NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Massive ordbokspalter i transparent presentasjon – NAOB på nett: samspillet mellom datastruktur, innhold og brukergrensesnitt

Forfatter: Petter Henriksen

Kilde: Nordiska Studier i Lexikografi 11, 2012, s. 251-262
Rapport från Konferens om lexikografi i Norden, Lund 24.-27. maj 2011

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive

© Nordisk forening for leksikografi

Betingelser for brug af denne artikel
Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.

Søgbarhed
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

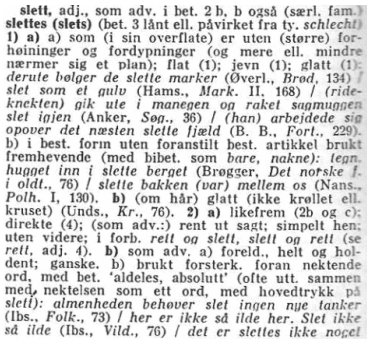
Massive ordbokspalter i transparent presentasjon – NAOB på nett: samspillet mellom datastruktur, innhold og brukergrensesnitt

Petter Henriksen

The *Norwegian Academy's Dictionary* (NAOB) is planned to be published digitally in 2014. The transformation of the massive 20th-century dictionary *Norsk Riksmålsordbok* into a user-orientated publication on the Internet poses special challenges to data structuring and presentation techniques. The NAOB project aims to provide the Norwegian public and research community with an easily accessible and authoritative lexicographical documentation of the majority language variants Bokmål and Riksmål, and, also, a suitable platform for future updating and development. Internet publishing offers a new world of opportunities for user-friendly presentation compared to last century's book publishing format – if one makes the right choices regarding data structure. In this article, the author gives a systematic presentation of the NAOB project's insights and solutions, hopefully of relevance for other extensive national dictionary projects in the future.

Det Norske Akademis Store Ordbok (NAOB) blir publisert i 2014, ikke på papir, men kun som en digital ordbok på Internett. Forrige gang dette innholdet ble publisert, var det kun på papir – det ærverdige gamle storverket Norsk Riksmålsordbok. Underveis har ordbokens innhold blitt radikalt transformert fra de massive bokspaltene som møtte brukerne i forrige århundre. Også vi selv i redaksjonen har blitt nokså transformert i løpet av prosessen, og forhåpentlig en del klokere. Vi vil gjerne dele våre erfaringer – kall det gjerne våre bedriftshemmeligheter – med de nordiske leksikografene i denne artikkelen.

Oppgaven er velkjent blant forlag som utgir ordbøker. Man begynner med (A) grafiske filer for bokutgivelse – eller enda mer primitivt som i vårt tilfelle – bare bokspalter (figur 1).



slett, adj., som adv. i bet. 2 b, b også (særl. fam.) **slettes (slets)** (bet. 3 lånt ell. påvirket fra ty. *schlecht*) **1) a)** a) som (i sin overflate) er uten (større) forhøininger og fordypninger (og mere ell. mindre nærmer sig et plan); flat (1); jevn (1); glatt (1): *derute bølger de slette marker* (Øverl., *Brød*, 134) / *slet som et gulv* (Hams., *Mark*. II. 168) / *(rideknekten) gik ute i manegen og raket sagmuggen slet igjen* (Anker, *Søg*., 36) / *(han) arbejdede sig opover det næsten slette fjæld* (B. B., *Fort*., 229). b) i best. form uten foranstilt best. artikkel brukt fremhevende (med bibet. som *bare, nakne*): *tegn. hugget inn i slette berget* (Brøgger, *Det norske f. i oldt*., 76) / *slette bakken (var) mellem os* (Nans., *Polh*. I, 130). **b)** (om hår) glatt (ikke krøllet ell. kruset) (Unds., *Kr*., 76). **2) a)** likefrem (2b og c); direkte (4); (som adv.:) rent ut sagt; simpelt hen; uten videre; i forb. *rett og slett, slett og rett* (se *rett*, adj. 4). **b)** som adv. a) foreld., helt og holdent; ganske. b) brukt forsterk. foran nektende ord, med bet. 'aldeles, absolutt' (ofte utt. sammen med nektelsen som ett ord, med hovedtrykk på *slett*): *almenheden behøver slet ingen nye tanker* (Ibs., *Folk*., 73) / *her er ikke så ilde her. Slet ikke så ilde* (Ibs., *Vild*., 76) / *det er slettes ikke noget*

Man skal ende med (B) en eller annen form for leksikografisk struktur som egner seg både for digital publisering og for fremtidig oppdatering og redigering. Hva er veien fra A til B og hvilke kritiske veivalg må man ta?

NAOB er en revisjon av den gamle Norsk Riksmålsordbok, enn så lenge Norges eneste komplette mangebinds ordboksverk. Norsk Riksmålsordbok ble utgitt fra 1937 til 1957, i fire bind med høyst varierende tykkelse, med det forvokste bind 4 som et ruvende monument over leksikografisk skaperfryd og forlagsadministrativ avmakt. I 1995 kom to nye bind med supplerende innhold, i et separat alfabet. På slutten av 1990-tallet startet prosjektet med å digitalisere verket med dét for øye å spleise de to alfabetene sammen i ett i en ny bokutgave til jubileumsåret 2014, samme år som Norge får sitt nasjonale ordboksverk for nynorsk – *Norsk Ordbok 2014*.

Underveis i prosjektet grydde erkjennelsen av at det ikke var et bokverk Norge ville ha, men en Internett-utgivelse. Og med det kom erkjennelsen av at dataenes struktur kom til kort, nå som man skulle publisere på den nådeløst avslørende arena som nettpublisering er. For ikke lenge siden fikk NAOB en fornyet datastruktur, og dataene gjennomgikk en storstilt konvertering. Resultatet er at vi nå har et redigeringsgrensesnitt og et grunnlag for et publiseringsgrensesnitt som fungerer godt.

Gjennom alt dette har vårt miljø fått finpusset sin metodikk for transformasjon av tunge ordbøker fra bokformat til Internett-format. Den presenteres her.

Strategisk vurdering

Første steg er en strategisk vurdering. Man må bestemme om målet for prosjektet er *dynamisk*, dvs. et verk som skal leve videre gjennom stadig nye oppdateringer, eller *statisk*, dvs. et verk som egentlig bare skal gjenskapes slik det var på papir, med noen punktvise oppdateringer. I fall målet er statisk, kan man

komme lettvint fra det, med en digital variant av faksimilen. Dette alternativet skal ikke drøftes her. Hvis målet er et dynamisk verk, må man åpne sinnet – og lommeboken – noe mer. Anbefalt metodikk er som følger, internt i prosjektet benevnt *drømmemodellen*:

- Man setter seg inn i *kildeverkets* innhold – hva som ligger der av artikkelinnhold å strukturere.
- Man legger så kildeverket bort og skaper sin idealversjon av det samme verket ("*drømmeverket*"), dvs. den ideelle oppbygning av verket og den enkelte artikkel.
- Man setter i gang med å omforme kildeverk til drømmeverk.

Med denne tilnærmingsmåten tar man konsekvensen av at komplekse dataprosjekter blir uoverstigelig vanskelige dersom man ikke frigjør seg fra forelegget. Drømmemodellen innebærer at man tar et oppgjør med klassiske motforestillinger i forlagsbransjen: Man ignorerer dem som sier at man ikke må miste informasjonen om det grafiske bildet i den gamle bokutgaven. Man ignorerer dem som sier at det blir mye å rette i de gamle dataene for å få dem til å konformere med idealet. Og hvis noen mener at man i utgangspunktet skal spare ressurser på ikke å implementere et system for lenkede henvisninger og automatisk betydningsnummerering – ignorerer man dem også.

Tiltak[1]

Med den strategiske vurderingen tilbakelagt går man over til konkrete oppgaver. Vi skal fra *kildeverk* til *målverk*, fra historisk verk til drøm. Veien går gjennom:

1. Analyse av kildeartikkel
2. Design av målartikkel
3. Strukturering av målartikkel ved hjelp av DTD
4. Konvertering av kildedata, inkl. evt. optisk lesning
5. Spesifisering av stilark
6. Redigering

I det følgende gjennomløpes stegene med NAOB som eksempel.

[1] Viss information i nedanstående figurer har gått förlorad på grund av svart-vitt tryck istället för färg (red:s anm.).

1) Analyse av kildeartikkel

Analysen av kildeartiklene viser at man har velkjente leksikografiske kategorier som oppslagsord, ordklasse, varianter av oppslagsordet, etymologi, betydninger med definisjoner, henvisninger, betydningsskillere, sitater, fraser og mange flere. Kategoriene has i mente.

2) Design av målartikkel

Her skal man tenke resultatorientert, dvs. på vegne av de fremtidige brukerne man ønsker å hjelpe og glede. Man tegner ut i ikke-kodet form – gjerne i Word – hvordan man vil at artikkelen skal ta seg ut visuelt og funksjonelt når avdukingen finner sted og brukeren møter den for første gang på nettet. Siden skal man bevege seg bakover fra dette i kjede i en form for teleologisk metode. Her er eksempler fra NAOB på aspekter som ønskes i målartikkelens design:

A) Det ønskes at artikkelen skal se luftig ut, med tydelig markering av de leksikografiske elementene, med intuitivt forståelige ledetekster, med hierarkisk innrykking og nummerering av betydningsnivåene (figur 2).

slett adj.
BØYNING -
UTTALE [slet:]
ETYMOLOGI av norr. *sléttr*; betydning 2 etter ty. *Schlichting*

BETYDNING OG BRUK

1 flat; jevn
SITATER
- *derute bølger de slette marker* (Øverl. *Brød* 134)
- *slet som et gulv* (Hams. *Mark. II* 168)

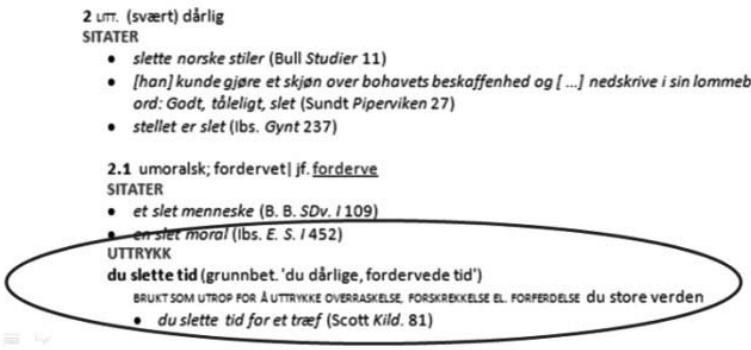
1.1 I BEST. FORM UTEN FORANSTILT BEST. ARTIKKEL BRUKT FREMHEVENDE glatt og bar; naken
SITATER
- *tegn, hugget inn i slette berget* (Brøgger *Det norske f. i oldt.* 76)
- *slette bakken [var] mellem os* (Nans. *Polh. I* 130)

1.2 OM HÅR glatt (ikke krøllete el. krusete)

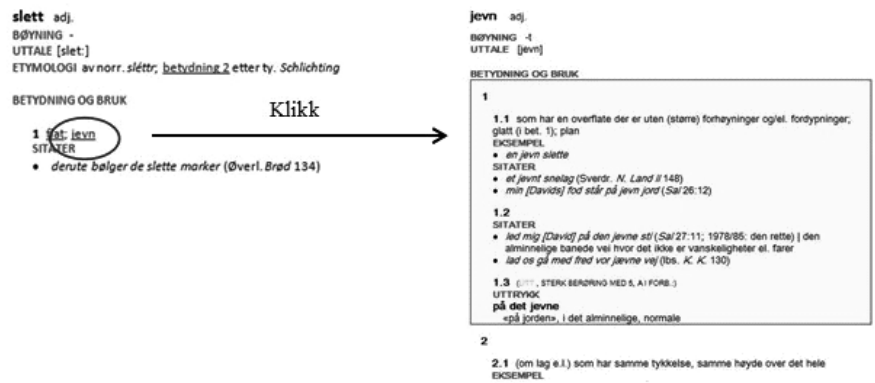
2 LITT. (svært) dårlig
SITATER
- *slette norske stiler* (Bull *Studier* 11)
- *[han] kunde gjøre et skjøn over bohavets beskaffenhed og […] nedskrive i sin lommebog et af de tre ord: Godt, tåleligt, slet* (Sundt *Piperviken* 27)
- *stellet er slet* (Ibs. *Gynt* 237)

2.1 umoralsk; fordervet| jf. forderve
SITATER
- *et slet menneske* (B. B. *SDv. I* 109)
- *en slet moral* (Ibs. *E. S. I* 452)
UTTRYKK
du slette tid (grunnbet. 'du dårlige, fordervede tid')
BRUKT SOM UTROP FOR Å UTTRYKKE OVERRASKELSE, FORSKREKKELSE EL. FORFERDELSE du store verden
- *du slette tid for et træf* (Scott *Kild.* 81)

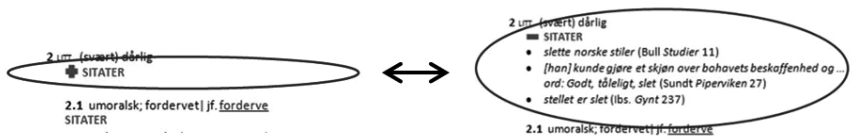
B) Det ønskes at idiomene (de såkalte uttrykkene), skal plasseres under sine respektive betydninger og danne små subartikler, sublemmaer (figur 3).



C) Det ønskes klikkbare henvisninger, som bringer brukeren til lenkemålet med markering av den delen av artikkelen som det vises til (figur 4).



D) Det ønskes at artikkelledd som sitatseksjonen skal kunne kontraheres og ekspanderes (figur 5).



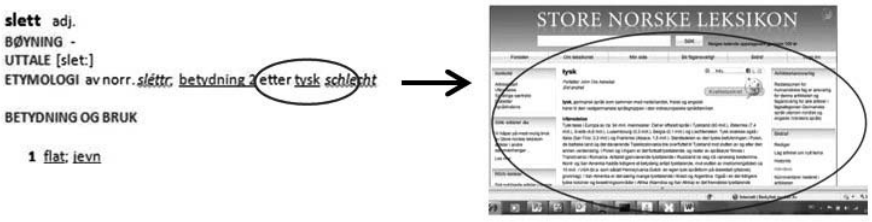
E) Det ønskes innholdsfortegnelser med lenkeklikk som fører til enkeltbetydninger og enkeltfraser (figur 6).

slett adj.

INNHOLD
Bøyning
Uttale
Etymologi
Betydning og bruk
1 flat; jevn
1.1 glatt og bar; naken
1.2 OM HÅR glatt (ikke krøllete el. krusete)
2 (svært) dårlig
2.1 umoralsk; fordervet
3 helt og holdent
3.1 aldeles; absolutt
3.2 BARNESPRÅK MED REFL. PRON.

BØYNING -

F) Det ønskes klikkbare lenker også til eksterne nettsteder, f.eks. for språkene omtalt i etymologiene (figur 7).



Med de ideelle ønskene ferdig designet i et tenkt brukergrensesnitt, utledes de forutgående stegene teleologisk av dette.

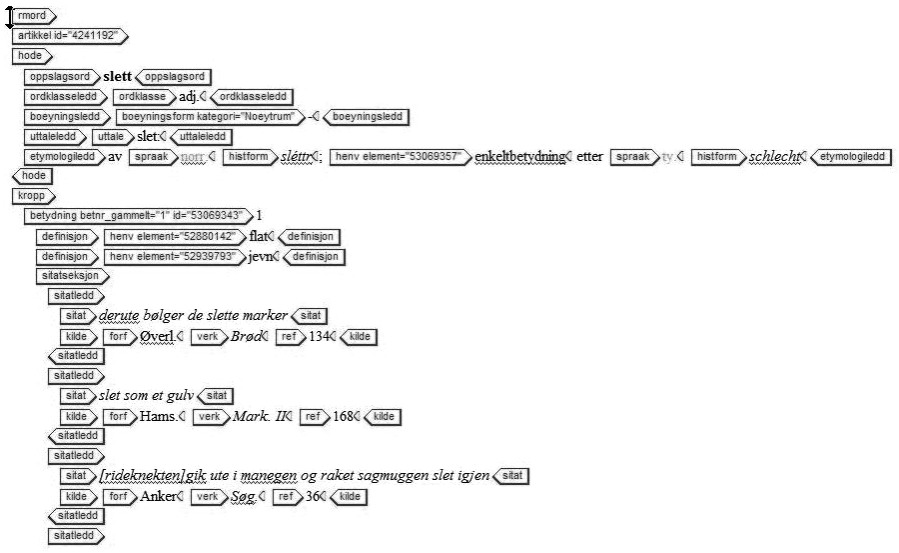
3) Strukturering av målartikkel ved hjelp av DTD

Neste steg er å bestemme hvordan dataene må være strukturert for å oppnå det ønskede visuelle og funksjonelle uttrykket. Verktøyet til dette er DTD-en, som er den formallogiske beskrivelsen av artikkelens innhold når innholdet har XML-format. DTD står for Document Type Definition, og brukes bak kulissene i redigeringsprogrammet til å regulere hvilke innholdstyper (leksikografiske kategorier) som er tillatt hvor i artikkelen. Figur 8 viser hele NAOBs struktur i pseudo-DTD-format.

artikkel = hode, kropp
hode = oppslagsord, ordklasseledd?, boeyningsledd?, uttaleledd?, variantledd?, etymologiledd?
oppslagsord = karakter-inlinetekst
ordklasseledd = redaksjonsspråk-inlinetekst / ordklasse
ordklasse = karakter-inlinetekst
boeyningsledd = redaksjonsspråk-inlinetekst / genus / boeyningsform
boeyningsform = karakter-inlinetekst
genus = karakter-inlinetekst
grammatikk = karakter-inlinetekst
uttaleledd = redaksjonsspråk-inlinetekst / uttale
uttale = karakter-inlinetekst
variantledd = redaksjonsspråk-inlinetekst / variant
variant = karakter-inlinetekst
etymologiledd = redaksjonsspråk-inlinetekst/ histform / grunnbetydning / appellativisering
spraak = karakter-inlinetekst
histform = karakter-inlinetekst
grunnbetydning = karakter-inlinetekst
appellativisering = karakter-inlinetekst
kropp = betydning*
betydning = betydningsskiller*, definisjon*, utdypning*, redeksseksjon?, sitatseksjon?, uttrykksseksjon?, betydning*
betydningsskiller = redaksjonsspråk-inlinetekst
definisjon = redaksjonsspråk-inlinetekst
utdypning = redaksjonsspråk-inlinetekst
redeksseksjon = redeksledd*
redeksledd = innledning*, redeks, utdypning*
innledning = redaksjonsspråk-inlinetekst
redeks = redaksjonsspråk-inlinetekst
sitatseksjon = sitatledd*
sitatledd = innledning*, sitat?, kilde*, utdypning*
sitat = typografi-inlinetekst
kilde = redaksjonsspråk-inlinetekst
uttrykksseksjon = uttrykksledd*
uttrykksledd = uttrykk?, variantledd?, etymologiledd?, uttrykksbetydning*
uttrykk = typografi-inlinetekst
uttrykksbetydning = betydningsskiller*, definisjon*, utdypning*, redeksledd*, sitatledd*

Her i NAOBs DTD finner man igjen karakteristiske strukturelementer som uttrykksseksjonen med idiomer (uttrykksledd med uttrykk) som subartikler, og det rekursive betydningshierarkiet (betydninger som kan inneholde betydninger).

Når artikkelen er strukturert som XML i henhold til DTD-en, omgis de ulike leksikografiske enhetene med tagger som forteller hva de er. Innenfor XML-terminologi kalles taggede enheter for *elementer*. Figur 9 viser NAOB-data i XML-format.



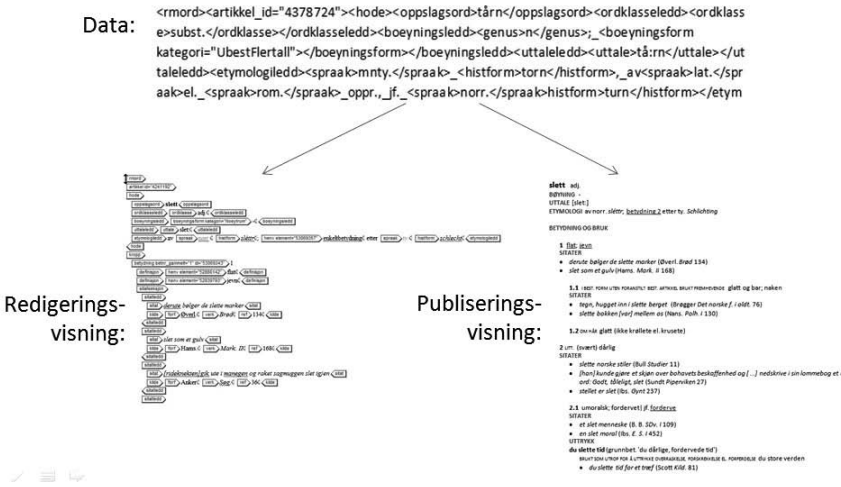
4) Konvertering av kildedata

Neste steg er det brutale møtet med virkeligheten – transformasjonen av eksisterende data til drømmestruktur, eventuelt innledet av optisk lesning dersom utgangspunktet er bokspalter. Til dette lager man en så god oppskrift som man kan, og forsoner seg med at den i mange tilfeller vil misforstå og feiltolke, og at man siden må foreta en full gjennomredigering for å fange opp feiltolkingene. I kildedataene for NAOB er f.eks. sitater, kollokasjoner og idiomer alle kursiverte og opptrer på samme plass i betydningen. Uansett hvor fintfølende program man lager, kan man være sikker på at ikke det bare er idiomer som ender opp i uttrykksseksjonen. Konverteringsprogrammet treffer kanskje bare i 85 prosent av tilfellene. Det er en kompleks oppgave å programmere datatransformasjoner av denne typen. Ikke mindre kompleks er oppgaven som faller på fagleksikografen – å spesifisere de utallige stegene i transformasjonen så leksikografien ivaretas og på en slik måte at de lar seg omsette av en IT-konsulent til et dataprogram.

5) Spesifisering av stilark

Den nest siste oppgaven er stilarkspesifikasjonene. Stilark er det regelsett som forteller hvordan datastrukturen skal vises frem, altså hvordan XML-koden skal

omsettes i grensesnittene hhv. i redigeringsprogrammet og i publiseringsvisningen for brukeren (figur 10).



6) Redigering

Med de fem første stegene tilbakelagt og med datakonverteringen gjennomført, er det bare redigeringen som står igjen. Den tar noen år.

Samspillet mellom datastruktur og visning

I vårt arbeid med NAOB-dataene ble vi overrasket over hvor begrenset kjennskapen var i dataprosesseringsmiljøet til grunnleggende sammenhenger mellom datastruktur og visning. Vi fant heller ikke noe om dette temaet i den leksikografiske litteraturen, så det er mulig at våre erkjennelser her ikke er allment tankegods. Derfor vies det plass i artikkelen til denne tekniske detaljen.

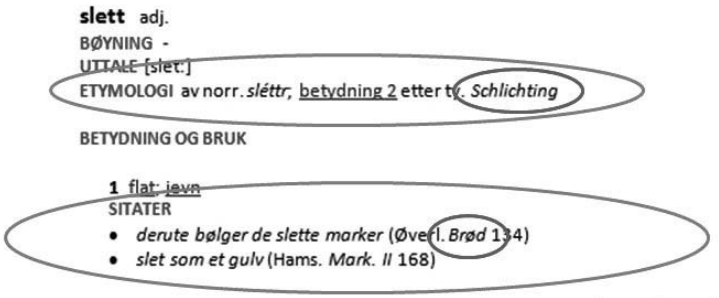
For å holde orden på en hvilken som helst ordbok, men særlig en kompleks, massiv en, må man håndheve strengt et dikotomisk skille mellom to elementtyper: *eskeelementer* og *innholdselementer*.

Eskeelementer er de strukturdannende elementene, metaforisk kalt esker fordi de liksom stables oppå hverandre og har innhold. Eksempler er *hode, kropp, etymologi, sitatseksjon, uttrykksseksjon*. I en luftig, digital publiseringsvisning kan eskeelementer gjerne vises med ny linje, kanskje også med en generert ledetekst så som ETYMOLOGI. Noen eskeelementer nær grunnplanet er slike som inneholder tekstlig innhold, slik tilfellet f.eks. er for *etymologi,* noen høyere

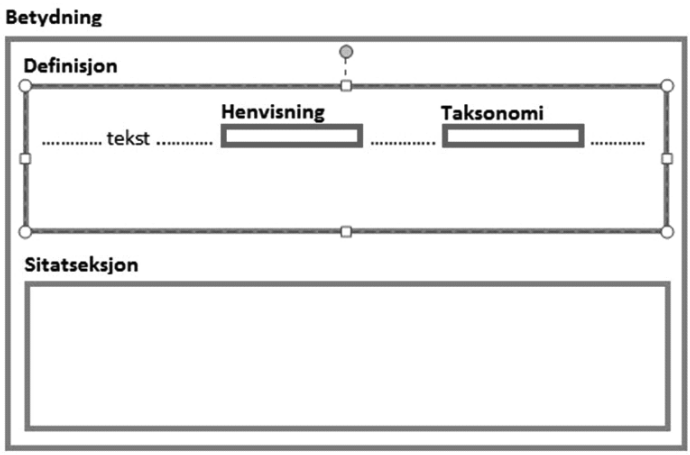
oppe i strukturhierarkiet har nye eskeelementer inni seg – "kinesiske esker" – slik tilfellet f.eks. er for *sitatseksjon.*

Innholdselementer opptrer på sin side inne i den løpende teksten som danner *innholdet* i eskene. Eksempler er *opphavsform, henvisning, utheving, stil* og *taksonomi.*

Figur 11 viser eksempler på eskeelementer ringet inn med grønt, innholdselementer med rødt.



I figur 12 er eske- og innholdsmetaforen tegnet skjematisk, igjen med eskeelementer i grønt og innholdselementer i rødt.



Håndhevingen av elementtypenes dikotomi medfører følgende regler:

- **DTD-plassering.** Eskeelementer må ikke tillates i DTD-en i en innholdskontekst, og innholdselementer må ikke tillates i DTD-en i en eskekontekst.

- **Skillere.** For eskeelementer skal omsluttende skillere (linjeskift, mellomrom, kommaer o.l.) ikke ligge i dataene, men genereres automatisk av stilarket. For innholdselementer skal omsluttende skillere ligge i dataene, dvs. de skal settes manuelt av redaktør (hardkodes).
- **Stilark for redigeringsgrensesnittet.** Eskeelementene og innholdselementene må vises ulikt i redigeringsstilarket for at distinksjonen skal være tydelig for redaktørene. Eskeelementene, som altså bygger opp artikkelen, vises på ny, innrykket linje og f.eks. med full slutt-tagg. Innholdselementer, som altså inngår i den løpende teksten, vises omløpende inne i linjen (ikke ny linje) og f.eks. med kort slutt-tagg.

Konsekvenser for innholdet

Nettpublisering med et transparent, brukervennlig stilark er nådeløst avslørende. Nettpublisering behøver selvsagt ikke være transparent. Det finnes mange eksempler på nettpubliserte ordbøker der de gamle, massive tekstspaltene er gjengitt mekanisk på skjermen, tro mot den gamle boka. Publiserer man ordboken sin på denne måten, blir man i det lange løp straffet, slik Internettpublisering nå utvikler seg. Med den ubegrensede plassen som Internett-publisering gir, forventer brukerne at grensesnittet skal være luftig, klart og lett å finne frem i. Uten dette uteblir trafikken, og det er det vanskelig å leve med i lengden.

Luftighet og klarhet er svakt strukturert leksikografis største fiende.

For NAOB gikk det som det måtte da tåken ble blåst bort – mye kronglete vegetasjon kom til syne. Vi måtte riktignok ta av oss hatten for den systematikken som de gamle leksikografer allikevel hadde klart å gjennomføre, med datidens verktøy og med 30 år mellom start og slutt. Men ja, det lot seg ikke lenger underslå at de gamle mestere begikk mange kunstneriske avvik. Vi kunne ikke lenger lukke øynene for fortidens dype og uforståelige betydningshierarkier, sitatløse sitater, blindhenvisninger og speilhenvisninger.

Transparent visning får frem struktur- og innholdsfeilene i avslørende relieff, men den nyetablerte logikken gir også muligheten til å skape de verktøyene som redaktørene trenger for raskt å reparere feilene. For eksempel kan betydninger flyttes med "drag-and-drop" hvor som helst i hierarkiet og betydningsnummereringen oppdateres automatisk. Henvisningsmålene følger med når artikler skifter stavemåte eller når artikkeldeler flyttes mellom artikler. Man har makroer som omformer uttrykk til eksempler og eksempler til uttrykk. Samt mange andre hjelpemidler.

Om publikum blir bortskjemt av det digitale mediets transparens, blir også redaktørene det. Har man en gang sett innholdets strukturer i skarpt relieff, vil man nødig tilbake til tåkehavet.

Petter Henriksen
Kunnskapsforlaget, Oslo
petter.henriksen@kunnskapsforlaget.no