Søknad om midler til kompetanse-hub for IT i forskning

Teksthub (med digital humaniora)

Søker:	Janne Bondi Johannessen
Vertsinstitutt (hub):	Institutt for lingvistiske og nordiske studier (ILN)
Samarbeidspartner (noder):	**Se 1.2. for en detaljert liste av samarbeidspartnerne (noder)** Søknaden representerer ansatte ved de nevnte instituttene, som har bekreftet at de ønsker å være aktive i teksthuben.
Søknadsbeløp og -periode:	Kr 3 500 000 for 2019-2021. Vi regner med å søke igjen for 2022/23.

1: Formål og gevinster

1.1: Hva er formålet?

Tekst, både som skriftlig uttrykk og i muntlig form, er i dag forskningsobjekt for de fleste fagområder på UiO, og særlig på HF. Teksttypene som brukes i og som gjenstand for forskningen er derimot svært heterogene og kan omfatte alt fra skriftlig tekst som litteratur, sakprosa og elevtekster, til muntlig tekst som videoopptak av samtaler, mer og mindre formaliserte intervjuer og oppleste lister. Å behandle disse teksttypene har derfor resultert i ulik kompetanse, ulike verktøy og ulike tilnærmingsmåter. Samtidig er det mye å hente på felles løsninger som for eksempel maskinlæring, felles infrastruktur for korpora og delt kunnskap rundt annotering og transkripsjon. Maskinlæring, AI og andre semi-automatiske språkteknologier bygger på manuelt arbeid med tekst. Dersom miljøene samarbeider om for eksempel å sette standardarder for annotering, vil miljøene lett kunne utnytte hverandres ressurser. Hensikten med teksthuben er derfor å samle og koordinere kompetanse og erfaringer fra tekst og digital humaniora på UiO, for slik å kunne bygge opp varig forskningsinfrastruktur som er nyttig for både forskning og undervisning, ikke minst for nye brukere. Dette er i henhold til målene som ble trukket fram i *Innspill til UiO strategi 2030 fra Det humanistiske fakultet* (mars 2019), særlig m.h.t. tverrfaglighet, kvalitet i undervisning og forskning på høyt nivå. Behovet for en teksthub er stort siden eksisterende kompetanse og verktøy ofte er spredt ut på forskjellige fagmiljø som til dels ikke kjenner til hverandres tilbud. På den andre siden er det også et økende behov for å lære opp nye brukere for å ta i bruk eksisterende infrastruktur og å utvikle nye verktøy, som møter utfordringer av en stadig mer digitalisert (forsknings)verden. Eksisterende infrastruktur og digital kompetanse må også integreres i mye større grad i undervisningen, som vil kreve opplæring og bevisstgjøring av både studenter og ansatte. Huben forener derfor samarbeidspartnere på tvers av UiO og skal satse tungt på kompetanseutvikling, koordinering og utvikling.

Tekstlaboratoriet ved ILN, HF har over mange år opparbeidet seg en unik kompetanse på tekst- og språkteknologi og har et godt etablert nettverk til andre forskningsmiljøer på UiO, i Norge og internasjonalt. Med sin lange erfaring som leder av flere store forsknings- og infrastrukturprosjekter (Forskningsrådet, EU, Nordforsk, Norway Grants) og som C-senter i den europeiske Clarin-strukturen, har det derfor vært naturlig at Tekstlaboratoriet tar initiativet til og søker om å få bli koordinatoren i teksthub-node-strukturen. Vertsinstitusjonen ILN har dessuten sterke, komplementære tekstfaglige miljøer som EDD, Ibsensenteret og MultiLing. Siden huben vil ha en tverrfaglig tilnærming og satse tungt på digital humaniora, forventer vi dessuten at den vil kunne bidra til metodeutviklingen og en økning av prosjektporteføljen på feltet.

1.2: Hvilket kompetanseområde skal dekkes?

Følgende kompetanseområder (noder) skal inngå i huben:
1. Litterær fjernlesing/masselesing og korpusanalyse: HF-ILN (Ibsensenteret, nordisk litteratur, EDD), HF-ILOS

2. Maskinlæring og automatisk transkripsjon: HF-ILN, MN-IFI, SV-STV, UV-eVIR, USIT

3. Transkripsjon/talespråk/korpus: HF-ILN (retorikk, lingvistikk, nordisk språkvitenskap og NOAS, NORINT, MultiLing), HF-ILOS (engelsk, fremmedspråkene), MED, UV-ILS, UV-eVIR, UV-ISP

4. Skriftspråkstekst/korpus: HF-ILOS (engelsk, fremmedspråkene), HF-ILN (EDD, nordisk språkvitenskap, norsk som andrespråk, lingvistikk, NORINT), HF-IFIKK, HF-IKOS, SV-STV, UV-ILS, UV-ISP, TF, UB

5. Annotering (tagging, parsing) og statistikk Alle: Gjelder all tekst og alle noder.

6. Nettdugnad (crowdsourcing): HF-ILN-MultiLing, MN-IFI, IFIKK, UV-IPED

7. Praterobot (chatbot): HF-ILOS (fremmedspråkene), HF-ILN (norsk som andrespråk)

+ 8. Koordineringsnode: HF-ILN (Tekstlaboratoriet): Koordinering av nodene, administrasjon og deltagelse i nodenes aktiviteter, personvern og GDPR, samt opphavsrett, TSD, nettside, datalagring, innkjøp av utstyr, organisering av kurs, møter og utstyr.

1.3: Hvilke ressurs-/eInfrastrukturbehov skal dekkes?

Ved å etablere en teksthub vil følgende behov blir dekket: **(1) Koordinering av nodene og markedsføring av eksisterende infrastruktur** på tvers av fagmiljø og utenom UiO, blant annet gjennom utvikling av felles møteplasser og nettportal. **(2) Kunnskapsdeling- og utvikling** gjennom danning av tverrfaglige dugnadsgrupper og eksterne kurs rundt temaer som avansert transkripsjon, annotering, nettdugnad og fjernlesing, som betegner en kvantitativ analyse av tekst ved hjelp av algoritmer. **(3) Utvikling og tilegning av nye digitale ressurser og kunnskap.** Tilby støtte, kompetanse og eksterne kurs og nettverk der det ikke finnes standardløsninger på markedet, som for eksempel for oppbygging av korpus, lister og databaser innenfor humaniora. Prosjekter på idéstadiet kan så seinere oppskaleres med annen finansiering.
(4) Brukerstøtte: Inkluderer veiledning, opplæring og oppfølging av både ansatte og studenter som ønsker å bruke hubens ressurser (som korpus, databaser, spesialisert programvare) i forskning og undervisning eller etablere nye. Støtten omfatter også veiledning rundt datahåndtering og personvern samt opphavsrett.

2: Organisering

2.1, 2.2., 2.3: Hvem skal være ansvarlig leder, datahåndteringsansvarlig og vertinstitusjon?

Leder: Professor Janne Bondi Johannessen (leder, Tekstlaboratoriet, kjernegruppemedlem MultiLing, ILN)
Datahåndtering: Senioringeniør Kristin Hagen (Tekstlaboratoriet, ILN)
Lokalisering: Institutt for lingvistiske og nordiske studier / Tekstlaboratoriet

2.4: Hvilke forskningsgrupper, fagmiljøer og lignende skal inngå som noder?

Teksthuben vil samle kompetanse fra seks fakulteter (HF, SV, UV, MN, MED, TF) og USIT. Nodene er beskrevet over (1.2). Når huben er etablert, vil derimot mye av tilbudet også være tilgjengelig for alle interesserte forskere og studenter ved UiO. Det skal dessuten legges til rette for at forskere og miljøer kan søke huben om midler til å gjennomføre hub-relevante kompetansehevings- og utviklingsprosjekter.

2.5: Hvordan skal arbeidet i organiseres?

Hubens noder skal administreres og koordineres av Tekstlaboratoriet, som har både vitenskapelige og teknisk-administrative stillinger knyttet til seg og som har god erfaring med å lede store infrastrukturprosjekter. Koordinatorens hovedoppgave vil være å sørge for en samkjøring av eksisterende infrastruktur, en felles profil utad og legge til rette for kompetanseheving- og deling på tvers av miljøene. Nodene internt setter seg mål og samarbeider med hubkjernen om aktiviteter og budsjett, men beholder et

selvstendig ansvar for det vitenskapelige innholdet knyttet til huben og for å knytte undervisningsrelaterte prosjekter, som bruk av korpus, fjernlesing eller nettdugnad i oppgaver, til hubens temaområder. Hver node vil også stå ansvarlig for et kompetanse-hevingstilbud på sitt kjerneområde. Nodene er organisert etter metode og materiale snarere enn institusjonell tilhørighet, og deltakere kan delta i flere noder. Aktivitetene, både kompetanseheving og utvikling, vil skje i tett samarbeid mellom teknisk personalet og forskerne i nodene. Det vil organiseres minst ett årlig møte for alle nodene for å utveksle erfaringer og justere kurs. Siden digital humaniora er et nytt felt i endring vil nye personer og noder kunne integreres etter behov.

2.6: Hvordan skal gevinstene videreføres etter avvikling?

Det forventes at tiltakene i huben, vil ha en varig effekt på både når det gjelder vitenskapelig og teknisk expertise i nodene, men også med tanke på samarbeidsflatene og nettverk på tvers av miljøene. Kompetansehevingstiltak, kurs og undervisningsopplegg vil være lett dubliserbare, slik at også nye ansatte og studenter vil ha nytte av dem i framtiden. Dataene og programvaren i huben vil lagres på UiOs interne servere og på Norstore (Nird), og være tilgjengelige via nettportalen. Tekstlaboratoriet har god kompetanse på lagring og vedlikehold av tekst- og språkteknologi og vil derfor også i framtiden være en pådriver for opprettholdelsen av eksisterende og ny infrastruktur.

3: Budsjett

Siden mye av infrastrukturen allerede er på plass, men er fragmentert, fordeler ressursbehovet for huben seg for det meste på lønnsmidler og kompetanseutviklingstiltak. Anskaffelser begrenser seg til noe maskin- og programvare (for eksempel til transkripsjon), databaser, korpus, lisenser og serverleie. Under Andre utgifter er det lagt inn et budsjett for transkribører og annen vitenskapelig assistance (f.eks. Amazon Mechanical Turk), som skal brukes både til å oppdatere eksisterende og etablere nye digitale ressurser.

Inntekter:	**2019:**	**2020:**	**2021:**	**Sum:**
Fra fagrådet	500 000 kr	1 500 000 kr	1 500 000 kr	3 500 000 kr
Egenfinansiering	125 000 kr	300 000 kr	300 000 kr	725 000 kr
Annen finansiering	kr	kr	kr	kr
Sum inntekter:	**625 000 kr**	**1 800 000 kr**	**1 800 000 kr**	**4 225 000 kr**
Utgifter:				
Lønnsmidler, frikjøp	450 000 kr	1 200 000 kr	1 200 000 kr	2 850 000 kr
Anskaffelser	50 000 kr	100 000 kr	100 000 kr	250 000 kr
Kompetanseutvikling	50 000 kr	300 000 kr	300 000 kr	650 000 kr
Andre utgifter	75 000 kr	200 000 kr	200 000 kr	475 000 kr
Sum utgifter:	**625 000 kr**	**1 800 000kr**	**1 800 000kr**	**4 225 000 kr**

4: Vedlegg

1) Støttebrev fra ILN
2) Personer ved Teksthubens noder