SLUTTRAPPORT

"Arkivet-i-dokumentet" og ikke "Dokumentet-i-arkivet"

31. Desember 2018

1. Prosjektorganisasjon

Prosjektets varighet: 1. januar 2018 til 31. desember 2018

Prosjekteier: Trondheim Kommune, Trondheim byarkiv

Styringsgruppe: Prosjektet etablerte ikke en formell styringsgruppe

Prosjektgruppe: Thomas Södring ved OsloMet og Jean-Philippe André Caquet, ved prosjektstart ansatt ved Trondheim Kommune

2. Prosjektets formål

Formålet med prosjektet var å se på muligheten for å lenke arkivrelaterte metadata direkte til et dokument i stedet for, eller i tillegg til, å ivareta disse dataene i en arkiveringsløsning. Prosjektet har flere potensielle bruksområder, men to har vært i hovedfokus. Disse er:

1. Se på muligheten for å benytte dette som et standard format for *utveksling av records* mellom offentlige (og potensielt også private) aktører
2. Se på muligheten for å benytte dette som et format for *frittstående arkivering* av dokumenter, uavhengig av arkiveringsløsninger.

"Arkivet-i-dokumentet" kan utredes og utvikles til å bli en fleksibel standard for kobling av kontekstuelle metadata til dokumentfiler, med muligheter for tilpasning og utvidelse ut fra behov.[1]

3. Prosjekthistorikk

Konseptet "Arkivet-i-Dokumentet" har lenge vært diskutert i Trondheim Kommune som en ny måte å tenke arkiv. Det ble bestemt å søke om midler for å gjennomføre prosjektet i 2018. I og med at det var stor usikkerhet omkring hvor mye prosjektet ville koste, ble det søkt om midler fra to steder: Arkivverket og Fylkesmannen i Trøndelag. Prosjektet fikk tildelt 200.000 for gjennomføring av Riksarkivet, men fikk avslag på søknaden til Fylkesmannen.

Prosjektet startet opp i januar 2018 og prosjektgruppen bestod av førsteamanuensis Thomas Sødring ved OsloMet og Jean-Philippe André Caquet, daværende rådgiver ved Trondheim Kommune.

Den 26. april var prosjektet svært nær sitt opprinnelige mål om å lage en enkel prototype, og det ble da avholdt et møte ved Trondheim Kommune hvor prosjektgruppen, samt Elin Harder fra Digitalt Førstevalg, Torgeir Kruke, prosjektleder for prosjektet "It-støtte for saksbehandling" og Kari Myhre fra Trondheim byarkiv var deltakere. I møtet ble det diskutert om rammebetingelser for prosjektet videre. En styringsgruppe ble ikke etablert. Den skulle tatt stilling til en liste av produkter foreslått av prosjektgruppen.

[1]Formålsbeskrivelsen har blitt oppdatert underveis, for å gjøres tydeligere. Følgende var opprinnelig beskrivelse: "Arkivet-i-Dokumentet er et konsept for å utveksle dokumentasjon med vedlagt kontekstinformasjon mellom det offentlige og kunder, offentlige og profesjonelle aktører, og profesjonelle og kunder. Arkivet-i-dokumentet er tenkt å være en standard som man kan forholde seg til helt eller delvis. Arkivet-i-dokumentet representerer også en mulighet til å lagre arkivmateriale uavhengig av dedikerte "arkivløsninger"."

Produktene som ble foreslått fra prosjektgruppen var som følger:

- *Utvidet Noark-basert metadata-modell.*
- *Eksponering av SystemIDer i ledger i Blockchain/Forslag til løsning for verifisering av data online*.
- *Pilotimplementasjon med en leverandør*
- *Forslag om utforming av løsning for sikkerhet og kryptering*

4. Konseptuell løsning

Det har, såvidt prosjektgruppen kjenner til, aldri vært en felles universell standard for strukturering av metadata i dokumenter. Prosjektet legger derfor frem hypotesen at en modifisert form av Noark-standarden kan være en slik standard. Vi utelukker ikke at det kan finnes tilsvarende innen spesifikke fagområder, men ikke noe vi kjente til i innledende fase av prosjektet, og heller ikke har fått kjennskap til siden.

Det er kun de delene av Noark-standarden som omhandler strukturering av metadata som er av interesse av prosjektet, og ikke det som beskriver funksjonalitet ift saksbehandling.

Metadatastrukturen er beskrevet i form av en datamodell og en metadatakatalog som sier hvordan et uttrekk av et avsluttet arkivsystem skal utformes. Et arkivsystem inneholder ofte store mengder dokumenter, samt store mengder metadata om disse dokumentene. Noark-standarden sier at ved avslutning (deponering) av arkivsystemer, skal det gjøres uttrekk av dataene i systemet etter denne modellen. Deretter skal de avleveres til et elektronisk depot i form av en eller flere mappe(r) som inneholder dokumentene, samt fem ulike xml-filer:

- *arkivstruktur.xml*
- *arkivuttrekk.xml*
- *endringslogg.xml*
- *loependeJournal.xml*
- *offentligJournal.xml*

Xml-filen vi tar utgangspunktet i prosjektet, er den som kalles «arkivstruktur.xml». I denne strukturen ligger alle metadata om alle dokumenter i arkivuttrekket, samt linker til de fysiske filene. Dette er en hierarkisk xml-fil som forenklet ser slik ut:

>Arkiv (Et organs arkivbeholdning under ett)

->Arkivdel (Organisering av arkivet etter serier)

-->Klassifikasjonssystem

--->Klasse (Kan være flere ting, f. eks. funksjonsområdet dette gjelder for)

---->Mappe (Samling av «registreringer» som hører sammen, f. eks. som er en del av samme sak)

----->Registrering (metadata om hendelsen som dokumentet knytter seg til, f.eks utsending av et brev.)

------>Dokumentbeskrivelse (metadata om selve dokumentet)

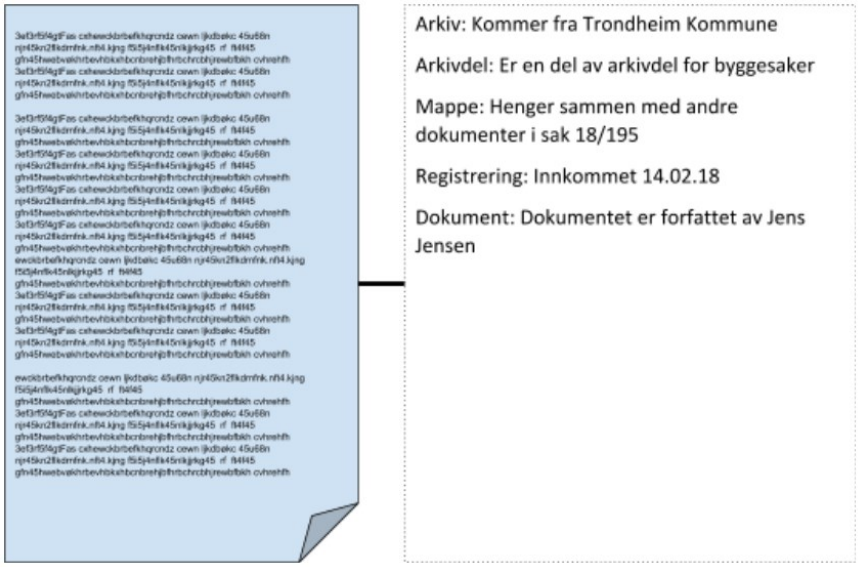
------->Dokumentobjekt (tekniske metadata om dokumentet)

-------->Elektronisk fil (selve filen i et arkivverdig format, f. eks. pdf/a, xml, tiff, jpeg)

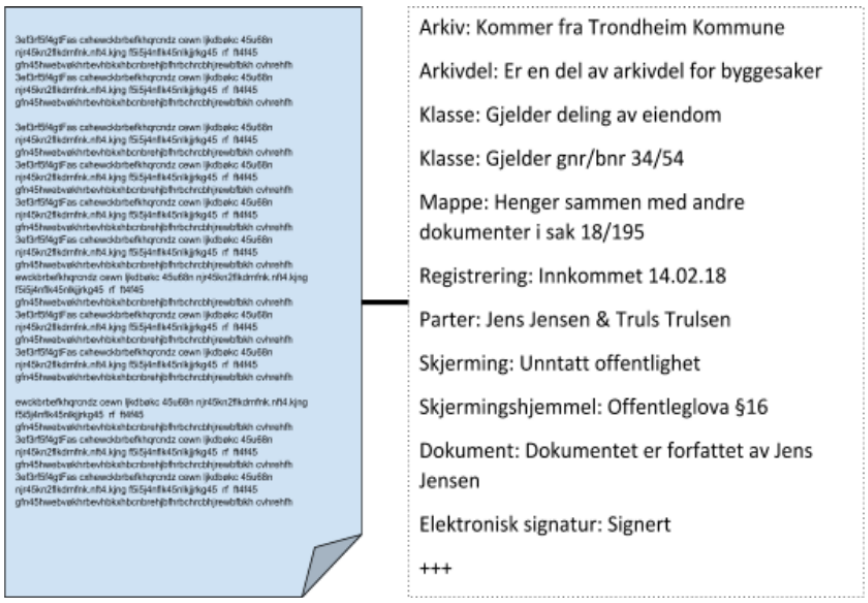
Et uttrekk kan inneholde flere objekter på alle nivåer, men typisk er det få objekter øverst i hierarkiet og mange nederst. I tillegg til data i den hierarkiske Noark-strukturen, kan xml-en inneholde andre metadata som kan legges på forskjellige nivåer, som f. eks. *skjermingsinformasjon*, *parter, bevaringsinformasjon, unike systemIDer* og *virksomhetsspesifikke metadata.*

Konseptet for "Arkivet-i-dokumentet" er å gjøre uttrekk av hele denne strukturen, *men kun for det gjeldende dokumentet*. Det vil i praksis si å legge dokumentets metadata i form av et "mini-uttrekk" av strengen med metadata fra *Arkiv* til *Dokument*-nivå. Det kan være aktuelt å trekke inn enkeltvise metadata fra et Noark-uttrekks øvrige xml-filer også, for eksempel endringslogg.xml for å spore endringer som er gjort, men dette forutsetter at man har en løsning for å linke disse dataene til hverandre.

Deretter må disse uttrekksdataene kobles til selve dokumentet, *slik at konteksten i praksis følger selve dokumentet*. Eksempler på data som følger med dersom man bare har minimum Noark-struktur kan vi se i tegningen nedenfor.



Dersom en tar i bruk fleksibiliteten som er implisitt i NOARK, vil en kunne utvide dette og gi enda mer kontekstinformasjon. I teorien er det ikke mange begrensninger.



Noen av disse dataene vil også kunne finnes som fritekst i selve dokumentet og kan i teorien hentes ut fra dette, men dette krever ofte manuell tolkning eller avansert AI. I tillegg knytter det seg en del usikkerhet til data hentet ut på denne måten, siden data kan leses og tolkes feil. Strukturerte data er i større grad maskinlesbare, entydige og inneholder færre potensielle feilkilder. Dette gjør dem også lettere å gjenbruke og dele. Eksempler er:

1. Tekniske løsninger kan skjerme dokumenter som i utgangspunktet ligger basert på underliggende skjermingsmetadata.
2. Partsinformasjon kan brukes til å vise skjermede dokumenter utelukkende til de som har et partsforhold (og derfor er unntatt skjermingskrav)
3. Eiendomsinformasjon kan koble sammen alle dokumenter som relaterer til en eiendom , slik at en finner tilbake til all informasjon gjeldende den eiendommen
4. Dokumenter som mottas kan automatisk mappes tilbake inn i Noark-baser (eller andre databaser)
5. Partsinformasjon gjør det mulig å sende dokumenter direkte til mottaker
6. Parter kan selv verifisere ektheten av dokumenter Nærmere beskrevet i kapittel 6)

Det meste av nytteverdien av dette konseptet realiseres av at det lages verktøy (som f. eks. innsynsløsninger) som kan nyttiggjøre seg disse dataene. Dette krever at en begynner å lagre metadata med dokumentene.

5. Noarks svakheter i denne modellen

Noark 5 v4 ble av prosjektgruppen vurdert til å ha enkelte svakheter som gjør den litt problematisk å benytte direkte/uendret som standard for “Arkivet-i-dokumentet”. Selv om Noark brukt riktig er god til tidfesting og til å vise organisasjons- og forvaltningstilhørighet, har den mangler iht stedfesting. Det er ingen standard for tilleggsinformasjon og partsinformasjon.

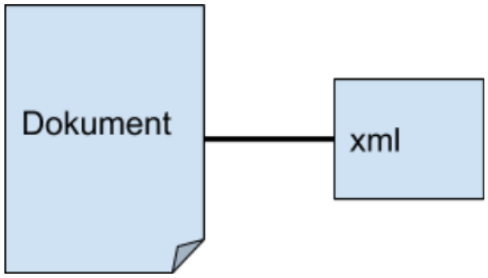
Det ble opprinnelig foreslått tatt med inn i arbeidet å lage et forslag til en tilpasset versjon av standarden. Prosjektet har ikke tilstrekkelig rammer for å ferdigstille dette arbeidet og arbeidet

finnes derfor bare i form av et utkast, men er likevel et forsøk på å både forenkle, gjøre mer fleksibelt og utvide med ofte brukte data.

Den nye versjon av standarden (Noark 5 v5 av 8. desember 2018) virker enklere og mer fleksibel. De viktigste endringene er at "Noark komplett" forsvinner som begrep, og at man i større grad enn før fokuserer på alternative arkivstrukturer, og at man har generalisert partsbegrepet i større grad. En fortsatt eksisterende mangel eller svakhet, er hvordan man skal arkivere eiendomsdata på en hensiktsmessig og gjenbrukbar måte. I tillegg er heller ikke standarden laget med det formålet vi foreslår å bruke den til for øyet.

6. Teknologisk løsning

En mulighet som ble diskutert før prosjektet ble iverksatt var å legge "mini-uttrekkene" av metadata som vedlegg til arkivdokumentene i form av xml-filer. Altså i form av "buddy"-filer eller "sidecar"-filer. Vi fikk etterhvert vite at med de nyere versjonene av pdf/a , var det mulig å legge metadatastrukturer direkte inn i filene.



Etter å ha diskutert ulike alternativer landet vi på Extensible Metadata Platform (XMP). Dette er en standard som allerede var mulig å benytte fra pdf/A2 og utover. Standarden er opprinnelig laget for å overføre metadata sammen med filer, altså som et utvekslingsformat på samme måte som vi beskrev et av formålene med "Arkivet-i-dokumentet". Standarden er opprinnelig utviklet av Adobe, men er nå ISO-standard 16684 Standarden er i utgangspunktet beregnet for *pdf, tiff, jpeg/jpeg2000, png, gif* og *webp.* Pdf (i form av pdf/a1 og 2), tiff og jpeg er godkjente arkivformater,

I tillegg kan den lagres som egne tilleggsfiler(såkalte sidecar-files) til dokumenter av andre formater, slik vi opprinnelig hadde tenkt. Dette gjør standarden brukbar for alle typer filformater.

I tillegg er standarden svært åpen og tilgjengelig. Det finnes allerede en 30-40 forskjellige freeware-løsninger og kommersielle løsninger som har støtte for å lese eller skrive denne typen metadata. En liste finnes beskrevet på Wikipedia. I tillegg har prosjektet laget en egen leser som befinner seg her[2], som viser hvor lett det er å få tilgang til disse dataene.

Som en del av Samdok prosjektet utførte Bouvet en analyse av deler av Noark arkivstruktur i RDF. XMP bruker RDF som format, så det bør være relativt enkelt å gjenbruke deler av Noark sin arkivstruktur i XMP.

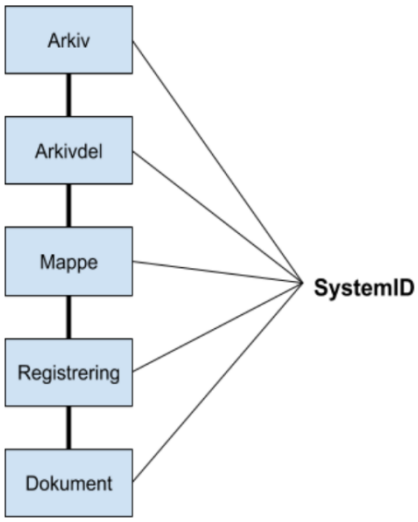
[2] https://gitlab.com/OsloMet-ABI/noark-xmp

7. Sensitivitet og sikring/kryptering

Noark-dokumenter og metadata kan potensielt inneholde gradert informasjon, sensitiv personinformasjon, eller personinformasjon som ikke nødvendigvis er sensitiv, men som allikevel må håndteres korrekt iht Personopplysningsloven/GDPR. Det er en svært forenklet regel at jo lavere ned i Noark-strukturen man kommer, jo større risiko er det for at data vil være sensitive eller av en annen grunn må skjermes. Det vil si at på *arkiv- og arkivdelnivå* er ingenting "sensitivt", mens det på *mappe-, registrerings- og dokumentnivå* vil kunne finnes f. eks. partsdata. I tilfellet med "Arkivet-i-Dokumentet" er dette noe man må vurdere på lik linje som i et annet arkiv, men med stor sannsynlighet vil et dokument som i seg selv er sensitivt (det fysiske innholdet i dokumentet) uansett måtte sikres og eller krypteres. Det finnes allerede muligheter for å legge krypteringsinformasjon inn i dokumenter, og det foreslås fra prosjektet at det også kan vurderes å standardisere måten dette gjøres på, men at det ligger *utenfor* rammene av dette prosjektet.

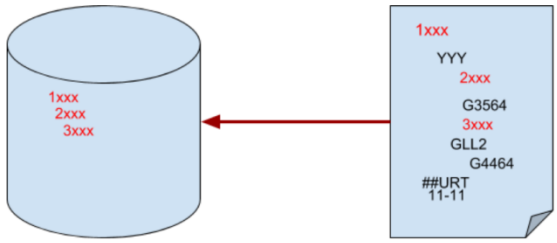
8. Verifisering

De fleste dataobjekter i Noark-standarden har en egen "systemID", som identifiserer de som unike objekter i Noark-strukturen. Hver mappe, hver registrering, hver arkivdel osv. har alle sammen en slik ID.



Innenfor et arkiv eller arkivsystem skal alle systemIDer være unike, men i teorien kan en og samme systemID kan dukke opp i to forskjellige systemer. Dette er veldig avhengig av regimet for tildeling av systemIDer i de enkelte systemene. Det som er mindre sannsynlig er at en konstellasjon av flere systemIDer skal kunne være identisk i flere systemer. I hvert fall ikke i nyere systemer som har et gjennomtenkt forhold til å lage unike identifikatorer.

Dersom systemIDene knyttes til et dokument kan de, sammen med sjekksum, og titler på høyere nivå i Noark-strukturen (f. eks. Arkiv, Arkivdel o.l, som heller ikke er sensitive data) utgjøre et dokuments unike "fingeravtrykk". Disse metadataene kan trygt eksponeres online som en del av en felles offentlig tjeneste. Denne tjenesten vil kunne benyttes av brukere til å selv verifisere om dokumentene de sitter på er genuine, ved å sammenlikne metadata om egne dokumenter med de strukturene av metadata som finnes eksponert gjennom denne tjenesten.



9. Online verifisering og Blockchain

Tidlig i prosjektet ble det vurdert om Blockchain (on-line ledgers) kan brukes til verifisering av arkivdokumenter online. Dette ble av prosjektgruppen sett på som en gyllen anledning til å kjøre prosjektet parallelt med, og som en pilot for prosjektet "Hvilken rolle blokkjede kan spille for forvaltningsarkivene". Derfor deltok også to av de samme personene (Thomas Södring og undertegnede) i begge prosjektgruppene.

Senere har vi vurdert at Blockchain trolig er for utfordrende å benytte i denne sammenheng - i lys av offentlig forvaltningsmulighet, samt teknologiens modenhetsgrad.

Vi foreslår for øvrig at man ser på andre muligheter for sikring av validitet, som f. eks "trusted timestamps" (Jf eIDAS) eller liknende teknologier.

10. Forslag til videre oppfølging

På møtet 26. april foreslo prosjektgruppen at prosjektet skulle resultere i en del konkrete produkter, disse ble også presentert på KDRS' seminar 13. og 14. juni (se vedlagte presentasjon). Siden det aldri ble opprettet noen styringsgruppe, har det heller aldri blitt konkludert med om det var disse punktene prosjektet skulle gå videre med.

- Utvidet Noark-basert metadata-modell.
  Det ble i løpet av prosjektet laget en skisse kalt "Åpen arkivstandard 0.2" som er basert på Noark-standarden men strippet kraftig ned. Denne er foreløpig på et skissestadium og ikke et ferdig utkast
- Eksponering av SystemIDer i ledger i Blockchain/Forslag til løsning for verifisering av data online.
  Eksperimentet blir videreført i det parallelle prosjektet som er nevnt i kapittel 9. Dette prosjektet anbefaler dog ikke Blockchain, og foreslår at en ser på andre løsninger for å oppfylle behovet.
- Pilotimplementasjon med en leverandør.
  Dette er noe som absolutt bør utredes nærmere, men siden deler av prosjektgruppen er blitt inhabil ønsker den ikke å komme med forslag.
- Forslag om utforming av løsning for sikkerhet og kryptering.
  Prosjektet foreslår at dette tas med inn i et eventuelt arbeid med en pilotimplementasjon.