



# Ordbokshotellet

## Varig lagring og formidling av norske ordsamlinger

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 ordsamlinger og ordbøker.
- Andre språk- og kulturmiljø.





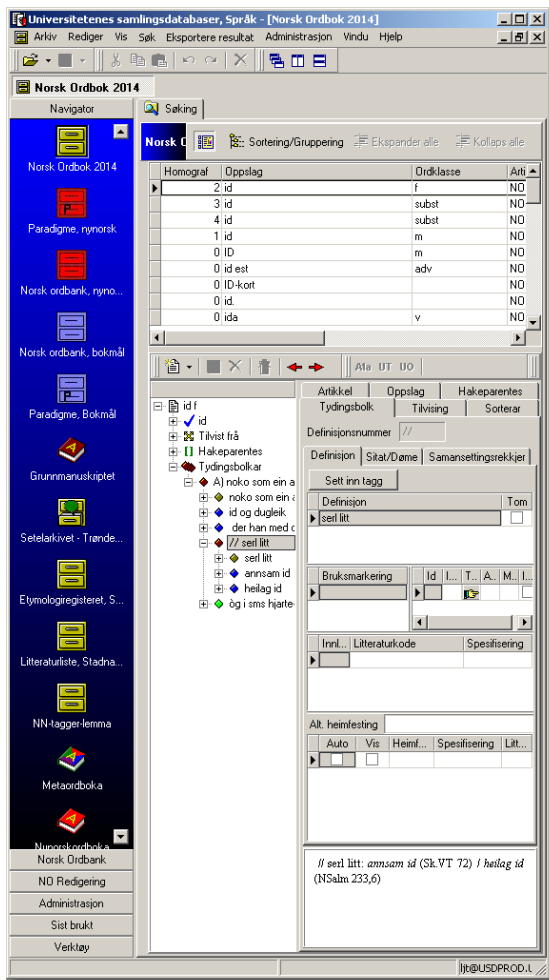
# Ordboksarkiva

Norsk Ordbok har eit omfattande kjeldemateriale på papir:

- Ordboksmanuskript
- Setelsamlingar
- Tekstsamlingar
- Ordbøker
- Ordsamlingar
- Ordlistar



# Digitalisering



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



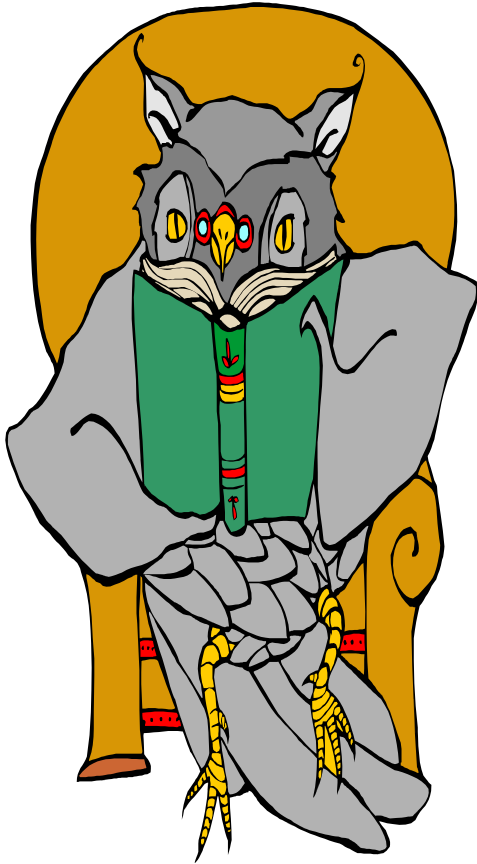


# Ordsamlingar er viktige kjelder

- 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.



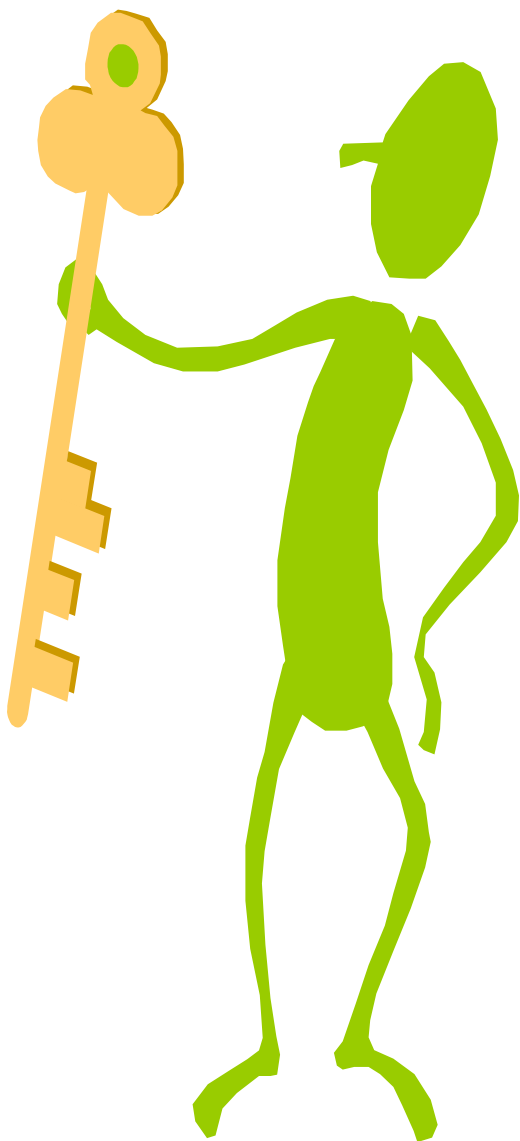
# Utfordringane



- 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

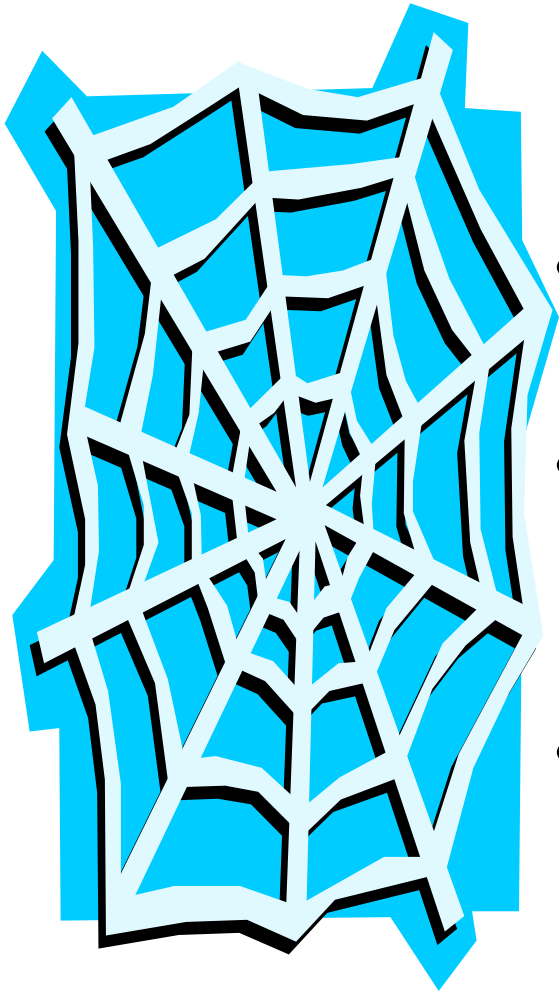
Dialekt	Ordform	Artikkel-id
Normalform		
+  Normalform : attre		Antall: 1
+  Normalform : attræ		Antall: 1
+  Normalform : attsitjande		Antall: 1
+  Normalform : attvæd		Antall: 2
+  Normalform : attægloymæ		Antall: 1
+  Normalform : attækjænnan		Antall: 1
+  Normalform : attåt		Antall: 3
+  Normalform : attåvær		Antall: 1
+  Normalform : au		Antall: 3
+  Normalform : aud		Antall: 1
+  Normalform : audegard		Antall: 1
-  Normalform : auga		Antall: 7
▶ Eidsvoll	Aue	76
Skien	aue	4780
Grenland	aue	5682
Grenland	øye	9130
Flå	GØTT AUGA TE	9652
Flå	ØUGU	10870
Flå	ØUGUN Æ STØRRE EI	10871
+  Normalform : auge		Antall: 7
+  Normalform : augemål		Antall: 1
+  Normalform : augestein		Antall: 1
+  Normalform : augestikker		Antall: 1
+  Normalform : augnelok		Antall: 1
+  Normalform : auka		Antall: 2

Aue n øye. Gn: auga.  
[EidsvollLjodal](#)  
[\(Forkortingsliste\)](#)

- 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 målføregrensene.



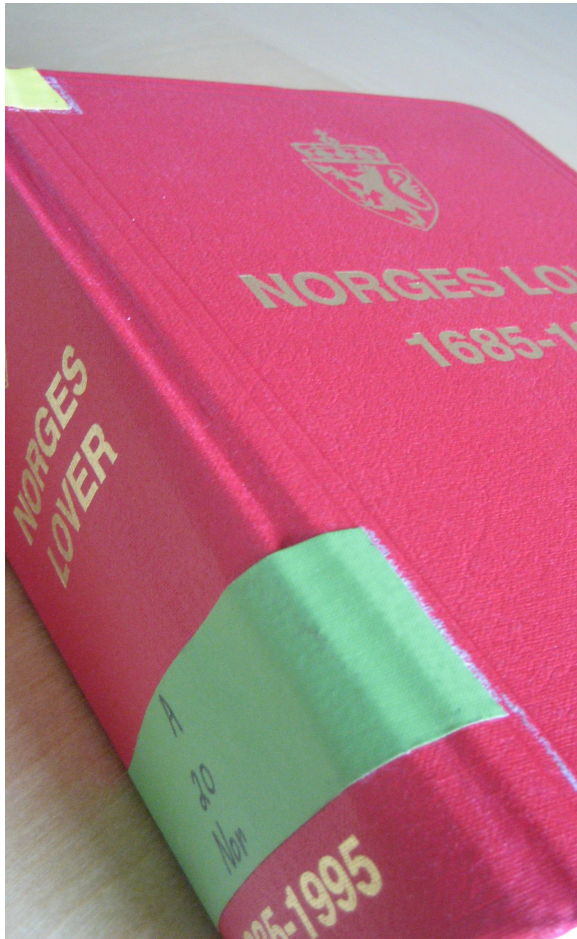
# Publisering på web



- 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.



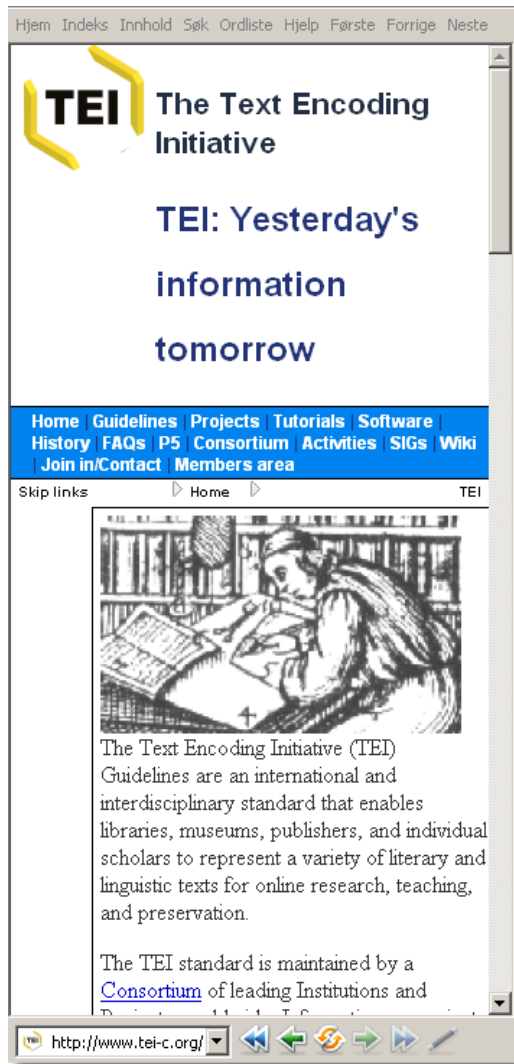
# Tekniske løysingar



- 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.





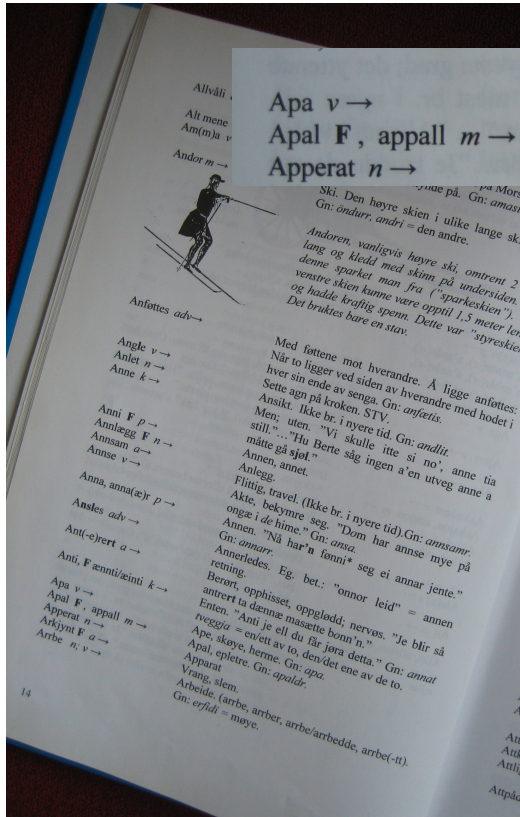
# TEI-dokumentet



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



# 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

```
tei-taggs_ordsml_Trysilflod.prt.txt - Notepad
File Edit Format View Help

chomp;
s/ +$/ /;
++$teller;

if (/^A[ \t ](asee)+$/ || /o[ \t ](osee)+$/ || /^B-V4600A[ ]$/ ) {
    $overskrift = $_;
    $overskrift =~ s/ \t //g;
    $overskrift =~ s/ / /g;
    $ut_tefordsml .= q{
</div> } unless ($overskrift =~ /AA/);
    $ut_tefordsml .= q{
<div type="chapter">
<head> . $overskrift . q{</head>
};

} else {
    if (/^N/ && /-/) { # inneholder oppslag, 11khetstegn, def.,
eksempler
        $_ =~ /A([\n=]+)(=)([ \n=]+)\n(.+)$/s;
        $oppslag = $1;
        $er_11k = $2;
        $def = $3;
        $eks = $4;
    } elsif (/^(halvtjog kilo)(=)( 10 kilo \t(jog = 20\))/) {
        $oppslag = $1;
        $er_11k = $2;
        $def = $3;
    } elsif (/^/ && ! ^N/) { # inneholder oppslag, 11khetstegn,
def.
        $_ =~ /A([=]+)(=)([=]+)$/;
        $oppslag = $1;
        $er_11k = $2;
        $def = $3;
    } elsif (/^([gjell]) (se lovgjell)/) {
        $oppslag = $1;
        $def = $2;
    } elsif (/^(\att \d\q\ fram) (er like langt)/) {
        $oppslag = $1;
        $def = $2;
    } else {
        print "Problem: $_\n";
    }
    $oppslag =~ s/ +$//;
    $def =~ s/A +//;
    $def =~ s/ +$/ /;
    if (defined $eks) {
        $eks =~ s/A +//;
        $eks =~ s/ +$/ /;
        $eks =~ s/\n //g;
    }
    $ut_tefordsml .= q{
<entry title="trysilflod_1" . $teller . q{">
<form type="simple">
<orth>chi rend="bold"> . $oppslag . q{</hi></orth>};
    $ut_tefordsml .= q{=} if (defined $er_11k);
    $ut_tefordsml .= q{
</form>
<def> . $def . q{</def>};
    $ut_tefordsml .= q{
<eg> . $eks . q{</eg>
</eg> } if (defined $eks);
```

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



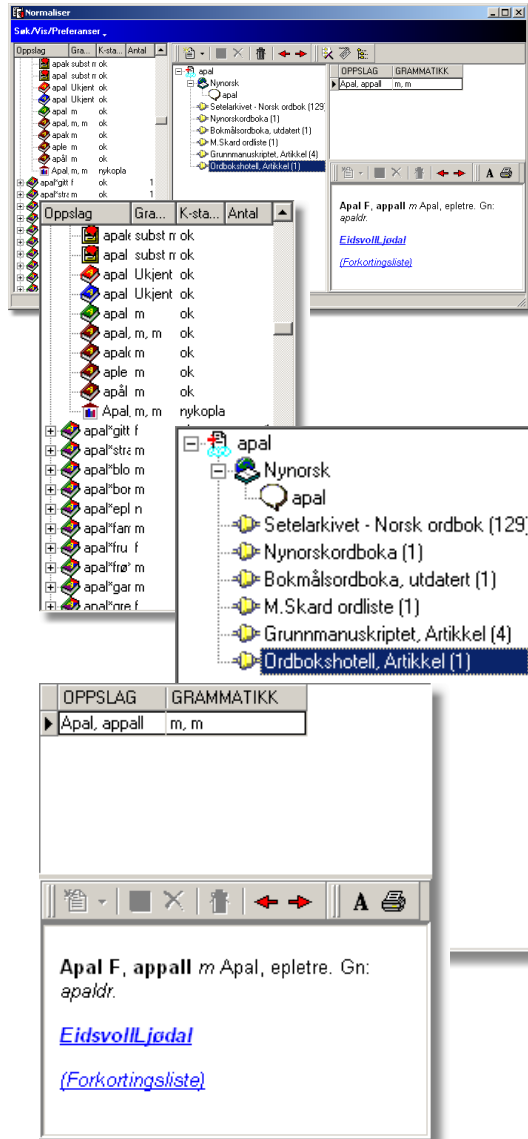
# Normering i ordsamlingsfila

```
Eidsvollljoedal(A-M_ferdig_tagget og N-AA_ferdig_tagget).t...
File Edit Format View Help
<ORDF NORM="" KL="" GRM=""
N=>A</ORDF><b>ns1</b></b>es <i>adv</i> <hpi1>
<tab>Annerledes. Eg. bet.: "onnor leid" = annen
retning.
<ORDF NORM="" KL="" GRM=""
N=>Ant(-e)re</ORDF><b>rt</b></b> <i>a</i> <hpi1>
<tab>Berørt, opphisset, oppglødd; nervøs. "Je
b<b>l</b>ir så antre<b>rt</b> ta dønnæ masatte
b<i>o</i>nn'n."
<ORDF NORM="" KL="" GRM="" N=>Antj,</ORDF>
<b>F</b> <b>ænti</b>æinti <i>k</i> <hpi1> <tab>Enten.
"Antj je e11 du får jøra detta." Gn: <i>annat
tveggjå</i> = en/ett av to, den<i>/</i>det ene av
de to.
<ORDF NORM="" KL="" GRM="" N=>Apa</ORDF>
<i>v</i> <hpi1> <tab>Ape, skøye, herme. Gn:
<i>apa.</i>
<ORDF NORM="" KL="" GRM="" N=>Apal</ORDF> <i></i>
<b>F</b> <b>appall <i>m</i> <hpi1> <tab>Apal,
epletre. Gn: <i>apaldr.</i>
<ORDF NORM="" KL="" GRM="" N=>Apparat</ORDF>
<i>n</i> <hpi1> <tab>Apparat
<ORDF NORM="" KL="" GRM="" N=>Arkjynt</ORDF>
<b>F</b> <i>a</i><b> <hpi1> <tab>vrang, slem.
</i>
</i>
arr
<ORDF NORM="" annleis" KL="adv" GRM=
N=>A<b>ns1</b></b>es</ORDF> <i>adv</i> <hpi1>
<tab>Annerledes. Eg. bet.: "onnor leid" = annen
retning.
<ORDF NORM="" alterert" KL="adj" GRM=
N=>Ant(-e)re<b>rt</b></b></ORDF> <i>a</i> <hpi1>
<tab>Berørt, opphisset, oppglødd; nervøs. "Je
b<b>l</b>ir så antre<b>rt</b> ta dønnæ masatte
b<i>o</i>nn'n."
<ORDF NORM="" anten" KL="konj" GRM= N=>Antj,</ORDF>
<b>F</b> <b>ænti</b>æinti <i>k</i> <hpi1> <tab>Enten.
"Antj je e11 du får jøra detta." Gn: <i>annat
tveggjå</i> = en/ett av to, den det ene av de to.
<ORDF NORM="" apa" KL="v" GRM= N=>Apa</ORDF> <i>v</i> <hpi1>
<tab>Ape, skøye, herme. Gn: <i>apa.</i>
<ORDF NORM="" apal" KL="m" GRM= N=>Apal</ORDF>
<b>F</b> <b>appall <i>m</i> <hpi1> <tab>Apal,
epletre. Gn: <i>apaldr.</i>
<ORDF NORM="" aparat" KL="n" GRM= N=>Apparat</ORDF>
<i>n</i> <hpi1> <tab>Apparat
<ORDF NORM="" argkyndt" KL="adj" GRM=
N=>Arkjynt</ORDF> <b>F</b> <i>a</i> <hpi1>
<tab>vrang, slem.
<ORDF NORM="" arbeid" KL="n" GRM= N=>Arrbe</ORDF>
<i>n</i> <hpi1> <tab>Arbeide. (arrbe, arrber,
arrbe/arrbedde, arrbe(-tt)).
<ORDF NORM="" arbeida" KL="v" GRM= N=>Arrbe</ORDF>
<i>n</i> <hpi1> <tab>Arbeide. (arrbe, arrber,
arrbe/arrbedde, arrbe(-tt). Gn: <i>erfid</i> =
```

- Normeringstagger 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.



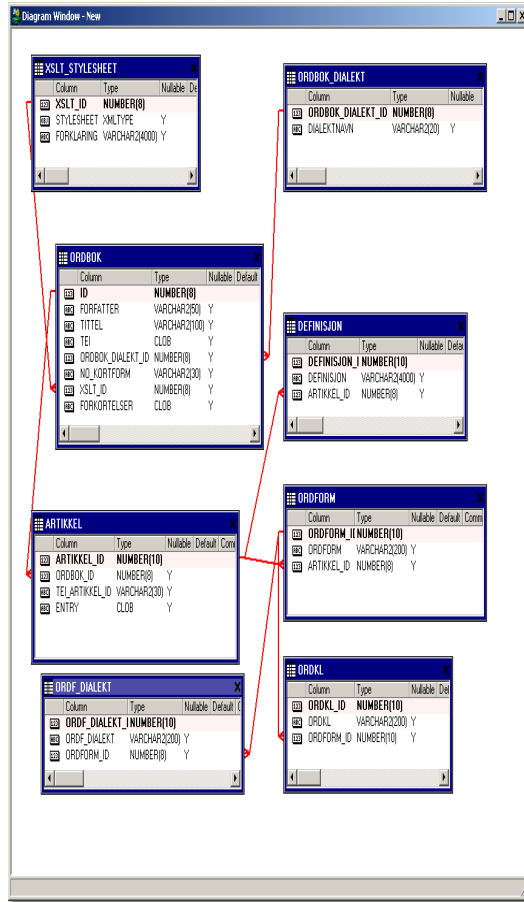
# Normering i Metaordboka



- 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.



# Databasformat



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



# Presentasjon

The screenshot displays the EDD web application interface. At the top, there is a navigation bar with the University of Oslo logo and various menu items. Below this, a search result is shown for the term 'Apal F, appall m Apal, epletre. Gn: apaldr.'. The result includes a link to 'EidsvollLjodal' and a note '(Forkortingsliste)'. A vertical navigation menu is overlaid on the right side of the page, listing various features and tools such as 'Norsk Ordbank', 'NO Redigering', 'Administrasjon', 'Sist brukt', 'Norsk Ordbok 2014', 'Metaordboka', 'Grammatikk, Eksternkopling', 'Person', 'Manglar kopling NO-setelarkiv <> ...', 'Koplingslogg', 'Ordbokshotell, Artikkel', 'Bolkinnndeling', 'Rapport', and 'Verktøy'.

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



# Status

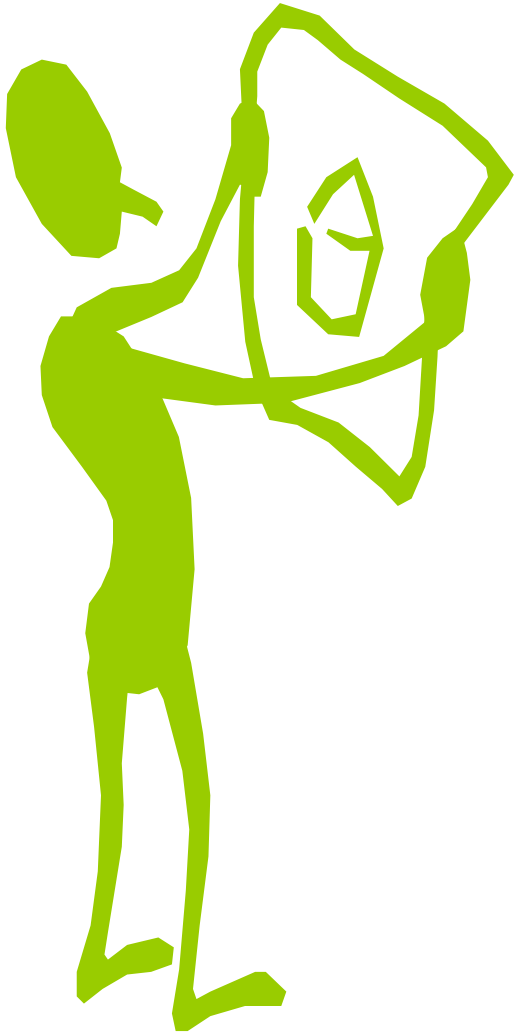


- Fungerande programvare er utvikla.
- 60 samlingar innsamla.
- 38 samlingar normert.
- 5 normerte samlingar lasta inn.
- 2 unormerte samlingar lasta inn.
- 17.000 artiklar lasta inn.
- Eksperimenterer med normeringsrutinar.



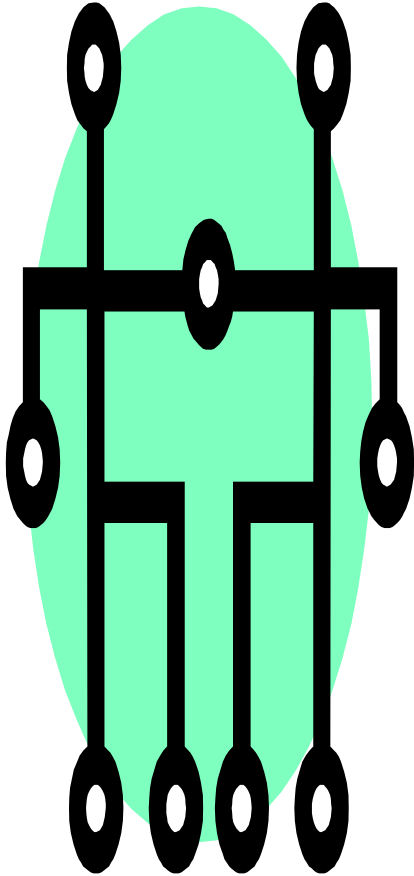


# Planar



- 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](mailto:oyvind.eide@muspro.uio.no)

+47-22854988

**Elisabeth Lien**

[elisabeth.lien@iln.uio.no](mailto:elisabeth.lien@iln.uio.no)

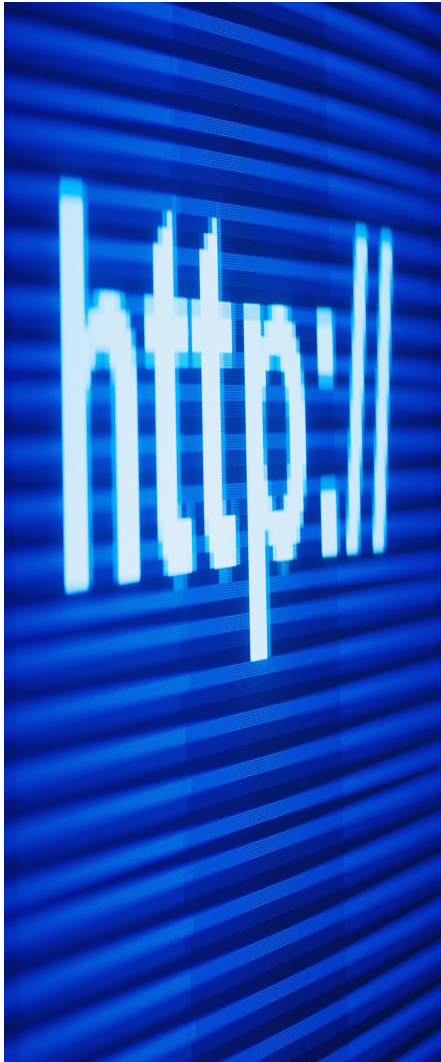
+47-22857016

**Lars Jørgen Tvedt**

[l.j.tvedt@edd.uio.no](mailto:l.j.tvedt@edd.uio.no)

+47-22854984





# 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/>

