

Wilfred Østgulen, IKT-direktør i Nasjonalbiblioteket
5. mars 2024

**Magnus Lagabøtes landlov
i Codex Håndbergianus**
Midten av 1300-tallet

Kong Magnus Lagabøtes landlov av 1274 er et stort verk i europeisk rettshistorie. Dette var den tredje rikskodifiserte lovboken i Europa, og ble først avløst av Christian Vs norske lov i 1687. Landloven omformet det norske samfunnet, etablerte forpliktelse som styreinstrument, ga de enkelte rettigheter og skapte utgangspunktet for en norsk statsmakt.

Det er bevart 29 manuskript og over 50 manuskriptfragmenter av Landloven, Codex Håndbergianus er det vakreste av dem alle. Lovmanuskriptet ble skrevet i Bergen på romant på midten av 1300-tallet, og har 11 dekorene og illustrerte bokstaver (såkalte illuminasjoner) som formidler essensen i teksten.

Codex Håndbergianus er en sammenbinding av flere fortekster. Landloven er den største og viktigste, men bindet inneholder også det som kalles rettsbøker, senere lover, som utfyller eller erstatter regler i Landloven, og erkebiskop Jon Raudes kirkelov.

Den mest betydningfulle arven etter Landloven av 1274 er at vi innvester oss etter skuffete lover, fremfor å bli møtt med avgjørende. Landloven la grunnlaget for at vi som lever i landet i dag, assosierer ordet lov med frihet og rettigheter, ikke med maktmisbruk og undertrykkelse.



KI i Nasjonalbiblioteket – Prosjekt Mímir

Generativ KI i offentlig sektor



Nasjonalbiblioteket – vår felles hukommelse

- Samfunnsoppdraget til Nasjonalbiblioteket er å **sikre avlevering og bevaring av publisert materiale** fra alle publiseringsplattformer og slik være **den fremste kilden til kunnskap om Norge og norske forhold**
- Nasjonalbiblioteket **formidler og gjør kulturarven tilgjengelig** og er dermed en kilde til forskning, læring og språkutvikling, og vi **bidrar til å skape identitet og tilhørighet**
- Nasjonalbiblioteket er både **infrastruktur for norsk forskning** og godkjent som en **selvstendig forskingsinstitusjon**
- Gjennom **Språkbanken** tilbyr Nasjonalbiblioteket **norske datasett for tekst og tale** til utvikling av språkteknologi for norsk
- ++



Nasjonal- bibliotekets samling

- Vi har en enorm samling av fysisk og digitalt materiale
 - Store fjellmagasiner i Rana der det fysiske materialet bevares på faglig vis
 - Splitter nytt datasenter i fjellmagasinet i Rana der det digitale materialet bevares på faglig vis
- Fysisk samling i Oslo av det mest dyrebare og sårbare materialet
 - Her er det også bibliotek, utstillinger, lesesaler og arrangementer for å formidle og gjøre kulturarven tilgjengelig for et bredt publikum
 - Akkurat nå er det Landslovsjubileum, med en fantastisk utstilling av gamle manuskripter fra middelalderen
 - Utstilling, podkaster, bokutgivelser og arrangementer
- Nasjonalbiblioteket har blitt et moderne medie- og kulturhus med egne utgivelser på alle formater – dette var en stor satsing gjennom korona-perioden
- Den systematiske digitaliseringen av alle samlingene startet i 2006

Alt har utspring fra samlingen vår

Manifest av 2005

—

«digitalisere samlingen»

DET DIGITALE NASJONALBIBLIOTEK

Strategimanifest 2005

Vedtatt av Nasjonalbibliotekets styre 23. februar 2005

Biblioteket kommer til brukeren

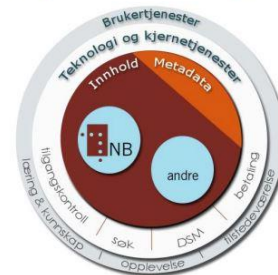
Moden informasjons- og kommunikasjonsteknologi og Internett har gitt biblioteket nye muligheter for profesjonelt samvirke og koordinert ressursutnyttelse. Det som likevel preger endringene mest er at biblioteket har fått mulighet til å gi tilgang til et rikt og variert innhold på alles skrivebord og i alles hjem. Nasjonalbiblioteket skal gjennom en brukerfokuset innsats på digitale bibliotek bidra til at det norske biblioteksamfunnet kan yte nye og bedre tjenester til samfunnsdeltakeren.

Norsk Digitalt Bibliotek står sentralt i satsingen, og både innhold og tjenester i NDB vil styrkes gjennom NBs satsing på digitale bibliotek. NB representerer et tyngdepunkt i utviklingen av et sterkt norsk digitalt bibliotek. Det norske biblioteksamfunnet vil styrkes som konsekvens av at NB etablerer tjenester som andre bibliotek kan bygge på i sin aktivitet, og de enkelte bibliotek får bedre muligheter til å tilby og utnytte digitale brukertjenester.

Kunnskap og opplevelse

NB skal gi brukere tilgang til et vell av innhold på digital form. Gjennom å fokusere innsats på digitale bibliotek skal NB legge til rette for bruk og gjenbruk av innhold fra NB i opplevelse og læring, i kunnskapsproduksjon og i næringsutvikling. Samlingene skal være tilgjengelig der brukeren er når brukeren vil ha tilgang. NB har valgt å fokusere innsatsen langs tre hovedakser:

- **Tilstedeværelse:** Metadata og innhold må presenteres i de arenaer som brukeren foretrekker og i en kontekst som er tilpasset brukeren. Brukeren skal kunne søke i NBs metadata og innhold i sitt fortrukne verktøy¹. NB skal gjøre det mulig å regulere tilgang for brukere på eksisterende tjenester i samfunnet².
- **Læring og kunnskap:** Det er behov for virkemidler for bedre å legge til rette for at ressurser i NB kan brukes i læring og for forskning. Slike virkemidler inkluderer f eks læringsobjekter for bruk i elæringsystemer og tilgang til høykvalitets digitale kopier av originalmateriale i NB. Dette vil også legge til rette for kommersiell utnyttelse av innhold fra NB.
- **Opplevelse:** Informasjon og kunnskap skal fortsatt settes inn i en sammenheng som gir brukerne nye opplevelser. Tjenester som gjør informasjon og kunnskap tilgjengelig som en del av en helhet, i form av utstillinger og opplevelsessturer på nett, må utvikles kontinuerlig.



Felles løft

Satsingene over krever nye og forbedrede teknologiske plattformer, og at volumet av tilgjengelig digitalt innhold økes. NB mener at det best skjer gjennom en koordinert innsats, og at NB representerer et tyngdepunkt gjennom så vel innhold

¹ Undersøkelser i andre land viser at også forskere foretrekker å bruke generelle søkemotorer på Internett i stedet for dedikerte tjenester når de søker etter informasjon på Internett.

² Som eksempel har Uninett etablert FEIDE som kan brukes av bibliotekmiljø for å regulere tilgang til innhold og tjenester for utdanning og forskning.



Noen tall fra den digitale samlingen – 19 år senere

Bøker:
ca. 630.000

Aviser:
ca. 4,5 millioner

Tidsskrift:
ca. 200.000

Brev og
manuskripter:
ca. 30.000

Notehefter:
ca. 6000

Musikk-
manuskripter:
ca. 10.000

Bilder:
ca. 2,2 millioner

Plakater:
ca. 10.000

Kart:
ca. 3000

Radio:
ca. 2,2 millioner
filer

Musikk:
ca. 285.000 filer

Film:
ca. 17.000 filer

Fjernsyn:
ca. 3800 filer

Programrapport
fra NRK:
ca. 36.000

Dette er digitale data om norsk språk, kultur, kunst, historie, geografi og samfunnsliv – samlet på ett sted, som kan brukes til å trene datamaskiner til å forstå «Norge og norske forhold»

NBs AI-lab

- Bakgrunn

Vi står på kjempers skuldre

- 150+ på digitalisering
- 70+ på IT
- Språkbanken
- brei og djup samlingskompetanse

- Bidrag til forskning, egen forskning
- Samarbeid med eksterne miljø, i

- Folkene

- 4 personer på full tid

uksjon og

ratorer

Vi deler

- kunnskap
- data
- modeller
- programvare

A historical map of Europe and the Baltic region, showing cities like Stockholm, Götterbarå, and Hamburg, and bodies of water like the Baltic Sea and the Gulf of Bothnia. The map is in a sepia tone and serves as a background for the text on the left.

Oppdrag fra KUD til Nasjonal- biblioteket

«Departementet ber med dette Nasjonalbiblioteket sette i gang et **koordinert forsknings-/utviklingsprosjekt** for om mulig å **undersøke verdien av opphavsrettslig beskyttet materiale** i trening av norske generative språkmodeller. **Relevante norske forskningsmiljøer** skal inviteres til å delta i prosjektet. **Forfatternes og forleggenes organisasjoner** inviteres til å følge prosjektet.

Vi ber om at Nasjonalbiblioteket, på bakgrunn av resultatene fra forskningsprosjektet, **vurderer grunnlaget for en eventuell kompensasjonsordning** for norske rettighetshavere, og eventuelt **utarbeide forslag til en slik ordning.**»



Samarbeids- partnere

- NorwAI – Norwegian Research Center for AI Innovation
 - NTNU med samarbeidspartnere
- Universitetet i Oslo, Instituttet for Informatikk
 - Language Technology Group
- Sigma2
 - High Performance Computing og storskala lagring til forskere i Norge

A historical map of Europe, showing various countries and cities. The map is in a sepia tone and includes labels for 'DENMARK', 'BALTIC SEA', 'HAMBURG', and 'DANTZIG'.

Gjennomførings- strategi for prosjekt Mímir

Vi skal:

1. Videreutvikle Nasjonalbibliotekets digitale norske tekstkorpus (NCC/NCC+)
2. Trene et sett av norske generative språkmodeller på ulike uttrekk av tekstkorpuset
3. Evaluere ytelsen til de ulike språkmodellene
4. Dokumentere funn og observasjoner

Avhengig av tilgang på datakraft til å trene modeller ønsker vi å gjøre pkt. 1-3 i flere iterasjoner

Planen er å gjøre dette før sommeren

1

Videreutvikle Nasjonal- bibliotekets digitale norske tekstkorpus (NCC/NCC+)

- Korpuset (NCC/NCC+) består av tekst fra en rekke kilder
 - Digitaliserte bøker
 - Digitaliserte aviser
 - Offentlige nettsteder – både tekst og dokumenter
 - Stortinget – både tekst og bøker
 - NRK
 - Lovdata
 - Wikipedia
- I tillegg bruker vi tekst fra store internasjonale datasett som er høstet fra internett, når vi trener modeller
 - Disse dataene inngår ikke i NCC/NCC+

Nasjonalbiblioteket koordinerer denne aktiviteten

A historical map of Northern Europe, showing Scandinavia, the Baltic Sea, and parts of Central Europe. The map is detailed with coastlines, rivers, and city names. The text 'NCC og NCC+' is overlaid on the map in a large, white, sans-serif font.

NCC og NCC+

- NCC → Mimir Base
 - **Åpent datasett** for fri tilgang uten begrensninger
 - Aviser og bøker i det fri eller etter avtale
 - Publikasjoner fra det offentlige
 - Tilrettelagte data fra Språkbanken, f eks fra Web
 - Utvidet beskrivelse ligger på <https://huggingface.co/datasets/NbAiLab/NCC>
- NCC+ → Mimir Extended
 - **Internt datasett** som brukes av NB AI-lab
 - Inneholder NCC og i tillegg materiale under opphavsrett

2

Trene et sett av norske generative språkmodeller på ulike uttrekk av tekstkorpuset

I prosjekt Mimir vil vi trene modeller i tre grupper:

- NorwAI vil trene modeller på Idun
- UiO vil trene modeller på LUMI
- NB vil trene modeller på Google TPU Research Cloud

- Vi vil trene norske språkmodeller opptil 7B-størrelse (kanskje 15B) basert på programvare/modeller fra Mistral og Llama2 (kanskje Bloom)
- Til sammen vil vi trene 10-15 norske språkmodeller som så skal evalueres på prestasjon
- Modellene vil trenes på ulike datasett av norsk tekst, med og uten opphavsrettslig beskyttelse

NorwAI koordinerer denne aktiviteten

3

Evaluerer ytelsen til de ulike språkmodellene

- Vi skal teste modellene på «språkoppgaver» for å se om modellene med opphavsrettslig beskyttet materiale presterer bedre enn de uten
- Både maskinell og menneskelig testing – NorwAI og UiO rekrutterer 20+ studenter til evalueringsarbeid
- Språkoppgaver kan være
 - Leseforståelse – lese en tekst og svare på spørsmål fra teksten
 - Eks. Bergentesten, Norskprøven
 - Generering av tekst
 - Oppsummering av tekst
 - Sentimentanalyse – positiv, negativ eller nøytral
 - Maskinoversettelse
 - Oversette mellom bokmål og nynorsk
 - Oversette mellom norsk (no/nn) og engelsk
 - Kanskje også til og fra samisk
 - Svare på flervalgs-quizer fra NRK
 - ++

UiO koordinerer denne aktiviteten

4

Dokumentere funn og observasjoner


- Vi vil oppsummere våre funn og observasjoner om verdien av det opphavsrettsrettslig beskyttede materialet fra evalueringen av de trente språkmodellene
- Disse funnene og observasjonene vil så inngå i vurderingen av om det er grunnlag for en kompensasjonsordning for norske rettighetshavere
- De vil også bidra til å utforme en eventuell kompensasjonsordning


Tale til tekst – NB-Whisper


NB-Whisper


updated 15 days ago


Models based on Whisper from OpenAI, and trained on data from Språkbanken and the digital collection at the National Library of Norway.

 NbAilab/nb-whisper-large


 Automatic Speech Recognition • Updated 7 days ago • ↓ 151 • ♥ 5

 NbAilab/nb-whisper-medium


 Automatic Speech Recognition • Updated 15 days ago • ↓ 49 • ♥ 1

 NbAilab/nb-whisper-small

 Automatic Speech Recognition • Updated 15 days ago • ↓ 55

 NbAilab/nb-whisper-base

 Automatic Speech Recognition • Updated 15 days ago • ↓ 4

 NbAilab/nb-whisper-tiny

 Automatic Speech Recognition • Updated 15 days ago • ↓ 39

- Trent på 8 millioner lydklipp på 30 sekunder, til sammen 66.000 timer norsk tale som er transkribert
 - Et talekorpus fra Stortinget tilrettelagt av Språkbanken
 - Et talekorpus laget av Nordisk Språkteknologi
 - Lydbøker
 - NRK undertekster
- NB-Whisper ligger åpent tilgjengelig på Hugging Face

A large, modern brick building with a prominent white tower section. The building features multiple rows of windows and a curved glass facade on the right side. In the foreground, there is a well-maintained green lawn. A tall white flagpole stands on the right side of the lawn. The background shows a steep, forested hill under a clear blue sky.

Takk for oppmerksomheten!

Spørsmål?