Ordbokshotellet

Varig lagring og formidling av norske ordsamlingar

Øyvind Eide (EDD)
Elisabeth Lien (ILN)
Lars Jørgen Tvedt (EDD)

Foredrag på 8. nordiske dialektologkonferanse på Aarhus Universitet, 15. august 2006

UNIVERSITETET I OSLO

Ordbokshotell:

”Ei teneste for elektronisk *lagring*, *indeksering* og *publisering* av innhaldet i ordsamlingar og ordbøker.”

Målgrupper

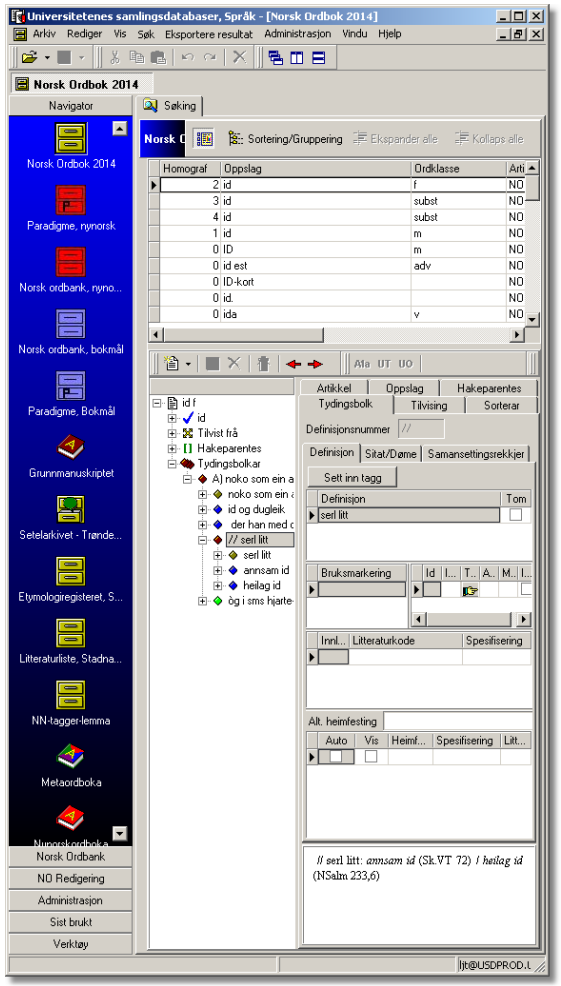
- Primært for Norsk Ordbok 2014.
- Forfattere av ordsamlingar og ordbøker.
- Andre språk- og kulturmiljø.

Ordboksarkiva

Norsk Ordbok har eit omfattande kjeldemateriale på papir:

- Ordboksmanuskript
- Setelsamlingar
- Tekstsamlingar
- Ordbøker
- Ordsamlingar
- Ordlister

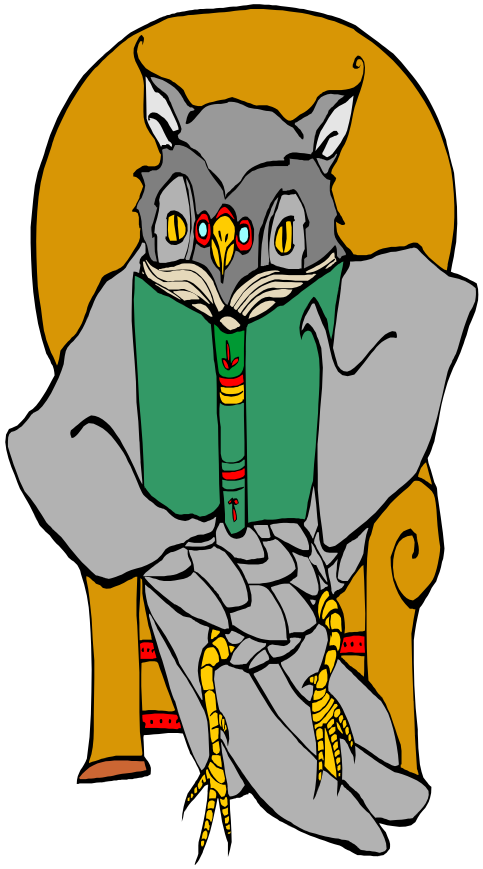
Digitalisering



- Digitalisering av kjeldar tok til på 1990-talet.
- NO 2014 har brukt datamaskin-baserte redigeringsmetodar sidan 2002.
- Dei digitaliserte kjeldene er integrerte med redigeringsprogrammet.

Ordsamlingar er viktige kjelder

- Norsk Ordbok dokumenterer både nynorsk skriftspråk og norske dialektar.
- Lokale ordsamlingar utfyller dialektmaterialet i setelsamlinga.
- NO 2014 har samla inn om lag 60 elektroniske manus til nyare ordsamlingar så langt.



Folkeleg interesse

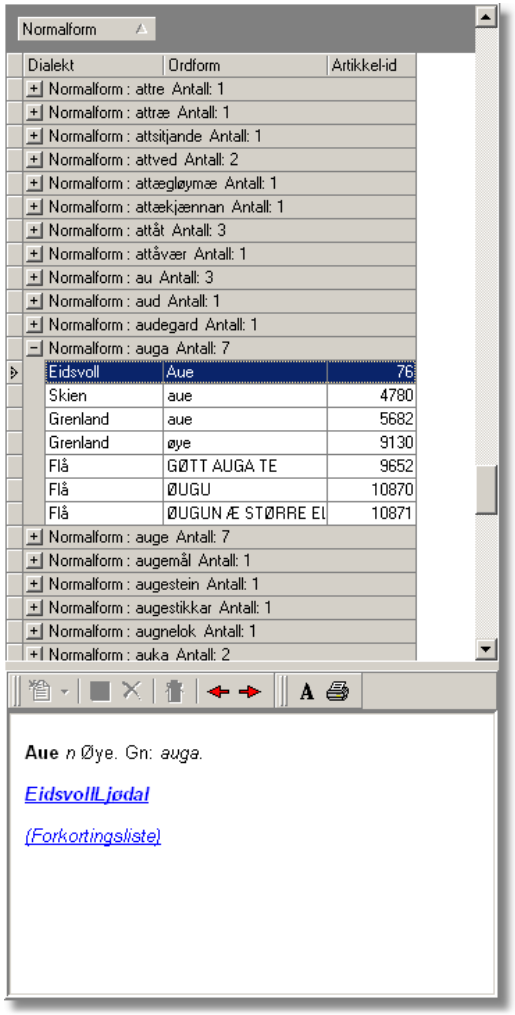
- Stor interesse for språk og dialektar i Noreg.
- Ordsamlingar er viktig dokumentasjon av lokal språk- og kulturhistorie.
- Enkel tilgang vil sannsynlegvis føre til ytterlegare interesse.

Utfordringane

- 60 ulike samlingar.
- Ulike digitale format.
- Ulik normalisering.
- Ulike ordklassemarkeringar.
- Ønskjer kopling til artiklar i Norsk Ordbok.

Løysinga

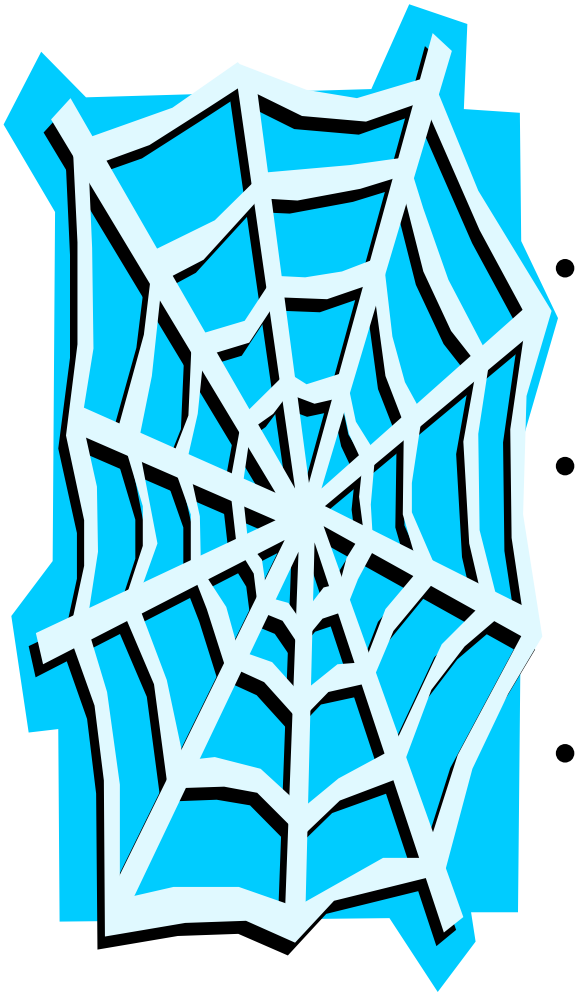
- Konvertere alle artiklar til eit felles ope tekstkodingsformat.
- Normalisere oppslagsord og ordklasseinformasjon.
- Indeksere på normerte former.
- Integrere i redigeringsprogrammet.



Resultat

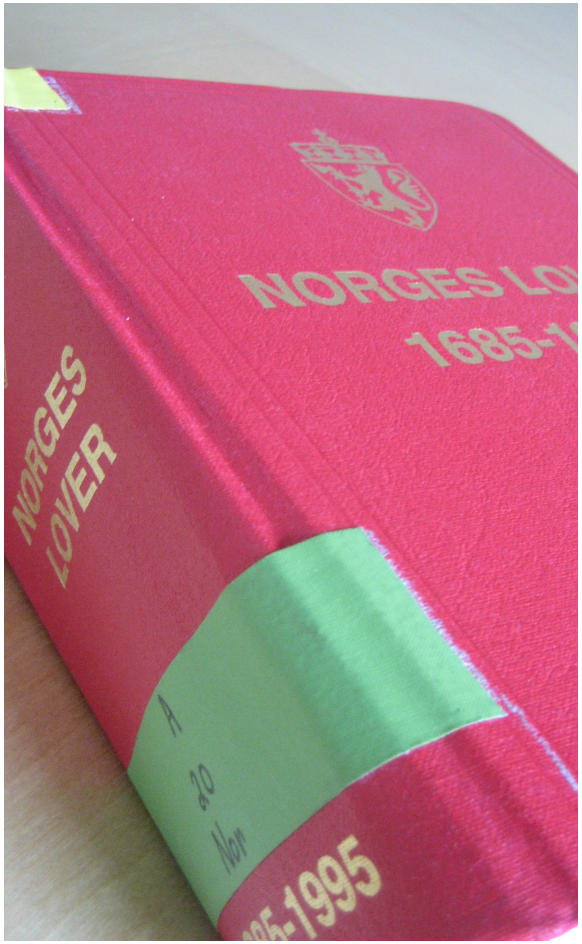
- Starten på ein stor norsk målføre-database.
- Vitskapleg kvalitetssikring av lokale samlingar via normaliseringa.
- Sikra sentral lagring av målføreinformasjon.
- Normaliserte oppslagsord gjer det mogleg å søkje på tvers av målføre-grensene.

UNIVERSITETET I OSLO



Publisering på web

- Norsk Ordbok og kjeldematerialet bør vere tilgjengelig på nettet.
- Pr. i dag ligg mykje av kjeldematerialet på NO 2014 si nettside.
- Ein kan ikkje utan vidare leggje ut ordsamlingane pga. *opphavsretten*.



Åndsverklova

- Skaparen av eit verk har einerett til å bestemme over verket.
- Det er lov å *sitere* eit verk som er verna av åndsverklova.
- Norsk Ordbok kan sitere ei ordsamling i artiklane.
- Nettpublisering av heile artiklar må godkjennast av forfattaren.

Beskytta ordsamlingar

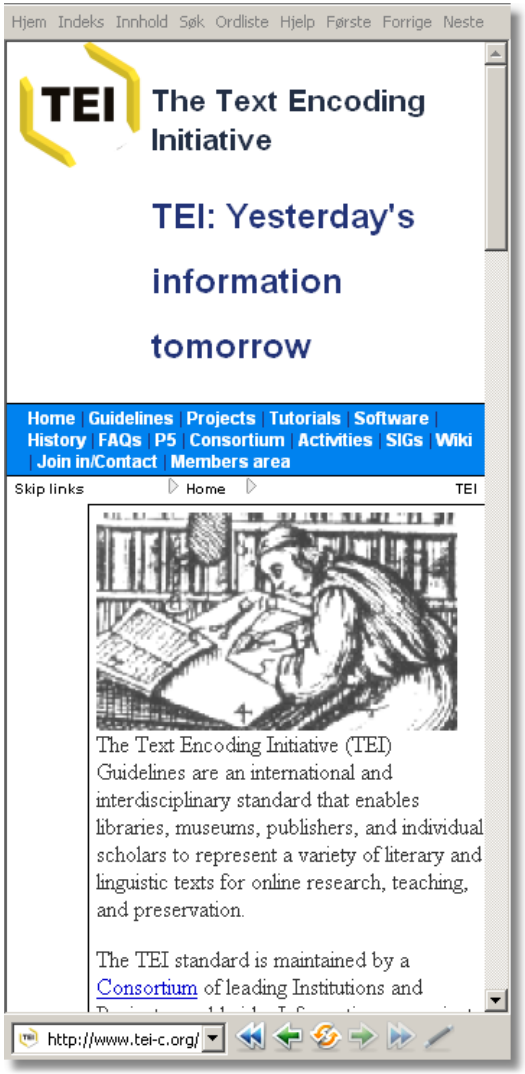
- Viser berre normert form og dialektform av oppslagsord.
- Viser enkel referanse til ordsamlinga.
- Viser full tekst for forskarar.

Frie ordsamlingar

- Vise full artikkeltekst til alle.
- Tilby marknadsføring av ordsamlinga via hotellet.

Tekniske løysingar

- Val av tekstformat.
- Normering.
- Databaseformat.
- Søking og presentasjon.



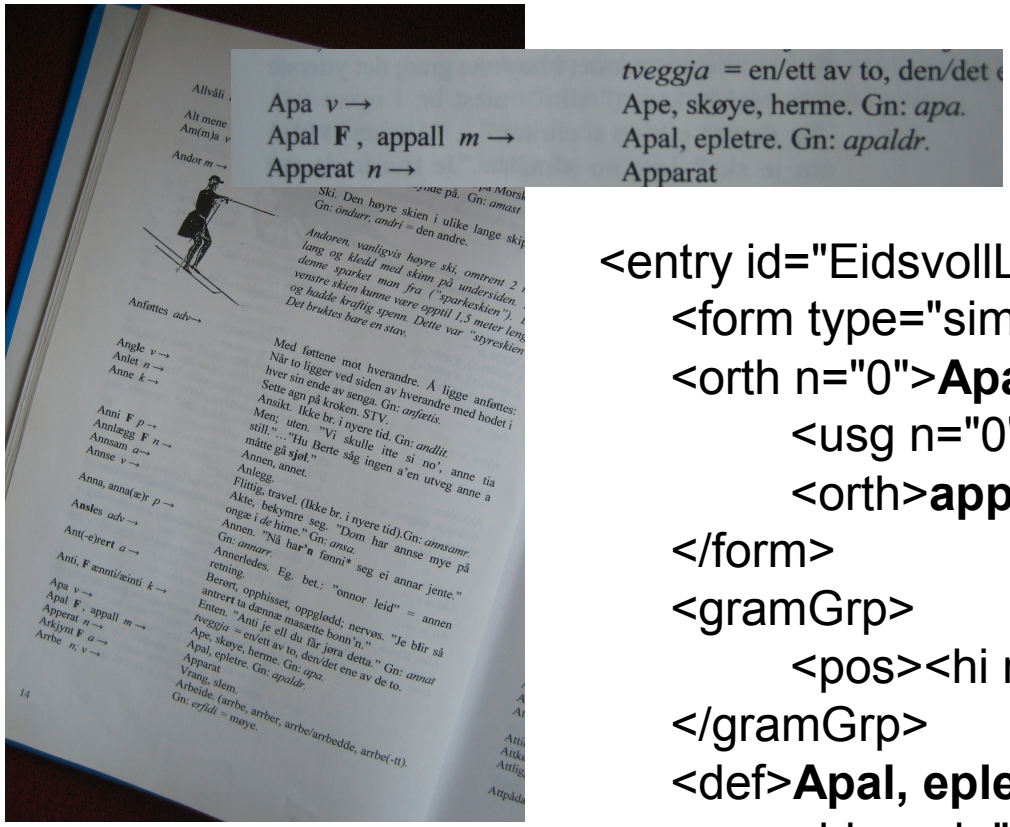
Val av tekstformat

- Ordsamlingar kjem i mange ulike dataformat.
- Treng eit format som
  - ikkje er avhengig av spesiell programvare (eit ope format),
  - som er nytta av mange og
  - som er godt dokumentert.
- Text Encoding Initiative (TEI) har utvikla eit format for koding av dokumenter innafor språk- og kulturfag. Bygger på XML-standarden.

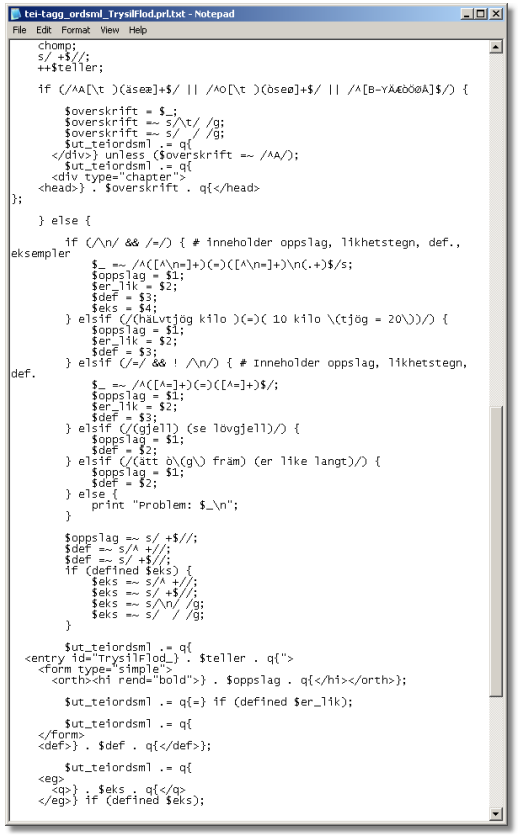
TEI-dokumentet

- TEI har eigne retningsliner for koding av ordbøker og ordlister.
- Eit TEI-dokument består av to delar:
  - Eit hode med bibliografiske opplysningar.
  - Ein tekstkropp der sjølve ordsamlinga er koda.
- Elementa i ein artikkel blir markerte med *taggar.* Kodinga skildrar *innhaldet* i artikkelen framfor *utsjånaden* i papirutgåva.

Døme




```
<entry id="EidsvollLjoedal_orig34">
    <form type="simple">
    <orth n="0">Apal</orth>
            <usg n="0"><hi rend="bold">F</hi></usg>,
            <orth>appall</orth>
    </form>
    <gramGrp>
            <pos><hi rend="italic">m</hi></pos>
    </gramGrp>
    <def>Apal, epletre. Gn:
            <hi rend="italic">apaldr.</hi>
    </def>
</entry>
```

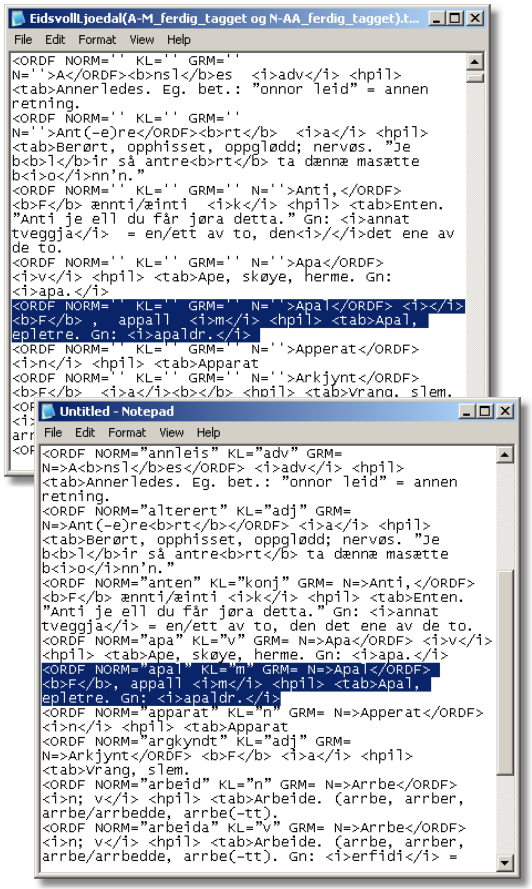


Prosess

- Tekstbehandlingsformat til tekst.
- (Normering.)
- Analyse av struktur.
- Utvikle konverteringsprogram for TEI-konvertering.
- Manuell gjennomgang.
- Innlegging i database.
- (Normering.)

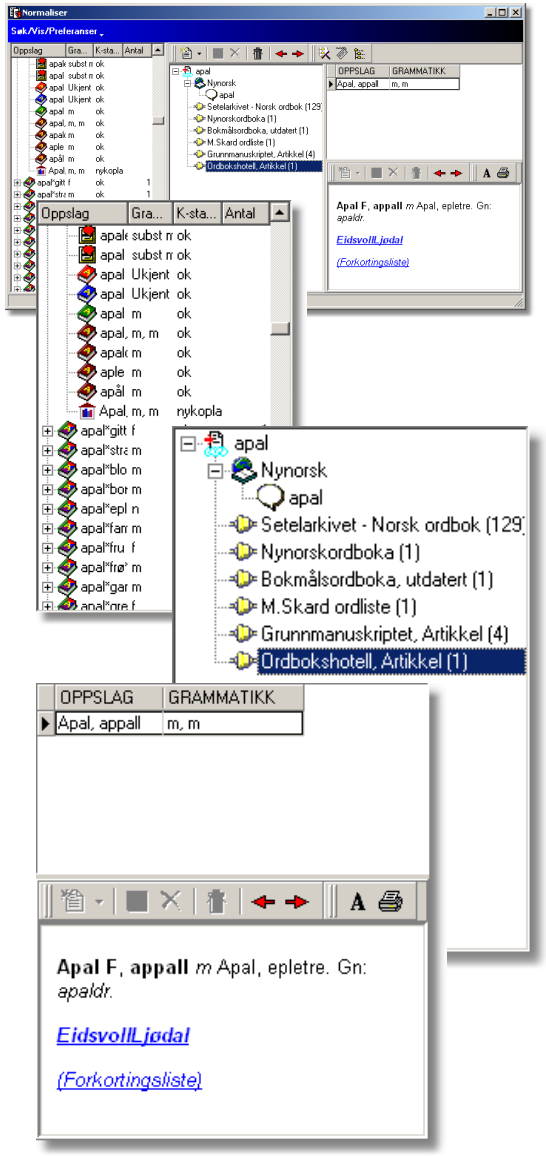
UNIVERSITETET I OSLO

Normering i ordsamlingsfila



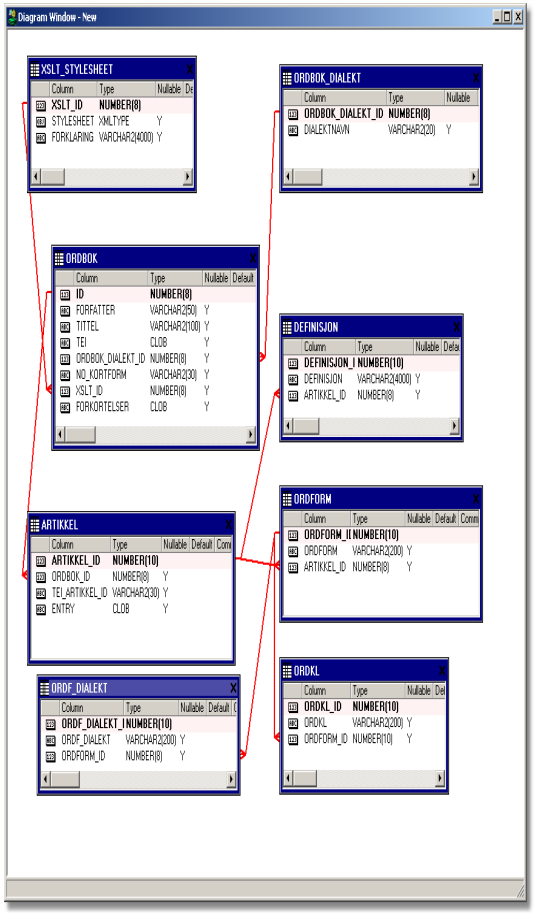
- Normeringstaggar i ordsamlingsfila.
- Fagperson normerer direkte i fila utan anna støtte enn ordbøker.
- Den ferdig normerte ordsamlinga blir konvertert til TEI og lagt inn i hotellet.
- Tidkrevande, lett å gjere feil, vanskeleg å administrere normeringa.

Normering i Metaordboka



- Metaordboka er ei lemmaliste med peikarar til kjelder.
- Ordsamlingsartiklane blir knytta til Metaordboka.
- Normert med eit normaliseringsverktøy.
- Forventa fordelar
  - Normeringa går fortare.
  - Mindre rom for feil.
  - Tilgang på alt kjeldematerialet.
  - Enklare å administrere arbeidet.

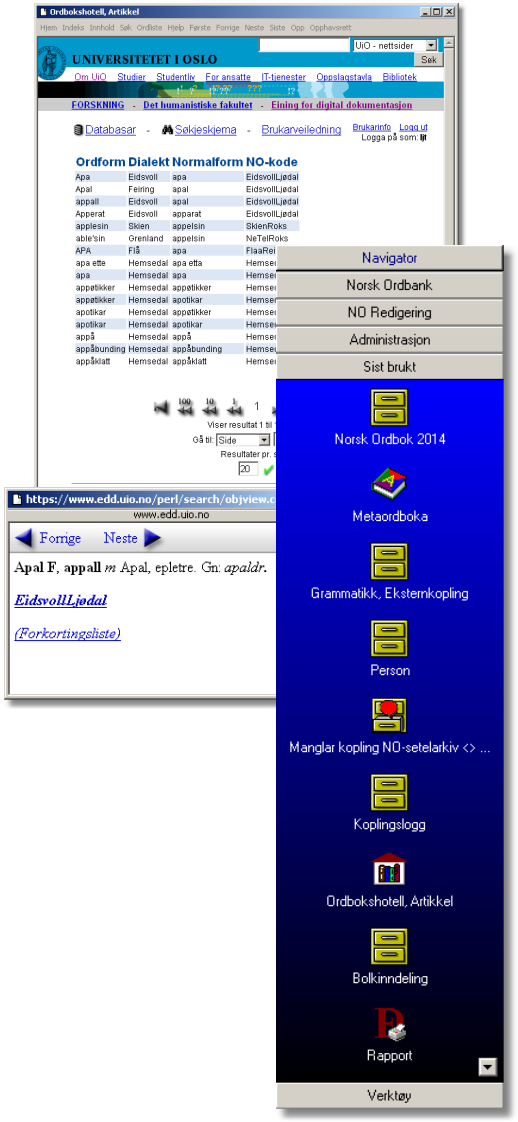
UNIVERSITETET I OSLO



Databaseformat

- Implementert i relasjonsdatabase (Oracle).
- Lagrar XML for
  - Heile ordsamlinga.
  - Kvar einskildartikkel.
- Lagrar i tillegg
  - Oppslagsformar.
  - Ordklasse.
  - Normerte data.

UNIVERSITETET I OSLO



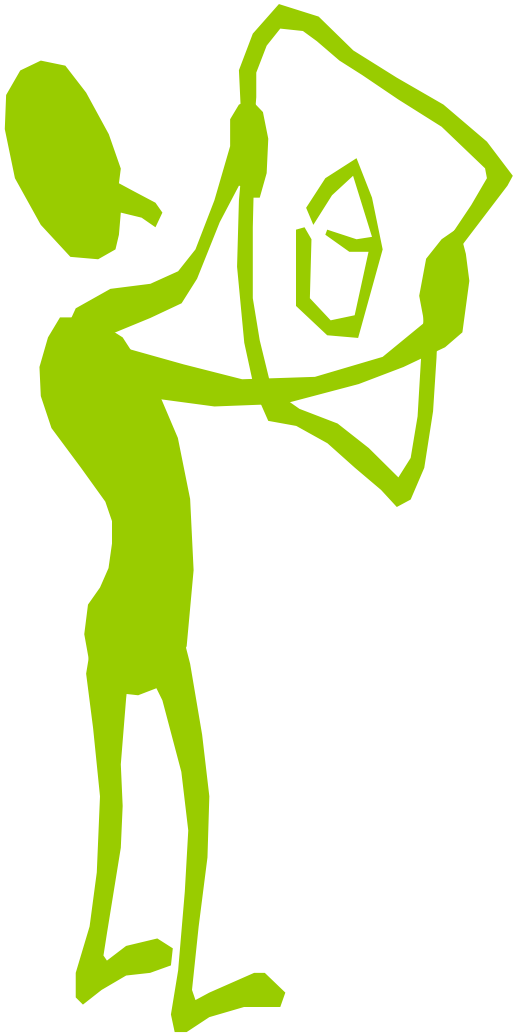
Presentasjon

- Del av felles rammeverk utvikla ved EDD.
- Integrert med redigeringssystemet for NO 2014.
- Klargjort for direkte web-publisering.

UNIVERSITETET I OSLO

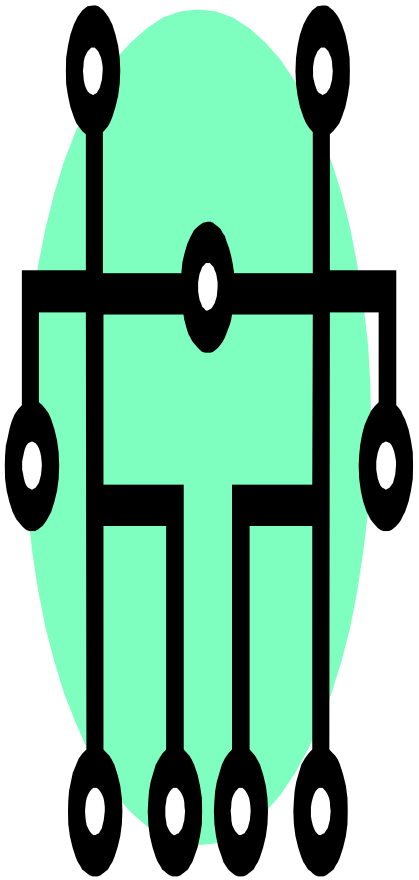
Status

- Fungerande programvare er utvikla.
- 60 samlingar innsamla.
- 38 samlingar normert.
- 5 normerte samlingar lasta inn.
- 2 unormerte samlingar lasta inn.
- 17.000 artiklar lasta inn.
- Eksperimenterer med normeringsrutinar.



Planar

- Halde fram innlegging og normering av samlingar.
- Utvikle web-presentasjonen for meir profilering av målføresamlingane.
- Vurdere teknologien for andre typar ordsamlingar (terminologilister, ...)



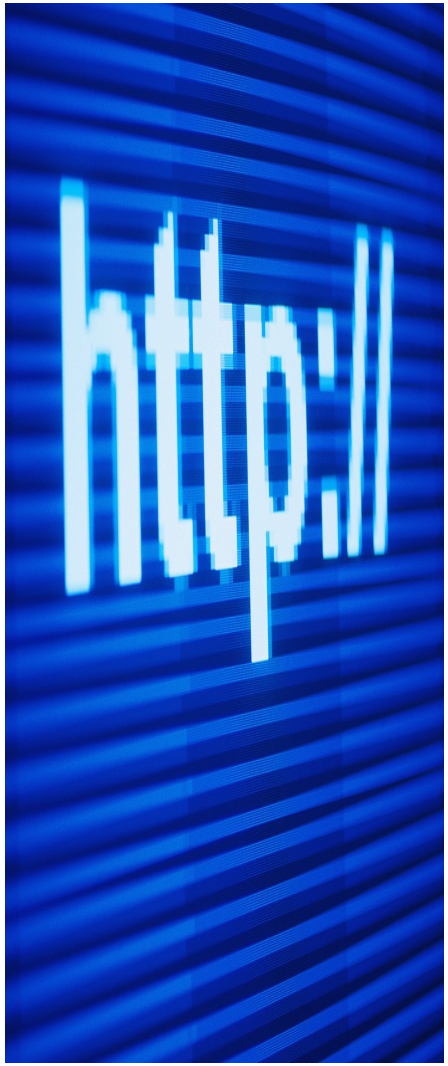
Kontaktpersonar

Øyvind Eide
oyvind.eide@muspro.uio.no
+47-22854988

Elisabeth Lien
elisabeth.lien@iln.uio.no
+47-22857016

Lars Jørgen Tvedt
l.j.tvedt@edd.uio.no
+47-22854984

UNIVERSITETET I OSLO



Peikarar

- Norsk Ordbok 2014
  http://no2014.uio.no/
- Eining for digital dokumentasjon (EDD)
  http://www.edd.uio.no/
- The Text Encoding Initiative
  http://www.tei-c.org/