Utkast: Prinsipper for utlevering av data

Dokumentet er en oppdatert versjon av det Styringsgruppen til NIB fikk fremlagt 10. oktober, med innarbeidede kommentarer og presiseringer.

# Innledning

Formålet med NIB er å gjøre bibliometriske data tilgjengelig. Dette dokumentet legger opp til en diskusjon rundt hvordan dette skal utføres

NIFU og andre forskere vil få tilgang til alle detaljerte rådata, mens alle andre i de nesten 200 institusjonene som NIB-konsortiet omfatter skal få tilgang til definerte datasett.

NIB har begrensede ressurser og er ikke den som kjenner behovene. Vi foreslår derfor en modell som kan forenkle utvikling av tjenesten ved å engasjere NIB-konsortiet direkte i utviklingen. Samtidig ønsker vi å fremme viktige prinsipper i bruken av bibliometri, slik som at alle bruker like data og i prinsippet kan sammenligne sine analyser. Vi ønsker å stimulere til samarbeid rundt utviklingen av bibliometri.

Dette må sees i sammenheng med utvikling av verktøy og analyseløsninger. Når dataene har lik utforming, vil det være enklere å utveksle ulike analyseverktøy og visualiseringer. Dette omtales nærmere i dokumentet «Prinsipper for fremtidig utvikling av NIB-tjenester».

En forutsetning for å utvikle datasett er at dataene blir tilrettelagt og vasket. Dette utføres i et eget løp som startet med en planleggingsrunde i Bibliometriutvalget 2. oktober, hvor Bibliometriutvalget ble enig om å samles til en todagers workshop i januar. NIB etablerer en liten arbeidsgruppe som dypdykker i dataene og forbereder workshopen.

# Datasett

Med datasett menes det her en samlet fremstilling av bibliometriske data fra NIB. I dette dokumentet skiller vi på tre former for datasett: spesialuttrekk, felles standardiserte datasett og testsett. I tillegg gir lisensen oss rett til å sette opp en web-service som kan gi tilgang til bearbeidede datasett for de som er utenfor konsortiet, men det behandles ikke her. NIB-konsortiet består av alle forskningsinstitusjoner i Norge som har en offentlig finansiering, samt departement, direktorater og andre statlige organ.

## Tilpassede datasett

Forskningsprosjekter kan få full tilgang til alle detaljerte rådata. Dette kan gjøres på to måter. Enten kan forskerne få direkte tilgang til rådataene i basen, gitt at de har den nødvendige kunnskapen (SQL), eller ved at de får et uttrekk tilpasset det formålet og de ønsker forskerne har. I lisensen er det en begrensning på 10 samtidige brukere av rådataene, for å holde prisen på dataene lavest mulig. Dette gjelder tilgang direkte på NIB-serveren og slik lisensen er innrettet anser Styringsgruppen at dette ikke vil ha noen praktiske konsekvenser i overskuelig fremtid.

## Felles standardiserte datasett

Lisensen på WOS-dataene gir alle institusjoner fri tilgang til datasett som har en eller annen form for aggregering eller er et utvalg av dataene. Aggregering og utvalgskriterier vil i de fleste sammenhenger representere en fordel. Spesielt ved at selektering, beregninger og aggregeringer vil være utført på server. Det påvirker også størrelsen av datasettet og kan forenkle behandlingen av dataene.

Ved å sikte mot godt gjennomarbeidede datasett vil vi kunne sørge for at lisensvilkårene utnyttes maksimalt til vårt beste. Ved å legge vekt på målet om å skape gode datasett til reelle formål, vil vi ikke behøve å diskutere grensene for hva det vil si å aggregere i henhold til lisensen. Det er formålet som blir styrende og ikke en teoretisk øvelse på hvor langt vi kan strekke strikken.

Gode datasett som ikke blir for store kan behandles i excel eller andre lett tilgjengelig programvare, slik at terskelen for å ta i bruk dataene kan gjøres lav som mulig.

Hvert datasett kan ha flere filer delt opp på land, institusjoner, år og andre relevante inndelinger ut fra formålet, i tillegg til en masterfil hvor alle data er samlet. Felles standardiserte datasett vedlikeholdes av NIB.

*Hovedpoenget i dette dokumentet er å argumentere for en modell hvor vi utformer disse datasettene i fellesskap.*

## Testsett

Med testsett menes det et datauttrekk som er egnet til å utvikle et datasett som beskrevet ovenfor. Et testsett skal ikke brukes til analyser som publiseres og dataene vil ikke bli vedlikeholdt. Det jobbes med å lage en god beskrivelse av datamodellen. En liten introduksjon vil bli gitt på Bibliometriforum 16. oktober.

## Utlevering av datasett

Datasett utleveres via en enkel innlogget filserver/nettside, hvor alle datasett ligger tilgjengelig klar til nedlastning for alle institusjoner i hele konsortiet. Filformatet for datasett antas inntil videre å bli .csv, som kan leses av de aller fleste aktuelle verktøy for å behandle dataene.

Det å legge ut datasett kan håndteres manuelt i starten, mens prosessen med å utvikle en god automatisert løsning kan skje parallelt. Her prioriteres tilgang på datasett øverst på listen, siden det uansett vil kreve noe tid å få på plass en god løsning. Ferdig løsning vil kunne inkludere API og søkefunksjoner, med mer.

Ut fra budsjettet til NIB vil det ikke være kapasitet til å kunne levere ut tilpassede datasett i stor skala basert på enkelthenvendelser. Utlevering av «tilpasset datasett» til enkeltinstitusjoner vil både kreve mye kompetanse hos bestiller og i Unit. Det vil alltids være behov for dialog og ved stor pågang vil spørsmål i seg selv kunne ta opp hele kapasiteten og ordningen vil dermed bli sårbar.

Det å sette opp en web-tjeneste som fortløpende kan levere ut tilpassede datasett direkte fra basen vil være ressurskrevende med tanke på serverkapasitet og vil være en tjeneste som det vil ta lang tid å utvikle.

# Utforming av felles standard datasett

Den som skal utføre en analyse er den som er best egnet til å definere databehovet. Samtidig vet vi at det er stor sannsynlighet for at det er flere som skal gjennomføre tilsvarende analyse. Det å definere et datasett er både kompetansekrevende og ressurskrevende, og vi risikerer dobbeltarbeid. Vi risikerer også at «like analyser» ikke kan sammenlignes fordi de bruker ulike datagrunnlag.

På sikt kan NIB bygge standard datasett basert på de erfaringer vi får med ulike henvendelser. Det vil forenkle utleveringen, men løsningen drar med seg ulempen som følger av et lite organisert samarbeid.

Fordelen med å bygge felles standard datasett sammen, hvor alle aktuelle institusjoner kan bli involvert, er mange. Den viktigste er kompetansebygging. Ved at man starter på de enkleste og mest opplagte datasettene, vil alle kunne høste nytten tidlig. De med minst kapasitet innen bibliometrifeltet får da også en mulighet til å henge med i utviklingen, slik at et avansement til mer komplekse og eventuelt mer «sære» datasett får større anvendelse enn det ellers ville fått.

En ulempe med å satse på å utvikle felles standard datasett er at NIB vil kunne fremstå som lite fleksibel. Dette kan vi kompensere med å være åpen i prosessene og gi alle tilgangen til «testsett» for å kunne utvikle forslag til felles datasett.

En enkel «portvokter» for å gi en sunn utvikling av testsett er å stille noen enkle krav til forespørselen for å få utlevert et spesialtilpasset testsett. Dette går på det å be bestiller om å beskrive formålet og andre opplysninger som skal med til endelig datasett. For det andre kan vi gi koordinerte henvendelser prioritet, for å stimulere til samarbeid.

# Arbeidsfordeling

## Bibliometriutvalget

Forslaget er at Bibliometriutvalget blir de som vedtar hvilke datasett som skal utvikles, prioriterer utviklingen og kvalitetssikre datasettene.

## Bibliometriforum

Bibliometriforum er foreløpig bare løst organisert en happening for bibliometriinteresserte i Norge, med en programkomite som arrangerer treffene. Forumet er initiert av NIB ut fra den interessen som ble vist da NIB arrangerte testseminar i forbindelse med anbudskonkurransen for kjøp av bibliometriske data og bibliometriske verktøy. Behovet for informasjon var stort og mange hadde innspill til arbeidet i NIB.

Styringsgruppen for NIB har tidligere gitt et mandat til å formalisere Bibliometriforum som en del av NIBs aktivitet. Det siktes mot at formalisering kan skje våren 2020.

Bibliometriforums rolle i utviklingen av felles standard datasett er å bli et verksted for det store bibliometriske felleskapet, hvor ideer unnfanges, møtes og materialiseres. Dette vil i første fase være en arena for samarbeid og kompetanseheving, med fokus på å kunne utvikle felles standard datasett, men også andre formål innen bibliometrien.

En preferert metode for å utvikle datasett kan være at små arbeidsgrupper eller enkeltpersoner utarbeider forslag som presenteres på Bibliometriforum og diskuteres der. Det kan også arrangeres egne workshops hvor disse gruppene samles for å jobbe sammen og hvor Unit stiller med kapasitet til å jobbe direkte i NIB-basen for å justere på små og store utfordringer og Bibliometriutvalget stiller med kompetanse til å utarbeide ferdige forslag og kvalitetssikre arbeidet.

## Kvalitetskontroll

Felles standard datasett vil kunne underlegges en viss kvalitetskontroll. Funksjonen bør tillegges Bibliometriutvalget, men her får Dag Aksnes og NIFU en sentral rolle i å gi veiledning.

Bibliometriutvalget må selv utforme hvordan kvalitetskontrollen skal utformes og hvilket «godkjenningsstempel» som skal gis. Det er allerede fremmet noen tanker rundt dette. Ut over det at datasettet teknisk sett er riktig og lar seg bruke, bør dokumentasjonen bli viktig del av godkjenningen. Spesielt det å få tydeliggjort hva dataene er egnet til å fortelle noe om og hva dataene *ikke* er egnet til å fortelle noe om, er viktig.

Det å gi tydelige veiledninger til de som skal bruke bibliometrien vil kunne bidra til å utvikle sunne holdninger til resultater fremskaffet fra bibliometriske analyser.

## Felles bibliometrisk rapport

Samarbeid for å utvikle datasettene vil gi en naturlig inngang til å samarbeide om analysene. Dag Aksnes «kongstanke» på å utvikle en felles bibliometrisk rapport i forskningssektorene, vil kunne inkludere samarbeid i alle ledd av prosessen. Sammenhengen fra det å utvikle datasett, gjennomføre en felles analyse og publisere funnene vil bli tydelig.

Fra NIBs perspektiv vil et samarbeid om en felles rapport i forskningssektorene gi store fordeler. Behovene som vi skal ta tak i vil kunne fremstå som entydige, koordinerte og helhetlig, og det blir lettere å identifisere hva som bør prioriteres. Institusjonene vil sammen kunne identifisere behov for vask og tilrettelegging av data. Videre vil institusjonene kunne samarbeide rundt vask av dataene. Hvis alle institusjoner vasker «egne data» vil institusjonene i stor grad kunne sies å ha bidratt til fellesskapet på like vilkår og kunne høste på like vilkår, ut fra institusjonens størrelse. Dette vil gjøre NIB bedre og det vil gå raskere enn om dette skulle gjøres fra sentralt hold.

Felles rapport vil også bety at det vil være flere arenaer for samhandling og kompetansebygging.

# Diskusjon av prinsipper for utvikling av felles standard datasett

For å sette alt fokus på utvikling av felles datasett er det ønskelig med en hovedregel som sier at det ikke utleveres spesialtilpassede datasett, med unntak til forskningsprosjekter. Etter hvert som det kommer på plass flere gode felles datasett er det grunn til å tro at flere forskningsprosjekter vil få dekt sitt behov gjennom disse datasettene. Dette kan tale for at definisjonen av forskningsprosjekt bør tolkes meget strengt i starten (ut 2020?), slik at de som vil forske styres mot det å styrke arbeide med å utvikle felles datasett, mens det så raskt som mulig etter dette åpnes for en større fleksibilitet og lekenhet.

NIB utleverer testsett med data fortløpende i henhold til Bibliometriutvalgets prioriteringer og dernest koordinerte henvendelser. Et «testsett» med data skal ikke brukes til «skarpe analyser» og vil ikke holdes oppdatert av Unit.

Felles standard datasett vil som hovedregel holdes oppdatert av Unit og kunne brukes til analyseformål når som helst.

Arbeid som er organisert gjennom Bibliometriforum og av Bibliometriutvalget gis prioritet for henvendelser til NIB.

Bibliometriutvalget gis ansvaret for å koordinere og prioritere innmeldte behov for datasett. Behovene tas imot i alle relevante fora, men det oppfordres til at Bibliometriforum brukes som den prefererte kanalen. NIB fører logg over registrerte behov, status, steg i prosessen og prioritering.

Bibliometriutvalget behandler og godkjenner endelige datasett med dokumentasjon. Bibliometriutvalget definerer krav til dokumentasjon og oppdaterer eksisterende dokumentasjon ved behov.

NIB legger ut datasettet med dokumentasjon og vedlikeholder datasettene.

# Andre forhold til videre diskusjon

Det er flere forhold som helt klart grenser opp til de prinsipper som drøftes i dette dokumentet og må drøftes i forlengelsen av dette dokumentet. For eksempel det å ta på seg oppdateringsansvar for nye kilder vil kreve flere ressurser. Nedenfor belyses to forhold, som henger sammen ved at de begge gjelder en forpliktelse til fremtidig oppdatering av nye kilder og koblinger og vil kreve flere ressurser.

## Kobling av data

Den jobben som NIFU nå gjør med å koble WOS-data mot Cristin-data er ressurskrevende, men det er opplagt at det er NIB som for fremtiden skal holde en slik kobling oppdatert og gjøre den tilgjengelig for hele konsortiet. Her kan det være flere åpenbare og like situasjoner, men også tilfeller hvor dette blir spørsmål om behov og nytte opp mot ressurser, kompetanse, eierskap til data og andre forhold.

Kanalregisteret tar seg i dag av en oversikt over Gold Open Access-tidsskrifter, men det finnes flere kilder som opererer på artikkelnivå. Hvis slike kilder er aktuelle for flere institusjoner bør for det første NIB kunne høste av det «foregangsinstitusjoner» gjør av arbeid, men også bidra tilbake ved å overta oppdateringsansvaret der det er gunstig på vegne av fellesskapet.

## Data til forskning

Det bør drøftes om datasett til forskning skal legges ut tilgjengelig for hele konsortiet og om det skal stilles krav om dokumentasjon, slik at vi bidrar til åpenhet i forskningen. Videre må det drøftes verdien av å sikre fellesskapet tilgang til «ferdige datasettet» når forskningen er publisert og en eventuell forpliktelse fra NIB til å vedlikeholde disse dataene. Det vil si at det tilrettelegges i NIB-databasen for å kunne ta imot tilpasninger som forskere gjør i dataene, herunder koblinger til nye kilder.