

Skjemainformasjon

Skjema	Samarbeids- og utviklingstiltak for arkiv og museum 2012 (bokmål)
Referanse	487659
Innsendt	15.10.2011 14:03:45

Opplysninger om søker

Søker

Navn på organisasjonen	Høyskolen i Oslo og Akershus
Institusjonens leder	Liv Gjestrum
Postadresse	Pb 4 St. Olavs plass
Postnummer / Poststed	0130 Oslo
Tlf. til organisasjon	22452671
Prosjektets kontaktperson	Thomas Sødning
Tlf. til kontaktperson	99570472
E-post til kontaktperson	thomas.sodring@hioa.no
Nettside	http://
Org. Nr	997058925

Opplysninger om prosjektet

Prosjekt

Søknaden gjelder Forvaltning og tilgjengeleggjøring av samlinger

- 6 Prosjekter og prosesser som fører til utvikling av fellestjenester, fellesmagasiner og tilgjengeliggjøring av samlinger

Tittel på prosjektet Kvalitetsvurdering av avlevert materiale

Sammendrag av prosjektet

Dette er et FoU prosjekt som forener forskningsfeltene arkiv og datakvalitet. Prosjektet vil rapportere hvilken datakvalitetsdimensjoner er relevant for Noark 4 avleveringer gjennom en analyse av et Noark 4 base/uttrekk samt en datakvalitetsanalyse av et Noark 4 uttrekk. Partnere i prosjektet består av det norske kommunal miljøet, arkivforskning ved HiOA og en internasjonal ekspert på datakvalitet. Prosjektet bygger på et tidligere samarbeidsprosjekt om datakvalitetsanalyse for Noark 5 kjerne.

Prosjektbeskrivelse

Hva vet vi om kvaliteten til elektroniske avleveringer til depot? Vår erfaring er at de kommunale depot institusjonene ikke har kapasitet til å vurdere kvaliteten på elektroniske avlevering samtidig har de bare forholdt seg til sentrale føringer om avleveringer. Datakvalitet som et eget forskningsfelt har eksistert siden 1970 tallet, men det var etter 2000 tallet at flere og flere ble interessert i fagfeltet. Dette er pga en eksplosjon i mengden elektronisk data som ble generert, og hvordan dataen ble (og fortsatt blir) håndtert på en ustrategisk måte av mange selskaper. Dette fagfeltet og dets kunnskap vil være fundamentalt viktig for arkivfeltet, både gjennom danning og bevaring. Der datakvalitet møter arkiv er et lite utforsket område og dette prosjektet er et forsøk på å forene disse feltene.

En forenklet definisjon av datakvalitet er «Datakvalitet angir i hvilken grad data i et system er i overensstemmelse med det virkelige scenarioet dataen representerer». Manglende datakvalitet er ofte et resultat av dårlig system design eller forbundet med dårlige prosedyrer ved data innførsel.

Datakvalitet måles vanligvis i det som kalles for datakvalitetsdimensjoner. Det finnes mange datakvalitets dimensjoner og fram til dette prosjektet er gjennomført, ingen standard svar på hvilken som gjelder for arkiv. Vi grupperer datakvalitetsdimensjoner i

tre forskjellige kategorier; subjektiv, objektiv og prosess og det er viktig å vite når og hvordan du kan bruke en kategori av dimensjon. Når det gjelder de subjektive datakvalitets dimensjoner så er oppfatningen av hva som er god datakvalitet avhengig av perspektivet til de som bruker dataen. Slike datakvalitets vurderinger er ofte basert på intervjueteknikker og svarene lar seg ofte ikke generalisere. En saksbehandler kan ha en forståelse at dataen i saksbehandlingssystemet er av høy kvalitet utifra sitt perspektiv som bygger på mange års erfaring og bruk av systemet mens en annen saksbehandler kan se på de samme dataene og mene at kvaliteten er dårlig. Et eksempel av denne typen datakvalitet er forståelighets dimensjonen, dvs hvor enkelt det er å forstå informasjonen. Dette er en av de vanskelige dimensjonene utifra depot perspektivet. Objektive datakvalitets dimensjoner vurderes gjennom en uavhengig analyse av dataen i systemet (databasen) og er ofte basert på programvare som går inn i databasen og måler kvaliteten i tabellene. Et eksempel på en slik dimensjon er fullstendighet eller om all relevant data er innhentet og lagret. Det vi mener med uavhengighet er at hvis du kjører en måling to ganger over samme database vil du få samme resultater. Den tredje datakvalitetsdimensjon er prosessdimensjonen hvor egenskaper som feks responstid måles.

Dette prosjektet forholder seg kun til objektive vurderinger av datakvalitet. Målet er å kunne se på enhver Noark 4 avlevering og objektivt måle og angi en datakvalitets vurdering av avleveringen. Ideen bak prosjektet kommer fra et spørsmål partene har diskutert seg imellom, der vi har stilt spørsmål om det lar seg gjøre å måle kvaliteten til elektroniske avleveringer. Videre har vi lurt på om programvare leverandørene har hatt en fokus på datakvalitet under utviklingsarbeidet.

Hva vet vi om datakvalitet og arkiv? Høsten 2010 startet Høyskolen i Oslo og Akershus (HiOA) og Dublin City University (DCU) et selv-finansiert samarbeidsprosjekt for å se nærmere på hvordan disse to fagfeltene kunne dra nytte av hverandre. Dette samarbeidet resulterte i en analyse av datakvalitet for Noark 5 indrekjerne og vi implementerte en demonstrasjonsløsning av teknologien

(<http://ark1.hio.no:8180/n5.ui.dq/>). Denne løsningen kan måle åtte forskjellige datakvalitets dimensjoner med tilbakemelding i form av en grafisk representasjon. Løsningen er implementert til en høy standard og prosjektet utviklet et eget scripting språk som kan brukes til å videreutvikle datakvalitets dimensjoner. Vi lærte mye fra datakvalitets analysen for Noark 5 indrekjerne men vi har kun jobbet med selv generert fiktiv data. Intuitivt mener vi at historien vil vise at kvaliteten på Noark 5 avleveringer vil være betydelig bedre enn kvaliteten på Noark 4 avleveringer, noe vi mener gir et enda sterkere argument til å se nærmere på datakvalitet av Noark 4 systemer.

Selve prosjektet vil bestå av en datakvalitets analyse av et Noark 4 system ved avlevering til depot. Kilden prosjektet baserer seg på kommer fra en av kommunene som avleverer til IKA Kongsberg. Denne kommunene var en av de første kommunene i Norge til å ta i bruk fullelektronisk arkiv og i disse dager forbereder en Noark 4 avleveringer. I samarbeid med IKA Kongsberg har HiOA sett nærmere på database strukturen til denne basen. Denne basen implementerer ikke tabell strukturen til Noark 4 selv om den indirekte lar seg sikkert realisere som en Noark 4 løsning.

Arbeidet som utføres og leveransene i prosjektet beskrives her:

1. Hvilken teknologi kan brukes til å måle datakvalitet? Det finnes en del forskjellige datakvalitets måling verktøy men det er usikkert om de egner seg til Noark 4 avleveringer. Vi vil først se på de forskjellige teknologiene (inkludert vårt eget) og se hvilken teknologi egner seg best for prosjektet. Leveransen her vil være en rapport som beskriver forskjellige verktøy som kan brukes til å måle datakvalitet og begrunne valget prosjektet tar.
2. Hvilken datakvalitetsdimensjoner er relevant for Noark 4? I Noark 5 prosjektet fant vi fram til følgende dimensjoner; fullstendighet, nøyaktighet, konsistens, korrekthet, behandlings forsinkelse og kassasjon. Prosjektet vil se på i hvilken grad disse dimensjonene gjelder for Noark 4 og prøve å finne andre dimensjoner som er relevant. Leveransen her vil være en rapport som beskriver de forskjellige datakvalitetsdimensjoner som er blitt undersøkt og begrunne sin bruk for Noark 4.
3. Datakvalitetsanalyse av uttrekket. Dette arbeidet vil bruke det vi har lært fra de to forrige stegene og utføre datakvalitets analysen av uttrekket. Det finnes ingen datakvalitetsverktøy for Noark avleveringer så vi vil måtte utvikle noe som gjør et

eksisterende verktøy i stand til å analysere en Noark 4 avleveringen. Leveransen her vil være en rapport som beskriver funnene fra datakvalitetsanalysen og publisering av koden som er blitt utviklet for å gjennomføre analysen. All kode utviklet i prosjektet vil bli frigjort som åpen kildekode.

4. Hvordan påvirker selve uttrekksprosessen datakvaliteten til avleveringen? Fra før så vet vi at databasen vi skal bruke ikke følger Noark 4 tabell strukturen så det gjør det veldig interessant å bruke den. Her ønsker vi å se nærmere på om det er noe forskjell mellom datakvaliteten i databasen og datakvaliteten til avleveringen.

Vi vet lite om datakvalitet når det gjelder elektroniske avleveringer til depot. Datakvalitets analyser er ofte utført på data som er i daglig bruk og som ikke har en rigid arkivdanningsprosess som styrer den. Dårlig datakvalitet er ofte en resultat av dårlig prosesser som introduserer feil. Intuitivt vil kvaliteten i en Noark 4 database være relativt bra, men ingen kan vite det før vi forsøker å måle datakvaliteten. Det er også bare i nyere tid at det er blitt satt et søkelys rundt problematikken knyttet til elektroniske avleveringer. Noark 4 og 5 standardene og avleveringsformatene er utviklet utifra arkivfaglige hensyn og har lite innebygget datakvalitets egenskaper og fokuset depot har hatt på elektroniske avleveringer har vært mer på et syntaks nivå. Ved mottakelse så vil depot sjekke at XML filene i avleveringen oppfyller noen semantiske regler når det gjelder struktur. Det kan argumenteres utifra metadata katalogen til Noark 5 og bruken av XSD så sørges det for et minimum av datakvalitet i korrekthets dimensjonen, men det er langt verre for Noark 4. Her brukes det DTD for validering og det er liten innebygd muligheter til å sjekke kvaliteten på dataen som avleveres men kvaliteten kan måles separat. Analysen av Noark 5 som vi utførte med DCU viser at Noark 5 har en begrenset datakvalitets aspekt.

Det kan være at bruken av migrering som bevaringstrategi, eller selve arkivtradisjonen i Norge, der danning og bevarings fasene er adskilt, forsterker effekten av datakvalitets problematikken. Et problemstilling som blir mer relevant for depot nå er hvordan de skal stille seg til en avlevering som har dårlig datakvalitet. Det kan være vanskelig å avvise en avlevering som er i tråd med paragrafene til arkivloven selv om den har dårlig veldig datakvalitet. Hvem skal ta regningen for å rette opp kvaliteten? I verste fall kan depot bare fraskrive seg alt ansvar og skylde på arkivdanningen. Uansett bør depot ha verktøy som kan si noe om kvaliteten til avleveringen. Både OAIS og Trusted Repository Archive Checklist standardene anbefaler et kvalitetssjekk ved mottak av arkivmateriale. Feks TRAC kravet «B1.4 Depotets mottaksprosess sjekker hvert mottatt objekt (SIP) for fullstendighet og korrekthet». Den graden depot kan si noe om riktigheten vil avhenge av hva depot vet om avleveringen og hvilken verktøy som finnes for å verifisere riktigheten. Dette kravet kan tolkes veldig enkelt som feks å sjekke at filformatene faktisk er det de hevder å være (feks TIFF-filer er gyldige eksemplarer av TIFF-format), eller det kan innebære et nærmere sjekk på innholdet. Fullstendighet og korrekthet er ikke definert i TRAC og standarden på mange måter er unnlattende i si noe konkret.

Prosjektgruppen består av HiOA, IKA Kongsberg og KDRS. HiOA har et FoU fokus som strekker fra danning til bevaring som er både teoretisk og praktisk. HiOA jobber fortiden med ferdigstilling av en Noark 5 friprog kjerne og har alt avsluttet et prosjekt der de har sett på datakvalitet av Noark 5 indre kjerne (i samarbeid med Markus Helfert ved DCU) og har de ferdighetene som trengs for å lede og gjennomføre prosjektet. For HiOA sin del er det nå naturlig å se mer på datakvalitet fra depot siden og spesielt se på Noark 4 hvor vi mener de store utfordringene med datakvalitet ligger. HiOA har hatt et samarbeid med Datakvalitets gruppen ved DCU og det er naturlig at samarbeidet fortsetter i dette prosjektet. Helfert er en datakvalitets ekspert som vil bidra med sin ekspertise. IKA Kongsberg stiller med kunnskap om basene vi vil vurdere og representerer kommunenes interesse. KDRS som er fortsatt under oppbygging vil stille med kunnskap om hvilken behov vi har til datakvalitet på nasjonalt nivå og representer kai miljøet.

Ressurser:

Prosjektet vil ha møter på alle institusjoner. Prosjektet ber kulturrådet fullfinansiere Helfert sin bidrag og reiser og begrunner dette med at hans innsats vil være av stor betydning for gjennomføring av prosjektet. Helfert, som kanskje den eneste i verden, har en forståelse av hvordan arkiv og datakvalitet kan forenes og hans bidrag vil sikre en velykket prosjekt. Dette prosjektet er nyskapende og vil ha stor overføringsverdi til

fagfeltet samtidig som det vil generere diskusjon og refleksjon over nåværende praksis. De øvrige partnere finansierer selv sin 50 % andel av arbeidet, men har ikke muligheten til å finansiere Helferts bidrag.

Formidling:

HiOA vil bruke resultatene som en del av sin Bachelor i Arkiv hvor studentene møter temaet i Digital Depot med en teoretisk og praktisk undervisning i datakvalitet fra et depot perspektiv. Det er sannsynlig at dette også på sikt kan være interessant å tilby som en del av undervisningen på masters nivå. Arbeidet er av et slik natur at det er sannsynlig at det kan publiseres som tidsskrift artikler. IKA kongsberg og KDRS vil bruke resultatene til øke bevisstheten internt i organisasjonen og ute blant sine medlemmer. De vil også ta i bruk kildekode som prosjektet utvikler. Prosjektet vil blogge om sin progresjon på <http://www.arkivlab.no>. Denne siden er under oppbygging og er en publiseringsløsning bygget på Drupal.

Framdriftsplan

Framdriftsplan for hele prosjektperioden

Startdato	05.03.2012
Sluttdato	08.03.2013

Aktivitetplaner

Tittel	Fra dato	Til dato
Verktøyanalyse	05.03.2012	04.05.2012
Analyse av relevant datakvalitetsdimensjoner	07.05.2012	07.09.2012
Utførsel av datakvalitetsanalyse	10.09.2012	11.01.2013
Analyse av av uttrekksprosessen	14.01.2013	08.03.2013

Andre opplysninger / kommentarer om framdrift- / aktivitetsplan.

Samarbeidspartnere

Har du en samarbeidspartner?

Ja

Samarbeidspartner 1

Navn
Formell avtale?
Postadresse
Postnr / Poststed
Kontaktperson
Tlf. til kontaktperson
E-post til kontaktperson
Rolle

Kommunearkivinstitusjonenes Digitale Ressurssenter
Ja
Maskinistgata 1,
7042 Trondheim
Tor Eivind Johansen
94790707
tor.eivind.johansen@kdrs.no
KDRS har både bruker og formidlings rolle. De vil være en del av analysearbeidet men også ha en rolle til å hente inn innspill og viderebringe resultater til KAI miljøet

Samarbeidspartner 2

Navn
Formell avtale?
Postadresse
Postnr / Poststed
Kontaktperson
Tlf. til kontaktperson
E-post til kontaktperson
Rolle

IKA Kongsberg
Ja
Nansengate 1
3616 Kongsberg
Wenche Risdal Lund
99279615
wenche@ikakongsberg.no
IKA Kongsberg har både en bruker og formidlings rolle. De vil være en del av analysearbeidet men også bruke resultatene internt.

Prosjektbudsjett

Prosjekt

Ettårig eller flerårig prosjekt? Ettårig

BUDSJETT FOR FØRSTE ÅR

Utgifter

Hva	Beskrivelse	Sum
Personalkostnader (lønn/overhead)	HiOA lønn inkl overhead	kr 304 335,00
	KDRS lønn inkl overhead	kr 95 000,00
	IKA Kongsberg lønn inkl overhead	kr 87 500,00
Varer/utstyr		kr 0,00
Tjenester kjøpt av andre	Markus Helfert inkl moms	kr 81 250,00
Andre utgifter (spesifiser)	Prosjektmøte reise	kr 98 000,00
Totale utgifter		kr 666 085,00

Inntekter

Beskrivelse	Tilskudd bekreftet?	Sum
<u>Søknadssum Norsk kulturråd</u>		<u>kr 385 167,00</u>
HiOA egenfinansiering	Ja	kr 152 168,00
KDRS egenfinansiering	Ja	kr 47 500,00
IKAK egenfinansiering	Ja	kr 43 750,00
HiOA reiseegenfinansiering	Ja	kr 12 500,00
KDRS reiseegenfinansiering	Ja	kr 12 500,00
IKAK reiseegenfinansiering	Ja	kr 12 500,00
Totale inntekter		kr 666 085,00

Kommentar til budsjett neste år

Markus Helfert kommer inn under HiOA, dvs han fakturerer oss. Han er ikke en egen partner i prosjektet, men vi synes det er viktig å kjøpe inn hans erfaring så vi har lagt han til som "Tjenester kjøpt av andre"

Vedlegg

- Bekreftelse_KDRS_HIOA_datakvalitet.docx
- ikakonsberg_prosjektdeltagelse.pdf