DK-CLARIN:

METADATA FOR WP4 RESSOURCER

DK-CLARIN WP 4

Version 2011-02-01

Bolette S. Pedersen, KU, bspedersen@hum.ku.dk

Lene Offersgaard, KU, leneo@hum.ku.dk

Nicolai H. Sørensen, DSL, nhs@dsl.dk

Viggo Sørensen, JO-ÅU, jysvs@humau.dk

Anna Braasch, KU, braasch@hum.ku.dk

INDHOLD

1 INTRO

Formålet med dette dokument er at specificere CLARIN-metadata for WP4's ressourcer der består af leksikalske data; dels leksikalske ressourcer beregnet til almindelig (menneske)brug på nettet (Jysk Ordbog, jf. www.jyskordbog.dk), dels leksikalske ressourcer beregnet til anvendelse i sprogteknologiske værktøjer (SprogTeknologisk Ordbase, wordnettet DanNet samt en sammenkobling af de to ressourcer, STONet, jf. hhv. wordnet.dk og cst.ku.dk/sto_ordbase).

WP4 har besluttet at metadata tilskrives til de enkelte leksikalske ressourcer, og ikke til de enkelte leksikon-opslag som leksikonet består af. Dette betyder at der for hvert enkelt leksikon specificeres et antal metadata, der karakteriserer hele leksikonet. Oplysningerne omfatter derfor bl.a. oplysninger om hvilke informationstyper, der kan findes i det enkelte leksikon og hvor mange leksikon-opslag, der er i det samlede leksikon. Således anvendes det samme sæt af overordnede metadata for alle leksikalske ressourcer selv om de internt indeholder temmelig forskelligartede informationstyper. En tilskrivning af metadata til de enkelte indgange ville både være en voldsom stor opgave, men det ville også være svært at gøre det på en ensartet måde for de meget forskellige typer af leksikoner. For en nærmere redegørelse af disse informationstyper henvises derfor til detailspecifikationerne udviklet for de enkelte delarbejdspakker (4.1 om DanNet, 4.2.1 om Jysk Ordbog og 4.2.2 om STONet).

WP4 har besluttet at tage udgangspunkt i TEI-standarden, når metadata for leksikalske ressourcer skal specificeres. TEI-standarden er et resultat af Text Encoding Initiative's arbejde der har været i gang siden 1994 og også fremad vedligeholdes. Denne standard er primært udviklet til tekster og tekstkorpusser, men er også brugbar for tekstannotationer og leksikalske ressourcer, når fokus er at beskrive hele ressourcen. Man kan senere konvertere metadata-oplysningerne til andre standarder, da det er et overskueligt antal metadata, der er i spil for leksikalske data. Internt i DK-CLARIN konverteres et uddrag af disse TEI-metadata til OLAC(Open Language Archives Community)[1]-formatet og for visse centrale metadata også til DC(Dublin Core)[2]-formatet. Metadata vil også internt i DK-CLARIN blive konverteret til CMDI (Component MetaData Infrastructure)[3] formatet.

2 METADATA-ELEMENTER FOR WP4

2.1 METADATA FOR ALLE DKCLARIN RESSOURCER

Alle ressourcer i DK-CLARIN har et fælles obligatorisk sæt af metadata. Disse metadata skal specificeres ved deponering af ressourcer i DK-CLARIN, og er anført i følgende tabel.

1 http://www.language-archives.org/OLAC/metadata.html

2 http://dublincore.org/

3 http://www.clarin.eu/cmdi

Metadata	Beskrivelse	Tilladte værdier
DKCLARIN Type	Resource type	For all WP4 resources: Dictionary
Title	Name of the metadata resource representing the linguistic resources including info on version and format.	Literal value.
Language	ID (ISO 639) of the languages included in the resource or the languages of input resources for a tool	ID (ISO 639) of the languages included in the resource. More than one language can be specified.
Format	open vocabulary describing the medium used to exchange the resource (typically a MIME-type, eg. 'application/rdf+xml')	http://www.iana.org/assignments/media-types and other values
PublicationDate	date of the publication of the resource specified by the title	dcterms:W3CDTF, eg. YYYY-MM-DD or YYYY-YYYY
Publisher	A person or organization owning or managing rights over the resource. name of the person to be contacted for accessing the language resource / for copyright issues	Literal value with the names of ublishers/rightsholders
Rights	Information about rights held in and over the resource.	"Link"/Reference to licence agreement

DK-CLARIN har også et antal optionelle generelle metadata, som kan udfyldes i det omfang ressource-leverandøren er interesseret i det. For WP4 er der mulighed for følgende optionelle generelle metadata:

Metadata	Beskrivelse	Tilladte værdier
ContentProvider	Den eller de der deponerer/distribuerer ressourcen. Kan være den samme som "Publisher"	Fri tekst
ContentProvidersId	Ressourceleverandørens lokale id for ressourcen	Fri tekst eller url.
Creator	Den eller de har skabt ressourcen, fx forfatteren til en tekst. Udelades hvis den samme som ContentProvider	Nil for Lexicon
Description	Beskrivelse af ressourcen. Der kan angives beskrivelser på flere sprog, blot skal sprogkoden også specificeres.	Fri tekst
Subject	Emneområde for ressourcen. Kan være en reference til et klassifikationssystem eller fri tekst	Fri tekst
Sponsor	Bidragydere til ressourcen, fx bevillingsgivere	Fri tekst med navne for bevillingsgivere
ExternalURI	NB: Kun for eksterne ressourcer	Url/link
InfoAbout	En url som indeholder information om ressourcen. Fx documentation eller readme-fil.	Url/link
Contributor	Projektangivelse eller andre bidragydere end der er nævnt i "ContentProvider" og "Creator"	Fri tekst eller url.

Når ressourcen deponeres vil infrastrukturen tilføje nogle metadata til ressourcen. Disse behøver brugeren ikke specificere, da de fastlægges som en del af deponeringsprocessen. Disse metadata er derfor ikke relevante i denne sammenhæng, eksistensen nævnes blot her, sådan at ressource-leverandøren ikke bliver overrasket, når der i repositoriet er flere metadata-oplysninger end de importerede. Et eksempel er PID (Persistent Identifier), som er et unikt id til den deponerede ressource i DK-CLARIN repositoriet.

Der kan læses mere om de generelle metadata i rapporten "DK-CLARIN: Metadata for ressourcer", hvor den seneste version kan ses på http://intern.dkclarin.dk/?q=WP5.2_task5_metadata.

2.2 METADATA FOR LEKSIKONRESSOURCER

For leksika kan der ud over de generelle metadata specificeres følgende metadata:

Metadata	Forklaring	Datatype og tilladte værdier
Lex_lexiconType	Hvilken type leksikalsk ressource er der tale om? Dialektordbog, udtaleordbog, ordliste, wordnet, alm. Ordbog Obligatorisk information	Dictionary, Computational_dictionary, Wordnet, Dialectal_dictionary, Wordlist
Lex_mainLevelInformation	Hvilken type information er central for ressourcen? fx dialektoplysning, semantik, morfologi eller udtale Obligatorisk information	Morphology, Syntax, Semantics, Prononciation, Dialect_word
Lex_LexicalUnit	Hvad er den grundlæggende enhed i ressourcen, fx lemma, betydning, synset, fonem Obligatorisk information	Fri tekst
Lex_vocabularySize	Hvor mange enheder (units) har ressourcen? Obligatorisk information	Antal
Lex_noLanguages	Antal sprog som ressourcen behandler Obligatorisk information	Antal

Lex_ortography	Hvordan forholder ressourcen sig til ortografi? Fx ortografi fra en bestemt periode, ortografisk variation mv.	Fri tekst
Lex_dialect	Hvilken/hvilke dialekt/dialekter behandles i ressourcen, fx jysk, lollandsk mv.	Fri tekst
Lex_geographicalCoverage	Hvilke geografiske områder dækkes af ressourcen, fx Sønderjylland, Bornholm	Fri tekst
Lex_timeCoverage	Hvilken sproghistorisk periode dækkes i ressourcen	Fri tekst

2.3 METADATAREALISERINGER FOR JYSK ORDBOG

Metadata	**Forklaring**	***Jysk Ordbog***
Lex_lexiconType	Hvilken type leksikalsk ressource er der tale om? Dialektordbog, udtaleordbog, ordliste, wordnet, alm. ordbog	*Dialect_dictionary*
Lex_mainLevelInformation	Hvilken type information er central for ressourcen? fx dialektoplysning, semantik, morfologi eller udtale	*Dialect_word*
Lex_LexicalUnit	Hvad er den grundlæggende enhed i ressourcen, fx lemma, betydning, synset, fonem	*Lemma*
Lex_vocabularySize	Hvor mange enheder (units) har ressourcen?	*55.000*
Lex_noLanguages	Antal sprog som ressourcen behandler	*1*
Lex_ortography	Hvordan forholder ressourcen sig til ortografi? Fx ortografi fra en bestemt periode, ortografisk variation mv.	*Phonetic transcription (simple, slightly extended Dania)*
Lex_dialect	Hvilken/hvilke dialekt/dialekter	*Jutlandic*

	behandles i ressourcen, fx jysk, lollandsk mv.	
Lex_geographicalCoverage	Hvilke geografiske områder dækkes af ressourcen, fx Sønderjylland, Bornholm	*Jutland*
Lex_timeCoverage	Hvilken sproghistorisk periode dækkes i ressourcen	*Approx. 1700-1920*

2.4 METADATAREALISERINGER FOR DANNET

Metadata	**Forklaring**	***DanNet***
Lex_lexiconType	Hvilken type leksikalsk ressource er der tale om? Dialektordbog, udtaleordbog, ordliste, wordnet, alm. ordbog	*wordnet*
Lex_mainLevelInformation	Hvilken type information er central for ressourcen? fx dialektoplysning, semantik, morfologi eller udtale	*semantics*
Lex_LexicalUnit	Hvad er den grundlæggende enhed i ressourcen, fx lemma, betydning, synset, fonem	*synset*
Lex_vocabularySize	Hvor mange enheder (units) har ressourcen?	*61,000*
Lex_noLanguages	Antal sprog som ressourcen behandler	*2*
Lex_ortography	Hvordan forholder ressourcen sig til ortografi? Fx ortografi fra en bestemt periode, ortografisk variation mv.	*Nil*
Lex_dialect	Hvilken/hvilke dialekt/dialekter behandles i ressourcen, fx jysk, lollandsk mv.	*Nil*
Lex_geographicalCoverage	Hvilke geografiske områder dækkes af ressourcen, fx Sønderjylland, Bornholm	*Nil*
Lex_timeCoverage	Hvilken sproghistorisk periode dækkes i ressourcen	*1950-*

2.5 METADATAREALISERINGER FOR STONET (SAMMENKOBLING AF STO OG DANNET)

Metadata	**Forklaring**	***STONet***
Lex_lexiconType	Hvilken type leksikalsk ressource er der tale om? Dialektordbog, udtaleordbog, ordliste, wordnet, alm. Ordbog	*computational_lexicon*
Lex_mainLevelInformation	Hvilken type information er central for ressourcen? fx dialektoplysning, semantik, morfologi eller udtale	*morfology, syntax, semantics*
Lex_LexicalUnit	Hvad er den grundlæggende enhed i ressourcen, fx lemma, betydning, synset, fonem	*morphu, synu, synset*
Lex_vocabularySize	Hvor mange enheder (units) har ressourcen?	*11,000*
Lex_noLanguages	Antal sprog som ressourcen behandler	*1*
Lex_ortography	Hvordan forholder ressourcen sig til ortografi? Fx ortografi fra en bestemt periode, ortografisk variation mv.	*RO 2001 – accepted spelling + a small number of spelling variants that are still frequent*
Lex_dialect	Hvilken/hvilke dialekt/dialekter behandles i ressourcen, fx jysk, lollandsk mv.	*Nil*
Lex_geographicalCoverage	Hvilke geografiske områder dækkes af ressourcen, fx Sønderjylland, Bornholm	*Nil*
Lex_timeCoverage	Hvilken sproghistorisk periode dækkes i ressourcen	*Modern Danish*

2.6 METADATAREALISERINGER FOR STO (SPROGTEKNOLOGISK ORDBASE)

Metadata	**Forklaring**	***STO***
Lex_lexiconType	Hvilken type leksikalsk ressource er der tale om? Dialektordbog,	*computational_lexicon*

	udtaleordbog, ordliste, wordnet, alm. Ordbog	
Lex_mainLevelInformation	Hvilken type information er central for ressourcen? fx dialektoplysning, semantik, morfologi eller udtale	*morphology, syntax*
Lex_LexicalUnit	Hvad er den grundlæggende enhed i ressourcen, fx lemma, betydning, synset, fonem	*morphu, synu*
Lex_vocabularySize	Hvor mange enheder (units) har ressourcen?	*83,500 morphu* *45,000 synu*
Lex_noLanguages	Antal sprog som ressourcen behandler	*1*
Lex_ortography	Hvordan forholder ressourcen sig til ortografi? Fx ortografi fra en bestemt periode, ortografisk variation mv.	*RO 2001 – accepted spelling + a small number of spelling variants that are still frequent*
Lex_dialect	Hvilken/hvilke dialekt/dialekter behandles i ressourcen, fx jysk, lollandsk mv.	*Nil*
Lex_geographicalCoverage	Hvilke geografiske områder dækkes af ressourcen, fx Sønderjylland, Bornholm	*Nil*
Lex_timeCoverage	Hvilken sproghistorisk periode dækkes i ressourcen	*Modern Danish*

2.7 EKSEMPEL: DANNET-METADATA I TEI-FORMAT

I det følgende ses et eksempel på et udfyldt metadata-dokument for DanNet. Se bilag 1 for originalt xml-dokument.

```
- <TEI>
  - <teiHeader type="lexicon">
    - <fileDesc>
      - <titleStmt>
        - <title>
            DanNet 1.4, udgivet 1. september, 2010, owl-version
          </title>
          <sponsor>DK-CLARIN</sponsor>
          <sponsor>Forskningsrådet for Kultur og Kommunikation</sponsor>
        </titleStmt>
      - <extent>
          <num xml:lang="da" n="synset">65000</num>
          <num xml:lang="da" n="lemma">69900</num>
          <geogName>Danmark</geogName>
          <date>1950-</date>
        </extent>
      - <publicationStmt>
          <distributor>dsl.dk</distributor>
          <idno type="cpsid">http://wordnet.dk/dannet/DanNet-1.4_owl.zip</idno>
        - <idno type="infoAbout">
            http://wordnet.dk/dannet/dannetspecifikationer_v1.0.2.pdf
          </idno>
          <idno type="infoAbout">readme-1.0</idno>
          <idno type="format">application/rdf+xml</idno>
          <idno type="externalUri">http://wordnet.dk/dannet/dannet</idno>
        </publicationStmt>
      - <notesStmt>
        - <note xml:lang="da">
            DanNet er et forsknings- og udviklingsprojekt der går ud på at udarbejde et dansk leksikalsk-semantisk ordnet; dvs. en sprogresurse hvor ords betydni
            udtrykt i et formelt sprog og derved gjort anvendelige for IT-systemer der arbejder med intelligent informationshåndtering.
          </note>
        - <note xml:lang="en">
            DanNet is a research- and development project concerned with the development of a Danish lexical semantic wordnet; i.e. a language resource where tl
            expressed in a formal language and thereby made usable for IT systems dealing with intelligent information handling.
          </note>
        </notesStmt>
      - <sourceDesc>
        - <bibl>
          - <author n="0" ref="#n/a">
              <name>Det Danske Sprog- og Litteraturselskab (DSL)</name>
            </author>
```

```
            <publisher>cst.ku.dk</publisher>
            <publisher>dsl.dk</publisher>
          </bibl>
        </sourceDesc>
      </fileDesc>
    - <encodingDesc>
      - <projectDesc>
          <ab>DK-CLARIN WP4.1</ab>
        </projectDesc>
      </encodingDesc>
    - <profileDesc>
      - <creation>
          <date when="2010-09-01"/>
        </creation>
      - <langUsage n="2">
          <language ident="da">Danish</language>
          <language ident="en">English</language>
        </langUsage>
      - <textClass>
          <classCode scheme="">General</classCode>
          <catRef scheme="lexType" target="Wordnet"/>
        - <keywords scheme="levelInfo">
            <term>Semantics</term>
          </keywords>
        </textClass>
      </profileDesc>
    </teiHeader>
  - <text>
    - <body>
        <ab>Lexicon in separate file.</ab>
      </body>
    </text>
</TEI>
```

3 VALIDERING

Metadata-beskrivelserne, der er udtrykt i TEI-formatet kan valideres med et rng-skema vha. en xml-skema-validator. På denne måde har leksikon-leverandøren mulighed for at teste om metadata er specificeret på en sådan måde at infrastrukturen kan benytte de specificerede metadata. Skemaet vil kunne findes på: http://dkclarin.dk/schemas/WP4/TEIP5DKCLARIN_LEX.rng.

4 REFERENCER

DanNet – wordnet.dk

Jysk Ordbog – www.jyskordbog.dk

SprogTeknologisk Ordbase, STO – cst.ku.dk/sto_ordbase