

Saksdokument

Porteføljestyre for forskning og kunnskapsressurser

Til Porteføljestyret for forskning og kunnskapsressurser
Dato 28.08.2023
Saksnummer 16/23
Saksarkivnr. 23/00000
Sakstype

Saksansvarlig Vigdis Namtvedt Kvalheim
Saksbehandler Bodil Agasøster

MODANSA- en modernisering av Sikts infrastruktur for forskningsdata

Forslag til vedtak

Porteføljestyret tok saken til orientering.

Saksfremstilling

MODANSA står for Modernisering av Data om det Norske samfunn. Datasamlingene er et resultat av en stor og bevisst satsing fra Norges forskningsråd (NFR) gjennom mange tiår og et forpliktende samarbeid mellom forskningsmiljøene representert ved NSD og sentrale dataprodusenter, særlig Statistisk sentralbyrå. Det kanskje viktigste resultatet av denne satsingen er Kommunedatabasen, som var NSDs mest brukte datatilbud gjennom tidene. Kommunedatabasen er også et godt eksempel på hovedintensjonen med datatilbudene fra Sikt:

1. Mest mulig data samlet under samme inngang, definert som data over tid
2. Godt dokumenterte data
3. Åpent tilgjengelig for forskning og undervisning
4. Med kort vei til analyse.

I 1993 passerte Kommunedatabasen sine første 100.000 variabler, allerede i 2001 ble 200.000 variabler passert. Da NSD ble en del av Sikt ved inngangen til 2022, hadde databasen nesten 400.000 variabler for hver enkelt kommune. Enkelte strekker tilbake til Norges første folketelling i 1769.

Kommunedatabasen og de øvrige datasamlingene om det om det norske samfunn har en kompleks struktur og består av både kvantitative og kvalitative data. De er driftet i eldre



løsninger som i liten grad oppfyller krav til universell utforming og FAIR-prinsippene¹. Behovet for å modernisere datatilbudet og formidlingsplattformen for disse datasamlingene

er stort. Satsingen som beskrives i dette notatet vil gi brukervennlig tilgang til datatyper som ikke faller inn under surveydata og microdatasatsingen til Sikt, som har vært prioritert de siste 5-10 årene. Tekstdata og data organisert i databaser (registerdata, forvaltningsdata, andre typer stordata, etc.) blir stadig viktigere for enkel tilgang til å studere det norske samfunnet langsiktig og komparativt. Kunnskapsutviklingen er avhengig av at norsk forskning har god tilgang også til denne typen data.

Målet er å øke bruken av denne typen data i norske forskningsprosjekter. Virkemiddelet er å utvikle en helhetlig plattform som tilbyr nye koplingsmuligheter og avanserte spørringer ved hjelp av kunstig intelligens. Det vil gi stor verdiøkning av eksisterende data og legge til rette for nye dataområder og datatyper. Satsningen vil også gi bedre tilgang og kortere vei til analyse for forskere, studenter og andre brukere.

Bakgrunn

Data om det norske samfunn omfatter blant annet Kommunedatabasen, Forvaltningsdatabasen, Kirkedatabasen, og Data om det politiske system. Disse datasamlingene har en unik nasjonalhistorisk verdi. De strekker seg fra så tidlig som 1796 frem til i dag, noe som gir mulighet til å analysere samfunnsutviklingen på politisk, geografisk og individuelt nivå. Data er levende i den forstand at de brukes til forskning, undervisning og av studenter, samt at de oppdateres kontinuerlig.

Denne type data er særlig egnet til sammenstillinger med individdata fra intervjuundersøkelser og registre. I et nasjonalt perspektiv er disse datasamlingene helt unike og de eneste i sitt slag. Det innebærer at et tap av datasamlingene ville fått store konsekvenser på et nasjonalt nivå, både for forskningsmiljøene, og offentlig forvaltning. Særlig gjelder dette data som finnes i Kommunedatabasen² og Forvaltningsdatabasen³. Datasamlingene er mye brukt og har siden 2019 resultert i 184 publikasjoner*:

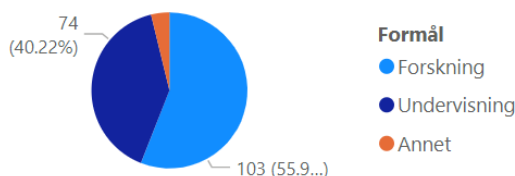
¹ FAIR står for Findable, Accessible, Interoperable og Re-usable og handler i all hovedsak om å håndtere og beskrive data på en måte som setter andre i stand til å forstå og bruke dem i fremtiden. På norsk kan man si at dataene skal være *gjenfinnbare, tilgjengelige, samhandlende og gjenbrukbare* ([Open Science | FAIR-prinsippene](#))

² Kommunedatabasen består av omfattende registeropplysninger om norske kommuner fra 1796 til 2018. Totalt inneholder Kommunedatabasen nærmere 400 000 variabler om hver kommune, om tema som demografi, økonomi, arbeidsliv, offentlige finanser, politikk, helse, utdanning, kultur, samferdsel og naturressurser

³ Forvaltningsdatabasen er en kartlegging av departement og ytre etater i Norge etter 1947. Den gir oversikt over den hierarkiske organiseringen og hvordan denne har endret seg over tid



Fordeling mellom data brukt til undervisnings og forskningsformål



Antall publikasjoner	Type
2	Bachelorgrad
6	Bok
12	Bokkapittel
9	Doktorgrad
11	Forskningsrapport
72	Mastergrad
7	Offentlig utredning
65	Tidsskriftartikkel
184	

*T.o.m mars 2023. Tallene er basert på manuell telling av publikasjoner. Bacheloroppgaver blir vanligvis ikke publisert og det reelle tallet på bruk er sannsynlig mye høyere.

I 2022 ble «forprosjektet» etablert for å lage et konseptuelt grunnlag for hvordan «Data om det norske samfunn» kan moderniseres. Dette var på bakgrunn av vedtak i styringsrådet for etablering av Sikt ([Referat styringsråd 2021-11-30, sak 6](#)): «Forprosjektet bør identifisere etterspørsel og brukerbehov, forprosjektet skal også lage en plan for finansiering av et gjennomføringsprosjekt». Vedtaket var en oppfølging og delvis stadfesting av vedtak i NSDs styre høsten 2021 om å prioritere MODANSA som en satsing finansiert gjennom NSDs egenkapital.

Forprosjektet kartla og vurderte datasamlingene, gjennomførte intervjuer og en spørreundersøkelse med brukermiljøene, og arrangerte en intern workshop. Prosjektet leverte «[Rapport om modernisering av data om det norske samfunn](#)⁴» som inneholder en rekke anbefalinger for hvordan datatilbudet og drift av «Data om det norske samfunn» bør moderniseres. Rapporten la grunnlaget for «Satsingsforslag for MODANSA» som ble vedtatt i ledergruppen til Sikt den 27. juni 2023.

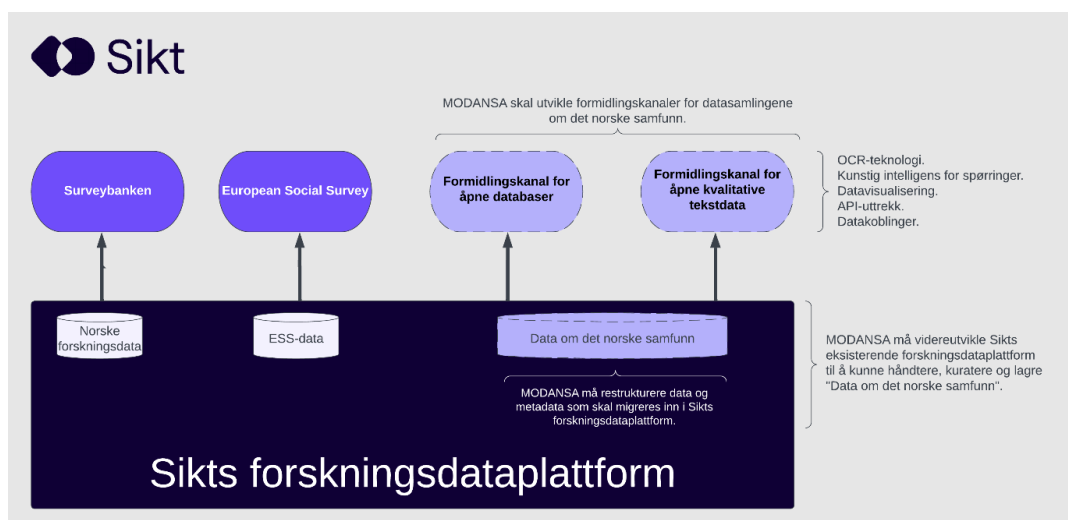
Satsingsforslag

MODANSA skal **utvikle Sikts forskningsdataplattform**, som i dag er innrettet mot kvantitative surveydata til å håndtere, kuratere og lagre ikke-surveybaserte data, dvs. kvalitative tekstdata og registerdata, offentlige data, stordata, etc.

Målet er å styrke Sikt sin datainfrastruktur generelt, slik at flere verdifulle typer data vil løftes opp. Det betyr at investeringen i MODANSA vil ha nytte utover datasamlingene under «Data om det norske samfunn». Sikt vil være i stand til å ta imot, gi merverdi til og tilgjengeliggjøre flere ulike sammensetninger av forskningsdata.

Satsningen vil utvide Sikt sin forskningsdataplattform med to nye datalagrings- og formidlingskanaler. Den ene vil formidle åpne kvantitative data organisert i databaser, mens den andre vil formidle åpne kvalitative tekstdata.

⁴ Se vedlegg



Formidlingskanalen for databaser vil åpne opp for koblinger på tvers av datasamlinger. For hver ny datasamling som legges inn, vil det oppstå et mangfold av nye relasjoner, som lar forskere konstruere datasett på tvers av datasamlingene. Dette er koblinger som tidligere ikke har vært tilgjengelig i Sikts forskningsdataplattform. Formidlingskanalen for tekstdata vil benytte seg av en språkmodell som lar brukere stille avanserte spørringer på tvers av store mengder tekstdata, som for eksempel: "Gi meg alle partiogram fra Høyre, Venstre og Arbeiderpartiet i perioden 1980 til 2000 som omhandler "klima".

Gevinster

MODANSA åpner opp for forskning og analyse på samfunnsmessig viktige data om det norske samfunn av høy kvalitet, et datagrunnlag som vil få økt relevans som kunnskapsgrunnlag for samfunnet i tiden som kommer. Moderniseringen av Sikts datainfrastruktur vil gi følgende gevinster:

- Tilgjengelighet og brukervennlige løsninger
- Universell utforming av tjenesten
- Sikrer FAIR-forskningsdata

Leveranser

Basert på anbefalingene har forprosjektet foreslått en liste over leveranser som kan inngå i satsingen. Leveransene definerer satsingens krav til kvalitet og omfang. Dette må sees i kontekst av tids- og kostnadsramme og justeres etter det som blir vedtatt som MODANSAs rammeverk.

Forslag til leveranser:

- Formidlingskanal for åpne databaser bygd på eksisterende forskningsdataplattform.
- Formidlingskanal for åpne kvalitative tekstdata bygd på eksisterende forskningsdataplattform.
- Omstrukturering og migrering av data fra "Data om det norske samfunn" til Sikts forskningsdataplattform.
- Videreutvikling av Sikts forskningsdataplattform til å håndtere kuratering og langsiktig lagring av nye type data.
- Standardisering av maskinlesbar metadata som beskriver "Data om det norske samfunn" og migrering til Sikts forskningsdataplattform.



- Verktøy for innhenting av data via API.
- Etablering av ny arbeidsflyt som er tilrettelagt for eksisterende løsninger.

MODANSA må levere løsninger som tilrettelegger for at "Data om det norske samfunn" og nye datasamlinger, får samme gjenbruksverdi og arbeidsflyt som Sikts øvrige datatilbud.

Finansiering

Under er en foreslått investeringsplan basert på det estimerte budsjettet i «Sikts satsningsområder innen forskning 2023-2026». Kostnadssommene er oppført i 1000 kr. Total budsjettert ramme for prosjektet er MNOK 35. Drift og utvikling skal etter prosjektslutt finansieres via bevilgning, men det vil være aktuelt å ta betalt fra bl.a. private dataprodusenter for tilrettelegging, dokumentasjon, kvalitetssikring og tilgjengeliggjøring av data. Det skal ikke tas betalt for *bruk* av data. På sluttbrukersiden skal data være åpnnet tilgjengelig for alle.

Leveranser	2023*	2024	2025	2026	2027	Totalt
Planlegging og kompetansebygging	500					500
Videreutvikling, migrering og standardisering		5250	4750	2 500	500	13000
Utvikle verktøy for innhenting av data via API		1 000	1 000			2 000
Utvikle formidlingskanal for åpne databaser			2 500	4 000	2 250	8 750
Utvikle formidlingskanal for åpne kvalitative tekstdata			2 500	4 000	2 250	8 750
Administrasjon, organisering og drift		250	250	500	1000	1000
Totalt	500	6 500	11 000	11 000	6 000	35 000

*Kvartal 4.

Veien videre og medfinansieringspartnere

Det legges opp til at Sikt dekker deler av prosjektkostnadene over sitt budsjett/risikokapital, mens resten av prosjektmidlene søkes dekket gjennom medfinansiering. Hvordan og hvem som skal dekke den resterende finansieringen er ikke avklart. Det er sendt et satsningsforslag til KD om finansiering fra 2024, men dette vil ikke bli avklart før i oktober 2023. Dersom satsningsforslaget til KD ikke går igjennom, må alternative finansieringsmuligheter utredes. Sikt vil gå i dialog med relevante aktører/brukere høsten 2023 om mulig medfinansiering.

Vedlegg

Rapport om modernisering av "Data om det norske samfunn"