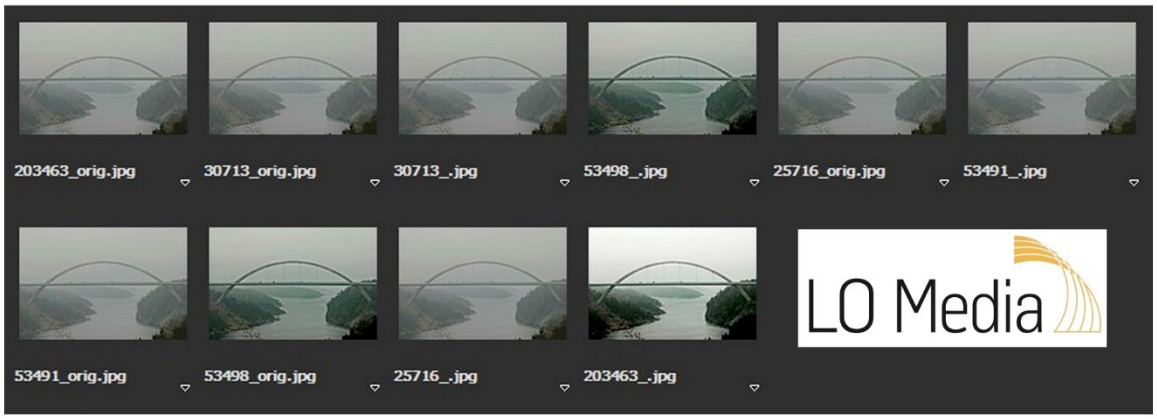
Prosjektrapport - Metodeutvikling for bearbeiding, tilgjengeliggjøring og langtidslagring av større digitalt skapte fotoarkiv

31.12.2021



Forfattere: Sølvi Bennett Moen og Odd-Rune Hansen

Med bistand fra: Miriam Idehen-Ohrvik, Dag Juvkam og Martin Ellingsrud

arb ark ARBEIDERBEVEGELSENS ARKIV OG BIBLIOTEK

Forsideillustrasjon:

Skjermdump fra FotoStation med eksempler på duplikater. Foto: Kristian Brustad / LO Media

Innhold

1. Innledning

Arbeiderbevegelsens arkiv og bibliotek (Arbark) har mottatt et digitalt skapt privat fotoarkiv fra LO Media bestående av nesten 600.000 bilder, på ca. 3,6 TB. Med bakgrunn i denne overleveringen har vi startet dette ettårige prosjektet, som har som mål å utforske og beskrive en metode for håndtering av mottak, analyse, bearbeiding, tilgjengeliggjøring og langtidslagring av større digitale fotoarkiver.

Materialet består av en nokså heterogen datamengde, skapt over en tidsperiode fra slutten av 90-tallet og frem til i dag. Bildematerialet har vært behandlet i flere DAM[1]- og fotoredigeringssystemer, deriblant Neo og Fotoware, og det er varierende kvalitet både på bildematerialet og bildemetadataene[2]. Forespørsel fra arkivskaper om avlevering av fotoarkivet ble gjort nettopp i forbindelse med overgang til nytt DAM-system hos dem.

Prosjektet har søkt å utforske maskinelle rutiner (algoritmer) for å kartlegge og analysere bildemateriale og metadata, for å fjerne duplikat informasjon, og for å automatisk oppdatere og komplettere metadatainformasjon der dette har vært mulig. Et sekundært mål har vært å utvikle støttesystemer for fotoarkivarer for masseoppdatering av metadata der det var mulig. Vi ønsket også å beskrive manuelle rutiner rundt mottak, håndtering og deponering av fotoarkiver, slik at vi kunne presentere en behandlingsløype som sikrer at fotoarkiver blir arkivmessig forsvarlig håndtert.

Målet var at disse rutinene kunne bli en verktøykasse som kan benyttes på mottatte fotoarkiv for å kvalitetssikre prosessene rundt håndtering av arkivene, og maksimere kvaliteten på og mengden informasjon man henter ut fra de tilgjengelige dataene.

Målgruppen for metoden er arkivinstitusjoner som ikke har hatt muligheten til å bygge opp egen kompetanse for å håndtere mottak, behandling, langtidslagring og tilgjengeliggjøring av digitalt skapte fotoarkiver. Vi har derfor valgt å inkludere informasjon om generelle problemstillinger ved digitalt skapt materiale i prosjektrapporten, spesielt i forhold til langtidslagring.

Vi har valgt å se på fotoarkivet fra et arkivståsted, der bevaring av originale metadata har hatt høy prioritet. Dette fordi det er enkelt å lage tilpassede uttrekk/visninger av metadata ved behov, slik at bevaringsprinsippet her blir det sterkeste. Dette har også vært en overordnet føring i de tekniske løsningene vi har utviklet; vi har søkt å implementere endringer i metadata som tilføyelser og rapporter/visninger i systemene heller enn ved modifisering av originalfilene.

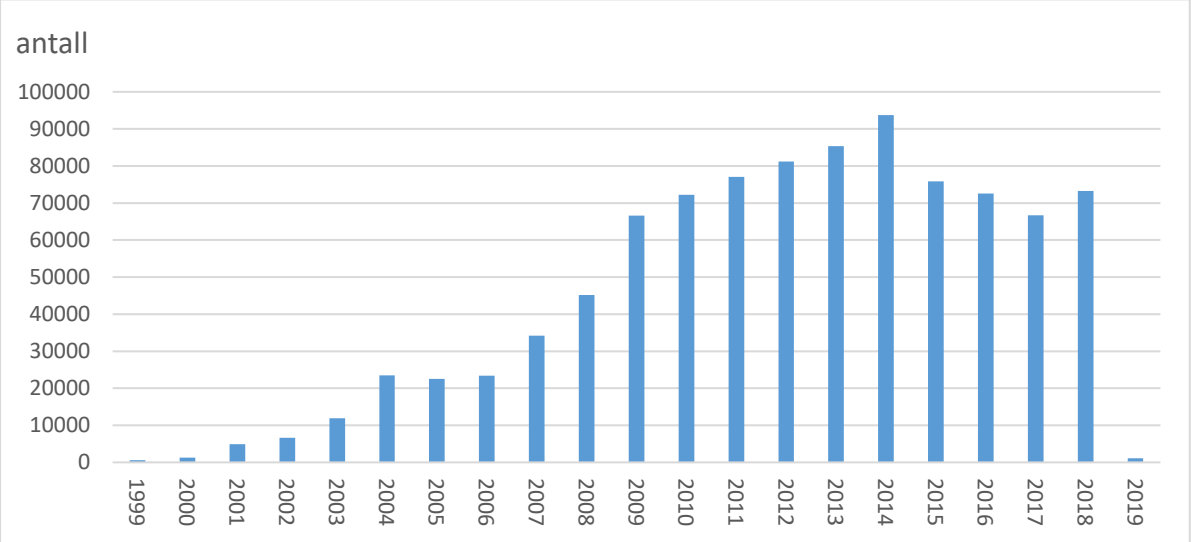
Vi har også valgt å forholde oss til OAIS- og DIAS-standardene og -terminologien ved behandling av digitalt skapt fotoarkiv.

[1] DAM – Digital Asset Management, datasystem for oppbevaring og håndtering av digitale ressurser
[2] Et bilde består av selve bildeinformasjonen, punktene som utgjør selve bildet, og metadata, som er tekstbasert informasjon om bildet, for eksempel kameratype, tidspunkt for fotografering, fotograf, lisens, stikkord, osv.

2. Utgangspunkt for prosjektet

LO Media er arbeiderbevegelsens mediehus og produserer innhold til 16 fagblader for ulike fagforbund/fagforeninger i tillegg til å drive nettstedet frifagbevelse.no. Fotoarkivet dokumenterer norsk arbeidsliv og arbeiderbevegelse fra de siste 20 årene. Det representerer dermed en unik kilde til informasjon og fremtidig forskning. Det er også det mest omfattende digitalt skapte privatarkivet Arbark så langt har mottatt.

Bildematerialet bærer preg av å være skapt over et tidsrom der det har skjedd en betydelig utvikling i teknologi for fotografering og bearbeiding av digitale bilder. Det gjenspeiler seg både i den tekniske kvaliteten på bildematerialet, men også i mengden og kvaliteten på metadata som ligger på det enkelte fotografi.



Antall mottatte bilder pr. år opprettet

2.1. Avvik fra prosjektbeskrivelsen eller i gjennomføringen av prosjektet

Vi har dessverre noe mangelfull informasjon om utgangspunktet for overføringen av fotoarkivet, da personellsammensetningen i organisasjonene har endret seg siden operasjonen ble avtalt og iverksatt.

Avtalen som opprinnelig var inngått med LO Media for overføringen omhandlet ikke videre deponering eller tilgjengeliggjøring. Vi ønsket å ta opp dette med arkivskaper underveis i prosjektet, for å få presentert lagringsløsningene til Digitalarkivet for dem, og for å avklare og skrive en ny avtale som omhandlet disse punktene. Dette fikk vi imidlertid ikke gjort.

Prosjektet ble dessverre kraftig hemmet av pandemien som berørte verden også i 2021, og som viste seg å bli mye mer langvarig enn noen hadde forutsett; dels fordi en del operasjoner ble vanskeligere å gjennomføre i en hjemmekontorsituasjon, og dels fordi møtevirksomhet ble mindre effektiv.

3. Utfordringer ved digitalt arkivmateriale

Digitalt skapte fotoarkiver har noen utfordringer som er spesifikke for digitalt arkivmateriale. Størrelsen på arkivet utgjør i så måte ingen vesentlig faktor, men det har praktisk betydning i forhold til både behandling, lagring og overføring av datamengden. Det er også noen fordeler med håndtering av digitalt materiale som vi skal komme tilbake til.

Tre viktige utfordringer som må adresseres er lagringsbestandighet, datasikkerhet og formatbestandighet.

(Se OAIS og DIAS-standardene for mer informasjon om dette)

3.1. Lagringsbestandighet

Intuitivt tenker man gjerne på digitalt skapt materiale som «evigvarende», men det viser seg at det på mange måter kan være mer sårbart enn analogt materiale. Avhengig av datamengde finnes det mange medier man kan oppbevare digitale data på:

- Spolebånd
- Disketter
- Optiske medier som CD, DVD, Laserdisc
- Minnepinner
- Eksterne harddisker av rotasjonstype
- Eksterne harddisker av brikketype (SSD)
- Backuptaper
- Aktive disksystemer uten redundans[3] (harddisk på PC, server, NAS-enhet)
- Aktive disksystemer med redundans (RAID-løsning på server, NAS-enhet)
- Skybaserte lagringsløsninger

Forventet levetid for et frakoblet datamedium (for eksempel en minnepinne eller ekstern harddisk) kan variere fra noen måneder til noen få år. Forventet levetid for en tilkoblet harddisk kan variere fra ett år til noen få år. Forventet levetid for en backuptape kan variere fra noen få måneder til 30 år, forutsatt lagring under optimale forhold. [4]

Det å legge digitale data på et eksternt medium og plassere dette i en arkivboks i et magasin er med andre ord ikke en permanent løsning, og materialet kan i verste fall være forsvunnet etter bare noen måneder, selv under perfekte lagringsforhold.

Det finnes også et fenomen som kalles «bitråte», der man kan få vilkårlige feil i datafiler over tid på grunn av degradering av ladningen til det magnetiske feltet på overflaten til lagringsmediet, eller degradering av det fysiske lagringslaget på optiske medier. Hvis en fil for eksempel består av 42000 tegn eller «bytes» kan feil som oppstår i en av disse bytene isolert sett være neglisjerbart, men over tid vil dette akkumulere seg og til slutt degradere innholdet i filen til et punkt der den ikke lenger er

[3] Redundans betyr at data lagres to eller flere steder samtidig. En redundant diskløsning består for eksempel av 6 harddisker, der 4 disker er satt av til datalagring, og 2 disker til redundans (sjekksummer). I en slik løsning vil 2 av 6 disker kunne feile uten at data går tapt.

[4] Tester av dette viser svært varierende resultater, de angitte tallene i denne rapporten er dels testresultater og dels produsentenes egne påstander.

lesbar. Noen deler av en datafil kan også være mer sensitiv for endringer enn andre deler, og hele filen kan ødelegges selv av et veldig lavt antall endrede bytes.
En løsning på dette kan være å ha data lagret på flere forskjellige lagringsmedier og gjerne flere forskjellige fysiske steder, og ha en migreringspolicy som sørger for at data kopieres fra gammelt til nytt lagringsmedium innen utløpet av den forventede levetiden til det fysiske lagringsmediet.

3.2. Datasikkerhet

Hvis man oppbevarer det digitale arkivet på en offline-løsning, for eksempel på eksterne harddisker plassert i et magasin, er man i teorien beskyttet mot dataangrep («hacking» og løsepengevirus som de to største farene), men møter da utfordringen med at slike lagringsmedier forringes over tid.

Hvis man oppbevarer det digitale arkivet på en online-løsning, for eksempel på en depotserver med en lagringsløsning med redundans og regelmessig backup, sikrer man seg mot uforutsett tap av data. Der vil man også kunne skifte ut og oppgradere komponenter etter hvert som de degraderes, slik at man alltid vil kunne holde løsningen kjørende, eller i verste fall vil kunne hente tilbake data fra backup om lagringsløsningen skulle svikte.

En onlineløsning vil imidlertid være sårbar for uautorisert tilgang, enten utenfra (hacking eller datainnbrudd) eller innenfra (uforsettlig eller villet handling av bruker, kryptovirus). Det finnes måter å sikre et digitalt depot på slik at man minsker risikoen for tap av data, og dette gjøres gjerne gjennom en kombinasjon av tilgangsbegrensning, lagringsløsning med redundans, og backup.
Det å sette opp og vedlikeholde et digitalt depot krever en viss IT-kompetanse. Slik kompetanse kan kjøpes fra IT-leverandører, men det kommer ofte med en ganske høy prislapp.

I forhold til disse to punktene kommer Digitalarkivet inn som en opplagt løsning for langtidsbevaring av det digitale materialet.

3.3. Formatbestandighet

Over tid utvikles og endres både maskinvare og programvare. Programfiler og datafiler henger tett sammen, slik at når programvare videreutvikles og endres, vil ofte formatet på de tilhørende datafilene endres.

Ofte søker programvareleverandører å holde programvaren «bakoverkompatibel», det vil si at gamle datafiler også skal kunne leses og benyttes av nyere utgaver av programvaren, men i en del tilfeller er ikke dette tilfredsstillende gjort, eller skjer ikke i det hele tatt.

Et eksempel er Microsoft Word, et svært mye benyttet tekstbehandlingsprogram. Der finnes det et titalls forskjellige dokumentformater, helt tilbake til Word 1.0 for DOS, via Word 95, Word 2000, Word 2003 og frem til dagens Word 2019. Dagens utgave av Word vil kunne lese datafiler mange versjoner tilbake, men kan ikke nødvendigvis garantere at de aller eldste formatene blir fremstilt på skjermen identisk med slik de ble fremstilt i sin opprinnelige programvareversjon. Det garanteres heller ikke at en slik *bakoverkompatibilitet* vil være implementert i alle fremtidige versjoner av programmene.

Formater for digitale fotografier er ganske stabile; JPEG og TIFF er de mest brukte formatene for disse. JPEG-standarden ble godkjent i 1992/1994, og har hatt en del tilføyelser til så sent som 2012. Siste revisjon av TIFF-formatet ble utgitt i 2004. Det har blitt utviklet en rekke nyere bildeformater som har fått en viss utbredelse for eksempel på web, men ingen av disse har i særlig grad blitt

adoptert av fotoverdenen. Det som imidlertid har endret seg er antall megapixels og fargedybden bildene er fotografert i, noe som har gjort den enkelte bildefilen vesentlig større.

For å holde digitalt arkivmateriale levende og tilgjengelig må man altså sørge for å til enhver tid forsikre seg om at formatet på datafilene som er arkivert, det være seg dokumenter, bildefiler, lydfiler, videofiler eller annet, fremdeles er lesbart med dagens programvare. Dette gjøres gjennom to strategier; konvertere de opprinnelige filene til et langtidslagringsformat ved opprinnelig arkivering, og ved å ha jevnlige gjennomganger av digitalt depot og rekonvertere filer hvis format står i fare for å gå ut på dato. [5]

[5] En tredje strategi er å vedlikeholde nødvendige maskin- og programvareplattformer til å åpne og lese alle gamle formater, men det er en krevende strategi, både kostnadsmessig og IT-logistikkmessig, og anbefales i utgangspunktet ikke.

4. Fordeler ved digitalt arkivmateriale

Arkivmateriale i digital form har også noen fordeler:

Det er fullt mulig å lage flere identiske kopier som kan oppbevares flere steder, uten at det skjer noen som helst forringelse av de originalt overførte dataene. Innhold kan gjøres tilgjengelig over internett raskt og enkelt, og uten risiko for at originalt innhold forringes ved bruk.

Det kan også innføres automatiske rutiner som overvåker dataene i lagringssystemene, og sender ut varsel og reparerer dem til opprinnelig tilstand hvis det blir oppdaget forringelser eller endringer.

Man kan også lagre svært mye materiale på liten plass – en normal harddisk på ca. 15 x 10 x 2,5 cm. kan nå lagre en datamengde på opptil 20 TB. Etter gammel regnemetode er en A4-side med tekst ca. 2 kB, og følgelig kan man lagre ca. 10 milliarder A4-sider med ren tekst på en slik harddisk. Et JPG-bilde kan variere svært i størrelse, men hvis man tar utgangspunkt i LO Media-fotoarkivet er snittstørrelsen på ca. 5.5 MB pr. bilde, noe som tilsvarer lagring av ca. 3.6 millioner bilder.

Digitalt materiale kan også, under visse forutsetninger, indekseres slik at man kan foreta fritekstsøk i det, og det kan benyttes metoder for analyse av innholdet, slik at man kan få automatisk genererte sammendrag og/eller stikkordslister som kan bidra til å gjøre gjenfinning av materiale enklere. For bilde- og filmmateriale kan automatisk analyse med kunstig intelligens-baserte systemer utføres, og disse kan automatisk merke blant annet objekter, ansikter og egenskaper ved mediet.

Tekstdokumenter som er avfotografert som bilder kan OCR-analyseres for å få ut tekstinnholdet. Det finnes også løsninger for automatisk tolking av tale til tekst, slik at man kan generere indekserbare tekstfiler fra lydopptak. Disse prosessene er avhengige av et godt datagrunnlag for å oppnå god presisjon, og det kan være problematisk i forhold til kvaliteten på originalmaterialet.

5. Metadata i materialet

For å håndtere bildesamlinger/fotoarkiver og tilhørende metadata benyttes såkalte DAM-løsninger. Disse består av et bildelager, en database eller indekser for metadata, et brukergrensesnitt der fotografer, fotoarkivarer og andre kan manipulere og redigere metadatainformasjonen, og en kobling mot fotoredigeringsprogrammer (for eksempel Adobe Photoshop). De kan også inkludere moduler for tilgjengeliggjøring av bilder, enten i form av album som deles til enkeltpersoner eller i form av «gallerier» som gjøres tilgjengelig på internett, moduler for kobling mot eksterne publiseringsløsninger som for eksempel Digitalt Museum, og annen funksjonalitet.

Bildematerialet fra LO Media har vært ivaretatt av flere forskjellige DAM-verktøy, som har brukt forskjellige måter og forskjellige «skjemaer» for å lagre metadata, noe som har bidratt til at mengden og kvaliteten på metadata som følger filene varierer. Det er også varierende hvor stor mengde og hvor detaljert informasjon den enkelte innholdsskaper har lagt på bildene.

Gjennom skifte av DAM-systemer og samkjøring av bildemateriale fra mange kilder har metadatainformasjonen blitt ytterligere sammenflettet og til dels duplisert over flere felter på hvert bilde.

En del metadatafelter med kamerainformasjon/teknisk informasjon har blitt fylt ut på fotograferingstidspunktet, som kameratype, blenderåpning, fokus, og så videre. Disse har for så vidt ganske konsistent informasjon, men det varierer likevel noe hvilke felter som er benyttet for å lagre disse opplysningene. Det viser seg imidlertid at en del av disse feltene inneholder feil informasjon; vi har for eksempel sett at opprettelsesdato og tidssone inneholder verdier som er helt inkonsistente eller inkorrekte. Vi antar dette skyldes at bildene er fotografert med kamerautstyr som ikke er konfigurert riktig.

Her er et utdrag av tidssoneverdier for å illustrere:

Tagname	**Tagcontent**
Time Zone	-01:00
Time Zone	-04:00
Time Zone	-05:00
Time Zone	-06:00
Time Zone	-08:00
Time Zone	-507:45
Time Zone	+00:00
Time Zone	+01:00
Time Zone	+02:00
Time Zone	+03:00
Time Zone	+04:30
Time Zone	+05:30
Time Zone	+09:00
Time Zone	+10:00
Time Zone	+12:00
Time Zone	+256:00

Nedenfor står noen eksempler på oversikter vi har hentet ut fra materialet. Disse er generert ut fra spørringer mot en SQL-database som alle metadata er importert inn i – mer om dette i avsnittet «Vurderinger og verktøyvalg». Her er en tabell over noen av de flest forekommende metadatafeltene (tags) som inneholder verdier:

Tagname	**Tagcount**
Create Date	941634
Focal Length	934450
Date/Time Original	902025
Modify Date	846276
Already Applied	573996
Application Record Version	573996
Bits Per Sample	573996
Color Components	573996
Color Mode	573996
Content Value	573996
Creator Tool	573996
Current IPTC Digest	573996
Directory	573996
Displayed Units X	573996
Displayed Units Y	573996
Document ID	573996
Encoding Process	573996
Exif Byte Order	573996
Exif Image Height	573996
Exif Image Width	573996

Det er interessant å observere at noen av feltene finnes flere ganger pr. bilde. Dette kan skyldes feil i tolkingen/innlesingen av det enkelte bildet, eller det kan være at disse feltene har blitt duplisert ved konvertering av bildene. Vi ser at de fleste av disse feltene enten inneholder dato eller tekniske egenskaper ved bildet.

Et eksempelutdrag av innholdet i «Create Date»-metadatafeltet indikerer at dataene er konsistente selv om det er duplikater av feltene, men at det er selve formatet på tidsfeltet som varierer:

fileID	Tagname	Tagcontent
335170	Create Date	2017:06:20 09:59:00
335170	Create Date	2017:06:20 09:59:00.73
335171	Create Date	2011:07:01 12:26:15+02:00
335172	Create Date	2014:09:27 19:34:01+02:00
335173	Create Date	2007:09:14 11:21:15+01:00
335174	Create Date	2016:06:10 14:29:39
335174	Create Date	2016:06:10 14:29:39.50
335175	Create Date	2009:10:07 22:55:37
335176	Create Date	2002:08:20 19:49:10
335176	Create Date	2002:08:20 19:49:10.11
335177	Create Date	2012:11:30 14:15:51
335177	Create Date	2012:11:30 14:15:51.60
335178	Create Date	2007:11:14 18:38:20
335179	Create Date	2016:06:21 20:15:25
335179	Create Date	2016:06:21 20:15:25.23
335180	Create Date	2013:11:21 10:55:50
335180	Create Date	2013:11:21 10:55:50.40

Her er en tabell over noen av de flest forekommende metadatafeltene som inneholder data, og som finnes i færre enn alle bilder:

Tagname	Tagcontent
Object Name	573970
Title	573970
IPTC Digest	573888
Thumbnail Image	573795
Thumbnail Length	573795
Thumbnail Offset	573795
Usage Terms	573795
Short Document ID	571322
Short Unique Id	571322
Y Cb Cr Sub Sampling	570845
Date Created	542187
Instance ID	534355
Release Date	529228
Color Space	516578
By-line	515493
Creator	515491
ICC Profile Name	509162
Caption-Abstract	499009
Description	498465
Exif Version	479017

Vi ser også her at det er mange tekniske metadatafelter, men felter som «By-line» og «Creator» er godt populert. (Merk: Det er ikke gjort noe forsøk på å verifisere verdiene i feltene for denne spørringen.)

En del metadatafelter har vært brukt om hverandre, for eksempel Caption-Abstract/Description/Image Description, og Credit/Artist/By-line.

Hvis vi ser på Description, Caption-Abstract og Image Description-feltene i sammenheng, kan vi trekke ut en oversikt som viser hvor mange av bildene som inneholder verdier i de respektive feltene:

Description:	498465
Caption-Abstract:	499009
Image Description:	73012
Antall bilder som ikke har verdi i noen av disse tre feltene:	14150

Dette indikerer at det for nesten alle bildene i arkivet finnes en beskrivelse av innholdet, og en spørring som kombinerer innholdet fra disse tre feltene bekrefter dette. Imidlertid viser spørringen også at en del av disse feltverdiene er ganske mangelfulle:

fileID	**Tagname**	**Tagcontent**
342887	Caption-Abstract	BYLINE
342961	Caption-Abstract	BYLINE
343060	Caption-Abstract	Tulipaner
343233	Caption-Abstract	el og it
343324	Caption-Abstract	valg 2013
343365	Caption-Abstract	app
343510	Caption-Abstract	RAD 25 a
343628	Caption-Abstract	Test
343698	Caption-Abstract	juni 2008
343720	Caption-Abstract	brudd
343762	Caption-Abstract	Band
343879	Caption-Abstract	RAD 10a
343976	Caption-Abstract	Mesta
344030	Caption-Abstract	merkefest

En del metadatafelter har vært definert som fritekst og utfylling av disse feltene har ikke alltid vært påkrevd. De inneholder også variasjoner i skrivemåten, siden de ikke er utfylt fra forhåndsdefinerte verdilister. Dette er felter som for eksempel inneholder stikkord, informasjon om fotograf, eller lisensinformasjon.

Her er et utvalg verdier fra feltet «Credit» for å illustrere problemet med forskjellige skrivemåter:

Sissel M .Rasmussen
Sissel M Rasmussen
Sissel M Rasmussen Sissel M Rasmussen
Sissel M. R.
Sissel M. Rarsmussen
Sissel M. Rasmusen
Sissel M. Rasmussen
Sissel M. Rasmussen Sissel M. Ra
Sissel M. Rasmussen Sissel M. Rasmussen
Sissel M. Rasmussen Sissel M. Rasmussen Sissel M. Rasmussen
Sissel M. Rasmussen, LO-Aktuelt
Sissel M. Rasmussen.
Sissel M. Rasmussen/LO Media
Sissel M. Rasmussen/LO-Aktuelt
Sissel M. Rasmussen/LO-Aktuelt Sissel M. Rasmussen/LO-Aktuelt
Sissel M. Rasmussen/Nicolaas Kip
Sissel M. Rasmusssen
Sissel M.Rasmussen
Sissel M.Rasmussen Sissel M. Rasmussen
Sissel M.Rasmusssen
Sissel Mary Rasmussen
Sissel Rasmussen
Sisssel M. Rasmussen

Fotoarkivet inneholdt også en mengde duplikater. Under den innledende overføringen ble alle thumbnails (miniatyrbilder) filtrert bort, og dette reduserte bildemengden med grovt sett 1/3. Vi visste at det både var en hel del helt identiske bilder i materialet, og mistenkte at det kunne være en viss mengde bilder som var helt like, men ikke identiske (der det for eksempel var gjort justeringer i lysstyrke, kontrast, fargemetning, etc.). Filnavnene på bildene (for eksempel 123456_.jpg og 123456_orig.jpg) indikerte at materialet inneholdt et nesten fullstendig sett dubletter, men vi valgte å ikke ta sjansen på å sortere ut bilder kun basert på filnavnet, da det ikke ville gi muligheten til å sammenstille metadata eller sammenligne bildeinnholdet for å verifisere at bildene faktisk var helt identiske.

Siden datamengden det var snakk om i LO Media-fotoarkivet var så stor, og vi ønsket å finne metoder som kunne håndtere større datamengder, måtte vi finne måter å analysere en slik datamengde på som var praktisk gjennomførbar ved bruk av enten kraftig PC- eller normalt kraftig serverhardware, og ved bruk av gratis programvare, overkommelig priset kommersiell programvare, eller en kombinasjon av disse.

Vi gjennomførte en serie forsøk for å forsøke å kvalifisere en rekkefølge av operasjoner som kunne redusere mengden duplikate data i størst mulig grad, og gi fotoarkivarene et sett med verktøy for å halvautomatisk eller manuelt behandle bildemengden videre. Det var også et mål at disse verktøyene skulle kunne brukes på både små og store fotoarkiver i etterkant.

6. Vurderinger og verktøyvalg

Utgangspunktet var å benytte DAM-verktøyet FotoStation fra Fotoware til å utføre all bilde- og metadatabehandlingen i prosjektet, da det er et verktøy vi bruker i det daglige og kjenner godt, og da dette var konsistent med løsningen LO Media benyttet på avleveringstidspunktet. Vi så imidlertid raskt at dette hadde noen begrensninger som kunne være uheldige:

- Fotowares programvareløsninger er relativt kostbare, og nytteeffekten av resultatene fra dette prosjektet ble begrenset hvis vi låste det mot en spesifikk programvareløsning
- Ved å benytte en ferdig programvareløsning mistet vi den detaljerte innsikten i materialet som vi kunne få ved å analysere enkeltelementer av hhv. bildedata og metadata med diskrete verktøy (Se punktet «Analyser bildedata med tanke på identifisering av «like» bilder»)
- Fotoware benytter en egen standard for feltoppsett på metadata. Denne kan tilpasses, men reduserte likevel muligheten for å manuelt kontrollere og konfigurere detaljerte metadataoppsett slik vi ønsket

Alternativet vi valgte var å basere oss på en åpen plattform og open source-programvare for å utforme et grunnsett med metoder til relativt lav kostnad, og så sammenligne operasjonene mot Fotowares programvare for å verifisere metodene, i den grad tidsrammen i prosjektet tillot dette, og vurdere hvorvidt det var store fordeler/ulemper ved å benytte den ene eller den andre verktøykassen.

Vi valgte å starte med en såkalt LAMP-plattform (Linux, Apache, MySQL, PHP), som er et kjent og fleksibelt rammeverk for å utvikle løsninger av lav til middels kompleksitet og kapasitet. Debian GNU/Linux ble valgt som operativsystem.

Se vedlegg B for spesifikasjon av maskinvare og programvare som ble benyttet i prosjektet.

7. Innhenting av materiale

Overføringen av fotoarkivet gikk som en løpende jobb fra LO Media til oss. De satte opp en prosess i sitt DAM-system som overførte bilder fortløpende til en FTP-server hos oss. Et script hentet så de mottatte bildefilene fra vår FTP-server og inn til vår fotoserver, som kjører programvaren Index Manager og Color Factory fra FotoWare AS. Våre fotoarkivarer benytter FotoStation 8.0 som klientprogramvare mot serveren.

Denne programvareløsningen hadde noen kapasitetsbegrensninger i vår versjon som vi støtte på med dette arkivet: I ett enkelt indeks kunne det kun behandles opptil 100.000 bildefiler, og total kapasitet for hele serverløsningen var 1.000.000 bilder. Dette ble løst mot slutten av dette prosjektet gjennom en oppgradering av lisensen, og oppdatering av programvaren. [6]

Alle de mottatte bildene ble lagt i en samlet mappe, og det ble ikke gjort noen prosessering/sortering av materialet for å foreta en automatisk inndeling med tanke på å holde hver enkelt mappe under maksimal indeks-størrelse. Dette medførte at serverprogramvaren vår ikke i utgangspunktet kunne prosessere fotoarkivet under ett.

I møte med LO Media fikk vi avklart at bildene var plassert i en mappestruktur hos dem, men der mappeinndelingen var helt uavhengig av innhold og struktur i fotoarkivet, og kun basert på behov for å holde antall bilder pr. mappe nede på et håndterbart nivå. Hvis det hadde vært en logisk sammenheng mellom mappeinndeling og bildeinnhold/metadatainnhold hadde det vært viktig å beholde den opprinnelige mappestrukturen ved overføringen.

Selve overføringsmetoden fungerte til dels tilfredsstillende, men vi ser at en overføring over lang tid har sine ulemper. Erfaringer vi har gjort oss er at det er vanskelig å ha en god, løpende kontroll med overføringen. En forutsetning for at dette skal fungere godt er en god dialog mellom arkivgiver og arkivinstitusjon på hva som til enhver tid er forventet overført. Noe rapportering av dette kunne vært satt opp til å genereres automatisk, som en tilleggsjobb til selve dataoverføringen.

Vi har også sett at endringer underveis i infrastruktur (servere og internettlinjer) har ført til uregelmessigheter i overføringene.

[6] Fotoware har valgt å gjøre store organisasjonsendringer mens dette prosjektet har løpt, og også endre lisensieringsmodellen sin vesentlig. Dette påvirket oss til en viss grad, og forsinket oppgraderingen av løsningen vår.

8. Metodeutvikling

Vi ønsket å se på metode i vid forstand i dette prosjektet; vi ønsket både å nedfelle en rutinebeskrivelse for overlevering, mottak, deponering og publisering av fotoarkiver, men også utforske metoder for analyse av bildeinnhold og fjerning av dubletter, og automatiske og halvautomatiske prosesser for å kvalitetssikre og ensrette datamengden uten at noe av det unike innholdet gikk tapt.

Den resulterende rutinebeskrivelsen er dokumentert i vedlegg A: «Rutinebeskrivelse for mottak av digitalt skapte fotoarkiver».

Arbark benytter i utgangspunktet Archivematica fra Artefactual Systems som løsning for å behandle og deponere digitale arkiver, men dette systemet er foreløpig i testfase og ikke produksjonssatt. Vi har likevel tatt utgangspunkt i at dette blir det endelige produksjonssystemet for vårt depot i rutinebeskrivelsen. Det er dog ikke noen nødvendighet eller forutsetning for å følge rutinen.

Vi vil nedenfor gå gjennom funnene vi har gjort, og utdype punktene av rutinebeskrivelsen der det er ønskelig. Noen elementer beskrives i detalj, mens noen vil være skissert på et overordnet plan.

Overordnet kan vi beskrive rutinen ved mottak av fotoarkiv slik:

1. Inngå avtale med arkivskaper, herunder avklaring av rettigheter og avtale om publisering
2. Avklare datamengdens beskaffenhet, og kvalitetssikre uthentingen av data fra arkivskapers bildehåndteringssystem
3. Avtale teknisk plattform for overføring av materialet mellom arkivskaper og arkivinstitusjon
4. Foreta selve overføringen
5. Kvalitetssikre mottak av data hos arkivinstitusjon
6. Sette dataene i karantene i en forhåndsdefinert periode
7. Hente data ut fra karantene for prosessering
8. Kontrollere filformater, fjerne identiske dubletter og slå sammen metadata
9. Analysere og fjerne «like» bilder, manuell gjennomgang av resultatet av dette
10. Analysere metadata, bruke automatiske og halvautomatiske verktøy for å ensrette og kvalitetssikre disse
11. Konvertere eventuelle bilder som ikke er i godkjent format for langtidslagring
12. Pakke ned arkivpakke med bilder og metadata for deponering i digitalt depot
13. Overføre en «arbeidspakke» til fotoserver for tilgang til fotoarkivarer

9. Metode – detaljert gjennomgang

9.1. Inngå avtale med arkivskaper

Et digitalt skapt eller digitalisert dokumentarkiv vil som regel ha en eller få rettighetshavere, mens et digitalt skapt fotoarkiv vil kunne ha mange rettighetshavere. Dette kan legge store begrensninger på mulighetene for fri publisering. Den enkelte fotograf kan ha individuell avtale med arkivskaper, og åndsverksloven regulerer også rettighetene til materialet.

Det må derfor avklares detaljert på hvilke måter arkivinstitusjonen kan bruke materialet, og hvilke kriterier som ligger til grunn for publisering av enkeltbilder og/eller hele arkivet enten via egne kanaler eller via Digitalarkivet.

Ved publisering til Digitalarkivet kan tilgang styres svært detaljert, og arkivskaper kan også være den som kontrollerer tilgangsstyringen og håndterer forespørsler om tilgang. Dette gir en stor grad av trygghet for arkivskaper for at materialet oppbevares og håndteres på en langsiktig og sikker måte.

Vår avtale med LO Media ble inngått før dette prosjektet ble definert, så den er innholdsmessig veldig enkel, og omhandler ikke videre overføring av fotoarkivet for deponering eller tilgjengeliggjøring.

9.2. Avklare datamengde og kvalitetssikre uthenting

Det er viktig å avklare rammene for datamengden som skal overføres på forhånd, hvilke tanker og ønsker arkivgiver har rundt overføringen av data, og å kvalitetssikre at mest mulig metadata blir overført. Dette kan for eksempel inkludere opprinnelige datostempler på bildefilene, og mappestrukturen filene er plassert i. Ved eksport fra et DAM-system kan det være satt opp et eksportfilter som kun henter et utvalg av metadata ved eksport, eller som samkjører metadatafelter på en måte som ikke er ønskelig for arkiveringsformål.

Likeledes er det viktig å beskrive en eventuell metadatamal som er benyttet, formater for datafelter (datoformater, tekstformater), og hvorvidt oppslagsfelter er fylt ut manuelt eller fra forhåndsdefinerte datalister (for eksempel fotografnavn).

Noe av dette ble gjort i avleveringsavtalen mellom LO Media og Arbark, men de tekniske aspektene ved metadatainnholdet ble ikke tilstrekkelig avklart og dokumentert.

9.3. Avtale overføringsmåte

Overføringsmåte vil være avhengig av flere faktorer: Størrelse på arkivet, hvorvidt det er et «levende» arkiv med ny tilvekst, og kompetanse og infrastruktur hos arkivgiver og arkivinstitusjon.

Hvis mengden totale data er under en gitt grense kan det ofte være greit å bruke et flyttbart medium for overføringen, både fordi det er praktisk å hente inn datamengden i en bolk, og fordi

utkopiering hos arkivgiver vil kunne gå såpass raskt at det ikke er til vesentlig ulempe. Det flyttbare mediet bør sikres mot tilgang for uvedkommende gjennom passordbeskyttelse og kryptering[7].

Hvis datamengden er (vesentlig) større kan det være formålstjenlig å vurdere en elektronisk overføring over internett, som kan kjøre som en automatisk oppgave over et stykke tid, uten å kreve særlig vedlikehold fra hverken arkivgiver eller arkivinstitusjon, og uten å belaste datasystemene i vesentlig grad. En slik overføring må i så tilfelle sikres mot uønsket tilgang og avlytting.

Det må også avtales om det er en engangsavlevering, periodisk avlevering, eller en løpende overføring som sørger for at ny tilvekst til arkivet automatisk overføres. Det bør avklares en ferdigstillingsdato for overføringen, eller eventuelt en oppfølgingsdato der man kan gjennomgå resultatet av overføring så langt, og fornye rammene eller definere nye rammer for videre overføring.

I avtalen vår med LO Media ble det ikke spesifisert noen parametre for overføringen, dette ble kun avklart muntlig.

9.4. Foreta overføring

Når parameterne for overføringen er satt, kan selve overføringen finne sted. Dette kan ta alt fra minutter til å gå over lang tid, avhengig av hvordan overføringen utføres.

Det bør føres jevnlige kontroller med kvaliteten på dataene som er overført og avtales en sluttgjennomgang etter endt overføring der arkivgiver verifiserer at datamengden som er overført er som forventet og i henhold til tidligere inngått avtale.

De mottatte dataene bør fortløpende flyttes fra mottaksområdet inn til et internt lagringsområde som ikke er tilgjengelig fra internett, der de ikke kan røres/åpnes, i påvente av å plasseres i karantene etter fullført overføring. Det er dog en fordel om det finnes en isolert arbeidsstasjon tilgjengelig som kan benyttes til å verifisere dataene som er mottatt, før eller mens de befinner seg i karantene.

9.5. Kvalitetssikre mottak av data

Etter fullført mottak av data bør det foretas en gjennomgang av de mottatte dataene sammen med arkivskaper. I første omgang dreier det seg om en opptelling og oversikt over data som er overført, og i den grad det er mulig en verifisering av innholdet av overføringen. Sett i forhold til karanteneperiode er det som nevnt ønskelig å ha en egen isolert arbeidsstasjon der denne verifiseringen kan utføres. Dette kan for eksempel være en virtuell arbeidsstasjon som nullstilles etter bruk.

Dette ble ikke gjort i forhold til overføringen fra LO Media.

[7] Kryptering av et lagringsmedium vil si at alle data som lagres blir kodet med en nøkkel og/eller et passord. For å kunne hente ut dataene må nøkkel/passord oppgis ved utkopieringstidspunktet. Kommer lagringsmediet på avveie vil ikke tredjepart kunne dekode og lese de lagrede dataene.

9.6. Sette dataene i karantene

Når overføring er ferdigstilt skal det foretas virussjekk på alle data, og de skal plasseres i karantene i en gitt periode. Anbefalt tidsperiode er 3 uker. Plassering i karantene vil si at alt mottatt materiale plasseres på et nedlåst lagringsområde slik at ingen har mulighet til å hverken uforvarende eller med hensikt åpne filene i et program. Dette gjøres for å beskytte arkivinstitusjonen mot virusangrep og såkalte nulldagssårbarheter, der filene kan være infisert med et virus som er så nytt at ingen antivirusprogrammer har kjennskap til det ennå.

Når karanteneperioden er ferdig kjøres det ny virussjekk på filene, og de leveres ut av karantene til et arbeidsområde.

Strengt tatt er det veldig sjeldent med virusinfeksjoner på bildefiler, men så lenge det er en risiko større enn null må man ta de nødvendige hensyn og forhåndsregler.

Overføringen fra LO Media gjennomgikk ikke et karanteneopphold.

9.7. Hente data ut fra karantene

Når filene har gjennomgått karanteneopphold, bestått oppdatert antivirussjekk og blitt levert ut til arbeidsområde er de klare til behandling.

Litt avhengig av størrelsen på fotoarkivet, og i hvor stor grad man har lagt opp en egen plattform for (automatisk) bearbeiding av bildemateriale, kan dette være et arbeidsområde inne på et dedikert lagringsområde for digitalt arkivmateriale, et område inne på en fotoserver, eller det kan være på arbeidsstasjonen til en fotoarkivar.

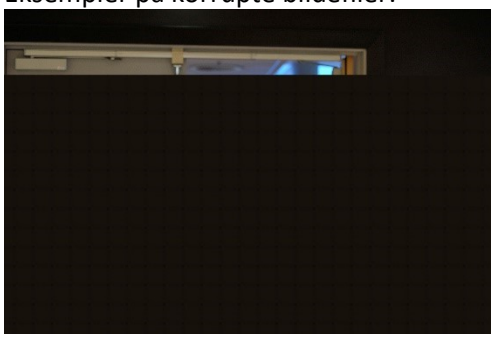
Vi valgte i prosjektet å hente ut alle metadata fra bildefilene og plassere i en SQL-database for videre behandling. Dette gir flere fordeler: Man kan bruke SQL-spørringer til å sammenstille og hente ut metadata, og til å foreta utvalg av bildefiler etter svært sammensatte kriterier, om ønskelig.

9.8. Kontrollere filer og fjerne dubletter

(Jfr. Delmål 1 i prosjektbeskrivelsen)

Den første operasjonen er å kjøre alle filene gjennom et verktøy som kan teste og verifisere at filene har riktig filformat. Hvis man har en andel korrupte bildefiler kan disse skape problemer på et senere tidspunkt i prosessen hvis de ikke blir separert fra fotoarkivet for manuell håndtering. Vi oppdaget for eksempel at av de 577.000 bildene vi prosesserte var det nesten 2000 korrupte filer. Feilene i disse spredde seg over et spekter fra der feilen ikke var merkbar i praksis til at filen var fullstendig uleselig. Ved å trekke disse filene ut til et eget område kan man behandle disse manuelt, og også melde en liste over filene tilbake til arkivskaper, slik at de eventuelt kan verifisere dette mot eget arkiv, og gjøre en separat overføring dersom dette er filer som har blitt skadet under overføringen. (Merk at filer i en suppleringsoverføring også må gjennom et karanteneopphold.)

Eksempler på korrupte bildefiler:




Neste trinn i prosessen er å fjerne identiske filer. Dette gjøres ved å beregne en sjekksum for hvert bilde, og så sammenligne sjekksummene etterpå for å finne dubletter. Vi forsøkte dette først på de originale bildefilene, men fant at det ga svært mangelfulle resultater, med bare noen få tusen dubletter.

For at dette skulle bli effektivt måtte vi separere bildedata og metadata, slik at sjekksummene kunne beregnes kun ut fra bildedataene. Vi valgte å lage en kopi av bildefilene der vi trakk bort metadata, kalkulere en SHA256[8]-sjekksum av hver fil, legge sjekksummene inn databasen i tabellen over alle filene, og så kjøre en sammenstilling av sjekksummene. Resultatet ble at vi satt igjen med ca. 260.000 unike bildefiler.

Å kombinere metadataverdier fra dubletter som fjernes fra systemet representerer en utfordring. Vi søkte å løse dette ved å utarbeide et regelsett som kan benyttes for å til dels automatisere denne prosessen. I fremstillingen nedenfor bruker vi A og B som betegnelser for de to identiske bildefilene:

- Lag en liste som inkluderer alle metadatafelter i både A og B
- Behandle listen over felter, trekk ut verdiene fra A og B
- Er begge feltene blanke? Gå i så fall videre.
- Er det verdi i bare ett av feltene? Bruk i så fall denne verdien.
- Er det verdi i begge feltene?
    - Er det et datofelt?
        - Hvor stort avvik er det mellom tidspunktene i feltene?
            - Er avviket mindre enn en neglisjerbar grenseverdi brukes det eldste tidspunktet. (Dette kan for eksempel være at ett av feltene inneholder en verdi for tusendels sekunder, mens det andre ikke gjør det.)
            - Er avviket større enn en neglisjerbar grenseverdi opprettes et nytt metadatafelt med feltnavn lik det opprinnelige med tillegg av «_2», verdien fra A beholdes i opprinnelig felt og verdien fra B legges i det nye feltet.
    - Er det et tekstfelt?
        - Er A en deltekst av B? Bruk i så fall B.
        - Er B en deltekst av A? Bruk i så fall A.
        - Er begge forskjellige? Slå dem i så fall sammen.
    - Er det et tallfelt?

[8] Det finnes en del kjente og mye brukte algoritmer for beregning av sjekksummer. De mest brukte er MD5, SHA-128 og SHA-256. En sjekksum er et kalkulert «tverrsnitt» av en datamengde/fil, som i utgangspunktet er helt unikt for enhver datamengde, og som endres ved endringen i datamengden. Den kan brukes for å verifisere at innholdet er urørt.

        - Er verdiene like? Bruk i så fall A.
        - Er verdiene forskjellige, men den ene er null? Bruk i så fall verdien som ikke er null.
        - Er verdiene forskjellige, og begge er ulike null? Opprett et nytt metadatafelt med feltnavn lik det opprinnelige med tillegg av «_2», verdien fra A beholdes i opprinnelig felt og verdien fra B legges i det nye feltet.
    - Er det andre dataformater i feltet?
        - Er verdiene like? Bruk i så fall A.
        - Er verdiene ulike? Slå i så fall sammen verdiene dersom feltet tillater det, eller opprett et nytt metadatafelt med feltnavn lik det opprinnelige med tillegg av «_2», verdien fra A beholdes i opprinnelig felt og verdien fra B legges i det nye feltet.

Denne metoden kan forbedres, men den sørger for at metadatainnhold ikke går tapt ved automatisk sammenslåing av feltene.

9.9. Analysere og identifisere «like» bilder

Dette punktet lar seg vanskelig gjøre å helautomatisere, men det er mulig å bygge et rammeverk som gir fotoarkivarene en mulighet for å kjøre en slik analyse, og manuelt velge å fjerne dubletter der forskjellene i bildeinnhold ikke tilsier at kopier bør beholdes.

Ved å sette en grenseverdi, kjøre et verktøy som prosesserer bildene og beregner et «fingeravtrykk», og så legge disse fingeravtrykkverdiene inn i databasen, kan vi kjøre en spørring som sammenstiller filer med like fingeravtrykk på samme måten som vi tidligere sammenstilte filer med like sjekksummer. Disse ikke-identiske dublettene vises i et grafisk grensesnitt for arkivarene, som gis mulighet for å krysse av for de tilfellene der de ønsker at de overflødige dublettene skal fjernes fra arkivet. Metadata fra dublettene vil i så fall slås sammen etter samme regelsett som beskrevet tidligere.

Eksempler på ikke-identiske dubletter som kan fjernes er for eksempel bilder der det er gjort justeringer av kontrast, lysstyrke, fargemetning eller andre parametere. Eksempler på bilder som blir feilaktig identifisert som ikke-identiske dubletter kan være bilder som har svært liknende «linjer» i motivet, eller bilder av svært enkle motiver, for eksempel logoer og andre illustrasjoner. (Bilder av samme motiv i forskjellig farge blir for eksempel identifisert som like.)

Her er eksempler på dette:

Foto: Kristian Brustad / LO Media

Foto: Erlend Angelo / LO Media

9.10. Analysere, ensrette og kvalitetssikre metadata

(Jfr. Delmål 2 fra prosjektbeskrivelsen)

I mange tilfeller kan bilder inneholde flere metadatafelter med identisk innhold. Dette kan for eksempel være datofelter (Created Date, Original Date), felter med fotografnavn (Artist, Creator, By-Line), eller felter med innholdsbeskrivelse (Description, Caption-Abstract, Content-Description). Første trinn er å fjerne slike «interne duplikatfelter». Utfordringen ved dette er at man helst bør gjøre det på en slik måte at sluttresultatet blir et så ensrettet metadatafeltoppsett som mulig på tvers av alle filene. Dette avdekker behovet for en metadatamal som man kan benytte for slik ensretting.

Vi kunne kanskje gjort en del antagelser om fellestrekk og innhold i metadatafeltene, men valgte å forholde oss veldig agnostisk til det, sett i forhold til at vi senere kan motta arkiver som kanskje har veldig avvikende metadataoppsett. Vi valgte heller å jobbe mot et verktøy som kunne la fotoarkivarene ta et ukjent metadataoppsett og bygge et regelsett for ensretting av det spesifikke fotoarkivets metadataoppsett inn mot en felles metadatamal. En relevant observasjon i så måte er at det blant metadatafeltene i dette fotoarkivet finnes nesten 100 «User Defined nn»-metadatafelter, noe som vanskeliggjør en kvalifisering av innholdet, hvis man ikke på forhånd har mottatt en oversikt fra arkivskaper over hva disse feltene inneholder.

I fotoarkivet i dette prosjektet var det totalt litt i underkant av 3500 unike metadatafelter i bruk. Svært mange av disse var metadatafelter som i bildeinnholdsmessig sammenheng ikke nødvendigvis var særlig relevante, som for eksempel informasjon om lukkertid, blenderåpning, kameratype, og andre tekniske data. Vi søkte derfor å finne en måte å analysere metadatafelter på slik at vi kunne få ut en liste over «prioriterte» metadatafelter for videre behandling.

Dette gjorde vi ved å sette opp et regelsett for programmatisk analyse av innhold, og så tildele hvert metadatafelt en prioriteringsverdi ut fra dette regelsettet, som eksempelvis kan se slik ut:

- Har feltet innhold? Hvis ikke får det verdi 0.
- Har feltet innhold, men det er likt på over 95% av bildene? Det gis verdi 1.
- Har feltet innhold, men det er bare tallverdier? Det gis verdi 2.
- Har feltet innhold, men det er likt på over 80% av bildene? Det gis verdi 3.
- Har feltet tallinnhold, og det er et antall unike verdier som er over 10% av totalt antall bilder? Det gis verdi 4.
- Har feltet tekstinnhold, og det er et antall unike verdier som er over 10% av totalt antall bilder? Det gis verdi 5.

Etter en generering av en slik prioriteringsverdi for alle metadatafeltene, så vi fort at antallet felter som ble trukket frem med de høyeste prioriteringsverdiene var ganske få:

Prioriteringsverdi	**Antall metadatafelter**
5	30
4	3
2	12
1	34
0	237
-1	3171

(-1 betyr at ingen av kriteriene for å tildele en prioriteringsverdi matchet.)

Her er listen over feltnavnene som scoret høyest:

ID	**Tagname**
429	Caption-Abstract
686	Create Date
688	Created Time
822	Date/Time Created
824	Date/Time Original
844	Derived From Document ID
846	Derived From Instance ID
847	Derived From Original Document I
850	Description
911	Document History
912	Document ID
977	Exif
979	Exif Camera Info
1113	File Name
1476	History
1479	History Instance ID
1483	History When
1564	Instance ID
1909	Metadata Date
1960	Modify Date
2020	Native Digest
2065	Object Name
2111	Original Document ID
2562	Raw File Name
2897	Slices Group Name

3079	Time Created
3090	Title
3158	Unique Document ID
3159	Unique Id
3460	XMP File Stamps

Vi ser her at datofelter, felter for unike ID-verdier, og tekstfelter blir trukket frem.

En annen problemstilling vi søkte å adressere er bruken av fritekstfelter der det hadde vært ønskelig med oppslagsfelter, for eksempel for fotografnavn. Dette valgte vi å løse ved å trekke ut en liste over unike feltverdier som fantes i det mottatte materialet, redigere listen slik at «aktive» elementer kun hadde korrekt stavemåte, og slik at feilaktige elementer ble koblet mot de aktive og korrekte verdiene. Slik kunne vi kjøre en oppdatering av databasen, og dermed få korrigert dette med en relativt sett liten innsats fra fotoarkivaren.

Dette kan nok automatiseres noe, ved å utvikle en algoritme basert på en kombinasjon av statistikk, ordlister og tekstmatching.

9.11. Eventuell formatkonvertering/normalisering

Som tidligere nevnt er det problemstillinger knyttet til lagring av digitalt materiale over lang tid, og dette gjelder også bildemateriale og bildeformatene dette er lagret i.

I hovedsak var materialet vi mottok fra LO Media i JPEG-format. Vi gjorde en vurdering av hvorvidt det var hensiktsmessig å konvertere dette til et ukomprimert/tapsfritt komprimert format før deponering i digitalt depot, men siden et annet format ikke ville tilføye noen kvalitet, og siden JPEG er et godkjent digitalt filformat i henhold til Riksarkivarens forskrift, valgte vi i dette tilfellet å ikke gjøre noen normalisering av bildene.

I tvilstilfeller anbefaler vi å samrå seg med Riksarkivarens forskrift, og gjøre nødvendige konverteringer i henhold til de godkjente formatene for avlevering av digitalt materiale som spesifiseres der.

Det er også verdt å merke seg at det kan være et skille mellom depotkopi og arbeidskopi – de originale filene som mottas bør deponeres i det mottatte formatet, men kan også gjerne normaliseres til et langtidslagringsformat som legges sammen med originalen i depotpakken dersom det mottatte formatet ikke fyller betingelsene for langtidslagring. Motsatt kan også gjerne filer i andre formater konverteres til JPEG med høy kvalitetsverdi når de skal overføres til arbeidskopi.

9.12. Opprette arkivpakke for deponering i digitalt depot

Se vedlegg A, «Rutine for mottak av digitalt skapte arkiver».

9.13. Opprette arbeidspakke for fotoarkivarer

Dersom det mottatte fotoarkivmaterialet skal behandles av fotoarkivar, vil det kunne være uhensiktsmessig å forholde seg til materialet slik det er deponert i digitalt depot. Vi genererer i de tilfellene også en «arbeidspakke» ut fra det ordnede materialet, og sørger for å lagre denne i et format som er hensiktsmessig å overføre til fotoserver for tilgang for fotoarkivarer. Dette vil for oss kunne medføre å trekke ut et spesifisert sett metadata i henhold til metadatamalen som er benyttet under ordningen, og lagre bildene i JPEG-format med høy kvalitet.

Disse arbeidspakkene vil fortløpende kunne oppdateres av fotoarkivarer, og vi vil tillate uttrekk av oppdateringer fra disse for tilbakeføring til digitalt depot for komplettering av det deponerte materialet. Disse oppdateringene vil lagres som tilleggspakker i depot, i et hensiktsmessig format, og vil ikke oppdatere opprinnelig depotpakke.

10. Bruk av Digitalarkivet

(Jfr. Delmål 3 i prosjektbeskrivelsen)

Arkivverket jobber for å utvikle Digitalarkivet til en felles, foretrukket og komplett løsning for mottak, langtidsoppbevaring og publisering av digitalt arkivmateriale, både for offentlige og private arkivskapere og arkivinstitusjoner.

På nettsidene til Digitalarkivet, https://www.digitalarkivet.no, kan man lese om både historien og intensjonene bak Digitalarkivet, se på utviklingsplanen og hvilke funksjoner som er i hvilket utviklingsstadium, og registrere seg for å ta i bruk de funksjonene som er publisert og i drift pr. i dag.

I skrivende stund er store deler av hovedrammeverket klart; sikker lagring, opplasting, tilgangsstyring og publisering er på plass. Det er også mange spennende funksjoner og moduler under utvikling, som vil øke funksjonaliteten og mulighetene for å utnytte informasjonen som ligger i innlevert arkivmateriale.

Som en del av prosjektet hadde vi et elektronisk møte med Digitalarkivet, der vi fikk en presentasjon av løsningene deres, og der vi kunne stille spørsmål til dem, og svare på spørsmål om prosjektet. Presentasjonen deres hadde spesielt fokus på tilgjengeliggjøringsløsning av bildemateriale, muligheter for oppdatering av metadata i denne løsningen, og mulighetene for styring av tilgang.

Vår vurdering i forhold til bildematerialet til LO Media er at den tjenesten som tilbys for tilgjengeliggjøring foreløpig ikke dekker behovet for en kompleks bildesammensetning. I dag vises kun overordnede metadata om samlingen hentet fra Asta/Arkivportalen. Materialet fra LO Media inneholder som tidligere vist mye og mangeartete metadata som ikke bør skilles fra bildene. Det som kan være aktuelt, dersom arkivskaper samtykker i dette, er å gjøre utvalg fra arkivmaterialet i mindre enheter som har en logisk samhørighet, og publisere disse separat. Da vil en del av metadataene som ikke pr. i dag vises pr. bilde i løsningen komme frem som en del av den overordnede strukturen og grupperingen.

Det vil være interessant å følge med på den videre utviklingen av Digitalarkivets løsning for håndtering av fotoarkiver, da spesielt med tanke på større og sammensatte arkiver, med komplekst metadatainnhold.

Når det gjelder langtidsdeponering av digitalt bildemateriale vil Digitalarkivet absolutt være den foretrukne løsningen, da man kan være trygg på at den bygger på de beste standarder og er basert på den beste kompetansen som kan fremskaffes på området. For arkivet i dette prosjektet vil vi utrede dette videre, men har foreløpig lagret arkivet i vårt eget digitale depot.

11. Sammenligning med tilsvarende funksjoner i FotoStation

Vi fikk dessverre noe begrenset tid til å kjøre gode sammenligningstester mellom egenutviklet plattform og funksjonaliteten i Fotowares programvare, men vi fikk sett på noen av de grunnleggende funksjonene.

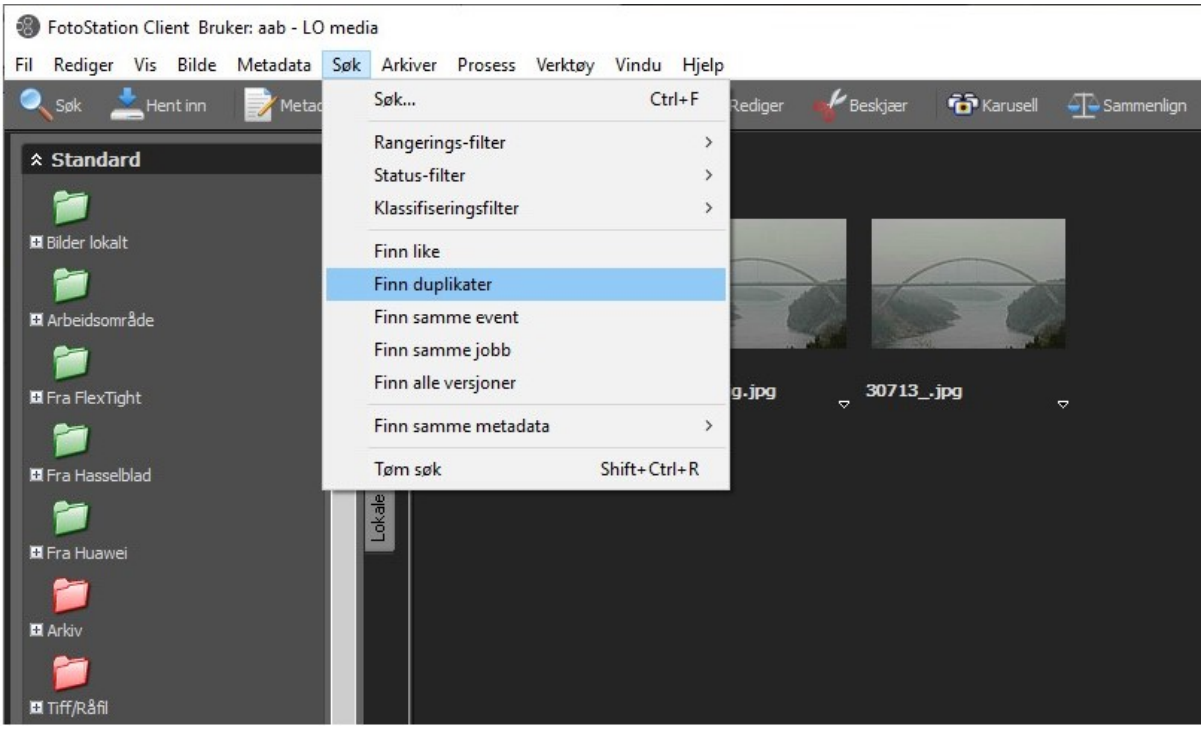
Som tidligere nevnt har Fotoware-løsningen/FotoStation en metadatamal som benyttes for å ensrette og redigere metadata tilknyttet bildene. Denne kan modifiseres, og man kan velge mellom ulike oppsett/ulike metadatamaler. I modellen vi bruker er det i noen tilfeller satt opp en samkjøring av felter som har sammenfallende feltinnhold, jamfør punkt 5, side 10. Vi har imidlertid ikke fått testet hvordan dette fungerer i tilfeller der de respektive samkjørte feltene inneholder forskjellig informasjon, eller nesten lik informasjon.

Det kan bygges opp et skjema i brukergrensesnittet for redigering av metadata, der man kan gjøre et utvalg av hvilke metadatafelter som skal være tilgjengelige for redigering. Dette er fleksibelt, men er også begrenset til feltene i metadatamalen FotoStation benytter.

Fotoware-løsningen benytter en programmodul som heter «Index Manager» til å analysere og indeksere bilder som legges inn. Denne kjører som en bakgrunnsprosess som vil stå og prosessere et mottatt arkiv inntil alle bildene er indeksert. Etter dette vil man ha muligheten til å benytte FotoStations programfunksjoner til å for eksempel søke i metadata eller identifisere dubletter.

Jo større arkivet er, jo lenger tid tar denne indekseringen, og for store arkiver på flere TB kan dette ta flere døgn. Dette så vi imidlertid også i vår egen løsning, der operasjoner som å kopiere filer, beregne sjekksummer og generere «fingeravtrykk» kunne ta lang tid.

Her er et skjermbilde fra FotoStation som viser noen av de valgene man har:



Ved å kjøre «Finn duplikater» på ett av bildene fra eksempelet i punkt 9.9. fikk vi ut dette resultatet:

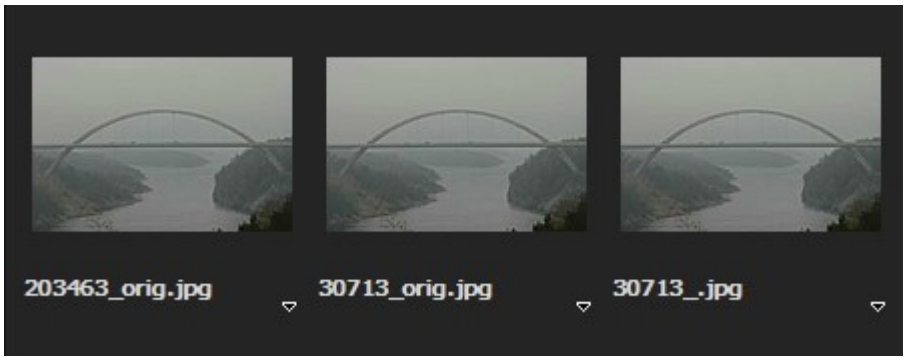


Foto: Kristian Brustad / LO Media

Dette viser bilder som er identiske i bildeinnhold.

Ved å kjøre «Finn alle versjoner» på samme bilde ble resultatet dette:

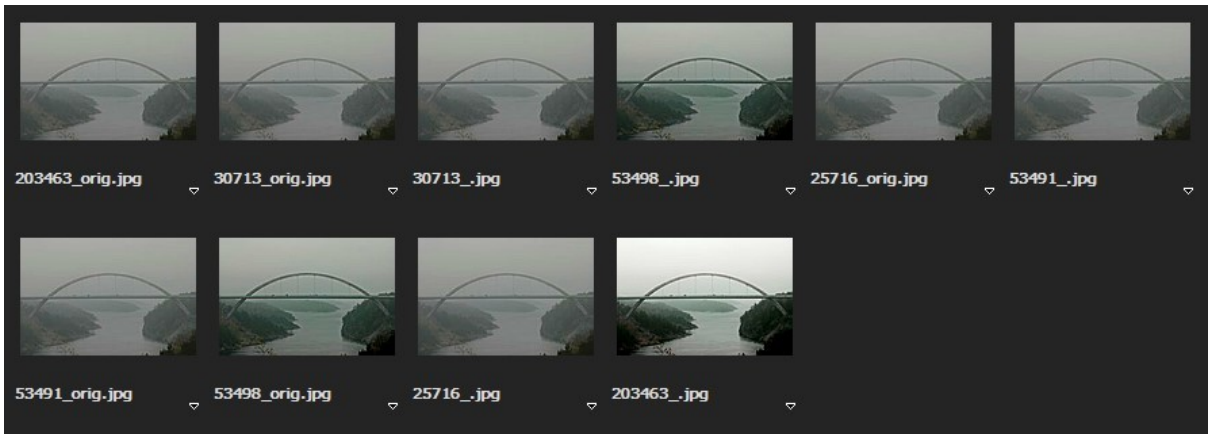


Foto: Kristian Brustad / LO Media

Her ser vi at FotoStation har klart å identifisere bilder med samme motiv, men der det er gjort justeringer på bildet.

Vi kjørte også et søk via funksjonen «Finn like» - men den resulterte i over 2000 treff, og ga et nokså stort utvalg av helt forskjellige bilder/motiver. Det er mulig at denne funksjonen kan tilpasses slik at man kan velge grenseverdi for når to bilder skal anses som «like», det rakk vi ikke å undersøke.

FotoStation har også kraftige søkemuligheter for søk i metadata, blant annet kunne vi søke på en kombinasjon av flere delstrenger med jokertegn og logiske operatorer, slik at vi for eksempel kunne få frem treff som inneholdt varierende skrivemåter av navnet «Sissel M. Rasmussen», jfr. punkt 5. Vi ser likevel at en del mer avanserte operasjoner, som for eksempel å generere en prioriteringsliste for videre analyse av metadata, basert på et regelsett for innholdet i de respektive metadatafeltene var utenfor mulighetene i programvaren.

FotoStation har et anerkjent og gjennomført brukergrensesnitt, som er ryddig og oversiktlig i bruk. Det gir fotoarkivarene et veldig godt verktøy for behandling av metadata i både små og store bildesamlinger. Søkefunksjonene er raske i bruk, og resultatene vises grafisk på skjerm på en måte som gjør det veldig enkelt å arbeide videre med dem. Det er også lagt til rette for batchregistrering/oppdatering av metadata, som kan være veldig effektivt.

12. Erfaringer og betraktninger

Prosjektet har gitt oss en del verktøy og erfaringer i håndtering av større digitalt skapte fotoarkiver. Vi har utarbeidet en metode som hjelper oss å redusere overflødig datamengde, og samkjøre de metadataene som finnes i bildematerialet slik at det er enkelt og mest mulig oversiktlig å forholde seg til det.

Vi har også erfart at den største utfordringen ligger i mangelfulle og ustrukturerte metadata, slik at det bare i begrenset grad er mulig å hente ut kontekst og informasjon om bildene. Her er det et videreutviklingspotensiale ved å se på forskjellige metoder for å øke metadatamengden, eksempelvis crowdsourcing der bildematerialet kan gjøres offentlig tilgjengelig, bruk av kunstig intelligens-systemer for å automatisk identifisere objekter, steder og personer, og utvikling av støttesystemer for fotoarkivarene der man ser på sammenhenger mellom de metadataene man har (for eksempel fotograf, tidspunkt, GPS-koordinater, stikkord, bildelikhet), og eventuelt koblinger mot eksterne datakilder, slik at fotoarkivarene gis enkle og effektive verktøy for å batch-registrere (massetilføye) metadata.

Ved å gi fotoarkivarene muligheten til å foreta fritekstsøk i alle metadatafelter maksimerer vi muligheten for å identifisere bilder ut fra de metadataene som finnes. Det kan også vurderes å utvikle en løsning for assistert fritekstsøk, der søkeord eller -begreper foreslås ut fra analyse av eksisterende metadata.

Link til websider for prosjektet: https://www.arbark.no/prosjekter/bildeprosjekt-2020-2021/fotoarkiv-prosjekt.htm

Vedlegg A: Rutinebeskrivelse for mottak av digitalt skapte fotoarkiver

(Jfr. Delmål 4 i prosjektbeskrivelsen)

Rutiner ved mottak av fotoarkiv hos Arbeiderbevegelsens arkiv og bibliotek.

Versjon 0.2, 08.11.2021.

Dette dokumentet beskriver rutiner som skal følges og dokumentasjon som skal opprettes ved mottak av fotoarkiv. Det baserer seg på standarden OAIS (Open Archival Information System)

Terminologi og begreper som brukes i denne rutinebeskrivelsen:

Mottaksområde: Et fil/mappeområde der fotoarkivet lagres ved mottak

Karantene: Et avlåst fil/mappeområde der fotoarkivet oppbevares i karanteneperioden

Ordningsområde: Et fil/mappeområde der fotoarkivet lagres under behandling/ordning

Overordnet prosess

Et mottatt digitalt arkiv skal gjennom et sett operasjoner før det kan deponeres:

- Inngåelse og signering av avtale med arkivgiver
- Dokumentasjon og registrering av mottak
- Forberedelse av plattform for bildeanalyse
- Analyse og rapportering
- Karantene i 3 uker
- Overføring til arbeidsområde
- Behandling og ordning
- Ny sjekksumkalkulering
- Klargjøring for pakking til arkivpakke
- Eventuell normalisering av filer til langtidslagringsformater
- Nedpakking til arkivpakke, sammen med produserte arkiv-metadata
- Deponering

For et fotoarkiv gjelder i tillegg følgende:

- En arbeidskopi av bildematerialet skal overføres til Arbarks system for fotobehandling, der dette er aktuelt. Dette er pr. november 2021 en løsning fra FotoWare. Rutinene er allikevel ikke avhengig av at man har produkter fra FotoWare, og skisserer noen generelle operasjoner basert på rutiner for mottak av el-arkiv ved Arbark.
- Rutine for samkjøring av depotoriginal og oppdatert arbeidskopi må være kjent av alle som arbeider med materialet fra arkivet

Detaljert beskrivelse av trinnene i prosessen

1. Forberedelse til mottak av digitalt fotoarkiv

Ved mottak av et digitalt fotoarkiv, må en avleveringsavtale signeres, og rettigheter ved videre bruk avklares med fotografen/e. Herunder kommer for eksempel videresalg og bruk i formidling på nett og i sosiale medier.

Mottak av materiale kan skje på flere måter: Enten ved direkte avlevering i form av et eller flere fysiske medier (CD, DVD, minnepinne, ekstern disk), eller ved elektronisk avlevering over internett (for eksempel FTP-overføring). Ved direkte avlevering er det viktig at arkivet så raskt som mulig enten blir duplisert over på et ekstra medium, eller aller helst blir kopiert over på en mer bestandig lagringsløsning. Hvis dette er en nettverksdisk er det viktig å avklare behovet med IT-personell, slik at nødvendige ressurser kan klargjøres spesifikt for dette.

Det er viktig å innhente informasjon om arkivet i forkant, og vi anbefaler å skaffe til veie og dokumentere disse detaljene:

- Eier av materialet
- Formål med avlevering
- Bruksramme for arkivinstitusjonen
- Arkivets størrelse i GB og antall bilder
- Filformater som finnes i arkivet
- En beskrivelse av innholdet, eventuell arkivnøkkel som arkivet er ordnet etter
- Eventuelle forberedelser/forhåndsbehandling av materialet arkivskaper kan gjøre før avlevering, for eksempel luke ut duplikater/thumbnails, gjennomgå metadata, evt. lage et separat metadatadokument eller regneark som vi kan bruke som støtte ved oppdatering av materialet i etterkant

2. Klargjøring av system for mottak og automatisk bearbeiding av fotoarkivet

For å kunne gjøre automatisk bearbeiding av bildematerialet må det klargjøres en plattform for dette. Det er en IT-oppgave.

Når plattform er klar, og arkivar har fått tildelt adresse, brukernavn og passord, overføres fotoarkivet til mottaksområdet

På mottaksområdet opprettes det en mappestruktur som tilsvarer nedenstående:

```
e-arkiv/
        arkivnavn/
                metadata/
                        arkivbeskrivelse
                        depotlogg
                        sjekksumrapport
                        rapport for fil-identifisering og virusskanning
                arkivmapper/
                        ARK-xxx_Uxxx/
```

Arkivnavn tilsvarer navnet som følger aksesjonsføringen i ASTA, som alltid gjøres ved mottak – for eksempel ARK-3188 LO Media. Arkivmappene blir da tilsvarende ARK-3188_Ud_L001 etc.

3. Dokumentasjon og registrering av mottak

Arkivbeskrivelse opprettes som et Word-dokument hvor hele arkivet beskrives, gjerne på mappenivå. Her kommer all informasjon om innholdet på de ulike lagringsmediene, for eksempel om det er snakk om CD/DVD/FTP og eventuell informasjon som er tilknyttet disse. Antall filer, antall MB og hvordan filene er navngitt oppgis også, sammen med eventuell informasjon om opphavsrett. Filer som eventuelt ikke overføres kan også beskrives her.

Depotlogg opprettes som et Word-dokument hvor dateringer for utførte operasjoner registreres. Disse operasjonene skal med:

1) Mottak og opprettelse av mapper
2) Opprettelse av arkivbeskrivelse
3) Opprettelse av sjekksummer
4) Utført virusskanning
5) Utført fil-identifisering
6) Utluking av duplikater
7) Karantene

Maler for disse dokumentene finnes på fellesområdet under \maler\arkivgruppa\mottaksdokumenter.

4. Analyse og rapportering

Når arkivet er kopiert inn på det angitte mottaksområdet kan første trinn av behandlingen starte. Arkivet går gjennom prosesser for fil-identifisering, virussjekk og sjekksumberegning, og det genereres en rapport som enten kan lastes ned fra web-grensesnittet til systemet, eller som kan hentes ut fra mottaksområdet til systemet.
Denne rapporten inneholder et sammendrag over filtyper som er funnet i arkivet, status fra virussjekk, og informasjon om når arkivet kan behandles videre (etter karantene). Den plasseres i metadata-mappen og operasjonene registreres i depotloggen.

Lage sjekksummer for filene:
Åpne programmet Checksummer. Velg sha256-algoritme og trykk «Kalkuler sjekksummer». Ved «Velg startmappe for beregning av sjekksummer», velg mappen der filene ligger (altså arkivets øverste nivå). I filnavn-feltet, skriv «sjekksummer», underscore pluss dato. Under filtype, velg «All files» og trykk «Lagre». Den plasseres i metadata-mappen og operasjonen registreres i depotloggen.

5. Karantene

Etter at arkivet er virussjekket og forhåndsanalysert, overføres filene til karantene-området. Dette er et internt område for prosesseringssystemet, der filene skal oppbevares utilgjengelig i 3 uker.

Formålet med dette er å gi antivirusprodusenter tid til å oppdatere sine løsninger, slik at en eventuell ukjent sårbarhet kan bli avdekket og løst før filene aksesseres av arkivar.

Etter karanteneperioden kjøres det på nytt automatisk virussjekk av arkivet, og det flyttes ut fra karantene og til ordningsområdet. Operasjonene registreres i depotloggen.

Ved overføring til ordningsområdet vil arkivet automatisk deles opp i undermapper, avhengig av størrelsen på arkivet. I hver undermappe vil det ligge maks. 10000 filer, og antall undermapper er da avhengig av antall bilder i arkivet. Dette gjør det mindre intuitivt å aksessere arkivet, men gjør det enklere for systemet både å håndtere filmengden, og i større grad utføre en del operasjoner i parallell.

Under denne reorganiseringen vil opprinnelig mappenavn legges inn som et metadatafelt i den enkelte bildefilen, slik at den opprinnelige strukturen bevares.

6. Behandling og ordning

Når bildefilene er overført til ordningsområdet skjer følgende operasjoner:

- Uttrekk av metadata til database, samkjøring av metadataene fra bildefilene inn i den felles metadatamalen vi bruker
    - Avvik som ikke dekkes av malen loggføres, og arkivar gis mulighet til å manuelt velge hva som skal skje med disse.
- Generering av sjekksummer for det rene bildeinnholdet i hver fil
- Kontroll av arkiv og flagging av korrupte filer
- Automatisk analyse av bildene, og flagging av duplikater
- Generering av «fingerprints» for forskjellige grenseverdier, som senere benyttes til å analysere resterende bilder for å finne ikke-identiske duplikater
- Generering av rapport fra disse trinnene av prosesseringen, som legges inn under «metadata», og som kan lastes ned av fotoarkivaren.

Alle endringene som gjøres i arkivet under dette punktet registreres i depotloggen.

Etter at dette trinnet er fullført gis arkivaren tilgang til materialet via FotoStation slik at manuell oppdatering og ordning kan gjennomføres.

Arkivbegrensing og kassasjon følger arkivinstitusjonens retningslinjer for dette. Arkivbegrensing innebærer å fjerne dokumenter som egentlig ikke hører hjemme i arkivet, altså dokumenter som ikke har verdi som dokumentasjon. Dette kan gjøres automatisk eller manuelt. Kassasjon registreres i depotloggen.

7. **Normalisering**

   Etter at eventuell manuell ordning er fullført behandles alle resterende bildefiler med tanke på konvertering til langtidslagringsformater, i henhold til et definert regelsett. Originalfilene blir beholdt, og eventuelle normaliserte kopier legges i samme mappe som originalfil, men med annen filendelse. Der det er mulig kopieres alle bildemetadatafelt fra originalfilen til langtidslagringskopien.

8. **Ny sjekksumberegning**

   Etter dette beregnes nye sjekksummer av filene i arkivet, og av arkivet under ett. Rapport fra dette legges inn under metadata og operasjonen registreres i depotloggen.

9. **Nedpakking for deponering**

   Når alle metadata er oppdatert og nye sjekksummer er generert, pakkes hele ordningsområdet ned til en 7z-fil. Denne filen kan da overføres til digitalt depot.

Vedlegg B: Beskrivelse av teknisk plattform for testing og utvikling

For dette prosjektet ble det definert opp en virtuell server på vår testplattform, med følgende spesifikasjoner:

Maskinvare for testplattform:

- Ryzen 9 3950X CPU
- 128 GB RAM
- 18 TB diskplass i RAID10
- VMWare ESXi 7.0

Testserver:

- 8 stk. VCPU
- 8 GB RAM
- 6 TB diskplass i en partisjon
- Debian Linux 10
- MariaDB 10.3.27 SQL-server
- PHP 7.3
- Apache 2.4

(Denne har blitt løpende oppdatert underveis i prosjektet.)

Disse verktøyene ble også brukt ved analysering og behandling av datamengden:

- findimagedupes (https://gitlab.com/opennota/findimagedupes) – for å finne visuelt like bilder
- sha256sum – for å kalkulere SHA256-baserte sjekksummer av bildefilene

Vi valgte å bygge opp metadatadatabasen på en måte som sikret størst grad av fleksibilitet.
Vi bygde i hovedsak opp to tabeller:

- En tabell over alle bildefiler, med felter for filnavn, sjekksum, fingerprint, status og original_ID.
- En tabell over alle metadatafelter, med felter for fil_ID, metadatafeltnavn og metadatafeltinnhold

Det ble opprettet flere indekser på begge disse tabellene.

Fordelen med å gjøre det på denne måten er at vi i praksis kunne prosessere alle bildefilene i sekvens og legge alle metadata rett inn i tabellen uten å tenke på hvilke metadatatags som fantes i filene og måtte tilpasse tabelloppsettet i forhold til det. Det ble også mye enklere å kjøre enkelte spørringer der man ville ha ut metadatafeltnavn som ett av resultatene fra spørringen.

Ulempen er at enkelte spørringer ble noe mer kompliserte å utforme og tok mye lenger tid å kjøre i forhold til om tabellen for metadata var utformet med et fast feltoppsett som korresponderte til de metadatafeltene som eksisterte i bildefilene.

Det var i underkant av 3500 unike metadatafeltnavn i fotoarkivet fra LO Media, samlet litt over 120 millioner metadatafelter i hele arkivet, og derav ca. 119 millioner som inneholder verdier.