# The DOT standard format header

*Description of the header used in DSL's corpora*

CLARIN WP 2.1 Technical Report
Jørg Asmussen

Version 0 – December 2008

**Resumé**

I dette dokument beskrives opbygningen af korpusheadere for korpustekster, som skal bruges til DOT/*ordnet*-formål.

## 1 Baggrund

I dette dokument beskrives opbygningen af korpusheadere for korpustekster, som skal bruges til DOT/*ordnet*-formål. Da der allerede blev defineret header-strukturer i forbindelse med opbygningen af Den Danske Ordbogs korpus (DDOC) og Korpus 2000, er det vigtigt, at oplysninger, som er indeholdt i disse headere, i et acceptabelt omfang kan konverteres til den header-struktur, der skal gælde for *ordnet*. Mere om konverteringen af tekstmateriale fra Den Danske Ordbogs korpus, Korpus 2000 og Infomedia kan læses i afsnit **??** på side **??**. Det er ikke alle oplysninger, der vil kunne bevares i den her beskrevne headerstruktur. Disse oplysninger vil dog stadig være knyttet til de oprindelige tekststykkers headere i leverandørformatet.[1] Således kan man i forbindelse med en evt. fremtidig udvidelse af headerstrukturen vende tilbage til disse teksters oprindelige headere og ekstrahere og konvertere yderligere relevant information fra dem.

DOT/*ordnet*-headeren er stærkt inspireret af Korpus 2000-headeren, som igen var stærkt inspireret af DDOC-headeren. Korpus 2000-headeren søgte at fjerne nogle af de uhensigtsmæssigheder, DDOC-headeren havde, først og fremmest en for fin detaljeringsgrad af mulige oplysninger – fx brugte man 131 forskellige genrer. Tilsvarende prøver DOT/*ordnet*-headeren at komme nogle af Korpus 2000-

[1] For DDOC's vedkommende betragtes dets oprindelige korpusformat som leverandørformatet (jf. dokumentationen om korpusformater), så den oprindelige header er bevaret i dette format.

headerens uhensigtsmæssigheder til livs, især ved at eliminere en række oplysningstyper, der i praksis ikke blev brugt/udfyldt alligevel.

En header kan opfattes som en liste over attribut-værdi-par, hvor attributterne er bestemte tekstuelle parametre og værdierne en nøje fastlagt udfyldning af dem. Hver tekstenhed (i hver sin fil) i tekstbank- og korpusformat begynder med en sådan header i XML-format. Attributnavnene (dvs. XML-elementbetegnelserne) og værdimængderne fremgår af tabellerne i denne dokumentation. Ingen attributter må udelades i headerne,[2] og de skal stå i den rækkefølge, som fremgår af tabellerne ndf. Attributterne falder i fire typer, nemlig

- administrative (`adminInfo`)
- tekstinterne (`tekstIntern`)
- teksteksterne om teksten (`omTeksten`)
- teksteksterne om værket (`omVærket`)
- sprogbrugerrelaterede[3] (`sprogbrugerInfo`)

Headerstrukturen er formelt skitseret i figur .1. Strukturen er med vilje holdt så flad som mulig for at sikre en forholdsvis problemfri processering også med simple midler.

```
header ::= adminInfo tekstIntern omTeksten omVærket sprogbrugerInfo+
adminInfo ::= (attribut værdi)+
tekstIntern ::= (attribut værdi)+
omTeksten ::= (attribut værdi)+
omVærket ::= (attribut værdi)+
sprogbrugerInfo ::= (attribut værdi)+
```

Figur .1: Formel beskrivelse af headerstruktur

DOT/*ordnet*-headeren svarer både med hensyn til attributter, attributnavne og udfyldning med få undtagelser til den obligatoriske del af Korpus 2000-headeren. I forbindelse med Korpus 2000-projektet blev der oprettet en header-database *etb*, der bruger den fulde Korpus 2000-header, selvom den fakultative del af oplysningerne (nok) aldrig er blevet udfyldt. Dokumentation af den fulde Korpus 2000-header findes under `univers/3korpus/01.007headerstruktur/k2000-baggrund/`.

[2] Dog gælder det, at den del af headeren, hvor sprogbrugeroplysningerne beskrives, gentages for hver enkelt sprogbruger, der har medvirket ved tekstfrembringelsen, jf. afsnit **??** på side **??**.

[3] Parallelt med *sprogbruger* anvendes betegnelserne *afsender* og *tekstproducent*.

# 2 Headerens enkelte bestanddele

Headeren består af én del med administrative, tre dele med tekstrelaterede og en med sprogbrugerrelaterede oplysninger. De enkelte dele fremgår ikke af selve XML-strukturen, der som nævnt bevidst er holdt så flad som mulig, men af elementbetegnelserne, der alle er præfigeret med et bogstav, der indikerer oplysningsgruppen. Udfyldningen af de enkelte oplysningstyper sker så vidt muligt med talværdier, der i en egentlig visning kan omsættes til de tekststrenge, man ønsker i en given kontekst.

Generelt gælder der for udfyldningen, at det for oplysningstyper, der skal udfyldes med en talværdi, er *0* – *nul* –, som er default-værdien, mens *-1* angiver, at pågældende værdi er udefineret. For oplysningstyper, der udfyldes med tekst, er default - (*bindestreg* med betydningen 'findes ikke', 'er ikke relevant'), mens den udefinerede værdi er *?*. Disse konventioner vil medføre en række konverteringer af udfyldningerne i Korpus 2000-materialet, jf. afsnit **??** på side **??**.

The DOT standard format header cotains four different groups of information:

**Administration:** Info on the particular text file as part of the data capture and processing by the collecting organisation

**Text-internal:** Info derived from the text contents

**Text-external:** Contextual, situational info which cannot be derived from the text proper

**Collection (værkoplysninger):**

**Author:** Info on the author(s) of the text

## 2.1 Admin info

**<aTextid>** Id of text unit. Values are defined in the corpus format documentation **?**

**BNC teiHeader:**

**<aFile>** Filename of text unit in standard format without extension (prior to textbank import), cf. **?**

**BNC teiHeader:**

**<aSrcfile>** Original sourcefilename including its extension. Only the filename proper without its path is given. Its path can be determined from other header elements.

**BNC teiHeader:**

**<aCollec>** Collection or collecting project or institution

**BNC teiHeader:**

**Values:**

| Value | Description |
|---|---|
| 0 | løbende DOT/ordnet-indsamling |
| 1 | løbende via InfoMedia |
| 2 | DDO-talesprog |
| 3 | DDO-skriftsprog |
| 4 | Korpus 2000 |

**<aAcqy>** Year of text acquisition given as a four-digit number or 0 if the precise year is not known

**BNC teiHeader:**

**<aSuppl>** Id of text supplier

**BNC teiHeader:**

**Values:**

Koder i form af talværdier er fastlagt i dokumentationen under univers/3korpus/01.007headerstruktur/dokumentation/leverandoeroplysninger/. Registrerede leverandøroplysninger findes i tabellen leverandoer i databasen leverandoer

**<aRestr>** Restriktioner på brugen af teksten

**BNC teiHeader:**

**Values:**

| Value | Description |
|---|---|
| 0 | kun udsnit (‘citater’) må vises |
| 1 | hele teksten må vises |

**<aAno>** Anonymiseringer i forbindelse med visning

**BNC teiHeader:**

**Values:**

| Value | Description |
|---|---|
| 0 | anonymisering ikke nødvendig |
| 1 | personnavne anonymiseres |
| 2 | stednavne anonymiseres |
| 4 | forfatternavn anonymiseres |
| 8 | teksttitel anonymiseres |

Kombinationer af anonymiseringsmuligheder udtrykkes ved at lægge pgl. koder sammen, fx *3*: person- og stednavne anonymiseres

## 2.2 Text-internal

De oplysninger, som direkte kan udledes af selve teksten, kaldes tekstinterne.

**<iTxttit>** Tekstens titel

**Values:**

Angives fuldt ud, hvis teksten har en titel; hvis den ikke har, indsættes en bindestreg

**<iGenre>** Tekstens genregruppe

**Values:**

Under hver af de nævnte genregrupper henregnes flere genrer. Overordnet skelnes mellem fiktive, ikke-fiktive og genremæssigt ubestemmelige tekster:

| Value | Description |
|---|---|
| 0 | ubestemmelig |
| 1 | fiktiv |
| 2 | ikke-fiktiv |

I de tilfælde, hvor genren kan specificeres nærmere, kan følgende talværdier komme til anvendelse for fiktive tekster:

| Value | Description |
|---|---|
| 11 | lyrisk |
| 12 | episk |
| 13 | dramatisk |
| 14 | diverse |

Og følgende talværdier kan komme til anvendelse for ikke-fiktive tekster:

| Value | Description |
|---|---|
| 21 | orienterende |
| 22 | vurderende etc. |
| 23 | påvirkende |

**<iDomain>** Fagområdet, som teksten vedrører

**Values:**

Når Drejebogens opgave 1.8 om automatisk domænetilordning er løst, bestemmes værdierne automatisk. Der vil blive anvendt samme værdimængde som for DDO's korpus, enten den fulde med 66 værdier eller den reducerede emnegruppe med 12, afhængig af hvilken af dem der giver de bedste tilordningsresultater. Indtil da vil værdien være udefineret, dvs. *-1*

**<iLang>** Oplysning om tekstens sprog. Det vil som regel sige dansk. Men med denne oplysning er der åbnet mulighed for også at samle tekster på andre sprog, fx som led i etableringen af parallelkorpora

**Values:**

| Value | Description |
|---|---|
| 0 | Danish |
| 1 | English |
| 2 | German |
| 3 | French |
| 4 | Spanish |

Der kan tænkes en yderligere uddifferentiering i sprogtrin eller dialekter efter de samme principper, som gør sig gældende for oplysningstyperne <iGenre> og <eMedium>

**<iTokens>** Antal løbende ord, som tekststykket består af

**Values:**
Beregnes automatisk, når tekstbankversionen af en tekst (og dermed headeren) oprettes

**<iTypes>** Antal forskellige ord, som tekststykket består af

**Values:**
Beregnes automatisk sammen med beregningen af token-værdien

**Kommentarer**

**<iGenre>:** Det er problematisk at definere en værdimængde for oplysningstypen *genre*. I Den Danske Ordbogs Korpus brugte man 131 forskellige genrer, der i praksis ikke lod sig administrere. I Korpus 2000 havnede man – efter en del diskussion – i den modsatte grøft med en (lidt uortodoks) genreklassifikation med kun tre værdier, nemlig *ikke-skønlitteratur*, *skønlitteratur* og *privat*. I DOT/*ordnet*-sammenhæng er denne klassifikation udvidet igen med inspiration fra wikipedia[4], som har et ganske fornuftigt bud på en nogenlunde differentieret, men alligevel administrerbar klassifikation. Overordnet skelnes stadig mellem fiktive og ikke-fiktive tekster, og det er muligt kun at nøjes med en angivelse af det uden videre differentiering.[5] Hvis der kan differentieres yderligere, bruges følgende klassifikation:

**Orienterende:** lærebog, nyhedsartikel, afhandling, monografi, biografi, selvbiografi, nekrolog, rejseskildring, reportage osv.

**Vurderende,** debatterende eller argumenterende: essay, partiprogram, kronik, anmeldelse, pamflet, partsindlæg, politisk tale osv.

**Påvirkende** eller adfærdsregulerende: lov, bekendtgørelse, regulativ, aftale, reklame, annonce, brugsanvisning, girokort osv.

**Lyrisk:** salme, digt, lejlighedsdigt, sang, vise osv.

**Episk:** roman, novelle, eventyr, saga, børnebog, ungdomsroman osv.

**Dramatisk:** tragedie, komedie, skuespil, parodi, revy osv.

**Diverse** litterære: anekdote, vittighed, gåde osv.

[4] Jf. http://da.wikipedia.org/w/index.php?title=Genre_.
[5] Det vil under konverteringen fx være tilfældet for K2000-tekster samt for Infomedia-materialet.

## 2.3 Text-external

Oplysninger, som ikke umiddelbart kan udledes af selve teksten.

**<ePrody>** Tekstens produktionsår

**Values:**

Udfyldes med et firecifret årstal. Hvis produktionsåret ikke er kendt, oplyses et skøn og <eProdyc> sættes til *1*

**<eProdyc>** Sikkerhed vedr. produktionsåret

**Values:**

| Value | Description |
|---|---|
| 0 | sikker |
| 1 | usikker |

**<eExpr>** Udtryksmediet (skrift, tale el. mellemstadier)

**Values:**

| Value | Description |
|---|---|
| 0 | skrift |
| 1 | tale |
| 2 | papirtale |
| 3 | talepapir |

**<eAspect>** Kommunikativt aspekt

**Values:**

| Value | Description |
|---|---|
| 0 | reception (‘professional’) |
| 1 | produktion (‘private’) |

**<eAgerel>** Aldersrelation mellem afsender og modtager

**Values:**

| Value | Description |
|---|---|
| 0 | ukendt |
| 1 | adult–adult |
| 2 | adult-child |
| 3 | adult–young |
| 4 | adult–elderly |
| 5 | child–adult |
| 6 | child–child |
| 7 | child–young |
| 8 | child–elderly |
| 9 | young–adult |
| 10 | young–child |
| 11 | young–young |
| 12 | young–elderly |
| 13 | elderly–adult |
| 14 | elderly–child |
| 15 | elderly–young |
| 16 | elderly–elderly |

*child* is aged 6–14, *young* 15–25, *adult* 26-60, and *elderly* 61–

**<eMedium>** Primært el. oprindeligt udgivelsesmedium

**Values:**

| Value | English | Danish |
|---|---|---|
| 0 | undetermined | ubestemmelig |
| 1 | newspaper | avis |
| 2 | periodical | blad/ugeblad |
| 3 | book | bog |
| 4 | internet | intenet |
| 5 | miscellanea | småtryk |
| 6 | journal | tidsskrift/fagblad |
| 7 | assignment | opgave |

Der kan etableres en yderligere differentiering, hvis der er behov for det. Den laves i lighed med oplysningstypen <iGenre> ved at de givne grundkategorier underinddeles, fx *41* for mail, *42* for blog etc.

**<eOlang>** Oplysning om tekstens originalsprog. Til markering af oversatte tekster

**Values:**

| Value | Description |
|---|---|
| 0 | Danish |
| 1 | English |
| 2 | German |
| 3 | French |
| 4 | Spanish |

Der kan tænkes en yderligere uddifferentiering i sprogtrin eller dialekter efter de samme principper, som gør sig gældende for oplysningstyperne <iGenre> og <eMedium>

**<eLoc>** Lokalisering af teksten. Hvis teksten er del af et værk, oplyses hvor i værket teksten er at finde

**Values:**

Udfyldes afhængig af værkets type på flg. måde (sidetal oplyses altid som den side, hvor teksten eller tekstudsnittet starter):

**Aviser:** *3:11* betyder *3. sektion, side 11*

**Blade:** *3:11* betyder *3. nummer, side 11*

**Bøger:** *311* betyder *side 311*

**Internet:** hele URL'en oplyses

**Småtryk:** *2-3* betyder side *2–3*

**Tidsskrifter:** der gælder det samme som for blade

Hvis der ikke kan gives nogen lokaliseringsoplysning, indsættes bindestreg

**<eUrl>** Tekstens (eller tekstleverandørens) URL. Kan i søgeværktøjer siden bruges som link

**Values:**

Hvis teksten findes på www, anføres tekstens URL, ellers anføres en URL, der kommer så tæt på teksten som muligt, fx indgang i en bogdatabase på et forlag, en nyhedstjeneste (fx Infomedia) e.l. Hvis der ikke findes nogen URL, indsættes bindestreg

**Kommentarer**

**<eAgerel>:** Denne oplysning vil i over 99% af tilfældene have værdien *0* (*voksen-voksen)*; den vil derfor aldrig være egnet til distributionsundersøgelser. Imidlertid kan den med tiden være interessant i forbindelse med udtræk af bestemte subkorpora. Derfor bibeholdes den, selvom den kan være vanskelig at administrere.

## 2.4 Collection

Denne gruppe af oplysninger vedrører det evt. værk, teksten stammer fra. Værket kan være en avis, en antologi eller en anden type tekstsamling.

**<vTit>** Værktitel, fx en avis' eller en antologis navn eller titel

**Values:**

Skrives fuldt ud, fx Berlingske Tidende. Der bør føres en fortegnelse over skrivekonventioner. Hvis der ikke er nogen værktitel, indsættes bindestreg

**<vPubly>** Værkets udgivelsesår

**Values:**

Oplyses som et firecifret tal, fx 2007. Hvis udgivelsesåret ikke kendes nøjagtigt, oplyses et omtrentligt udgivelsesår og <vPublyc> sættes til 1

**<vPublyc>** Sikkerhed vedr. udgivelsesåret

**Values:**

| Value | Description |
|---|---|
| 0 | sikker |
| 1 | usikker |

**<vPublm>** Værkets udgivelsesmåned; kun relevant for aviser og blade

**Values:**

Skrives som et tal mellem 1 og 12. Hvis irrelevant, udfyldes med 0

**<vPubld>** Værkets udgivelsesdato; kun relevant for aviser og blade

**Values:**

Skrives som et tal mellem 1 og 31. Hvis irrelevant, udfyldes med 0

## 2.5 Sprogbrugeroplysninger

Da en tekst kan være produktet af flere sprogbrugere, oprettes der i headeren om muligt en sprogbrugerdel for hver sprogbruger. I disse tilfælde sættes <uType> for hver sprogbruger til *sprogbrugerkollektiv*. Da korpussøgesystemer og tekstbanksystemer ikke nødvendigvis kan håndtere tekster med flere afsendere, bør man altid sørge for at give oplysningerne for den formodede 'primære' sprogbruger først, som så går ind som 'stedfortræder' for alle de øvrige. Kan der ikke fastlægges en 'primær' sprogbruger (fx ved wikipedia-tekster e.l.), sættes sprogbrugerens <uName1> og <uName2> til - (bindestreg) og <uType> sættes til *1* (sprogbrugerkollektiv). For oversatte tekster gælder, at oversætteren betragtes som primær sprogbruger.

**<uName1>** Sprogbrugerens fornavn(e)

**Values:**

Angives så nøjagtigt som muligt. For mere kendte forfattere bør der føres en fortegnelse over skrivekonventioner. OBS! Hvis <uType> er et afsenderkollektiv sættes <uName1> og <uName2> til - (bindestreg)

**<uName2>** Sprogbrugerens efternavn

**Values:**

Kun ét efternavn er tilladt. For mere kendte forfattere bør der føres en fortegnelse over skrivekonventioner. OBS! Hvis <uType> er et afsenderkollektiv sættes <uName1> og <uName2> til - (bindestreg)

**<uType>** Sprogbrugertype. Bruges til at skelne mellem individuelle og kollektive afsendere, som man har, hvis teksten er produceret af flere ligeværdige afsendere (fx wikipedia-tekster).

**Values:**

| Value | Description |
|---|---|
| 0 | individuell afsender |
| 1 | afsenderkollektiv |

OBS! Ved et afsenderkollektiv sættes <uName1> og <uName2> til - (bindestreg) og <uSex> til 0

**<uRole>** Sprogbrugerens rolle under tekstproduktionen. Bruges til at skelne mellem tekstens evt. oversætter og tekstens oprindelige afsender

**Values:**

| Value | Description |
|---|---|
| 0 | ophavsmand |
| 1 | oversætter |

**<uSex>** Sprogbrugerens køn

**Values:**

| Value | Description |
|---|---|
| 0 | blandet/usikker |
| 1 | mand |
| 2 | kvinde |

OBS! Hvis teksten er produceret af et afsenderkollektiv (af formodentligt blandet køn) sættes <uSex>-værdien til 0 -1 er den uspecificerede værdi

**<uBorny>** Sprogbrugerens fødselsår

**Values:**

Oplyses som et firecifret tal, fx 1963, -1 er den uspecificerede værdi

**<uBornyc>** Sikkerhed vedr. fødselsåret

**Values:**

| Value | Description |
|---|---|
| 0 | sikker |
| 1 | usikker |

-1 er den uspecificerede værdi