Sluttrapport

Forbedring av tjeneste for import av publikasjonsdata

Prosjektnummer: 521015	*Journalnummer:* ---	
Behandlet dato: 13.05.2022	*Behandlet av / Prosjekteier:* Sikt / Frode Arntsen	*Utarbeidet av* Terje Hellesvik (prosjektleder)
Beslutning: Behandlet i prosjektets styringsgruppe – godkjent for videresending til Digitaliseringsstyret.		
Signatur ved godkjenning (prosjekteier) Frode Arntsen (Sign)		

Innhold

ENDRINGSLOGG

Versjon	Dato	Endring	Produsent	Godkjent
0.9	11.05.2022	Første versjon av sluttrapporten til behandling i styringsgruppen	Terje Hellesvik (prosjektleder)	Frode Arntsen (prosjekteier)
1.0	20.05.2022	Endret etter behandling i styringsgruppen	Terje Hellesvik	Frode Arntsen

DISTRIBUSJONSLOGG

Versjon distribuert	Dato	Navn
0.9	10.05.2022	Styringsgruppen for Cris/NVA
1.0	20.05.2022	Saksfremlegg for Digitaliseringsstyret

1 Innledning

Dette prosjektet, organisert som et delprosjekt under Cris/NVA-prosjektet, har hatt som primærmål å rette feil og forbedre funksjonalitet for importfunksjonen knyttet til dagens Cristin-tjeneste samt legge til rette for bruk knyttet opp mot Cris/NVA-tjenesten der det er relevant og mulig. Denne rapporten omtaler kun den avgrensede delen av arbeidet med import av publikasjonsdata som er finansiert av de felles investeringsmidlene bevilget av Digitaliseringsstyret i juni 2021. I tillegg til arbeidet som er omtalt her utføres det omfattende oppgaver knyttet til import av publikasjonsdata i den nye Cris/NVA-tjenesten.

2 Bakgrunn for prosjektet

For at universitet og høgskoler (brukerinstitusjonene) skal oppleve en effektiv og rasjonell bruk av den nye Cris/NVA-tjenesten, er det en forutsetning at det foreligger en velfungerende importløsning for publikasjonsdata (sentralimport). En velfungerende importløsning vil gi gevinster i form av redusert årsverksforbruk ved institusjonene og bedre kvalitet på data i forskningsinformasjonstjenesten. Importløsningen som ble tatt i bruk sommeren 2020 opplevdes gjennom erfaringer fra NVI-rapportering for 2020 å være mangelfull. Dette medførte en betydelig mer ressurskrevende NVI-rapportering for 2020, enn om importtjenestene hadde fungert optimalt. Sak 25/21 til Digitaliseringsstyret gjorde rede for behovet og ba om en finansiering på inntil 2,5 mnok for å forbedre importtjenesten for publikasjonsdata. Finansieringen ble vedtatt i DS-møte den 9 juni 2021. Prosjektet ble gjennomført som et delprosjekt i Cris/NVA-prosjektet med samme prosjektleder og styringsgruppe som dette prosjektet. Dette ble valgt fordi tjenesten i fremtiden skal fungere som en integrert del av Cris/NVA-tjenesten.

Cristins brukere og deres brukerinstitusjoner er svært opptatt av at denne importfunksjonen fungerer så godt som mulig og automatiserer import av publikasjoner så langt det er mulig. Tidligere løsning innebærer mye manuelt arbeid, medfører lang ventetid på importerte publikasjonsdata og har delvis mangelfull kvalitet.

Tjenesten for sentral import av publikasjonsdata, som er en tjeneste adskilt fra, men integrert i dagens Cristin, ble satt i drift i versjon 2 sommeren 2020. Tjenestens formål er å hente kvalitetssikrede data til Cristin og kunne levere disse til den enkelte forsker/institusjon slik at de ikke trenger å egenregistrere alle poster. Data hentes fra tredjepartsleverandør, for tiden Scopus (Elsevier).

Ved planlegging av NVA/CRIS-prosjektet i 2019 ble det forutsatt at importtjenesten som skulle tas i bruk i 2020 var velfungerende og med begrensede ressurser kunne integreres med den nye løsningen til erstatning for dagens integrasjon med Cristin. Dette viste seg å ikke være tilfelle.

Basert på en kritisk gjennomgang og vurdering av tidligere tjeneste med tilhørende manglende funksjonalitet og uttrykte ønsker for hvordan tjenesten må fungere, ble det med grunnlag i en dialog med relevante brukere foreslått å:

- Etablere en mer automatisert arbeidsflyt i tjenesten, som da raskere leverer publikasjonsdata til bruker/brukerinstitusjonene.
- Utvide tjenesten til å benytte flere kilder for å heve kvaliteten (PubMed/WoS eller evt andre kilder) på den enkelte publikasjon.
- Benytte flere kilder (se forrige punkt) for å inkludere et større antall publikasjoner i importen til Cristin-tjenesten.

3 Prosjektets mål

Prosjektets mål var å gjøre forbedringer på eksisterende tjeneste for import av publikasjonsdata knyttet til dagens Cristin-tjeneste, slik at det kunne oppnås høyere kvalitet på hver metadatapost og redusert arbeidsinnsats per importpost. I tillegg var det et mål å øke dekningsgraden innen flere fagområder.

Prosjektets mål	*Vurdering av om det sannsynligvis vil oppnås*	*Forklaring på ev. avvik*
Raskere leveranse av publikasjonsdata til bruker/ brukerinstitusjonene.	Det er gjort en rekke feilrettinger og innført mer/bedre automatisering i applikasjonen. Dette vil gjøre at tiden på å prosessere hver importpost går ned. Forfattere som har blitt matchet blir automatisk gjenkjent ved senere importer. Det er gjort store forbedringer i brukergrensesnittet generelt.	
Heve kvaliteten på den enkelte publikasjons metadata	Det er gjort forbedringer i brukergrensesnittet som gjør det lettere å velge riktig forfatter.	
Øke dekningsgraden på publikasjoner som importeres	Ikke gjennomført	Det er gjort en vurdering av PubMed. PubMed inneholder ikke IDer på forfatter og institusjon. Det gjør det vanskelig med automatisert import.

3.1 Hastighet og kvalitet på importtjenesten

Med utgangspunkt i at prosjektet skulle rette feil og gjøre forbedringer i koden på den applikasjonen som benyttes for import av publikasjonsdata til dagens Cristin-tjeneste samt om mulig legge til rette for import av publikasjonsdata til den nye Cris/NVA-tjenesten er målet nådd. Det er gjennomført omfattende feilrettinger i eksisterende kode og gjort betydelige endringer og utvidelser av tjenestens funksjonalitet.

Ved gjennomføring av NVI-rapportering av 2020 data i 2021 var det et betydelig antall metadataposter fra 2020, 2019 og 2018 som fortsatt ikke var utført grunnet feil og mangelfull funksjonalitet i import-applikasjonen.

Ved rapporteringsdatoen 1 april 2022, der en rapporterte 2021-publikasjoner var samtlige publikasjoner fra foregående år, inkludert 2021 importert til Cristin. Dette har ikke skjedd tidligere.

Ettersom det normalt ikke har blitt importert nye publikasjoner, dvs årets (2022-utgivelser) til Cristin før etter 1 april (skyldes fare for duplikater og overskrivinger ifm rapporteringsperioden) har det tidligere vært et kontinuerlig etterslep gjennom resten av året og en har som nevnt ikke komet ajour før rapporteringsfristen påfølgende år.

Pr 12 mai 2022 forelå det totalt 7481 publikasjoner utgitt i 2022 med norske forfattere klare for import. Samme dato var det allerede importert 5324 publikasjoner til Cristin og Sikt forventet å være ajour før sommerferien 2022. Dette innebærer med all sannsynlighet at vi resten av året vil være så nær ajour med import av tilgjengelige publikasjoner i Cristin som mulig.

Dette er et resultat av feilrettinger og forbedringer som har øket hastigheten og kvaliteten på import av publikasjonsdata.

3.2 Dekningsgrad på importerte data

Prosjektet hadde også et mål om å kunne bedre dekningen innen alle fagfelt. Det har tradisjonelt vært lavere dekning i importerte publikasjoner innen humaniora og samfunnsvitenskap sammenlignet med øvrige fagfelt. Det har også vært et tydelig uttalt ønske om ytterligere forbedret dekning innen helse/medisin ved å gjøre data fra Pubmed tilgjengelig i Cristin. Til forskjell fra data fra Scopus som er kilden for dagens import benytter ikke Pubmed egne identifikatorer på personer og organisasjoner, men derimot generiske ID som for eksempel ORCID på personer. Dette er ideelt sett en bedre løsning enn Scopus som benytter en egen Scopus-ID. For Scopus-data har vi imidlertid over lang tid matchet Scopus ID mot dagens person-ID i Cristin lik at dette normalt fungerer relativt bra.

Utfordringen med ORCID er så langt at det fortsatt er et relativt høyt antall norske forskere som ikke har registrert seg i ORCID. Dette vil imdlertid avhjelpes de nærmeste årene når forlagene i økende grad krever ORCID av forfatterne samt at forfatterne bes om å registrere sin ORCID første gang de logger inn i den nye Cris/NVA-tjenesten. Når det foreligger god ORCID-dekning blant norske forskere innen Helse/Medisin og publikasjonene benytter ORCID på forfatterne ligger det godt til rette for å importere fra Pubmed.

Scopus har dårlig dekning for humaniora og samfunnsvitenskap. Innenfor dette prosjektets rammer har det ikke latt seg gjøre å utforske/etablere nye datakilder som kan bedre denne dekningen. Sikt har en forventning til at dataleverandøren *Dimensions* har bedre dekning innen disse fagområdene.

Det er planlagt møte med *Dimensions* i løpet av mai 22 for å avklare muligheten for en utprøving av API for import av data fra *Dimensions*, der hensikten er å avklare tekniske forhold relatert til vår importtjeneste samt denne leverandørens dekningsgrad innen de forskjellige fagområdene. Basert på de opplysningene vi ha pr i dag er det store forventninger til en bedre dekningsgrad enn med dagens datakilde. Avklaringer rundt dette forventes i løpet av høsten 2022.

Arbeidet med å forbedre importtjenesten for bruk i Cris/NVA-tjenesten videreføres som en del av det prosjektet.

4 Prosjektets økonomi

Det ble tildelt inntil 2,5 MNOK fra de felles investeringsmidlene for å gjennomføre prosjektet. Midlene er i sin helhet benyttet til utviklingsaktiviteter, inkludert løpende spesifiseringsarbeid. Beløpet er i sin helhet benyttet til lønnsmidler.

5 Prosjektets fremdrift

Etter avklaringer rundt utviklingsressurser ble arbeidet startet i uke 41/21. Det er rullet ut ny versjon av sentralimport annenhver torsdag, til sammen 10 utrullinger frem til midlene i sin helhet er benyttet ved utløpet av april 22.

6 Avvik i prosjektet

Samtlige planlagte aktiviteter er gjennomført, men grunnet forhold ved eksterne datakilder (PubMed) viste det på nåværende tidspunkt ikke mulig å importere data fra denne kilden.

7 Prosjektets anvendelse av IT-politiske prinsipper og føringer

Prinsipper for realisering av digitalisering	Den faktiske implementeringen i dette prosjektet
Sett brukeren i sentrum.	Behov og brukerhistorier kvalitetssikret med faktiske brukermiljøer som også ble involvert i testing. De samme brukerne tok løsningene fortløpende i bruk.
Tenk stort, start smått gjennom smidig utvikling. Prototyping og utprøving foran utredning.	Det ble brukt smidig utviklingsmetodikk og tverrfaglig team gjennom hele prosjektet.
Data lagres kun én gang og skal gjøres tilgjengelige for gjenbruk.	Ivaretatt. Gjenbruk av registrert informasjon er ett av de langsiktige målene for gevinst av prosjektet.
Bygg inn sikkerhet og personvern i løsningene.	Ja, implementert i form av tilgangsstyring.
Sikre kontroll på tilgang til data og ressurser.	Ja, det ble implementert tilgangskontroll til APIene.
Sky først:	Ja, Amazon Web Services benyttes der relevant.
Offentlig sektor skal i utgangspunktet ikke gjøre selv det som markedet kan gjøre bedre og mer effektivt	Det finnes ikke hyllevareprodukt i markedet som dekker prosjektets leveranser.

8 Evaluering av prosjektets styring

Prosjektet ble støttet av Sikts ledelse ved omprioritering av utviklingsressurser for å kunne gjennomføre prosjektet for å gi effekt allerede for NVI-rapportering av resultater for 2021.

Prosjektet ble organisert som et delprosjekt i Cris/NVA-prosjektet da det finnes kompetanse i styringsgruppen og prosjektgruppen for gjennomføring av delprosjektet samt at det har en nær tilknytning til funksjonalitet i Cris/NVA. Styringsgruppen er holdt løpende orientert om fremdrift og resultater.

Det er benyttet utviklingskapasitet som ikke har påvirket Cris/NVA-prosjektets fremdrift, men snarere gitt flere utviklingsressurser innsikt i- og forståelse for Cris/NVA-tjenesten og dens funksjoner.

8.1 Interessenter og behovskartlegging

Alle institusjoner og deres brukere som benytter dagens Cristin-tjeneste er interessenter. Deres klart uttrykte behov er at metadata for vitenskapelige publikasjoner importeres til Cristin så raskt som mulig og med så god kvalitet som mulig. I dagens versjon av importtjenesten er det ansatte ved Sikt som er direkte brukere av tjenesten, dette vil endres ved driftssetting av Cris/NVA-tjenesten i 2023, da også institusjonenes brukere kan benytte tjenesten.

8.2 Prosjektplan

Prosjektet ble planlagt for og gjennomført med en smidig tilnærming. Det ble laget og prioritert oppgaver i to-ukers sprinter, med påfølgende planlegging og utrulling av ny versjon av programvaren hver andre uke.

8.3 Utvikling- testing og produksjonssetting

Gjennomføringen har skjedd i nært samarbeid med de deler av linjeorganisasjonen som er brukere av importtjenesten. Disse har derved kunnet gi løpende innspill på hvordan ny funksjonalitet og feilrettinger har gitt effekt og påvirket målsettingene. Dette samspillet har vært helt avgjørende for det arbeidet som er gjort på importtjenesten.

Den importtjenesten som ble rullet ut sommeren 2020 var en minimumsversjon, og var forutsatt å kunne benyttes fullt ut også for den kommende Cris/NVA-tjenesten. Det oppsto imidlertid raskt misnøye med kvaliteten på importtjenesten ved gjennomføring av NVI-rapporteringen for 2020. Samtidig ble det ved nærmere gjennomgang usikkert om hvorvidt den kunne gjenbrukes i Cris/NVA.

Gjennom nært samarbeid mellom de som benytter tjenesten i det daglige, prosjektleder (samme som i Cris/NVA) og utviklerne ble det raskt identifisert vesentlige feil. Utviklerne gjennomførte omfattende feilrettinger og fikk god innsikt i applikasjonen.

Det viste seg mulig å gjenbruke større deler av eksisterende applikasjon etter feilrettinger og det er utviklet ny funksjonalitet som legger til rette for effektive og rasjonelle importprosesser i fremtiden.

8.4 Usikkerhet

Det ble ikke gjennomført egen ROS-analyse for delprosjektet, da det i hovedsak omfatte feilretting og introduksjon av ny funksjonalitet innenfor svært begrensede rammer. Der det var relevant ble det behandlet som del av usikkerheter i Cris/NVA-prosjektet.

8.5 Prosjektets rammebetingelser

Prosjektets behov for utviklingsressurser var innledningsvis en utfordring av samme grunn som de utfordringer øvrige utviklingsprosjekter er stilt overfor – manglende tilgang. Når dette løste seg ved å omprioritere interne ressurser i Sikt har betingelsene ligget godt til rette for gjennomføring av prosjektet.