Digitala personarkiv och digital forensics

Digitalisera – och sen då? **Nordiska museet 28.11.2014**

Lennart Stark, Göteborgs UB Mats Danielsson, Umeå UB

Allt mer digitalt material i arkiven

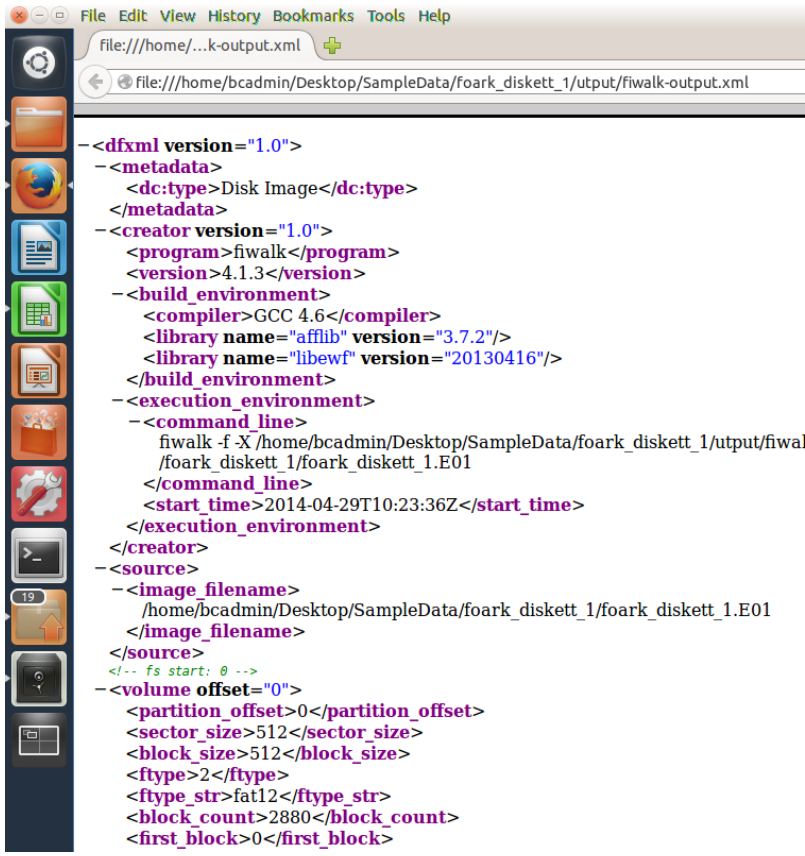
- Pappersarkiv
- Digitala arkiv
- **Hybridarkiv**

  Utmaningar med digitalt mtrl:

  - **Analysera, beskriva och förteckna materialet**
  - **Bevara materialet utan ändringar (utan konverteringar etc.)**
  - **Göra åtkomligt för forskning**
  - **Reglera användingen (vilka data får användas, jmf. “restinformation”)**

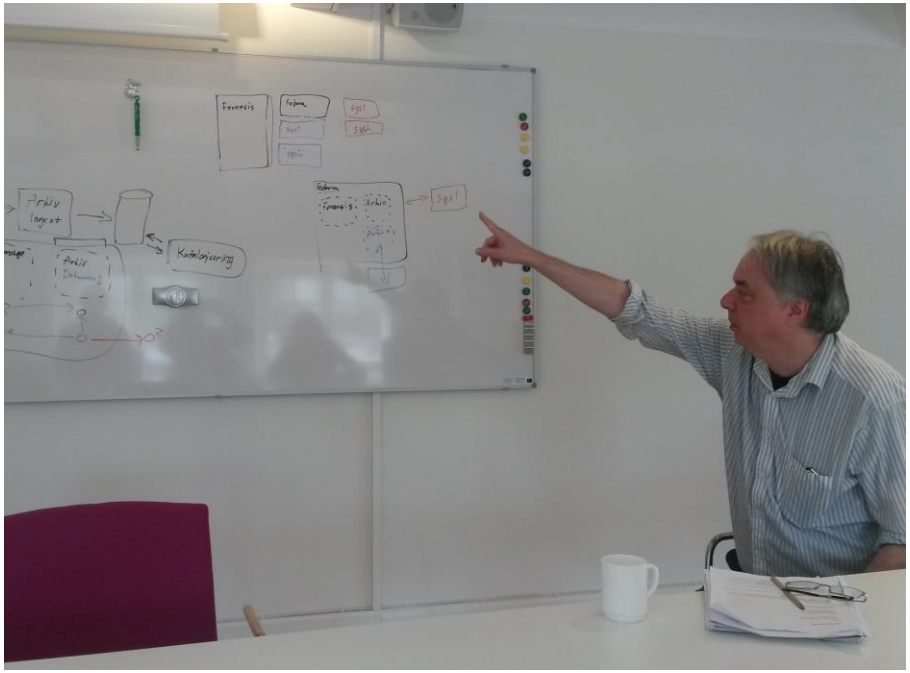
Ett vanligt personarkiv idag: manuskript, korrespondens, bilder mm. i både digital och analog form

Digitalt kräver nya metoder



Projektet digitala personarkiv och digital forensics

- KB finansierat projekt 2014
- Skapa rutiner
- Höja kompetensen inom digitala arkiv
- Bättre förstå utmaningar och omfattning av digital arkivering.



Första skiss till arkivsystem för skivavbildningar

Det digitala - en del av arkivet

- Hantera det digitala som en naturlig del av arkivet
  - Reglera förfogandet i donationshandlingen
  - Dokumentera proveniensen (använt av vem, till vad, när …)
  - Bestäm ambitionsnivå (vad är arkivets respektive forskarens uppgift)
  - Ta ställning till bevarande och gallring (kan delar gallras?, krävs ytterligare analys av innehållet?)
  - Bevara långsiktigt (fysiskt medium och/eller skivavbildning)

men också ett arkiv i sig

- Tänk på att
  - Ett medium (hårddisk t.ex) kan innehålla allt som ett pappersarkiv innehåller – utom papper
  - Ofta innehåller ”osynliga” data som kan upplevas som känsliga

Analys av en hårddisk från 2004

- En 10 Gb hårddisk, fylld till knappt 25%
  - Efter ca två timmar har en identisk diskkopia skapats utan att originalet påverkats (writeblockers 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

Dokument, bilder, raderade filer …

- Analysen visar på ett komplext material
  - En stor mängd system- och programfiler
  - Många dokument i word, excel, html, pdf-format och bilder som är direkt läsbara (ca 20% av totalen)
  - Många dupletter eller olika version av samma dokument
  - Många raderade dokument som helt eller delvis kan återställas

och oväntade data

*En hårddisk rymmer **alltid** oväntad information - som kan upplevas som känslig, t.ex*

- *Epostadresser*
- *Kreditkortnummer*
- *Besökta hemsidor*
- *Telefonnummer*
- *Rubriker på epost*
- *Sökningar i databaser*

79737	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.hotel-browser.org	2795497254
79738	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	24hourbooking.net	2795497724
79739	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.Frequent-Traveller.org	2795498171
79740	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	sv.wikipedia.org	2795499062
79741	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795499576
79742	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	en.wikipedia.org	2795499871
79743	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795500380
79744	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.jerusalem.se	2795500675
79745	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795501077
79746	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	susning.nu	2795501344
79747	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795501810
79748	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.jpost.com	2795502083
79749	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795502559
79750	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	jerusalem.usconsulate.gov	2795502820
79751	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.timeanddate.com	2795503385
79752	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795503907
79753	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.jewishvirtuallibrary.org	2795504240
79754	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795504790
79755	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.youtube.com	2795505297
79756	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	img.youtube.com	2795505436
79757	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.youtube.com	2795505560
79758	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.aish.com	2795506919
79759	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	66.102.9.104	2795507394
79760	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	images.google.se	2795507690
79761	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	images.google.se	2795507838
79762	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	tbn0.google.com	2795508128
79763	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.atlastours.net	2795508278
79764	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	www.atlastours.net	2795508336
79765	*Temporary Internet Files/Content.IE5/U8RWPL3C/search[2]	images.google.se	2795508413

Kvarvarande filer i webbläsaren antyder t.ex att denne person är mkt intresserad av Israel och kanske rent av planerade en resa dit

som kräver att användningen regleras

Processen


Donator
Arkiv
Skivavbildning
Metadata
Repositorium
Forskare

Bärare

SATA HDD
SATA SSD
IDE
SCSI (50-, 68-, 80-pin)
SAS (SCSI eller SATA generellt, serverdiskar)
FC Fiber Channel (Serverdiskar)
RLL (Äldre hårddiskar) – Mer komplext att analysera
MFM (Äldre hårddiskar) – Mer komplext att analysera
USB-sticka
USB-hdd (SATA/IDE)
FW-disk (400/800)
ESATAThunderbolt
CD-R
CD-RW
DVD-R + / - single/dual layer
DVD-RW + / - single/dual layer
DVD-RAM (Gammalt och ovanligt)
BD-R single/dual/quad
BD-RW single/dual/quad

Floppy 3,5”
Floppy 5,5”
ZIP-disk
JAZ
Superdrive (Apple) 120MB
Seagate 44/88 MB
VHS-band
DDS-band (Backupband)
Quik-band ca 150MB
Övriga backupband (DLT,LTO
Kassettband
MMC/SD/SDHC/SDXC standard/mini/micro
MemoryStick/M2 (Sony)
CompactFlash I/II
XD-minne
SmartMedia
SDHX

Filsystem

Stöd i BitCurator

ext 2,3,4 - Linux
FAT 12,16,32 - DOS och Windows
NTFS - Windows NT, Windows 2000, Windows XP, Windows Server 2003, Windows Server 2008, Windows Vista och Windows 7
HFS - **Mac OS**
HFS+ - **Mac OS** X

Ej stöd i BitCurator

FFS – Unix och AmigaOS
HAMMER – DragonFlyBSD
HPFS – OS/2
JFS – AIX, Linux,OS/2
ReFS Windows 8
ReiserFS – Linux
SFS – AmigaOS
USFS FS – UNIX
XFS – SGI IRIX och GNU/Linux
ZFS – Solaris

BitCurator

- Projekt 2011-2014
  - SLIS University Of North Carolina och Maryland ITH
  - Programsvit - open source, konsortium 2014-
- Skivavbildning, lågnivå bit för bit
- Analys av filsystemet, dfxml
- Identifikation av information i filer
- Filextraktion
- Teknisk metadata
- Linux, körs oftast som virtuell maskin

Exempel dfxml

•<fileobject>
<filename>RAW/report02-3.pdf</filename>
<id>19</id>
<filesize>1421998</filesize>
<partition>1</partition>
<alloc>1</alloc>
<used>1</used>
<inode>39</inode>
<type>1</type>
<mode>511</mode>
<nlink>2</nlink>
<uid>0</uid>
<gid>0</gid>
<mtime>1230764913</mtime>
<ctime>1230764913</ctime>
<atime>1230764978</atime>
<crtime>1230764978</crtime>
<seq>1</seq>
<byte_runs>
<run file_offset='0' fs_offset='241542144' img_offset='241542144' len='1421998'/>
</byte_runs>
<hashdigest type='MD5'>dede94f84fb2d00dc93ed00fda272a18</hashdigest>
<hashdigest type='SHA1'>3c078d039398c44611b6365e8afdeadeb61967d4</hashdigest
</fileobject>

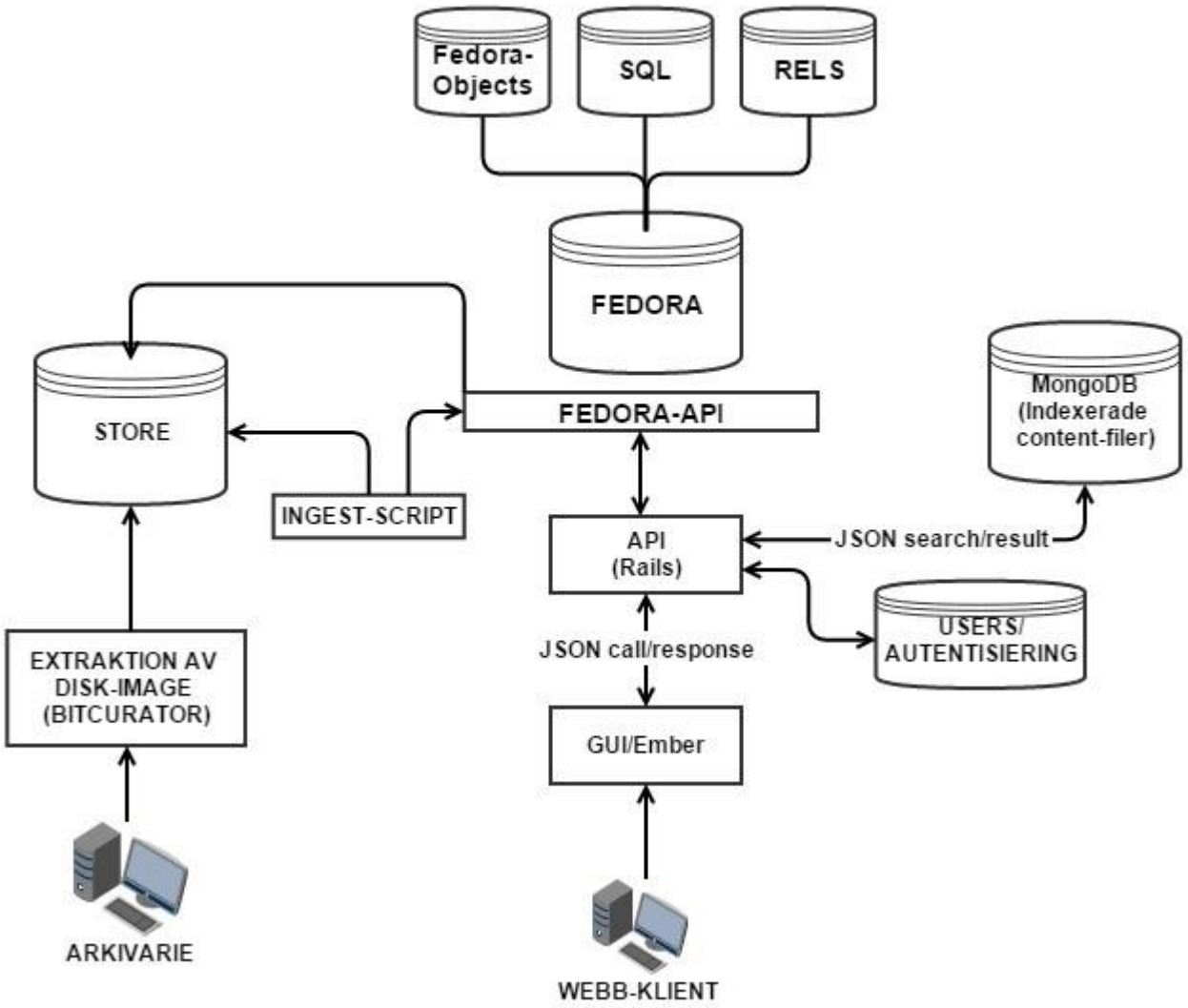
Repositorium

- ”dArc”, system för arkivering
- Ingest av skivabildning, metadata
- Byggs på Fedora Commons
- Bläddring i filsystem (contentfiler)
- Sök på filnamn, filtyp
- Indexerad dfxml för sökning
- Koden fri, Github

Repositorium

- Vidareutveckling:
  - Filtrering av contentfiler
  - Ingest av extraherade filer
  - Normalisering av filformat
  - Rättighetssystem
  - Extraktion av filer direkt mot skivavbildning
  - Fulltextsökning i extraherade filer

dArc


Fedora-Objects
SQL
RELS
FEDORA
STORE
FEDORA-API
MongoDB (Indexerade content-filer)
INGEST-SCRIPT
API (Rails)
JSON search/result
USERS/ AUTENTISIERING
JSON call/response
EXTRAKTION AV DISK-IMAGE (BITCURATOR)
GUI/Ember
ARKIVARIE
WEBB-KLIENT

Användbara länkar

- Göteborgs universitetsbibliotek, digitalisering http://www.ub.gu.se/samlingar/digital/
- Göteborgs universitetsbibliotek, handskrifter http://www.ub.gu.se/samlingar/handskrift/

- Umeå universitetsbibliotek, digitalisering http://www.foark.umu.se/digitalisering
- Umeå universitetsbibliotek, handskrifter http://www.foark.umu.se/samlingar/arkiv

- dArc https://github.com/ub-digit

- Bitcurator.net http://www.bitcurator.net/

- British Library, digital scholarship blog http://britishlibrary.typepad.co.uk/digital-scholarship/

- Bloggen Digitala personarkiv http://digitalapersonarkiv.wordpress.com/