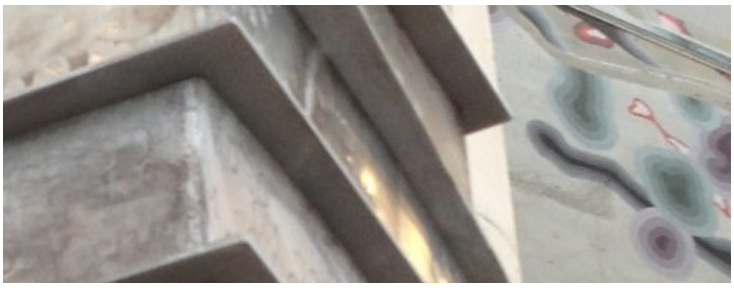
ARKIVVERKET
RIKSARKIVET

Samdok

samla samfunnsdokumentasjon

RAPPORT 2016



PRIORITERT OPPGAVE

7 Forprosjekt: Nasjonal publiseringsplattform for skanna arkivdokumenter

Delrapport 1b) teknisk beskrivelse og løsningsforslag

Utarbeidet av

Espen Tønnesen

Rapportdato

27.10.2016

ARKIVVERKET
RIKSARKIVET

SAMDOK delprosjekt:	Prioritert oppgave:	SAK (ePhorte):	Dato:	Side:
Kommunale arkiv	**7 Forprosjekt: Nasjonal publiseringsplattform for skanna arkivdokumenter**		**03.11.2016**	**1 av 9**
Forfatter: **Espen Tønnesen**	Tittel: **Delrapport 2016 – Teknisk beskrivelse og løsningsforslag**			

OPPGAVE	*7 Forprosjekt: Nasjonal publiseringsplattform for skanna arkivdokumenter*
Ansvarlig delprosjekt	*Kommunale arkiv*
Arbeidsgruppens leder	*Anette Skogseth Clausen (Arkivverket)*
Arbeidsgruppens medlemmer	*Espen Tønnesen (Arkivverket), Kristian Hunskaar (Arkivverket), Ingrid Jørgensen (Arbak), Aasta Karlsen (Arkiv i Nordland), Espen Sæterbø (Fylkesarkivet i Sogn og Fjordane), Øystein Eike (Oslo byarkiv), Ottar André Anderson (Ika Møre og Romsdal), Snorre Dag Øverbø (Aust-Agder Museum og Arkiv), Tove Wefald Pedersen (Norsk folkemuseum)*
Målformulering	*Målet for forprosjektet er en beskrivelse av hvordan nasjonal publiseringsplattform kan realiseres faglig, teknisk, driftsmessig, økonomisk og organisatorisk – både i en første versjon og på lengre sikt. Skal omfatte «back end» (for innholdsleverandører) og «front end» (for brukere).*
Sammendrag	*Rapporten omfatter den tekniske beskrivelsen og løsningsforslaget for en nasjonal publiseringsplattform. Ettersom arbeidsgruppen anbefaler å videreutvikle Digitalarkivet til å bli løsningen for publiseringsplattformen, har beskrivelsene tatt utgangspunkt i de eksisterende planer for videreutvikling av Digitalarkivet. Løsningen vil være basert på åpen kildekode og være utformet på en slik måte at det er lett å legge til ny funksjonalitet og forbedre eksisterende.* *De skannede arkivdokumentene skal publiseres i «media» i Digitalarkivet og framfinning til materialet, samt profilering for depotinstitusjonene blir gjennom «digark». Her vil det særlig vektlegges å gjøre Digitalarkivet optimalisert for at depotinstitusjoner skal kunne presentere materialet selv/profilere innholdet sitt gjennom API-er.* *For metadataarbeidet er det tatt utgangspunkt i at arkivene er registrert i ASTA/Arkivportalen og metadata hentes derfra. Det er likevel mulig å legge til metadata manuelt hvis det ikke finnes data i ASTA/Arkivportalen. For å lette framfinning i materialet for brukerne vil det også utarbeides et felles sett emneknagger som kan benyttes for å klassifisere materialet på tvers av arkivskapere.* *Administrasjonsverktøyet for Digitalarkivet skal utbedres slik at depotinstitusjoner kan følge arbeidsflyten fra skanning til ferdig publisert bilde i Digitalarkivet. Inntil administrasjonsverktøyet er ferdig utviklet, vil eksisterende administrasjonsløsning, «metaop», åpnes for depotinstitusjoner.* *Alt av bilder lagres i SAN og håndteres i et eksisterende system, «filbanken». Hver fil får et UUID til filnavn og alle filer vil dermed ha unike filnavn og være beskyttet. Denne UUID vil også bli brukt som permanent bilde-id på nettsidene og gi sluttbrukerne en trygg ID å ta vare på for referanse.* *Bildebehandling vil bli håndtert gjennom «prosesseringsriggen» hvor bilder konverteres til publiseringsversjon, lagringsversjon, her kan depotinstitusjoner benytte seg av ytterligere funksjonalitet for OCR-lesing og sladding. Dette reduserer behovet for programvare og kompetanse til bildebehandling hos depotinstitusjonene.* *Tilrettelegging av bildene før publisering, kalt indeksering, vil bli løst gjennom en egen modul for indeksering som er under utvikling. Depotinstitusjonene kan velge å publisere bildene uten indeksering, indeksere selv eller de kan legge det ut til sluttbrukere for at brukerne selv skal indeksere bildene. Indekseringsmodulen vil også håndtere klausulering av bildene.* *Det legges opp til gode verktøy for å hente ut statistikk for hver depotinstitusjon.* *Det eksisterer et godt verktøy for å kontrollere hvem som kan se bildene og dette vil også depotinstitusjonene benytte seg av i Digitalarkivet. Det kan tildeles adgang til bilder for en enkelt person og/eller grupper.*

SAMDOK delprosjekt:	Prioritert oppgave:	SAK (ePhorte):	Dato:	Side:
Kommunale arkiv	**7 Forprosjekt: Nasjonal publiseringsplattform for skanna arkivdokumenter**		**03.11.2016**	**2 av 9**
Forfatter: **Espen Tønnesen**	Tittel: **Delrapport 2016 – Teknisk beskrivelse og løsningsforslag**			

Oppfølging 2016	– *Utviklingsarbeidet for Digitalarkivet vil fortsette i regi av Arkivverket, men for å sikre involvering fra arkivsektoren, bør det etableres et hovedprosjekt i 2017 som følger utviklingen og tar funksjonaliteten i bruk så tidlig som mulig.* – *Ytterligere delrapporter skal leveres før forprosjektet kan avsluttes i 2016. Her vil bl.a. prosjektorganisasjonen bli avklart.*
Vedlegg	*1 delrapport*

Vedlegg:

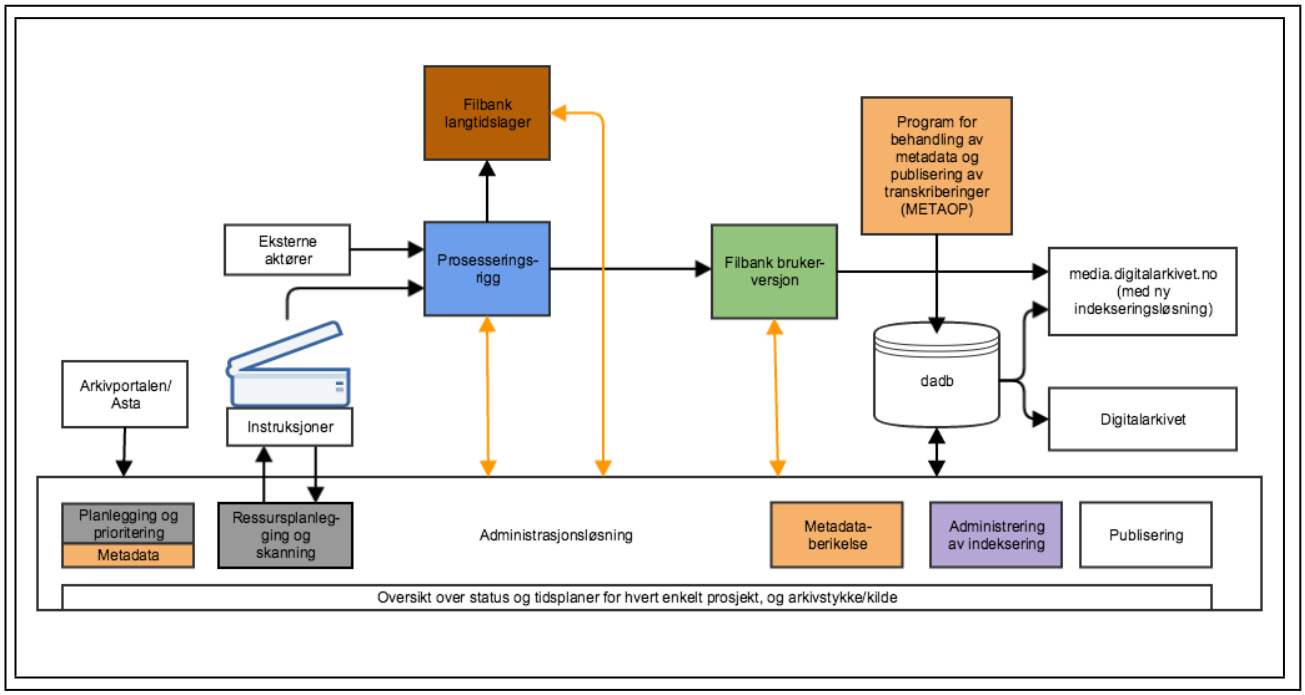
1b) Teknisk beskrivelse og løsningsforslag

Basert på prosjektgruppas anbefaling om å bruke Digitalarkivet som plattform for "Nasjonal publiseringsplattform for skanna arkiver" (NAPSA) beskrives Digitalarkivets løsninger for søk og visning og systemene bak, samt administrasjonsløsningene som er under utarbeidelse.

Arkivverket har lenge ønsket å tilby Digitalarkivet som en plattform for publisering av arkivmateriale. Ikke bare skannet materiale, men også transkribert materiale. NAPSA endrer dermed ikke de ambisjonene Arkivverket har hatt for utviklingen av tjenesten. Arkivverkets egne behov er såpass omfattende at de på de fleste områder allerede sammenfaller med de behovene som gjelder for NAPSA, men også utover behovene i NAPSA. Med NAPSA vil Arkivverket imidlertid ha et nettverk for samarbeid for å gjøre løsningene så bra som mulig også i forhold til de behovene som finnes hos andre arkivinstitusjoner. Dette inkluderer også støtte for lyd, video og foto i tillegg til skanna dokumenter.

Arkivverket er i ferd med en omfattende utskiftning av serverparken som brukes i Digitalarkivet. Denne serverparken er dimensjonert for å kunne håndtere den trafikkveksten som NAPSA måtte medføre.

Løsningsforslag



Figur 1: Oversikt over de forskjellige komponentene i løsningen

Brukertjenester

15 september 2016 stengte den siste søkeinngangen i de gamle løsningene for skanna materiale. Alt skanna materiale er nå tilgjengelig i den nye løsningen, samlet på ett sted. Denne løsningen heter inntil videre "media". Denne nye løsningen benytter samme database som ligger til grunn for Digitalarkivet, altså transkribert matriale. Denne kalles for "digark". Skanna materiale som publiseres i media vil i all hovedsak være tilgjengelig via søkeinngangen for "Bla i skanna materiale" i media og i Finn kilde i digark. I den grad en arkivinstitusjon har materiale som passer inn under noen av de andre søkeinngangene vil det være naturlig å legge materialet der. Alle søkeinnganger har søkefelt for oppbevaringssted. Det gjør at det enkelt kan søkes på materiale tilhørende en arkivinstitusjon.

Fremover vil vi også videreutvikle og løfte digark som en mer attraktiv tjeneste. Arkivverket vil jobbe med et helhetlig design og uttrykk for Digitalarkivets tjenester slik at Digitalarkivet vil fremstå som en frittstående tjeneste. Dette innebærer også en egen logo, enda bedre støtte for universell utforming, responsive sider osv. Vi ønsker å jobbe for at både transkribert og skanna materiale ligger i samme løsningen, med fokus på materialet i seg selv og brukeren. Dette skal inkludere fritekstsøk med automatiske søkeforslag osv. Det skal utvikles API mot alle søkeinngangene slik at dataene kan hentes ut og integreres med andre løsninger. Ved hjelp av API og egne landingssider for hver arkivinstitusjon vil det være mulig å promotere eget materiale, men det vil også bli muligheter for å lage egne spesialsider for materiale som krever det.

Metadata

En kilde i Digitalarkivet kan sidestilles med et stykke i Asta. For å publisere en kilde må kataloginformasjon importeres som et minimum. Dersom Asta ikke benyttes, eller arkivet ikke er beskrevet i Asta skal det være mulig å registrere arkivet manuelt. Kun kataloginformasjon vil gi redusert søkbarhet. Derfor bør kilden berikes med metadata som emneknagger, geografisk informasjon og annen kategorisering. Dette krever at det finnes effektive verktøy for å legge inn og administrere materialet. Denne berikelsen kan gjøres allerede før skanning og hele veien til etter publisering.

Det må være fokus på å samordne metadataene og utarbeide felles krav til metadataregistrering, f.eks KAISA, men også generelle standarder som f.eks dublin core eller FOAF. Emneknagger vil bli svært sentralt i løsningen fremover. Det er emneknaggene som vises i ordskyen på forsida av media i dag. Søk i søkeinngangen for skanna arkiver er avhengig av emneknaggene for å skille forskjellig materiale fra hverandre. Kategorisystemet som finne si Digitalarkivet i dag vil bli mer og mer underordnet, og bruken av emneknaggene må derfor være konsekvent for likt materiale på tvers av arkivinstitusjonene. Her må det gjøres en jobb for å se på hva som finnes av definerte standarder for emneknagger og hva som er egnet for bruk i Digitalarkivet.

Administrasjonsverktøy

Arkivverket og Digitalarkivet har behov for bedre verktøy for administrasjon av det som publiseres i Digitalarkivet. Arkivverket har behov for gode verktøy for administrering av skanningsprosessen, registrering av metadata, etterkontroll og korrigering av de skannede bildene og indeksering. Det trengs også større grad av automatikk i prosessene. Eksisterende programvare er under utskiftning, og Arkivverket har dermed mulighet til å ta hensyn til eksterne behov i denne prosessen.

Arbeidsflyten blir analog til skissen i figur 1. Man starter med å definere "prosjekter" for skanning og publisering. For skanning vil dette i praksis være et resultat av de prioriteringene som blir lagt for skanning. Fra administrasjonsverktøyet hentes metadata fra Asta og informasjon om plassering i magasinene, men også om størrelse og omfang, slik at man kan sette opp en tidsplan for arbeidet. Hvert prosjekt kan dermed få tildelt tid på de tilgjengelige skannerne i forhold til skannerens egenskaper i forhold til materialet som skal skannes.

Operatørene kan ved hjelp av informasjon på iPad, mobil eller PC hente materialet i magasinene og oppdatere status for tilrettelegging av materialet for skanning. Ansvarlig vil dermed hele tiden kunne følge med på hvordan arbeidet ligger an i forhold til ressursplanene, og eventuelt gjøre justeringer for å oppnå optimal utnyttelsesgrad.

Det finnes allerede en løsning for administrasjon av metadata (METAOP). Denne er ikke egnet til massebehandling av metadata og vil kun ha tilstrekkelig funksjonalitet så vi kan håndtere publisering av restansene Arkivverket har i en overgangsperioden fram til nytt administrasjonsverktøy er på plass. Inntil ny admnistrasjonsløsning er på plass vil METAOP i en oppstartsperiode gjøres tilgjengelig for et lite antall innholdsleverandører. Det vil gjøre at vi kan tilby publisering av arkivmateriale før administrasjonsløsningen er på plass i løpet av høsten 2017.

Lagring

Alle bilder lagres i et SAN (Storage Area Network). Dette er koblet opp mot et system som organiserer innholdet i san'et. Dette systemet kalles "Filbanken". All kommunikasjon mellom filbanken og digitalarkivet skjer via et API (Application Programming Interface). Filer organiseres i såkalte "Buckets", som er et slags virtuelt diskområde. Det gjør at innholdet i filbanken kan deles opp og samles. F.eks har alle kirkebøker en bucket, alle folketellinger en bucket, tinglysingsmaterialet sin egen osv. Her vil hver arkivinstitusjon få sin egen bucket. Filene lagres på disk i diskvolumer. Disse opprettes automatisk når et diskvolum når en viss grense. Buckets går på tvers av diskvolumene, og filene lagres derfor om hverandre på disk. Filbanken genererer UUID (Universally Unique IDentifier) som blir filnavnet. Eksempel på en UUID kan være "123e4567-e89b-12d3-a456-426655440000". Det betyr at alle filnavn er unike, og dermed er det ingen krav til at filnavnet på bildet som lastes opp må være unikt, men det bør være unikt innenfor en bucket. For hvert bilde som lastes opp lagres det en rekke metadata i filbankens database. De samme dataene legges også i en fil som får samme UUID som bildet. Det betyr at dersom databasen mot all formodning skulle gå tapt, vil databasen kunne gjenoppbygges basert på filene med metadata.

Filbanken inneholder informasjon om antall filer og brukt lagringsplass for hver bucket, som gjør at vi kan hente ut informasjon om bruk f.eks i administrasjonsverktøyet.

Arkivverket kommer til å ta i bruk programvaren som er for Filbanken også til langtidslagring av bilder. Det kan være aktuelt å tilby langtidslagring av skanna bilder som en tjeneste for NAPSA's medlemmer. Her vil også andre data relatert til bildet kunne lagres, f.eks. ocr-fil, transrkiberte data for et bilde, indekeringsinformasjon osv. slik at også denne informasjonen kan lagres videre for framtida.

Filbanken er operativ i dag, og dokumentasjon til denne er under utarbeidelse.

Bildebehandling

Etter skanning behandles bildene i et system for masseprosessering av filer. "Prosesseringsriggen" inneholder et sett med instruksjoner for hva som skal skje med de filene som kommer inn. Det kan f.eks være konvertering til bildeformater og størrelser, lagring i filbank og langtidslagring, legge pekere til bildet inn i Digitalarkivets database, kjøre OCR, sladde bilder osv. Disse oppgavene er konfigurerbare, og lagres i et instruksjonssett eller "prosesseringsprofil". Før skanning utføres bestemmer operatøren via administrasjonsløsningen hvilken prosesseringsprofil som skal benyttes, eller lager en ny. Deretter vil bildene bli prosessert etterhvert som de kommer inn til prosesseringsriggen. Det skal være et eksternt grensesnitt for opplasting av bilder, slike at skanning kan foregå utenfor Arkivverket sine nettverk, og det vil dermed være mulig å laste opp allerede skannede bilder. Det vil dermed ikke være noen strenge krav til bildenes størrelse og format før opplasting. Innholdsleverandører vil dermed ikke ha behov for kompetanse eller infrastruktur for bildekonvertering.

Prosesseringsriggen er beregnet å være klar til innfasing i løpet av april 2017.

Indeksering

Etter skanning kan en kilde publiseres umiddelbart dersom man vet at materialet ikke inneholder informasjon som ikke kan legges ut fritt tilgjengelig. Denne kilden vil i såfall ikke få noen innholdsfortegnelse, og brukeren må bla gjennom kilden for å finne fram. En ny indekseringsløsning er nå under utvikling. Den er integrert i løsningen for søk og visning av skanna materiale. Dette gjør at vi i stor grad kan velge hvordan kilden skal indekseres. Muligheten for indeksering kan legges ut fritt så hvem som helst kan utføre indeksering, eller man definere at brukere som tilhører en bestemt tilgangsgruppe kan utføre indekseringa. Det går an å forhåndsbestemme hvordan indekseringa skal utføres, eller det kan være opp til den som indekserer å velge hvordan det skal indekseres. På forhånd kan det bestemmes hvilke tilganger som skal gjelde for bildene i kilden. I en indekseringsprofil kan det defineres f.eks 100 års klausulering. Dette gir fleksibilitet i forhold til å øke graden av brukermedvirkning, og at man kan konsentrere ressursene om prioriterte oppgaver. Tilgang til bildet vil automatisk settes til 100 år etter et årstall gitt ved indeksering. Hvis det f.eks er 100 års klausulering og bildet som indekseres dateres til 1918, vil bildet automatisk bli tilgjengelig i 2019. Indekseringsprogramvaren vil også ha støtte for å endre rekkefølge og erstatte bilder. Når prosesseringsriggen er på plass vil det også lages funksjonalitet så man kan sladde bilder og generere sladda versjoner automatisk.

Indekseringsprogramvaren er under utvikling, med ferdigstilling i desember 2016.

Statistikk og rapportering

Via administrasjonsprogrammet skal det være mulig å hente ut statistikk. Hvilket materiale blir brukt? Hva kan vi lære av dette i forhold til hvilket materiale som skal prioriteres? Hvordan påvirker nyhetssaker bruken av arkivene?

Det kan være statistikk fra filbanken, men også bruksstatisikk. For hver visning av bilder skal det lagres statistikkdata som kan identifisere hvor mange ganger bilder i en enkelt kilder er blitt vist i løpet av en tidsperiode. Her kan det bli store datamengder, så statistikktallene vil bli aggregert etter en tid. Dette gjør at man kan lage rapporter etterhvert som behovene melder seg. Det vil også være akutelt å koble seg opp mot API'ene til Google Analytics.

Sikkerhet og tilgangsstyring

Digitalarkivet har i dag avansert tilgangsstyring ned på enkeltbilder. Tilgangsstyringen fungerer på to måter. Et bilde kan være blokkert, eller det kan være satt et år for når bildet frigjøres. Bilder som er blokkert er blokkert på ubestemt tid. I begge tilfeller kan man få tilgang ved at innlogget bruker tildeles en brukergruppe som har tilgang til kildene. I tillegg finnes det en dimensjon for sladding av informasjon på bildet. Det kan finnes en sladda og en usladda versjon av et bilde. Dersom man ikke har rettigheter til å se usladda versjon, vil sladda versjon vises. Hvis man har rettigheter og logger inn, vil det usladda bildet vises.

Filbanken ligger i et eget, isolert området i nettverket, som kun tillater at det kommuniseres med api'et. Applikasjonen som skal hente ut eller legge til bilder i filbanken må identifisere seg med en api-nøkkel. En applikasjon kan ha kun leserettigheter, eller både lese og skriverettigheter. Prosesseringsriggen vil kunne ha skriverettigheter, mens Digitalarkivet kun har leserettigheter. Disse rettighetene settes for hver bucket. Dvs. at en applikasjon kan settes til kun å lese innhold fra en gitt bucket men ikke en annen.

Det kreves også funksjonalitet for å styre brukerrettigheter. Arkivverket utvikler nødvendige administrasjonsverktøy for å håndtere dette. Det inkluderer også administrasjon av skanneressurser og metadata, men ikke minst skal hver arkivinstitusjon kunne administrere rettighetene til egne brukere og grupper for å kunne

gi tilgang til egne kilder. Selv om en bruker er administrator, gir ikke det automatisk tilgang til å se klausulert materiale.

Digitalarkivet har en egen brukerstyring som er integrert med brukerne i Arkivverkets forum-løsning. Funksjonaliteten i forbindelse med pålogging skal byttes ut for å ta i bruk OAuth. Det betyr at vi på sikt kan åpne opp for at man kan logge inn på Digitalarkivet med f.eks google- eller facebook-konto, minID osv. Uansett hvilken innloggingsmetode som brukes vil det alltid være mulig å sperre ute brukere og fjerne tilganger der hvor det er nødvendig.

For overføring av bilder vil alt foregå via krypterte forbindelser med innlogging. Alle filer som blir lastet opp vil gå via virussjekk. Disse sikkerhetsforanstaltningene vil også følge de kravene som utarbeides i forbindelse med MAVOD-prosjektet.

Robusthet

Dagens løsning baserer seg på minimum dobbelt sett med servere for de viktigste tjenestene. Dersom en server skulle ryke, trenge oppgradering eller annet vedlikehold vil systemet automatisk sørge for at all trafikk går mot den serveren som er igjen. Før omlegging til nye servere som gir bedre robusthet hadde vi en oppetid på 99,9%.

Permanente bildelenker, sidelenker og bilde-id'er

Fra den første løsningen for skanna kirkebøker ble lansert, ble det lovet permanente bildelenker og permanente sidelenker. Permanente bildelenker er basert på URN-standarden som administreres av Nasjonalbiblioteket i Norge. Permanente sidelenker var spesifikke for de forskjellige løsningene for søk og visning, og var basert på parametere i URL'en. Løsningene med permanente sidelenker var ikke gode og er til dels svært problematiske å omsette til nye lenker i ny løsning. Arkivverket er imidlertid forpliktet til å videreføre disse permanente sidelenkene.

En permanent bildelenke kan se slik ut: http://www.arkivverket.no/URN:NBN:no-a1450-rk20080922650164.jpg. Denne er knyttet til arkivverkets domene, og det er derfor ikke naturlig at dette brukes for andre innholdsleverandører. Lenkene baserer seg på URN-standarden med et system for meningsbærende filnavn (urn-id). Dette systemet har ikke fungert, så det er i praksis ikke mulig å benytte informasjonen i filnavnet maskinelt. Da faller poenget med meningsbærende navn bort. For de kildene Arkivverket allerede har publisert i Digitalarkivet må disse lenkene likevel videreføres under Arkivverkets domene.

For andre innholdsleverandører vil vi fraråde å legge opp til garantier for permanente lenker, for det gjør at man låses i forhold til ny teknologi slik Arkivverket opplever med den nye løsningen. I media-løsningen legges det derfor kun opp til permanent bilde-id. For de bildene som allerede er publisert i Digitalarkivet vil en permanent bilde-id være på denne formen: rk20080922650164. Fremover vil Arkivverket gå over til at den permanente bilde-id'en er UUID fra filbanken. Permanent bilde-id vil da bli slik: 123e4567-e89b-12d3-a456-426655440000.

For å lenke til et bilde i Digitalarkivet skal "Brukslenke for sidevisning" benyttes. Brukslenke for sidevisning blir https://media.digitalarkivet.no/123e4567-e89b-12d3-a456-426655440000. Arkivverket vil ikke garantere at adressen for all fremtid vil ha denne formen, men vi kan garantere at brukeren ved hjelp av bilde-id alltid vil kunne finne fram til rett ressurs. Dette innebærer naturligvis at Digitalarkivet maskinelt vil omsette lenkene til en eventuell ny form.

Teknisk beskrivelse

Programvare

Systemene i Digitalarkivet benytter Open source teknologi. Serverne kjører alle på Linux Ubuntu eller Linux CentOS.

Programmeringsspråket er PHP versjon 7. Alle tjenester benytter et PHP-basert MVC-rammeverk som heter Mako-framework. Databasen til Digitalarkivet er basert på MariaDB som er satt opp med Galera cluster. For filbanken brukes databasemotoren PostgresSQL.

Alle søk skjer via en søkeindex som heter Sphinx. Søkeindeksene oppdateres hver natt.

Automatiske jobber styres av Crontab som er en innebygget mekanisme i Linux eller ved hjelp av et køhåndteringsstytem som heter beantalkd.

Installasjon av applikasjonene, dvs. makoframework og selve programkoden for Digitalarkivet håndteres av et pakkesystem kalt Composer. Composer hånderer avhengigheter og versjoner, og henter inn nødvendig programvare bl.a. fra packagist.org eller direkte fra github som brukes som koderepository. Med en slik metodikk sikrer vi at installasjon og oppgraderinger av programvaren skjer på en trygg måte med mulighet for å rulle tilbake hvis det skulle være nødvendig.

Serveroppsett

Serveroppsettet er satt opp slik at det innes minst to servere for de sentrale komponentene. Disse overvåkes av et system (haproxy) som sørger for at trafikk kun går til servere som er tilgjengelige. Dersom det skulle oppstå en feil på en server, eller serveren tregner oppgradering, vil haproxy automtisk sørge for at trafikken går til en av de andre likeverdige serverne. Haproxy er også duplisert. Hele oppsettet overvåkes av nagios.

Frontend: To servere som begge kjører nginx. All webtrafikk går inn via disse serverne, og nginx er satt opp til å håndtere lastbalansering.

Applikasjonsservere: Det er fire applikasjonsservere som kjører makoframework. Alle serverne har alle applikasjoner installert og i normal driftsituasjon brukes alle serverne til å betjene forespørsler. Her vil også administrasjonsverktøyene ligge.

Sphinx kjører et master/slave-oppsett. Master-server er kraftigere enn slaveserveren. Alle søk kjøres mot master. Slaven står som hotspare, dvs. at den brukes kun når master ikke er tilgjengelig.

Filbanken består i alt av 4 servere. To applikasjonsservere med programkoden for API'et og to databaseservere med postgresql. Lagringsløsningen er et SAN hvor alt av strømforsyninger, diskhyller, controllere, fibergrensesnitt osv. er duplisert.

Prosesseringsriggen vil i utgangspunktet bestå av en eller to servere (master) som kun skal operere som api og for å administrere en rekke servere som skal utføre selve prosesseringen (slaver). Disse slavene vil bestå av gamle servere som har gått ut av garanti, men som likevel fungerer. Prosesseringskapasiteten skal kunne utvides kun ved å koble en server til nettverket. Det vil deretter automatisk installeres nødvendig programvare, og den vil automatisk bli operativ. Servere som skulle slutte å fungere vil automatisk ekskluderes, og eventuelle påbegynte jobber vil automatisk bli fullført av en annen server.

/eof