras utan en fysisk informationsbärare till exempel över internet via FTP.

Dokumentation av innehållet

I bästa fall kan en genomgång av materialet göras i samarbete med donatorn eller med delgiven information av donatorn före en leverans. En direktkommunikation med donatorn kan innebära att donatorn kan leverera ett välstrukturerat material som blir betydligt lättare att hantera för mottagande arkiv.

Dokumentera så långt det är möjligt innehållets historik. Har det kopierats från någon annan informationsbärare tidigare? Hur har media lagrats under tid hos donatorn? Inkom det utan informationsbärare som bilagor till e-post eller via FTP? All information om innehållets historia och donatorns datorvanor kan hjälpa arkivinstitutionen bevara och sätta in informationen i rätt kontext.
Säkra informationen så snart som möjligt genom att göra skivavbildningar så man inte blir beroende av bärarnas livslängd.

I nästa steg behövs grundläggande information om innehållet. Hur är innehållet strukturerat? Gör en kortfattad beskrivning av innehållet. Notera arkivbildare om möjligt. Finns det någon typ av undertitlar? Vilken relation har det digitala materialet till ett eventuellt analogt material i samma samling? Finns till exempel det digitala innehållet utskrivet på papper?

Kontrakt/överenskommelse

Skriv ett kontrakt eller en överenskommelse mellan donatorn och arkivinstitutionen som definierar vilket material som ingår. Hur gör man med material som finns på informationsbärarna men inte ingår. Vem är ägare till filerna? Har donatorn en kopia av materialet kvar hos sig?

Urval av mediaformat

Då det finns hundratals potentiella typer av bärare som skulle kunna ingå i ett hybridarkiv, så behövde ett urval göras för att nå ett relevant resultat inom ramen för projektet. De typer som sedemera låg till grund för tester var:

- Disketter (3½”, 5”)
- IDE hårddisk
- USB-minne
- CD-skiva

Dessa är de de format en arkivarie troligast kommer att stöta på.

Val av programvaror

Inom projektet ingick inte något utrymme för egenutveckling av programvaror för att extrahera data från bärare, så för detta syfte behövdes programvaror från tredje part.

Bakgrund

Digtal forensis har sitt ursprung inom kriminaltekniken vilket det innebär att de programvaror som existerar i stor utsträckning syftar till att tillgodose kriminalteknikens behov. Till många delar sammanfaller de med arkivens, men det finns vissa skillnader.

- Inom kriminaltekniken utförs analyserna av specialutbildad personal med hög datateknologisk kompetens , medan man kan förvänta att analysen inom arkivvärlden måste kunna utföras av arkivarierna själva, vilket ställer högre krav på att programvarorna är begripliga och lättanvända.
- Arkivarier har ett behov av att kunna arkivera skivavbildningen, extraherade filer och metadata för ”all framtid”
- Arkivarier måste kunna förteckna och visa upp innehållet för andra. Det innebär att man dels måste kunna visa önskvärd data, men också att man ska kunna hindra att känslig, personlig eller hemlig data exponeras.
- Arkivarier behöver kunna hantera äldre filtyper och filsystem som idag inte används aktivt längre.

För att tillfredsställande kunna använda programvarorna i ett arkivsammanhang bör följande kriterier vara uppfyllda

- Möjlighet att analysera många typer av filsystem, både nya och äldre.
- Möjlighet att göra en säker utläsning av data, dvs att innehållet inte påverkas av läsningen.
- Möjlighet att beräkna kontrollsummor både för skivavbildningen och enskilda filer för att bibehålla datintegriteten.
- Möjlighet att exportera information om diskinnehållet på ett standardiserat sätt.
- Möjlighet att snabbt kunna bilda sig en uppfattning om vad disken innehåller.
- Möjlighet för personer med relativt liten teknisk kunskap att hantera programvarorna.

För att utföra alla eller delar av de processer som krävs, finns det idag en hel rad programvaror att tillgå, både fria och kommersiella.

Kommersiella programvaror

Följande exempel på kommersiella programvarusviter innehåller fullständiga verktygslådor för att utföra kriminaltekniska undersökningar.

EnCase Forensics	från Guidance Software
P2 Commander	från Paraben
FTK	från AccessData
Forensic	från X-Ways
Evidence Center	från Belkasoft
ProDiscover	från TechPathways

De kommersiella programvarorna har nackdelar ur ett arkivperspektiv eftersom de tenderar att ha en sämre bakåtkompatibilitet när det gäller äldre filsystemstyper, då det är mest troligt att polisutredningar rör relativt moderna datorer.
En annan nackdel är att de ofta har proprietära metadataformat vilket också innebär att de inte har någon gemensam metadatastandard för att beskriva diskinnehållet. Det är alltså mycket svårt att utbyta information eller jämföra resultat mellan systemen.
De är ofta stängda monolitiska system som är svåra att integrera i ett större flöde.
Priserna varierar mellan 1000-4000$ plus årliga supportkostnader.

Fria programvaror

De programvaror som är upptagna här är alla ramverk som i sig integrerar ett antal olika verktyg för att utföra olika specifika uppgifter. I flera fall ingår samma fria verktyg i olika ramverk.
Gemensamt för alla dessa ramverk, med undantag för BitCurator, är att de riktar sig mot kriminaltekniker.

The Sleuth Kit - samling kommandoradsverktyg som gör det möjligt att analysera filer och skivavbildningar.

Autopsy - grafiskt gränssnitt byggt ovanpå The Sleuth Kit för enklare hantering.

Digital Forensics Framework (DFF) - kommandoradsverktyg och grafiskt gränssnitt.

Open Computer Forensics Architecture framework - utvecklat av den holländska polisen, främst för att automatisera processen vid stora volymer av digitalt bevismaterial.

*BitCurato*r - ramverk bestående av fria programvaror, bla inkluderar den Sleuth Kit, och som specifikt vänder sig till arkivarier, Vmware.

I valet mellan fria eller kommersiella programvaror beslöts att de kommersiella inte skulle användas. De är mindre flexibla när det kommer till frågan om integration i ett större flöde och de proprietära formaten kan ge framtida problem med datautbyte och migration av metadata.

För att gå vidare med projektet valdes BitCurator då det var det enda ramverket som hade verktyg för alla steg i processen, från skapande av skivavbildning till export av metadata och rapporter och som dessutom kunde leverera metadatat i ett öppet format.

BitCurator

Ramverket riktar sig specifikt mot arkivarier och försöker adressera de specfika problem och behov som finns i arkivvärlden.

En av de stora fördelarna med BitCurator är att det stödjer DFXML (Digital Forensics XML) som är ett xml-språk för att beskriva innehållet på en skivavbildning. DFXML ger möjlighet att bl a beskriva filers namn, typ, storlek, position och tidsstämpel. Språket kan också hålla annan teknisk information, t ex vilka processer som använts i analysen. Vidare kan den hålla kontrollsummor för filerna. Användning av DFXML gör det alltså möjligt att exportera metadata från skivavbildningen och förenkla den vidare analysen av innehållet.

Ramverket körs i Ubuntu och man kan välja att installera det som operativsystem eller att köra det som en virtuell maskin i exempelvis Windows. Det vanligaste är att ramverket körs som virtuell maskin vilket gör att installationen blir relativt lätt och det är då också enkelt att installera nya uppdateringar.

BitCurator stödjer följande filsystemsformat fullt ut:
FAT16, FAT32, NTFS (Windows)
HFS, HFS+ (Mac)
ext2, 3, 4 (Linux)
Dessa är de mest troliga filformat som en arkivarie kommer att stöta på. Ytterligare ett antal mindre vanliga format stöds delvis.

De viktigaste ingående programvarorna:

GuyMager - används för att skapa en lågnivå bit-för-bit-kopia av den fysiska skivan.

BitCurator Reporting Tool - grafiskt gränssnintt för att köra flera av verktygen i sekvens och producera rapporter av skilda slag.

FiWalk - används för att traversera filsystemet och skapa en xml-representation av innehållet, bl a position, filnamn, filstorlek och för att räkna ut checksummor på filerna.

Bulk-extractor - gör en lågnivåsökning i systemet för att identifiera fördefinierade mönster, ex personnummer, epostadresser, webbadresser etc.

BitCurator Disk Image Access - grafiskt gränssnitt för att genomsöka skivavbildningar, exportera filer och raderade filer.

Safe Mount mjukvara för att hindra oavsiktlig skrivning vid läsning av skivavbildningar.

Analys av skivavbildningar utifrån arkivperspektiv

Skivavbildningar innebär ofta att en oerhört stor mängd information blir tillgänglig. För att hantera den stora mängden behöver man hjälp sortera ut det relevanta innehållet. Vi ska titta på hur man gör det med hjälp av BitCurator.

Bakgrund

Pappersarkiven, våra "klassiska" arkiv, innehåller ofta en blandning av material: handskrivna manuskript, korrespondens, fotografier (glasplåtar, påsiktskopior, negativ o positiv) men kan även innehålla artefakter som konst, priser, kläder etc. Det är arkiven som vi är vana att se dem och som vi ordnar och förtecknar och gör tillgängliga för forskning.

Ett digitalt arkiv är ett arkiv som enbart består av digitala dokument. Det kan vara worddokument, power point presentationer, excelblad. Det kan lika gärna vara samma typ av dokument skrivet i Wordperfect, Indesign, Page maker, Pages, Notepad. Det kan vara fotografier i jpeg, tiff, png, eller ett råformat. Det kan vara ljudfiler i wav, mp3, aiff osv. osv. Det som är gemensamt för alla dessa är att de kräver både en anpassad hårdvara (en dator av de format som vi använder idag) och mjukvara för att kunna visas och det är naturligtvis en oerhörd utmaning för framtiden.

Slutligen har vi hybridarkiven, dvs de som både består av traditionella fysiska arkiv men med allt större inslag av digitalt material på olika medier. Det är den typ av arkiv som vi står inför att hantera. Utmaningen med dessa är att det krävs både teknisk kompetens och arkivkompetens för att kunna ordnas upp, bevara och tillgängliggöra dessa. Dessutom så måste vi ta fram någon form av gemensam arkivstruktur som kan användas för att hitta i båda arkiven.

Återskapa och läsa digital information

När man gör en skivavbildning blir den exakt kopia av disketten med alla filer som finns på skivan men också det utrymme som är tomt på skivan. Tomt utrymme på en skiva kan innehålla filer som tidigare har tagits bort från skivan. Det är nämligen så att när man raderar en fil så finns den fortfarande kvar på skivan tills dess att den skrivits över med nya filer. Därför kan en skiva innehålla mycket dolt utrymme som man tror att man har tagit bort men som egentligen finns kvar. Digital forensics handlar också om att analysera gammalt borttaget material.

Analys av skivavbildning

Vi har använt en hårddisk ur ett personarkiv i Umeå universitetsbibliotek. Det är en IDE hårddisk tillverkad 1998 och är 10 GB stor men innehållet fyller 2,3 GB.

Processen

- Vi använder en plattform med fria verktyg – BitCurator
- Efter ca två timmar har en identisk diskkopia skapats utan att originalet påverkats då skrivblockerare används.
- Efter ytterligare några timmars automatisk process har allt diskinnehåll analyserats (även raderade filer) och redovisats i en 42 Mb stor informationsfil samt dussintalet specifika rapporter.
- Efter ytterligare **några dagars** automatisk process har alla filer extraherats

Det är fullt möjligt att se detta som olika etapper. Till exempel att som ett första steg se till att det skapas identiska diskkopior av alla hårddiskar, disketter etc som kan lagras på ett säkert sätt. I ett senare steg kan sedan diskkopiorna analyseras, filer extraheras vid behov mm. Som framgår ovan är analys- och steget där filer extraheras väldigt tidsödande.

I detta fallet består disken av 4 partitioner:
Partition 1 innehåller 27521 filer, 14712 borttagna filer, 6516 tomma filer
Partition 2 innehåller 9171 filer, 1763 borttagna filer, 61 tomma filer
Partition 3 innehåller 472 filer, 451 borttagna filer, 67 tomma filer
Partition 4 innehåller 384 filer, 387 borttagna filer, 7 tomma filer

Skivavbildningen är lika stor som innehållet (2,3 GB) med en info-fil. Till denna finns en xml-fil (fi.xml) 42 Mb stor.

Till dessa produceras en rad rapporter:
T ex statistik över filsystem och filer, samtliga filformat, de tjugo vanligaste filformaten, borttagna filer

Analysen visar på ett komplext material med en stor mängd system- och programfiler. Många dokument i word, excel, html, pdf-format och bilder som är direkt läsbara. Många dupletter eller olika version av samma dokument. Många raderade dokument som helt eller delvis kan återställas.

Som en del i analysen identifieras alla förekomster av data – oavsett var på disken dessa data lagras och oavsett om denna information finns i ett raderat dokument, en besökt hemsida, en använd epost-adress etc. Det görs med hjälp av ett slags mönsterigenkänning, dvs allt som ***ser ut*** som ett telefonnummer, en e-postadress, ett kreditkortsnummer, en url-adress identifieras. Det krävs sedan ytterligare analys för att avgöra om det också är det. En hårddisk rymmer **alltid** oväntad information - som kan upplevas som känslig. Förutom ovanstående kan det till exempel gälla rubriker på epost, webbhistorik eller sökningar i databaser.

Filinnehåll

En listning av filerna finns i filen fiwalk-output.xml.xlsx

En snabb uppskattning säger att det är ca 20 % av innehållet som kan vara av värde undersöka vidare. Resten är systemfiler och programfiler och annat som inte är relevant.
Här verkar de intressanta filerna utgöras av word-dokument (.doc), excel-dokument (.xls), bildfiler (.gif och .jpg) och pdf-dokument. Dessutom kan en del textfiler (.txt) och htm-filer behövas ses över. Utöver dessa finns det filer från programmet ”graph in the box” med ändelsen .giw av intresse.
Även bland dessa filer är väldigt mycket ointressant. Se exemplet med .doc filer nedan:

I detta fall kan man identifiera 1494 .doc filer. Antalet filer av intresse är dock betydligt mindre. Samma fil kan ligga på flera olika ställen.
En fil med ändelsen .doc kan förekomma i flera filformat, exempel filen Kyrkos_pres.doc finns som sju träffar i samma bibliotek. Filformaten är (med filstorleken inom parentes):
1. ASCII text (92672)
2. Composite Document File V2 Document, corrupt: Can't read SAT (126464)
3. Composite Document File V2 Document, Little Endian, Os: Windows, Version 4.0, Code page: 1252, Title: KYRKOSTUDIER - Kris och förnyelse , Author: Anders Lindberg, Template: Normal, Last Saved By: Egil, Revision Number: 7, Name of Creating Application: Microsoft Word 8.0, Total Editing Time: 01:46:00, Last Printed: Mon Jul 21 12:55:00 2003, Create Time/Date: Thu Jul 10 14:37:00 2003, Last Saved Time/Date: Mon Jul 21 12:58:00 2003, Number of Pages: 1, Number of Words: 4375, Number of Characters: 24938, Security: 0 (92672)
4. Data (92672)
5. Data (92672)
6. GIF image data, version 89a, 1 x 1 (92672)
7. GIF image data, version 89a, 14 x 22 (91648)

I rapporten Deleted files finns filen med sex gånger. (Filen ~$rkos_pres.doc finns med elva gånger)

Kommentar

Processen resulterar i stora mängder information, inte minst av information som en gång raderats och som man helt eller delvis kan läsa igen. Den allra största delen av informationen har normalt ingen relevans ur ett arkivperspektiv. I vårt exempel ovan försvann till exempel 80 % omgående för att sedan reduceras ytterligare pga ett flertal kopior av samma filer.

Digitalt repositorium – 'dArc'

Efter att materialet har tagits emot och skivavbildningar skapats, behöver dessa lagras på ett säkert och långsiktigt sätt. Det finns i dagsläget ingen känd programvara som inriktar sig på lagring och tillgängliggörande av skivavbildningar samt deras innehåll, därför ingick inom ramen för projektet utvecklandet av ett rudimentärt arkiveringssystem, kallat **dArc**.

Bakgrund

Då skivavbildningar har skapats behöver dessa lagras på ett beständigt sätt. Förutom en långsiktig lagring av den faktiska skivavbildningen så fanns det även en önskan att kunna lagra och visa information om innehållet på disken (DFXML). Det fanns under projektets gång inget känt alternativ som har stöd för att tillgängliggöra denna typ av information, vilket var den anledningen till att ett eget system utvecklades.

Behovsanalys

Med begränsningar i både tid och resurser har det varit viktigt med rätt fokus i utvecklingen. För att åstadkomma detta hölls två separata workshops, där den ena handlade om de arkivtekniska delarna, och den andra om tekniska lösningar. Bland deltagarna fanns både arkivarier och utvecklare för att ta tillvara på relevant kompetens.

Avgränsningar

Synpunkterna från de två workshoptillfällena samt de tillgängliga resurserna satte fokus på följande funktionalitet:

- Möjlighet att skapa arkiv och tillhörande arkivbildare
- Möjlighet att lagra en skivavbildning med tillhörande filinformation (DFXML)
- Möjlighet att visualisera och söka bland DFXML-informationen

Teknisk plattform

Som lagringsplattform valdes fedora-commons (http://www.fedora-commons.org) som är en populär lösning för lagring av fildata.

Till affärslagret av applikationen valdes Ruby on Rails (http://rubyonrails.org), och för presentationen EmberJS (http://www.emberjs.org).

Resultat

Utvecklingen resulterade i tre separerade repositorier som tillsammans skapar en användbar produkt. Observera att systemet är **rudimentärt** och inte ska ses som en 'out-of-the-box' produktionsfärdig lösning.

dArc_store (www.github.com/ub-digit/dArc_store) - Innehåller en konfigurerad Fedora installation med tillhörande script för datahantering.

dArc (www.github.com/ub-digit/dArc) - Innehåller affärslogiken och levererar ett API utåt som pratar med **dArc_store** som datakälla.

dArc_ember (www.github.com/ub-digit/dArc_ember) - En Frontend-applikation som nyttjar APIer från **dArc** för att hantera data som lagras i **dArc_store**.

Samtliga repositorier är helt fria att använda eller vidareutveckla.

Resultat och slutsatser

Mallavtal och checklistor

Liksom vid leveranser av pappersarkiv bör det upprättas en överenskommelse/kontrakt vid digitala arkivleveranser som definierar vilket material som ingår och annan information om förvärvet. Det gäller till exempel vilka rättigheter som följer med förvärvet, behåller donatorn en kopia av hela materialet, hur ska kringinformation hanteras, medföljer hårdvara osv.

Se Bilaga 1: Donationshandling

Är det många bärare av digital information i leveransen kan en enkel registrering vara praktisk med information om arkivbildare, år, media, PC/mac mm.

Se Bilaga 2: Registreringsblad elektroniska lagringsmedia

Se Bilaga 3: Checklista för digitala arkivleveranser

Guide för ”Best Practice”

En bra guide för hantering av digitala arkiv är “Born digital: Guidance for Donors, Dealers, and Archival Repositories” där en rad experter medverkar. Inom projektet gjordes en svensk sammanfattning av denna.

Se Bilaga 4: Sammanfattning av ”Born Digital: Guidance for Donors, Dealers, and Archival Repositories.

Kunskapsspridning

För att få största nytta från projektets håll har ett antal insatser gjorts för att sprida projektets innehåll och resultat.

BitCurator Workshop

Under två dagar i slutet av juni 2014 genomfördes en workshop i ämnet digital forensis på Kungliga biblioteket och Riksarkivet. Ett fyrtiotal arkivarier från hela Sverige deltog .
Första dagen inleddes med en presentation av föreliggande projekt och därefter höll Porter Olsen från BitCurator en längre föreläsning om digital forensis för arkiv med tonvikt på utläsning och analys av data från bärare..
Dag två var avsatt för praktiska övningar där deltagarna under Porter Olsens ledning fick använda de verktyg som utvecklats i BitCurator-projektet för att göra skivavbildningar och extrahera metadata på eget material .

Presentationer

Projektet har presenterats vid:

- Biblioteksdagarna - Umeå, 8 maj 2014
- BitCurator Workshop - Kungliga biblioteket och Riksarkivet 23-24 juni 2014 https://github.com/ub-digit/digital-forensis/raw/master/bilagor/an_introduction_to_digital_forensics_for_archivists.pdf
- Konferensen ”Digitalisera, men sen då?” - Nordiska Museet, 28 november 2014 https://github.com/ub-digit/digital-forensis/raw/master/bilagor/digitala_personarkiv_och_digital_forensics.pdf

Vitbok

En Vitbok om ämnet Digital Forensis har författats inom projektet, och finns publicerad via gitbook på adressen http://digital-forensis.ub.gu.se .

Vitboken är öppen för samtliga att bidra med uppdateringar till via github-repositoriet http://www.github.com/ub-digit/digital-forensis .

Slutsatser

Även om digitala medier (”bärare”) under lång tid har inkommit till arkiv både i Sverige och internationellt , så är digital forensics för ABM-sektorn en relativt ny företeelse. I bästa fall har man tidigare lyckats skriva ut de mest angelägna dokumenten och sparat dem i pappersform, men i många fall har man bara arkiverat den fysiska bäraren utan att veta vad den innehåller eller hur man bär sig åt för att extrahera datat. Säkert har man ibland även avstått från den här typen av material då man känt sig osäker på hur och om det går att hantera.

Så länge accession av hårdvara var en sällsynt företeelse har man kanske inte sett detta som ett stort problem, men i takt med att antalet bärare i arkiven ökar blir det uppenbart att det måste finnas rutiner för hur man går till väga.

Projektet har velat medverka till att skapa sådana rutiner alltifrån förvärv till lagring utan behov av specialkunskaper inom juridik eller IT.
Den tvådagars workshop som projektet genomförde i Stockholm där många arkivarier deltog var ett tydligt tecken på att skapande av skivavbildningar och analys av data inte är ett oöverstigligt hinder. De programvaror som finns framtagna är enkla att använda och ger möjlighet att få en god uppfattning om innehållet. Att lagra den utvunna datan kan göras både mer eller mindre sofistikerat. I projektet valdes att bygga ett repositorium från grunden där man också kunde bläddra och söka i det extraherade metadatat, men det går att lösa lagringen med enklare metoder, exempelvis bara genom lagring på en filserver. Det viktigaste är att framtidssäkra innehållet i bäraren genom att skapa en identisk kopia av det digitala innehållet, en så kallad skivavbildning.

Självklart finns det fortfarande många problem som behöver lösas.
Det krävs resurser både i tid och utrustning för att analysera och lagra en skivavbildning som kan vara 2TB eller större för moderna hårddiskar. Om man bestämmer sig för att spara hela

skivavbildningen till skillnad från att bara lagra de filer från den som bedömts som relevanta, kan lagringskostnaderna bli höga om man har många diskar att hantera.
Gamla eller ovanliga hårdvaror och filsystemstyper är svåranalyserade och ställer större krav på programvaror och utrustning.
Det krävs också system för att lagra enskilda extraherade filer från avbildningen. Dessa filer behöver i många fall normaliseras, dvs att man har parallella versioner i öppna och mer framtidssäkra format.
System behövs också för att presentera dessa filer för slutanvändarna.

Kopplingen mellan det befintliga fysiska arkivet och de digitala delarna i det, är fortfarande en underutvecklad del i processen. Hur ser en gemensam arkivstruktur ut för fysiskt och digitalt material? Hur ska man beskriva innehållet i en skivavbildning med arkivens nuvarande metadatastandarder?

Inte heller de juridiska frågorna är helt oproblematiska eftersom det i många fall saknas rättslig praxis. Vad gäller för visning/lagring av programvaror, epost och bilagor författade av annan part? Hur hanterar man personlig och känslig information? Vem och i vilken utsträckning får ha tillgång till materialet?

De etiska frågorna har också otydliga svar, är det exempelvis rätt att återskapa och läsa filer som donatorn själv medvetet har raderat? Man kan även här få problem med känslig information eller information som rör tredje part.

De etiska och juridiska kraven på materialet gör det också nödvändigt att kunna sätta restriktioner på de enskilda filerna för att försäkra sig om att bara rätt målgrupp får tillgång till det aktuella materialet.

I dagsläget finns ändå goda möjligheter att säkra det digitala material som gömmer sig i arkiven. Programvarorna för att göra det finns och är fria, med relativt låga inlärningströsklar. Det kan ändå vara ett problem att vara insatt i vad som händer på området, ett litet arkiv kanske inte behöver göra skivavbildningar särskilt ofta och det kan då vara svårt att hålla sig uppdaterad på de tekniska delarna av processen. Det borde därför vara en klar fördel att samverka i regionala eller nationella nätverk där olika aktörer kan ha skilda specialiteter.

Utvecklingen i omvärlden går fort på detta område och man kan förvänta sig att det inom de närmsta åren finns solida lösningar på åtminstone de tekniska delarna, dvs från avbildning till analys, lagring och visning.

Referenser

Redwine, Gabriela. Barnard, Megan, Donovan, Kate. Farr, Erika, Forstrom, Michael. Hansen, Will. Leighton John, Jeremy. Kuhl, Nancy. Shaw, Seth. Thomas, Susan; Born Digital: Guidance for Donors, Dealers, and Archival Repositories; CLIR Publication No. 159; Washington; 2013
http://www.clir.org/pubs/reports/pub159/pub159t.pdf

Kirschenbaum, Matthew G. Ovenden, Richard. Redwine, Gabriela; Digital Forensics and Born-Digital Content in Cultural Heritage Collections, CLIR Publication No. 149; Washington; 2010
http://www.clir.org/pubs/reports/reports/pub149/pub149.pdf

Leighton John, Jeremy; Digital Forensics and Preservation; DPC Technology Watch Report 12-03 November 2012
http://dx.doi.org/10.7207/twr12-03

Lee, Christopher A. Woods, Kam. Kirschenbaum, Matthew. Chassanoff, Alexandra; From Bitstreams to Heritage: Putting Digital Forensics into Practice in Collecting Institutions; BitCurator Project; 2013
http://www.bitcurator.net/wp-content/uploads/2013/11/From-Bitstream-to-Heritage-S.pdf

Carroll, Laura. Farr, Erika. Hornsby, Peter. Ranker, Ben; A Comprehensive Approach to Born-Digital Archives; Archivaria volume 72; 2011; 61-92
https://open.library.emory.edu/publications/emory:cksgv/

Shein, Cyndi; From Accession to Access: A Born-Digital Materials Case Study; Journal of Western Archives: Vol. 5: Iss. 1; 2014; 1-42
http://digitalcommons.usu.edu/westernarchives/vol5/iss1/1

Jansson, Ina-Maria; Ett förlorat kulturarv? Digitala personarkiv – problem, lösningar och framtid; Institutionen för ABM, Uppsala universitet; Uppsala; 2012
http://www.diva-portal.org/smash/get/diva2:533206/FULLTEXT01.pdf

Lee, Christopher A; Archival Application of Digital Forensics Methods for Authenticity, Description and Access Provision; International Council on Archives Congress; Brisbane, Australia; 20-24 August 2012
http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00290.pdf

Woods, Kam. Lee, Christopher A. Misra, Sunitha; Automated Analysis and Visualization of Disk Images and File Systems for Preservation; Archiving2013, Washington; April 2-5 2013; 239-244

http://ils.unc.edu/callee/p239-woods.pdf

Kam Woods' web site
http://digpres.com/index.php?title=Main_Page

Garfinkel, Simson; Digital forensics XML and the DFXML toolset
http://www.sciencedirect.com/science/article/pii/S1742287611000910

Bilagor

Bilaga 1: Donationshandling

https://github.com/ub-digit/digital-forensis/raw/master/bilagor/donationshandling.pdf

Bilaga 2: Registreringsblad elektroniska lagringsmedia

https://github.com/ub-digit/digital-forensis/raw/master/bilagor/elektroniska_lagringsmedia_registerblad.pdf

Bilaga 3: Checklista för digitala arkivleveranser

https://github.com/ub-digit/digital-forensis/raw/master/bilagor/checklista_digitala_arkivleveranser.pdf

Bilaga 4: Sammanfattning av ”Born Digital: Guidance for Donors, Dealers, and Archival Repositories.

https://github.com/ub-digit/digital-forensis/raw/master/bilagor/born_digital_sammanfattning.pdf