UNIVERSITETET
I OSLO

Linking large lexical materials

Åse Wetås and Oddrun Grønvik

CLARIN, Solstrand, Bergen

15. December 2008

Norsk
Ordbok
2014

The Nynorsk collections - goals

Obtain coverage of the Nynorsk period of Norwegian language history (1600 – until today) for all dialects and for written Nynorsk (from the Landsmål period onwards).

To digitise as much as possible of

- dialect materials before 1900
- the earliest Nynorsk materials, with emphasis on fact literature
- the Nynorsk literary canon
- Nynorsk text in adequate samples at regular intervals now and in the future

What kind of materials?

Non-digital research materials UiO 1900 - 2000

Original collected materials from the 1930s onwards

- Paper (slip archives, maps, hand written and typed manuscripts etc)
- Sound conservation items (wax rolls, large tapes, cassette tapes)

Books and from ca 1800 onwards (representing text from 1600 – 2008) – ca 4000 i all (so far)

Bound copies of newspapers (some in poor condition)

A lot is digitised, the rest waiting for digitising

Original digital research materials

Electronic texts (excerpts, full texts)

Databases

What is digitised?

In databases

Nynorsk **Slip Archive** of 3,2 mill slips

”**The Dictionary Inn**” (Ordbokshotellet) : Ca 20 dialect dictionaries

Dictionaries: NOB/BOB, Grunnmanuskriptet, Torp, Skard

Ca 300 titles (signatures) for the Nynorsk corpus – 36 mill words

- 10 years of newspaper text
- 17 years of journal text

The **Dialect Synopsis** (images)

Not in databases

A lot of text (some hundred books, 50 older dictionaries etc) digitised but not searchable through databases or the corpus

Norsk dialektatlas - 550 maps with legends (images and text)

Characterised by diversity

Large lexical materials characterised by diversity in

- Time (1600 – 2008)
- Genre (from short utterances of transcripted speech to encyclopedias)
- Size of document (from ca 100 word text files to huge images)
- Inherent categorical complexity (from plain modern text to multi-layered philological editions)

Standards for digital conservation

Text materials

- Sign system: Unicode (except Norvegia-font)
- Txt-files (structured)
- Tagging sgml – xml (TEI)

Images (including images of maps)

- TIFF for long term storage

Digital maps

- GML vector maps (new international standard)
- Norwegian national standard
- Looking at web solution
- Software: ArcMap

Sound

- WAV for conservation
- Linking of utterance and sound in xml (TEI)
- (software Transcriber)

Programming earlier in Delphi, moving over to Java

All databases in Oracle

Access

The public has

- **Free** access to products that are published over the web
- **Controlled** access to most of the materials for further use in research (ask for permit, get access to relevant materials)
- **Restricted** access to materials covered by copyright legislation (must get permission from copyright holders)

Access through FEIDE – not yet

Possible, but not yet discussed (would take capacity)

Linking resources

To link and index resources is an integral part of our practice!

Norsk Ordbok 2014

- Hundreds of sources can be linked within one entry
- All materials will be linked and interpreted through the finished dictionary

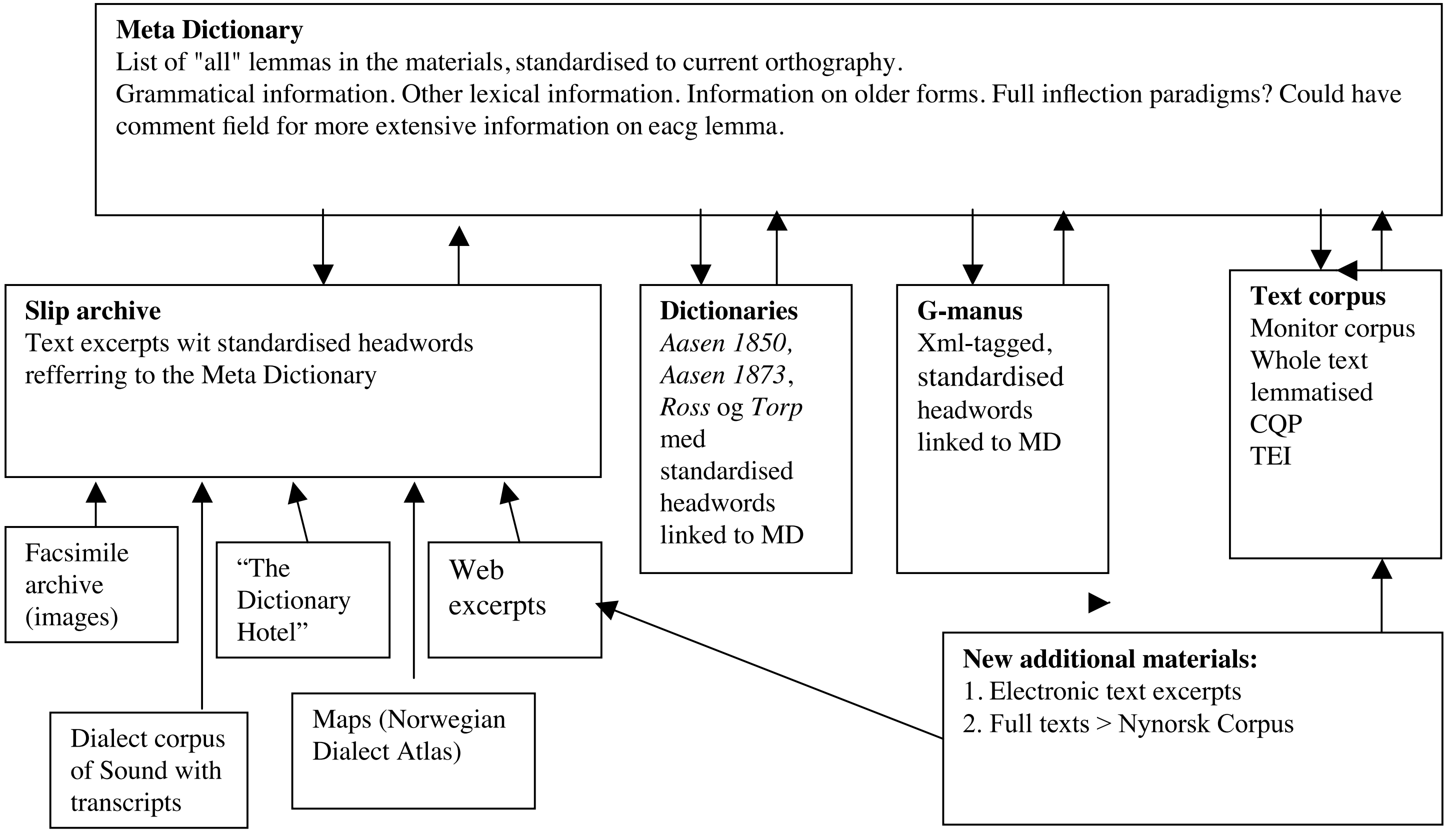
The Meta dictionary

- Links all raw materials (including NOB and BOB) under standard nynorsk as the index language

The Nynorsk Corpus

- Monitor corpus linking Nynorsk texts from 1866 to 2008

A model for the organisation of UiO’s Nynorsk resources



Meta Dictionary
List of "all" lemmas in the materials, standardised to current orthography.
Grammatical information. Other lexical information. Information on older forms. Full inflection paradigms? Could have comment field for more extensive information on eacg lemma.

Slip archive
Text excerpts wit standardised headwords refferring to the Meta Dictionary

Dictionaries
Aasen 1850, *Aasen 1873*, *Ross* og *Torp* med standardised headwords linked to MD

G-manus
Xml-tagged, standardised headwords linked to MD

Text corpus
Monitor corpus
Whole text lemmatised
CQP
TEI

Facsimile archive (images)

“The Dictionary Hotel”

Web excerpts

Dialect corpus of Sound with transcripts

Maps (Norwegian Dialect Atlas)

New additional materials:
1. Electronic text excerpts
2. Full texts > Nynorsk Corpus

Frå *Sluttrapport for Delprosjekt nynorsk* (1999)

Linking to resources outside

Two ways

Others link themselves to our resources (at macro or micro level)

- How will changes that we make, affect the outside links?

We link our materials to other nodes (grids, networks …)

No objection in principle!

Question of capacity and funding