Utkast: Prinsipper for fremtidig utvikling av NIB-tjenester

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

# Innledning

Dette dokumentet trekker opp noen forslag til hvordan vi kan få mest mulig nytte av dataene i NIB ved å stimulere til bruk av fri programvare og deling av kode (verktøy). Siden NIB-dataene vil være like for alle i konsortiet, vil det ikke være noe som teknisk sett hindrer en slik utvikling.

I dette dokumentet brukes begrepet «verktøy» oftest synonymt med «programkode», men er ikke avgrenset til den betydningen. Når begrepet «visualisering» brukes eksplisitt er det med tanke på at noen datasett kan lastes direkte inn i et åpent verktøy for visualisering (f.esk. VOSviewer) og behandles der.

Alle i konsortiet kan i prinsippet ta ut datasettene og bruke de akkurat slik de vil. De minste datasettene vil fungere helt greit i Excel, men dataene kan ha større informasjonsverdi i godt tilpassede analyseverktøy med gode visualiseringer, slik som InCites og SciVal tilbyr.

I budsjettet til NIB finnes det ikke rom til å sette opp et eget utviklingsteam som kan skape nye tjenester på NIB-dataene. Det er samtidig lite realistisk å tro at vi kan skape tjenester som kan gjøre de profesjonelle bibliometriske verktøyene overflødige, bare vi får penger nok. Det vi derimot har mulighet til er å stimulere til at institusjonene i fellesskap utvikler verktøy og visualiseringer som kan løse viktige, vanskelige eller hyppig forekomne oppgaver for institusjonen, i fritt tilgjengelige programvarer.

## Ny modell

Nedenfor finner dere to figurer. Den til venstre er en tidlig modell på hvordan vi tenkte at NIB skulle bli tilgjengelig for brukerne. Den til høyre er en revidert modell som illustrerer konseptet slik det beskrives i dette dokumentet. Fokus dreies bort fra egenutvikling av analyseserver, egne verktøy og faste rapporter, til å fokusere rendyrket på datasett og legge til rette for at institusjonene utvikler verktøy sammen, i R eller andre tilsvarende fri programvare.



Den åpne serveren/web-løsning (nederst i figuren) som skal være tilgjengelig for alle kan benytte både de samme datasettene og de samme R-løsningene som NIB-konsortiet gjør. Vi plukker bare de beste bitene og lager nettsider av det. Datasett må bare aggregeres på et høyere nivå enn det vi gjør internt i konsortetiet.

## NIBs koordinerende rolle

NIBs rolle i en utvikling med bruk av åpen programvare vil være det å stimulere til denne type utvikling og bli et koordinerende ledd. Vi vet at flere jobber med relativt like problemstillinger, men i ulike verktøy og gjør egne tilpasninger i data. Dette gir mye unødig dobbeltarbeid og det er få som kommer seg videre til å løse større oppgaver. Den koordinerende rollen handler om å skape et fellesskap hvor det er naturlig å dele ressurser og løsninger, mer enn å styre utviklingen. De som har en problemstilling er de som best vet hvordan den løses, men vi kan sørge for at løsningen blir kjent slik at andre både kan bruke den og bygge videre på den.

Prosjektet med datadeling i kunnskapssektorene vil på sikt kunne gi et annet tilbud enn det som omtales her. Dette gjelder spesielt kobling til andre datakilder. Hvordan dette påvirker arbeidet i NIB er uklart, men det er ingen grunn til å tro at arbeidet der vil gjøre utviklingen som skisseres i dette dokumentet uaktuell. Bibliometrien vil uansett ha behov for sine løsninger for å bruke dataene.

# Kort om fri programvare

## Valg av verktøy

I NIB-sammenheng er det for tiden mye snakk om R, som er en programpakke beregnet for behandling av statistikk og har gode visualiseringsløsninger. Det er ikke ønskelig å begrense bruken til bare R-løsninger, men det er den programvaren vi ser vokser frem også innenfor bibliometrien.

Karolinska Institutet bruker også R-løsninger for å gi egne forskere tilgang til analyser fra egen server. De har lovet å dele sin løsning med algoritmebasert fagfeltinndeling med oss. Den kan implementeres direkte hos oss uten videre seremoni, gitt det at vi nå har de samme dataene og databaseløsning å forholde oss til. Derfor velger vi å fokusere på R i dette dokumentet, men det kan i de fleste sammenhenger være andre programmer som brukes og deles, så lenge vi snakker om fri programvare.

*Det presiseres at NIB har ikke bestemt at R skal brukes. De som jobber med bibliometri kan velge det de selv vil. Datasettene vil fungere med stort sett hvilke som helst statistisk verktøy. NIB vil i samarbeid med institusjonene legge til rette for opplæring i de verktøyene som institusjonene foretrekker. Det har kommet initiativ til å sette opp R-kurs og det vil forhåpentlig vis annonseres på førstkommende Bibliometriforum.*

Det vil være fullt ut mulig å bruke datasett fra NIB i Excel, så lenge størrelsen på datasettet ikke blir for stor.

## Visualiseringsverktøy

Det finnes også «et hav» av ulike visualiseringsløsninger og VOSviewer er spesiallagd av CWTS i Leiden til å vise samarbeidsanalyser. Vi går ikke nærmere inn på disse mulighetene ut over å «konkludere» med at de er lett tilgjengelige og terskelen for å ta de i bruk er lav.

## Terskelen for bruk av fri programvare

Terskelen for å ta i bruk fri programvare virker generelt å ikke være spesielt høy, gitt at bibliometrimiljøene er vant til å forholde seg til store mengder med data. Mange forskere bruker allerede R og vi ser en trend med at flere publiserer artikler hvor R er brukt. Det er også etablert spesielle miljøer for bruk av programmeringsverktøy i ulike fag. TecFest er et initiativ innen bibliotekene for tekniske og økonomiske fag, hvor de jobber med bruk av Python. Terskelen kan også bli lavere når man får tilgang på gode relevante eksempler som man kan bruke i egen jobb.

# Oppgavene

Den som skal løse en oppgave er den beste til å finne ut hvordan oppgaven skal løses. Samtidig vet vi at det er stor sannsynlighet for at flere skal løse lignende oppgaver. Med standardiserte data i bunn er det stor sannsynlighet for at mange vil utvikle rimelig like løsninger og det ligger med andre ord til rette for mye dobbeltarbeid. Den beste medisinen mot dobbeltarbeid er at institusjonene kjenner til hverandres arbeid, samarbeider om løsninger og deler disse løsningene.

Typiske oppgaver som skal løses er benchmarking, samarbeidsanalyser, finne trender, med mer. Hvor godt de løses vil avhenge av kompetansen til de som skal løse oppgaven og ressurstilfanget. Det kan ikke være spesielt radikalt å hevde at vi kan få bedre løsninger hvis flere jobber sammen om løsningene. Samtidig kan vi få på plass flere løsninger hvis flere deler på oppgavene.

# Deling av løsninger

Det som skisseres her er R-pakker som kan brukes direkte på de datasett som NIB skal tilby. Det vil si at det som trengs er et datasett fra NIB og en programpakke som er utviklet spesielt for et slik datasett og som viser den ønskede analysen.

Deling av løsninger kan gjøres på GitHub eller tilsvarende plattformer. NIB kan også legge ut løsningene sammen med datasettene. Det viktige er å gi en oversiktlig fremstilling, gi lavest mulig terskel for å komme i gang og være oppdatert.

# Organisering av samarbeid

I motsetning til utvikling av datasett vil det her ikke være et like stort behov for organisert samarbeid med prosedyrer (slik det legges opp til i utviklingen av datasett) og initiativene kan komme fra enkeltpersoner med riktig kompetanse. Det kan imidlertid være behov for opplæring og mulighet til å få demonstrert løsningene, for å få med nok mange og de riktige personene.

Det å utvikle en løsning sammen vil i dette landskapet heller ikke bety at det bare blir utviklet én løsning på ett problem, men at flere kan finne sammen om ulike løsninger. Det kan med andre ord også bety noen parallelle løsninger og «litt dobbeltarbeid», men det er langt fra en situasjon hvor hver institusjon har sin egen løsning. Situasjonen vil heller tendere mot at løsningene bygger på hverandres løsninger.

## Forsterke kvalitet

Det er ønskelig med en aktiv tilrettelegging for å fremme kvalitet, brukervennlighet og utvikling av løsninger som dekker prioriterte behov. Det bør diskutere om det har en verdi å innføre en form for systematisk anerkjennelse og eventuelt premiering av godt arbeid. Videre må vi sikre at vi ikke går glipp av gode innspill fordi vi ikke tilrettelegger godt nok.

Det å utvikle løsninger i et fellesskap innebære vil ha noen former for innebygd konkurranse. Den som første presenterer en løsning vil kunne få et stort eierskap til endelig løsning. Ved parallelle løsninger vil det kunne bli konkurranse om å få gjennomslag. Den «beste løsningen» kan fort tape for den løsningen som er enklest å bruke, eller en løsning fra de som er mest aktive eller som nyter mest tillit i fellesskapet.

De som er mest aktive kan bli brukt for ofte slik at vi får en «Matheus-effekt». Det vil si at de som i utgangspunktet har de beste muligheter til å bidra får enda flere muligheter på bekostning av andre. Det å sikre god kvalitet kan derfor handle om å være bevist på at alle skal slippe til. Ikke minst for å sikre en viss lojalitet til det konseptet som velges.

## Opplæring

Hvis vi velger en modell hvor analyseverktøyene utvikles i fellesskap som R-løsninger, vil kursing være sentralt. Gitt at alternativet er å etablere et utviklerteam hos Unit, vil det å være kursorganisator fremstå som lang billigere og mer effektiv bruk av ressursene i Unit. Samtidig opplever vi en interesse fra enkelte institusjoner til å ta ansvar for kurs, hvor NIBs rolle blir å gi kursholder tilgang til data som de kan bygge kursinnholdet på.

Det å holde kurs kan ha to opplagte fordeler. Den ene er å øke kunnskapsnivået og den andre er å bring folk sammen. Sistnevnte fremstår som et av de viktigste prinsippene for den modellen som skisseres i dette dokumentet.

Opplæring vil kunne foregå i flere ulike fora, på ulike kunnskapsnivå og de kunnskapsområder som det er behov for i denne sammenheng. Det sentral er behovet.

## Utvikle et bredere miljø

Workshopen på bibliometriske data og bruk av fri programvare hadde som formål å øke samarbeidet i Skandinavia og det var flere som knyttet nye kontakter. Det er ingen tvil om at vi får tilgang på spisskompetanse ved å samarbeide med KI og KTH. Det samme gjelder DTU, men da en annen kompetanse. De har lenge hatt sitt OA-barometer og utvikler nye tjenester «på løpende bånd». Hvis vi satser aktivt, vil vi kunne være en attraktiv samarbeidspartner i fremtiden.