FAGFELLEVURDERT ARTIKKEL

Herbjørn Andresen

FREMFINNING OG BRUK AV DIGITALT SKAPT MATERIALE I ARKIVDEPOTENE

Sammendrag

Artikkelen sammenligner to metoder for fremfinning av digitalt skapte arkivuttrekk i depoter. Begge metoder baserer seg på velprøvd og utbredt teknologi, og bør ikke by på større teknologiske utfordringer. Digitalt arkivmateriale er organisert på litt ulike måter, og det kan variere hvor treffsikre forskjellige fremfinningsmetoder vil være for ulike typer digitale arkiver. Valget av en bestemt fremfinningsmetode har imidlertid ikke bare betydning for mulighetene for å finne det man leter etter, det kan også i noen grad ha betydning for hvilke forkunnskaper som er nødvendig for å bruke arkivene, og for hvordan man kan evaluere resultatene av fremfinningsbestrebelsene. Spørsmålet om fremfinningsmetoder *bør diskuteres videre ut fra arkivfaglige premisser, og ikke ses kun som et spørsmål om teknologivalg.*

Innledning

Denne artikkelen sammenligner to metoder for fremfinning av digitalt skapte arkivuttrekk i depoter. Den ene er å finne frem ved hjelp av hierarkiske arkivbeskrivelser, der brukeren orienterer seg i de enkelte uttrekkene fra generelle til mer konkrete beskrivelser av dokumentasjonen. Den andre er å søke i fulltekst innenfor hele eller store deler

av depotets samling av digitale dokumenter, der søkeresultatet er en rangert liste over dokumenter som inneholder søketermene. Begge metoder baserer seg på velprøvd og utbredt teknologi, og bør ikke by på større teknologiske utfordringer. Likevel er det noen forskjeller mellom disse metodene som dreier seg om behovet for ekspertorganisering av arkivmaterialet, om hvilke forkunnskaper brukerne har behov for, og mulighetene for en kritisk vurdering av søkeresultatene. Metode for fremfinning av digitalt skapt materiale er derfor ikke bare et teknologivalg, det er også et valg som har betydning for samspillet mellom arkivinstitusjon og bruker.

Om «bruk av digitalt skapt materiale» som emne i arkivundervisningen

Denne artikkelen er basert på et innlegg holdt på Arkivarforeningens vårseminar i 2018. Temaet var *Kompetansebehov for fremtidige arkivarer*. I bachelorgraden Arkivvitenskap ved OsloMet dekker undervisningen både arkivdanning og depot, med en liten overvekt av emner knyttet til arkivdanning.

Et nylig opprettet emne i undervisningen, gjennomført for første gang høsten 2017, handler om fremfinning og bruk av digitalt skapt arkivmateriale i arkivdepotene. I mangel av et bedre navn heter dette emnet «Digitalt depot II», mens det litt eldre undervisningsemnet som omhandler mottak, innlemming og bevaring av digitalt skapt materiale er døpt om til «Digitalt depot I». Det nye emnet er plassert sent i bachelorgraden, i 5. semester. Det utgjør 15 studiepoeng, omtrent et halvt semesters arbeidsomfang ved normal studieprogresjon.

De fleste av emnene i undervisningen kombinerer teoretisk og praktisk kunnskap. Studentene skal kjenne til hvordan ulike oppgaver

løses i norske fagmiljøer, kunne reflektere kritisk over praksiser, og delta konstruktivt i den videre utviklingen av faget.

Et emne om bruk av digitalt skapt materiale har foreløpig lite praksis å bygge undervisningen på. Læringsmålene går mer direkte fra teoretiske perspektiver til å kunne delta konstruktivt i den videre utviklingen av faget. Til erstatning for kunnskap om praksis, bygger undervisningen på enkelte hypoteser om hvordan man kan finne frem i, tilrettelegge og bruke digitalt skapt arkivmateriale. Hypotesene tar utgangspunkt i noen eksisterende teknologier og metoder med ulike opphav, som alle har det til felles at de er drøftet med større eller mindre entusiasme i internasjonal arkivteoretisk litteratur. Forskjellige tilnærminger til bruken av digitalt skapt materiale har noen konsekvenser både for arkivarenes kompetansebehov og for hvilken forkunnskap eller forståelse som kreves av brukerne. Konsekvenser for arkivarens rolle og kompetansebehov er et tema som tas opp i undervisningen, men som det egentlig er beskjeden oppmerksomhet om i den pensumlitteraturen som dreier seg om digitalt skapt materiale.

Innlegget på Arkivarforeningens vårseminar tok for seg en sammenligning mellom fremfinning basert på hierarkiske arkivbeskrivelser, og fremfinning basert på fulltekstsøk. Disse to tilnærmingene til digitalt skapt materiale er kanskje de mest nærliggende for tilgang til og bruk av digitalt skapt materiale i den nærmeste fremtid, og egner seg godt til å drøfte ulike konsekvenser både for arkivaren og for arkivbrukeren. Emnet «Digitalt depot II» i bachelorgraden tar i tillegg opp noen andre tilnærminger til fremfinning, tilgjengeliggjøring, bruk og formidling av digitalt skapt arkivmateriale enn de som har fått plass i denne artikkelen, særlig knyttet til ulike strategier for å koble eller lenke data fra ulike kilder sammen.

Arkivpakker, innholdet i et digitalt depot

Store deler av det digitale arkivmaterialet som faktisk er i bruk så langt, er dokumenter, bilder, lyd og film som er skannet inn eller konvertert fra fysiske eller analoge medier. Digitalt skapt materiale er det foreløpig lite bruk av. Manglende bruk av digitalt skapt materiale skyldes dels at det er relativt ungt, og dermed omfattet av ulike restriksjoner. En del slike restriksjoner begynner etter hvert å gå ut på dato. En annen årsak til lite bruk av slikt materiale er at depotene i liten grad har lagt til rette for å bruke det. Det er imidlertid ikke nødvendigvis noe veldig stort sprang fra tilgjengeliggjøring av digitalisert materiale til digitalt skapt materiale. Mange av betraktningene om fremfinning og bruk i denne artikkelen kan også anvendes både på digitalisert og digitalt skapt materiale. Ulike former for digitalt materiale har mye til felles. Digitalisert materiale er også sårbart for teknologiske endringer, og må forvaltes videre på betryggende og gjennomtenkte måter. I denne artikkelen er likevel digitalisert materiale holdt utenfor, for å unngå flere varianter og forbehold enn strengt nødvendig.

Begrepsmodell for digitale depoter

Innen digital langtidsbevaring og tilgjengeliggjøring finnes det en toneangivende begrepsmodell, OAIS-rammeverket, som angir et vokabular og noen grunnleggende prinsipper for overføring, bevaring, fremfinning og bruk av digitalt materiale (ISO 14721:2012). OAIS-rammeverket er ikke spesifikt avgrenset til arkivinstitusjoner og arkivmateriale, det er også relevant for andre typer virksomheter der materialet må bevares og holdes tilgjengelig over tid, blant annet digitaliserte museums- og biblioteksamlinger og gjenbrukbare forskningsdata.

OAIS-rammeverket består grunnleggende sett av en prosessmodell og en informasjonsmodell. Det er informasjonsmodellen som er sentral i

denne artikkelen. Informasjonsmodellen er imidlertid så nær knyttet til OAIS-prosessene at det er nødvendig å gi et kort riss av prosessmodellen også.

Prosessmodellen går ut på at en produsent, eller en arkivskaper etter våre begreper, overfører digitalt materiale til en depotvirksomhet. Depotvirksomheten innlemmer dette materialet i sin samling, noe som samtidig innebærer at depotvirksomheten aksepterer at det de har mottatt er noe de påtar seg å forvalte videre. Spørsmålet om hvorvidt råderetten blir overdratt til arkivinstitusjonen eller ikke kan man sette en parentes rundt i denne sammenhengen, OAIS-prosessene dreier seg om den praktiske håndteringen av materialet. Depotvirksomheten sørger for den videre bevaringen av materialet, noe som kan innebære et behov for å flytte materialet over til nye systemplattformer, eller å endre filformater eller representasjonene av arkivmetadata. De rutinene depotvirksomheten etablerer og følger i bevaringsprosessen, er grunnlaget for at man opprettholder tilliten til det digitale materialets autentisitet, til tross for at det med ujevne mellomrom kan være nødvendig å endre de digitale representasjonene. I tillegg til bevaring av det autentiske materialet, er depotvirksomheten også ansvarlig for å etablere og vedlikeholde fremfinningsverktøy, herunder arkivbeskrivelser. Siste ledd i prosessmodellen er tilgang og bruk. I OAIS-terminologi er «consumer» fellesbetegnelsen for ulike arkivbrukere. En arkivbruker kan hente frem og se på arkivmateriale som depotvirksomheten har publisert og gjort allment tilgjengelig, eller sende inn en forespørsel og få tilgang til et særskilt tilrettelagt utsnitt av de opplysningene han har spurt etter. Tilgjengeliggjøringsprosessen omfatter også depotvirksomhetens rutiner for å overholde tilgangsrestriksjoner av ulike slag som måtte gjelde for det materialet som etterspørres. OAIS-prosessene omfatter ikke det som ligger forut for overlevering, altså arkivdanningen, og heller ikke hva arkivbrukeren gjør med materialet etter å ha fått tilgang til det.

Informasjonsmodellen i OAIS-rammeverket består i hovedsak av informasjonspakker og informasjonsobjekter. En informasjonspakke er en identifiserbar digital enhet som inneholder ett eller flere informasjonsobjekter. Informasjonsobjektene kan være alt som omfattes av arkivlovens dokumentbegrep i sin videste forstand – fra tradisjonelle innholdstyper som brev eller notater, til datatabeller, sammensatte koblinger av rader fra ulike tabeller, bilder, lyd, video eller tredimensjonale digitale modeller av bygninger og landskap og mye annet. Det er mange ulike former for informasjonsobjekter som det foreløpig er vanskelig å vite hvordan man bør bevare og tilgjengeliggjøre for fremtidige brukere på en hensiktsmessig måte, men slike problemer går ikke denne artikkelen nærmere inn på.

I OAIS informasjonsmodell er det tre ulike typer informasjonspakker, og det er her sammenhengen med prosessmodellen kommer til syne. Det som produsenten/arkivskaperen sender inn er en overføringsinformasjonspakke, kalt SIP eller «submission information package» i OAIS-rammeverket. En overføringspakke vil ofte være et periodisert uttrekk fra en arkivdanningsprosess, og omtales gjerne som «arkivuttrekk» i norsk praksis. Etter at depotvirksomheten har innlemmet overføringspakken, danner depotvirksomheten en ny type pakke, en arkivpakke, kalt AIP eller «archival information package». Det er opp til depotvirksomheten hva de anser som en hensiktsmessig måte å organisere sine arkivpakker på. At depotvirksomheten har godkjent en overføringspakke, innebærer ingen plikt til at den bevares videre i samme form. Depotvirksomheten er ansvarlig for å sørge for at innholdet i arkivpakkene forblir tilgjengelig, og at de overlever de teknologiske endringer og den forvitring av lagringsmedier som uunngåelig rammer alt digitalt materiale over tid. Noen tiår etter at en arkivpakke er dannet, vil det nødvendigvis ha skjedd endringer i den digitale representasjonen, eller i utstyret som leser og tolker den digitale representasjonen, eller begge deler. Den tredje typen informasjonspakke

er en formidlingspakke, kalt DIP eller «dissemination information package». En formidlingspakke settes sammen av innhold fra én eller flere arkivpakker, i en form som er lesbar og tolkbar på brukstidspunktet. Innholdet i en formidlingspakke er basert på de til enhver tid vedlikeholdte og bevarte arkivpakkene. Dersom depotvirksomheten har utført bevaringsprosessen godt nok, vil ikke arkivbrukeren bli rammet av de eventuelle teknologiske endringene som har inntruffet mellom overføringstidspunkt og brukstidspunkt.

I tillegg til de tre typene informasjonspakker, omfatter informasjonsmodellen også beskrivende informasjon, «descriptive information». Det tilsvarer i hovedsak det vi kjenner som arkivbeskrivelser, eller kataloger, men det er ingen begrensninger i OAIS-modellen mot å bygge ut beskrivelsene med flere alternative søkeinnganger til det samme materialet. Den beskrivende informasjonen brukes til å finne frem i og administrere arkivpakkene. Beskrivende informasjon trenger ikke være avgrenset til å gjengi det som ble overført fra produsenten, den kan bygges ut og endres ved behov, eller når man får nye kunnskaper om materialet. For eksempel kan rettslige rammer for tilgangsbegrensninger endres, slik at beskrivende informasjon må tilpasses for å overholde restriksjonene på riktig måte. Beskrivende informasjon er ikke en del av det som depotinstitusjonen må bevare som autentisk. En viss form for bevaringsaktivitet er likevel nødvendig; for at beskrivelsene skal gjøre nytten, må de holdes ved like på en slik måte at de er egnet til å finne frem til og administrere den til enhver tid gjeldende, lesbare og tilgjengelige versjonen av arkivpakkene.

Arkivmateriale betraktet som dokumenter, og betraktet som data

Selv om digitale arkivpakker i prinsippet kan bestå av svært ulike former for informasjonsobjekter, kan man for enkelhets skyld konsentrere seg om to hovedformer som er rådende blant det digitalt skapte

arkivmateriale man i hovedsak beskjeftiger seg med i norske depotinstitusjoner. Den ene er de generelle Noark-uttrekkene, der informasjonsobjektene hovedsakelig er korrespondanse og notater som har oppstått i en virksomhets saksbehandling. Typiske informasjonsobjekter er dokumentfiler, pakket sammen med metadata som representerer arkivstrukturen, og viser hvor i strukturen de ulike dokumentene hører hjemme. Noark bidrar med visse kjente opplysningstyper for å finne frem i klassifikasjonskoder, saksmapper, datoer og rekkefølgen av dokumenter innen en sak med videre, og en viss trygghet for at man kan etterprøve dokumentasjonens autentisitet.

Den andre hovedformen er digitale uttrekk representert som en samling av datatabeller. Tabelluttrekk er en vanlig avleveringsform for registre og fagsystemer som ikke følger Noark-standarden. Både samlingen av tabeller, den enkelte tabell i uttrekket, eller for den saks skyld enkelte rader innenfor én eller flere tabeller, kommer helt greit inn under arkivlovens dokumentdefinisjon, og byr derfor ikke på noen rettslige avgrensningsproblemer. Bruken av tabelluttrekk vil ofte forutsette en kunnskap om sammenhenger mellom tabellene, for eksempel slik at man finner frem til ulike bekymringsmeldinger som gjelder samme barn i et fagsystem fra barnevernet ved å koble sammen en rad fra tabellen over personer i systemet med en annen tabell som dokumenterer bekymringsmeldingene. Dokumentasjonen av en bestemt hendelse eksisterer ikke nødvendigvis i form av et dokument som kan leses direkte, man må ofte koble sammen informasjonselementer fra ulike tabeller på det tidspunktet informasjonen skal brukes for å få det samlede bildet av hvordan hendelsen ble dokumentert.

Man kan si det slik at deler av det digitalt skapte arkivmateriale er organisert som dokumenter, mens andre deler av materialet er organisert som strukturerte datatabeller. Dette er for så vidt en forenkling, det finnes kombinasjonsformer og gråsoner i mange av de digitale

uttrekkene. Likevel kan den forenklede dikotomien mellom arkivmateriale organisert som dokumenter og organisert som data være hensiktsmessig for formålet i denne artikkelen. Forskjellen på et dokumentperspektiv og et dataperspektiv har betydning både for hvordan man beskriver arkivbestanden og for hvordan man forstår og vurderer resultatene fra et informasjonssøk.

Ulike kvalitative egenskaper ved digitalt arkivmateriale

En tredje knagg for forståelsen av det digitalt skapte arkivmaterialet er hvordan, og hvor komplett, informasjonsobjektene representerer «det som egentlig skjedde», altså graden av samsvar mellom dokumentasjon og det som dokumenteres. Med et litt upresist begrep kan man kalle dette kvalitative egenskaper ved materialet.

En type arkivmateriale som kan sies å stå i en særstilling er grunndataregistre og andre sektorspesifikke registre som er designet for å kunne foregi en negativ troverdighet. Grunndataregistrene omfatter Folkeregisteret, Enhetsregisteret og Eiendomsmatrikkelen. Eksempler på sektorspesifikke registre er AutoSys, som inneholder registrerte kjøretøy, og Helsepersonellregisteret. Den negative troverdigheten innebærer at man kan belage seg på at et 9sifret tall *ikke* er et gyldig organisasjonsnummer dersom man ikke gjenfinner dette tallet i Enhetsregisteret. Slike registre er ikke nødvendigvis fullstendig feilfrie, men de har tilstrekkelig kvalitet til at de kan brukes som om de var det. Kvaliteten i et slikt register kan uttrykkes som graden av samsvar mellom registeret og de gjenstandene registeret representerer. Registrene som sikter på en negativ troverdighet er gull verdt for å koble sammen kilder og vurdere påliteligheten i annen dokumentasjon. Det er imidlertid ganske beskjedne mengder opplysningstyper i slike registre, og de utgjør en liten andel av arkivmaterialet.

Dokumentasjonen som er organisert i en arkivstruktur, dannet gjennom løpende journalføring eller registrering av transaksjoner, kan i det minste sies å ha en betydelig grad av positiv troverdighet. Normene for arkivdanning, både de som arkivtjenesten følger som faglige normer og de normene som er kodet inn i systemene, bidrar til kvalitet i form av at dokumentasjonen er autentiske uttrykk for handlingene. Begrepet kvalitet, eller representasjonens samsvar, er likevel litt snevrere enn det arkivfaglige begrepet autentisitet. Arkivmaterialet kan være autentisk i arkivfaglig forstand selv om samsvaret mellom det teksten foregir og den hendelsen som dokumenteres skulle være tvilsom.

Selv om dokumentasjonen som er organisert i en arkivstruktur ikke har negativ troverdighet, gir arkivstrukturen en viss støtte til å forstå arkivskaperens vurderinger og handlemåte, og en kunnskap om typiske dokumentasjonsmønstre. Sammenhengene i et arkivuttrekk har en selvstendig informasjonsverdi. Den kunnskapen om arkivenes tilblivelse som ligger implisitt i organiseringen av materialet kan, ifølge Furner & Gililand (2016), gi et grunnlag for å trekke slutninger om meningsbærende mangler («meaningful absences») i dokumentasjonen. Slutninger basert på hvordan man vurderer huller og uoverensstemmelser i dokumentasjonen er en svakere kvalitativ egenskap enn negativ troverdighet, men sterkere enn bare å basere seg på det man kan lese direkte ut av dokumentene.

For dokumenter og data som søkes opp helt uavhengig av arkivstrukturen, vil man i utgangspunktet måtte vurdere kvalitet og etterrettelighet ut fra egenskaper ved det enkelte dokument. Det kan sammenlignes med klassisk diplomatikk, vurdering av ekthet ut fra dokumentets ytre og indre egenskaper. Det å utstyre digitale dokumenter med proveniensinformasjon og andre arkivmetadata som gjør det enklere å vurdere dem utenfor den konteksten de ble til i, kan betraktes som

en egen skole innen digital arkivering, «digital diplomatics», særlig fremmet av Luciana Duranti (1989). Selv om man kan oppnå en god del ved å tilføre gjennomtenkte metadata til dokumentene, vil likevel grunnlaget for å vurdere kvaliteten av helt frittstående digitale dokumenter være en del svakere enn for dokumenter som befinner seg innenfor en kjent arkivstruktur.

Grunnleggende fremfinningsmetoder

Organisert kunnskap, som man finner i arkiver og biblioteker og andre typer samlinger, har man tradisjonelt funnet frem i ved hjelp av surrogater, eksempelvis ulike typer sorterte kort eller lister med kortfattet informasjon som representerer enhetene i samlingen. I flere tiår har man hatt gode elektroniske verktøy for å registrere og finne frem i surrogater. Elektroniske katalogsystemer har lenger historie i bevaringsinstitusjoner enn det digitale kildematerialet. Elektroniske katalogsystemer har i hovedsak de samme egenskapene som sine forgjengere i papp elle papir, men med den fordelen at det er enklere å dele og slå sammen elektroniske kataloger for flere samlinger i en felles database. Både fremfinningsverktøyene og kildematerialet er organisert av bevaringsinstitusjonen. Man trenger en del forkunnskaper om de aktuelle bevaringsinstitusjonene både for å finne frem *til* informasjonen, og å finne frem *i* den.

Etter hvert har de fleste vendt seg til å søke mer direkte etter ord og uttrykk som gjenfinnes i dokumentene i den samlingen de søker i. Søkemotorer som Google eller Bing gir som regel treff i dokumenter som vi oppfatter som relevante for den informasjonen vi er ute etter. Denne formen for informasjonsgjenfinning kalles gjerne «IR», en forkortelse for Information Retrieval. IR har noen svært praktiske og tiltalende egenskaper. Det stilles få krav til brukerens forkunnskaper for å finne noe, det går som regel svært raskt å få ut et resultat, og rekkefølgen resultatene blir presentert i vil ofte oppfattes som en

hensiktsmessig rangering. Når man kan søke gjennom hele den samlingen som en eller flere bevaringsinstitusjoner forvalter, får også skillet mellom å finne frem til og å finne frem i informasjonen mindre å si.

Det er fordeler og ulemper ved begge disse grunnleggende fremfinningsmetodene. Noen forskjeller mellom disse metodene finner man diskutert i arkivfaglig litteratur. I stor grad står den arkivfaglige litteraturen godt plantet i tradisjonen for arkivbeskrivelser, der IR innimellom presenteres som en utfordrer. I annen informasjonsfaglig litteratur, som i mindre grad er knyttet til en arkivfaglig tradisjon, tas ofte fordelene med IR mer for gitt. Den litteraturen om IR som det vises til i denne artikkelen, er særlig knyttet til begrepet relevans og algoritmer som sørger for at søkeresultater blir rangert på en måte som man ofte oppfatter som fornuftig.

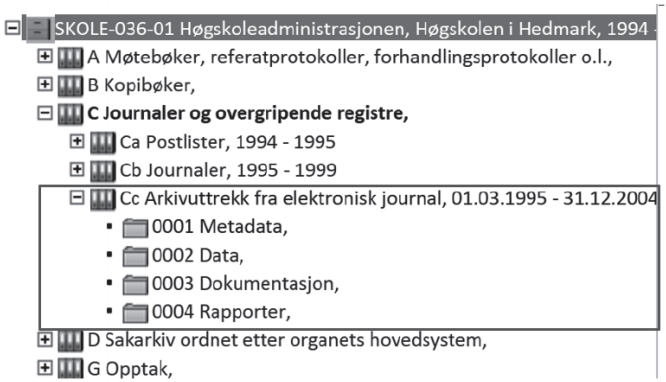
Arkivbeskrivelser som speiler arkivrepresentasjonen

Arkivrepresentasjon betegner hvordan det overleverte materialet blir ordnet og organisert i arkivdepotet (Zhang, 2012). Organiseringen av materialet følger visse faglige normer om å opprettholde proveniens og indre sammenheng, og innimellom noen mer pragmatiske valg om for eksempel å holde bestemte grupper av formater eller innholdstyper eller tidsperioder samlet. Arkivrepresentasjonen rommer kunnskap om materialet, og gir holdepunkter for å vurdere og forstå det. Mens fysisk arkivmateriale bare kan stilles opp på én måte, kan man i prinsippet arrangere flere alternative representasjoner av det samme digitalt skapte materialet.

Arkivbeskrivelser er opplysninger som kan brukes til å finne frem i en arkivrepresentasjon. En arkivbeskrivelse som er i tråd med den veletablerte faglige standarden ISAD(G), fra International Council on Archives (ICA) er hierarkisk inndelt, slik at beskrivelsens detaljeringsgrad er tilpasset den konkrete arkivenheten som blir beskrevet.

Hvis man skal ta stilling til hvor godt egnet arkivbeskrivelser vil være for å finne frem i digitalt skapt materiale i norske arkivdepoter, vil det virke nedslående å ta utgangspunkt i de beskrivelsene som er skrevet inn så langt. Mange mottatte arkivuttrekk har ingen tilgjengelig arkivbeskrivelse i det hele tatt, og der en slik arkivbeskrivelse finnes er den oftest av ganske lav kvalitet. Det finnes noen få beskrivelser av tabelluttrekk, og noen flere av elektronisk journal som følger med et papirarkiv, slik tilfellet er med Noark 3-uttrekk og en god del Noark 4-uttrekk.

Figuren nedenfor viser et eksempel på et Noark 3-uttrekk, der journalen er ført opp i arkivbeskrivelsen, og er organisert sammen med et fysisk arkiv.



Hvis man ser nærmere på hvilken informasjon som finnes om dette uttrekket, viser den fire punkter på «mappenivå» i beskrivelsessystemet. Innholdet er som følger:

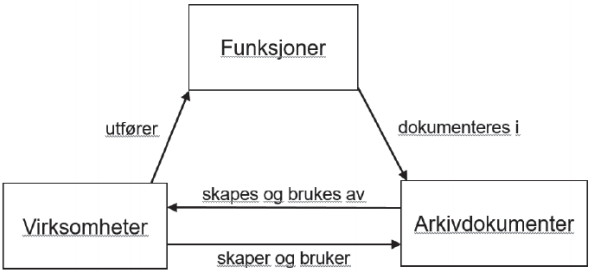
0001 Metadata – som omtaler to tekstfiler som beskriver hva som er avlevert
0002 Data – som inneholder filene med «selve» journalinnholdet

0003 Dokumentasjon – som inneholder testrapport fra innlemmingen av uttrekket i depotet
0004 Rapporter – som inneholder to obligatoriske rapportutskrifter fra Noark 3-standarden

Disse beskrivelsene, så langt de finnes, bidrar dermed bare med opplysninger som kan hjelpe en med å finne frem *til* informasjonen, og ikke til å finne frem *i* informasjonen. Det man kan lese seg til er at «noe elektronisk er avlevert».

Mangelfulle arkivbeskrivelser i praksis kan både ses som et uttrykk for at arbeidsomfanget er undervurdert, og som et uttrykk for usikkerhet om hvordan en nyttig eller hensiktsmessig beskrivelse bør utformes. Man yter likevel ikke arkivbeskrivelser som fremfinningsmetode full rettferdighet om man bare ser på hvordan dette er gjort i praksis i dag.

Fra ICA, som står bak den toneangivende beskrivelsesstandarden ISAD(G), har det kommet flere standarder for arkivbeskrivelser som kan settes sammen slik at de gir flere alternative søkeinnganger til det samme materialet. Ved siden av ISAD(G) som er en hierarkisk beskrivelse av arkivstrukturen som gir en søkeinngang til dokumentene, har det senere kommet til en standard for å identifisere virksomheter som er knyttet til materialet (ISAAR(CPF)), og en standard for å identifisere funksjoner og prosesser som er årsaken til at dokumentasjonen ble skapt (ISDF). Samspillet mellom disse standardene for beskrivelser, som gir ulike søkeinnganger til materialet, omtales med samlebetegnelsen «Records-in-Context».



Records-in-Context tilfører en omfattende ontologi for å utvikle søkbare representasjoner av arkivmaterialet. Tankesettet har nokså bred tilslutning i arkivfaglige miljøer. I prinsippet kan arkivbeskrivelser som følger disse standardene utveksles, deles og søkes i på tvers av bevaringsinstitusjoner. En innvending mot å gjøre beskrivelsene så omfattende, er at det er arbeidskrevende. Det er imidlertid store muligheter for å automatisere beskrivelsene, ved å gjøre beskrivelsesinformasjon basert på ICA-standardene til en del av de elektroniske uttrekkene. Likevel kommer man ikke unna at «noe i toppen» av beskrivelseshierarkiet må vurderes og beskrives. Et arkivsystem som det lages et uttrekk fra vil kunne produsere beskrivelsesinformasjon som egner seg til å finne frem *i* materialet, men det vil fremdeles være behov for å tilføre noe informasjon som kan bidra til å finne frem *til* materialet. Det er altså først og fremst de underliggende detaljene i arkivbeskrivelsene som egner seg for å generes automatisk.

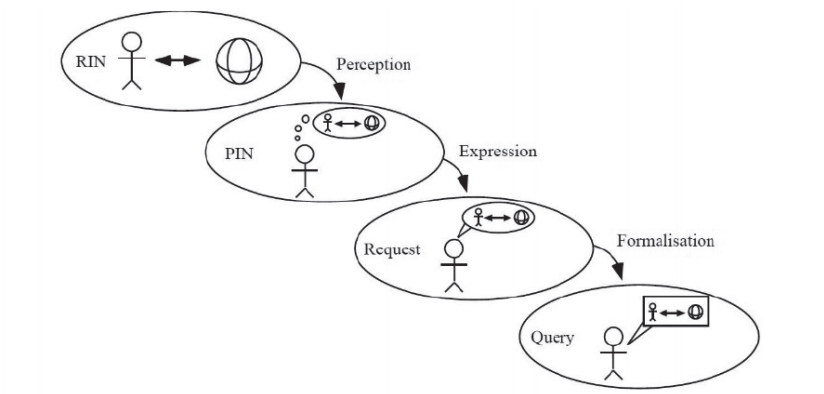
Både eksisterende arkivbeskrivelsesverktøy i Norge og de internasjonale standardene for arkivbeskrivelser fra ICA bygger på en grunnleggende forutsetning om en hierarkisk arkivstruktur som materialet er ordnet etter. Denne formen for arkivbeskrivelser er i mindre grad egnet til å presentere oppbygningen av et arkivuttrekk i form av en samling datatabeller. Man har egne måter å dokumentere slike tabelluttrekk og sammenhenger mellom tabellene i uttrekk på, i form av

Arkivverkets ADDML-skjema, eller alternativt ved å dokumentere innholdstyper i et standardisert verktøy som genererer automatiske tabelluttrekk fra databasesystemer, eksempelvis SIARD. Foreløpig er det gjort lite for å få til noe effektivt samspill mellom arkivbeskrivelser som representerer en arkivstruktur og beskrivelser av tabelluttrekk. Det som kan være praktisk mulig å få til med dagens beskrivelsesverktøy er derfor hjelp til å finne frem *til* et eksisterende tabelluttrekk, mens man må gå mer konkret inn i dokumentasjonen av det enkelte tabelluttrekket for å kunne finne frem *i* informasjonen.

IR, eller fulltekstsøk, informasjonsgjenfinning på dokumentnivå

Informasjonsgjenfinning, eller Information Retrieval (IR) betegner et sammenhengende forløp der brukeren presenter sine søkeord for et gjenfinningssystem, gjenfinningssystemet tolker eller bearbeider søkeordene, finner de dokumentene i en samling som inneholder søkeordet eller den tolkede kombinasjonen av søkeord, og leverer tilbake en liste over dokumentene som er rangert etter hvor relevante de er. Indeksering av ord i samlingen, og fremfinning av dokumenter som inneholder ord som samsvarer med søketermene, er utbredt og ganske greit tilgjengelig teknologi.

Det sammensatte forløpet for informasjonsgjenfinning har vært illustrert slik, i en figur av Mizzarro (1998), som det ofte vises til. Det reelle informasjonsbehovet en bruker har fortolkes først av brukeren selv, til et oppfattet informasjonsbehov. Brukeren kan uttrykke sitt oppfattede informasjonsbehov som en spørring, og eventuelt formalisere spørringen på en måte som er tilpasset det som gjenfinningssystemet krever. Kvaliteten på svaret man får, vil for brukeren dreie seg om hvor godt det samsvarer med reelt eller oppfattet informasjonsbehov.



Magien i et godt IR-system ligger i rangeringen av resultatene. En treffliste med hundrevis eller tusenvis av treff vil være nærmest verdiløs hvis de blir presentert i en rekkefølge som fremstår som tilfeldig for brukeren. En algoritme beregner hva som er de mest relevante treffene i samlingen for den søketermen man har oppgitt. Hva som er en god relevansalgoritme for arkivmateriale og arkivbrukere er ikke nødvendigvis det samme som for en bibliotekar, en avisleser eller en bruktbilkjøper. En typisk ingrediens i en relevansalgoritme kan være at ord som forekommer ofte i en tekst gir høyere rangering enn et ord som forekommer sjelden. Dersom for eksempel ordet «spesialundervisning» forekommer én gang i et enormt dokument, havner det lenger ned i resultatlisten enn i et mindre dokument der det gjentas flere ganger. Relevansalgoritmer kan også innstilles slik at treff i metadata gir høyere rangering enn treff i brødteksten. På den måten vil for eksempel et arkivdokument som har saksnummeret som metadata kunne få et høyt rangert treff dersom saksnummeret er brukt som søketerm, selv om saksnummeret kanskje ikke forekommer mer enn én gang i brødteksten.

Relevansbegrepet, brukt i forbindelse med IR, har historie tilbake til 1960-tallet (Borlund, 2013). Det er et eget forskningsfelt, som

kanskje foreløpig er litt fremmedartet i vårt fag. Relevansbegrepet tar utgangspunkt i brukerens søkeadferd, og omsettes til generelle algoritmer som kan evalueres ved hjelp av velprøvde forskningsmetoder. Den vanlige bruker av moderne søkeverktøy, som ikke kjenner til forskernes evalueringsresultater, vil likevel være tilbøyelig til å «stemme med føttene», og velge de verktøyene som inngir størst tillit til rangeringen av resultatene. De som har vært lenge på internett, vil kanskje huske hvilke søkeverktøy de brukte før Google i løpet av kort tid ble den ledende søkemotoren. En vesentlig faktor i Googles gjennomslag var en relevansalgoritme som blant annet tok hensyn til hvor mange eksterne lenker som peker inn til dokumentet. Det er imidlertid ikke nødvendigvis slik at en relevansalgoritme som er godt egnet for å søke på internett, også vil være godt egnet for den som søker etter dokumenter i et arkivdepot.

Relevans er et sammensatt begrep, og det er et gradsspørsmål. De dokumentene som matcher brukerens søkeord er ikke nødvendigvis fullstendig relevante eller fullstendig irrelevante. En nestor innen forskning på relevansbegrepet i IR, Tefko Saracevic, har beskrevet ulike dimensjoner ved relevans (Saracevic, 1975, 2007):

Affektiv relevans	Det brukeren egentlig vil med informasjonssøket
Situasjons-relevans	Hvilen situasjon/oppgave søket er en del av
Kognitiv relevans	Hva er det brukeren allerede vet?
Emne-relevans	Forholdet mellom spørsmålets og dokumentets emne
Algoritmisk relevans	Hvilke dokumenter systemet «tror» er relevant for søket

Mange sider ved relevansbegrepet dreier seg om brukerens oppfatning av at det man finner gjennom søket, er det man var ute etter. Et spørsmål i forlengelsen av det, er hvem man anser som arkivbrukere, og hvilke forkunnskaper man forutsetter. En arkivbruker som har erfaring fra tradisjonell organisering av arkiver og fremfinning ved hjelp av arkivbeskrivelser, vil ha forutsetninger for å justere og tilpasse søketermene på måter som den mindre kyndige brukeren ikke har.

Sammenligning av arkivbeskrivelser og IR som fremfinningsmetoder

Selv om det er flere vesentlige forskjeller på disse to tilnærmingene til å finne frem til og i digitalt arkivmateriale, kan det være greit å starte med noe de har til felles. Både arkivbeskrivelser og IR er i hovedsak innrettet slik at informasjonsobjektene betraktes som og forstås som dokumenter. Dersom man har et arkivuttrekk som består av store tabeller, som kanskje bare gir mening som dokumentasjon når et utvalg rader fra flere tabeller kobles sammen, vil hverken arkivbeskrivelser eller IR gi noen direkte tilgang til den samlede informasjonsmengden som utgjør dokumentasjonen av en bestemt transaksjon eller hendelse. Et IR-system kan for så vidt finne ordforekomster i store tabeller, i og med at hver tabell vil bli sett som og håndtert som et dokument. Relevansalgoritmene kan imidlertid komme til kort i en stor tabell. Logikken bak det å utlede hva som er vesentlig i et dokument er ikke nødvendigvis direkte overførbar fra en løpende tekst til strukturerte dataelementer. Hvis man for eksempel søker etter termer som tilfeldigvis forekommer i ligningstabellen for et bestemt år, kan fødselsnummer gi svært lav rangering fordi det bare forekommer én gang i en enorm tabell, mens kommunenummeret for en stor kommune kanskje vil gi høy rangering av den samme tabellen fordi det gjentas mange ganger. Den som skal bruke et tabelluttrekk til noe, vil først og fremst trenge konkret dokumentasjon av tabellene og

sammenhenger mellom dem – hverken arkivbeskrivelser eller IR kan ventes å bidra i vesentlig grad til å finne frem i slik informasjon.

Ellers er det helst forskjellene mellom arkivbeskrivelser og IR som det vil være relevant å sammenligne.

Fremfinning via arkivstruktur innebærer ordning og beskrivelser som er basert på faglige vurderinger. Informasjonen er ekspertorganisert, det er en viss kompetanseterskel for å finne frem i den. Ekspertorganisering er samtidig noe som gir et selvstendig grunnlag for tillit til dokumentasjonen.

Bruk av arkivbeskrivelser, og det å forutsette kompetanse og forkunnskaper hos den som skal finne frem i materialet, kan gi innfallsvinkler til materiale som man ikke finner ved å søke etter bestemte ord. Hvis man for eksempel interesserer seg for hvordan klageordninger har fungert i utkontrakterte kommunale tjenester i en bestemt tidsperiode, vil antakelig kunnskaper om hva slags arkiver dette har generert (eller ikke generert) være en stødigere vei til å finne informasjonen enn å gjette seg til hvilke søketermer som kunne ha satt deg på sporet.

Det kan også i noen sammenhenger ses som en fordel ved arkivbeskrivelser at de bidrar med et hierarkisk overordnet nivå der man for eksempel kan angi klausuleringer som begrenser tilgangen. Det vil være lettere å avskjære søk i arkivenheter som man likevel ikke vil få tilgang til. I et IR-basert søk vil man kunne få mange treff som systemet vurderer som relevante, men som man likevel ikke får mulighet til å undersøke nærmere.

Et moment som særlig er trukket frem i Furner & Gililand (2016), er at arkivbeskrivelser inneholder informasjon om struktur og kontekst som er en støttende kilde til kunnskap. Kontekstinformasjonen kan

gi bedre belegg for å vurdere negativ troverdighet eller meningsfulle fravær i kildematerialet.

IR, og relevansrangerte treff i metadata og innhold har kanskje som sin fremste fordel at det senker terskelen for å kunne søke etter informasjon. Søketermer som gjenfinnes i arkivmaterialet vil bli presentert for brukeren, i en rekkefølge som det er ganske høy sannsynlighet for at brukeren vil oppleve som relevant. Kontekstinformasjonen vil ofte lide noe under det. Selv om man kan presentere noe kontekstinformasjon sammen med søkeresultatet, vil treff nummer en, to og tre i resultatlisten kunne stamme fra svært forskjellige arkivenheter, uten noen sammenheng som nødvendigvis er mulig for brukeren å ta stilling til. IR vil stille mindre krav til brukerens forkunnskaper om aktører og funksjoner. Det vil være lettere å finne noe, men kanskje også vanskeligere å vurdere egen dømmekraft, og «forstå hva det er man ikke forstår».

Et konkret problem med relevansalgoritmer, som ikke er like påtrengende for arkivbeskrivelser, er at det er vanskelig å integrere tidsbegreper i generelle relevansalgoritmer. For arkivmateriale som man finner frem til gjennom arkivbeskrivelser, vil det være enklere å orientere seg i en rekkefølge av transaksjoner, eller å vite hvilke dokumenter som var på plass før det dokumentet man leter etter. Relevansalgoritmer kan ta hensyn til dateringer som er påført i dokumentenes metadata som del av en søketerm, men vil ha problemer med å rangere dokumenter som ligger nær hverandre i tid ut fra for eksempel den rekkefølgen de har oppstått i.

I samfunnet rundt arkivinstitusjonene er det denne måten å finne frem i informasjon på som de ikke-arkivkyndige brukerne er mest vant til. For så vidt kan man vel også spørre om ikke arkivarer også har blitt vant til IR som fremfinningsmetode gjennom det livet de

lever utenfor arkivinstitusjonen. Brukeres vaner og forventninger er et moment som det er rimelig å tillegge en betydelig vekt, selv om det kanskje også vil kollidere noe med den forståelsen av faglighet som følger med ekspertorganisering. Relevansbegrepet innen IR er mer brukerstyrt, og følgelig også mer subjektivt, enn den representasjonen av arkivene som nedfelles i arkivbeskrivelser. Furner & Gililand er inne på en fare for at IR og relevansrangering basert på sannsynligheter kan lugge litt på et begrepsmessig plan, men likevel er praktisk anvendelig: «Nevertheless, once we have persuaded ourselves that even if this poses a problem for the theorist, it does not for the practitioner» (Furner & Gilliland, 2016 s. 604).

Det er også visse teknologiske egenskaper som kan ha betydning ved sammenligningen av fremfinningsmetoder. IR er vesentlig mer automatiserbart og skalerbart enn å utarbeide systemer for arkivbeskrivelser. Et system for arkivbeskrivelser må forvaltes og tilpasses omfang og bruksmønster av en virksomhet som kjenner sitt brukermiljø og deres behov godt. IR-systemer kan i større grad tres nedover en dokumentsamling og brukes til å tolke søketermer og finne frem dokumenter uten at det kreves inngående kjennskap til særtrekk ved bruksomfang og brukermiljø.

Er kombinasjonsformer mellom arkivbeskrivelser og IR veien å gå?

Et elektronisk katalogverktøy har gjerne noen muligheter for å søke i teksten, så en viss grad av hybridformer kan man si at eksisterer allerede. Det er likevel ikke egentlig noe man kan betrakte som et IR-system, det vil fremdeles være et søk innenfor arkivrepresentasjonen og på arkivbeskrivelsenes premisser.

Som et tankeprodukt kan man likevel tenke seg nye former for verktøy som kombinerer arkivbeskrivelser og IR. Et postulat hos Furner

& Gililand er at «archival IR» kan og bør være en slags hybrid mellom hierarkisk struktur som inngang til materialet og et mer tekst- og relevansbasert tilslag på søketermer. Dette innebærer en tettere kobling mellom beskrivelse og materiale. Det er noe som for eksempel det XML-baserte arkivbeskrivelsesskjemaet EAD åpner for. Det er et skjema som har stor praktisk utbredelse, blant annet som søkeinngang i portalen www.archivesportaleurope.net/. Et slikt skjema påtvinger alt som registreres en hierarkisk struktur, og gir hierarkisk organiserte søkeresultater også ved fulltekstsøk. Det vil imidlertid være mindre brukervennlig enn klassisk, relevansbasert IR, og ligge fjernere fra de erfaringer brukerne har med IR fra andre områder enn arkiv. En fordel de ser for seg, er at en «archival IR»-hybrid vil gi bedre muligheter for å vurdere hva man faktisk finner opp mot hva man mener man kunne ha ventet å finne, enn ved relevansbasert IR.

Når man stiller spørsmålet om en kombinasjonsform, eller hybrid, bør være veien videre, kan det være ryddig og redelig også å ta spørsmålet et skritt videre og spørre om IR vil kunne overta fullt og helt for arkivbeskrivelsene. En tenkbar, og ikke helt usannsynlig, utvikling kan være at arkivbeskrivelsene etter hvert vil bli oppfattet som irrelevante, eller økonomisk uforsvarlige, dersom IR gir resultater de fleste er fornøyde med og opplever som relevante.

Det er neppe fornuftig å avvikle arkivbeskrivelsene fullt og helt. Det vil være behov for «finne frem til»-informasjon for arkiver, noe i toppen av hierarkiene, som kan danne et meta-arkiv over bestanden. Man vil trenge å vite hvor arkivpakkene kom fra, av bevaringshensyn. Det er noe større grunn til å tvile på om tradisjonelle arkivbeskrivelser vil fortsette å ha samme betydning for det å finne frem *i* arkiver. Det gir visse fordeler, særlig i form av kunnskap om arkivmaterialets kontekst, som kan ha en egenverdi og være nyttig for kildekritiske vurderinger av materiale. På den annen side er det vanskelig å argumentere

mot en suksess, brukernes vaner og preferanser kan bli svært tungtveiende. Noe man kan si med en rimelig grad av sikkerhet, er at arkivbeskrivelsene for digitalt skapt materiale må prioriteres høyere og gjøres bedre dersom man skal argumentere for at de fortsatt har livets rett.

Oppsummering

Man kan finne frem til, og finne frem i, digitalt skapt arkivmateriale på ulike måter. Forskjellige metoder har sine fordeler og ulemper. Fremfinning kan baseres på arkivbeskrivelser, på IR, eller på hybridvarianter. Alle tre posisjoner har sine talsmenn blant arkivteoretikere, og det vil være vanskelig å si noe sikkert om hvordan dette blir om noen år.

En diskusjon av hvorvidt man bør satse på den ene eller den andre metoden for fremfinning kan være vanskelig å føre på en strukturert og ryddig måte. Det er en sammensatt vurdering som omfatter profesjonsrollen, faglighet og brukeres vaner og forventninger, samtidig som det er ulike teknologiske forutsetninger for og virkninger av disse metodene. Det er verdt å merke seg at ulike fremfinningsmetoder er et modent og velutviklet forskningsfelt innenfor bibliotek- og informasjonsvitenskapen. Det betyr at man ikke trenger å lete i blinde. Effektene av og aksepten for ulike måter å finne frem på, kan undersøkes, evalueres og diskuteres.

Referanser

Borlund, Pia. (2013). Interactive Information Retrieval:
An Introduction. *Journal of Information Science Theory and Practice, 1*(3), 12-32. https://doi.org/10.1633/JISTaP.2013.1.3.2

Duranti, Luciana. (1989). Diplomatics. New uses for an old science. Part I. *Archivaria, 28*, 7-27.

Furner, Jonathan & Anne J Gilliland. (2016). Archival IR: Applying and Adapting Information Retrieval Approaches in Archives and Recordkeeping Research. I Gilliland, Anne J, Sue McKemmish & Andrew J Lau (Red.), *Research in the Archival Multiverse* (s. 581-631). Australia: Monash University Publishing.

ISAAR(CPF). *International Standard Archival Authority Record for Corporate Bodies, Persons and Families*. Ottawa: International Council on Archives.

ISAD(G). *International standard archival description (general)*. Ottawa: International Council on Archives.

ISDF. *International Standard for Describing Functions*. Ottawa: International Council on Archives.

Mizzaro, Stefano. (1998). How many relevances in information retrieval? *Interacting with computers, 10*(3), 303-320.

Saracevic, Tefko. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for information science, 26*(6), 321-343.

Saracevic, Tefko. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology, 58*(13), 2126-2144.

ISO 14721:2012. *Space data and information transfer systems – Open archival information system (OAIS) – Reference model*. Geneve: International Organization for Standardization.

Zhang, Jane. (2012). Archival Representation in the Digital Age. *Journal of Archival Organization, 10*(1), 45-68. https://doi.org/10.1080/15332748.2012.677671