Ver. 1.5, 2008-09-30
Lene Offersgaard og Sigfrid Lundberg

DK-CLARIN Metadata for ressourcer - forslag

DK-CLARIN Metadata for ressourcer - forslag 1
Metadata 2
Generelle metadata for alle ressourcer og værktøjer 2
Generelle DK-CLARIN metadata 3
Øvrige metadata 4
Kategorisering i ressourcetyper 4
Basisressourcer og annoteringsressourcer 5
Basisressourcer og applikationer 5
Basisressourcetyper 5
Annoteringsressourcetyper 6
Værktøjsressourcer og services 6
Oversigt over ressourcetyper 6
Metadata for tekstsamlinger 8
Metadata for tekstenheder 9
Metadata for eksisterende tekstenheder 9
Metadata for DK-CLARIN tekstenheder 10
TEI-header for DK-CLARIN tekstenhed 10
Dublin Core-header for DK-CLARIN tekstenhed 11
Metadata for leksikalske ressourcer 12
Metadata for tale- og videoressourcer 12
Metadata for billeder og billedsamlinger 12
Metadata for annoteringsressourcer 12
Metadata for værktøjer og webservices 13
Opsummering angående typer af ressourcer 14
Status og videre arbejde 14

Metadata

Metadata angiver oplysninger om data – ”data om data”. Metadata bruges til at karakterisere og administrere og levere data. Metadata udnyttes normalt til at foretage en effektiv datalagring, data-styring og datasøgning. Hvilke metadata, der er nødvendige, afhænger af typen af data og i hvilken kontekst data ønskes anvendt. I DK-CLARIN benyttes metadata blandt andet til at karakterisere og administrere ressourcer. Brugerne har vha. metadata et redskab til at beskrive de enkelte ressourcer og ressourcernes afhængigheder. Når disse beskrivelser foretages på en ensartet måde for mange ressourcer er det lettere for den enkelte bruger og for applikationer at tilgå disse beskrivelser. Metadata kan udnyttes ved søgning, sådan at man kan vha. metadata kan finde de ressourcer der matcher brugerens ønsker eller som en applikationen kan bruge som data.
I det følgende vil der pga. brug af internationale standarder og af hensyn til samarbejdet internationalt bl.a. i EU-CLARIN og blive foreslået engelske betegnelser for metadataoplysninger.

Generelle metadata for alle ressourcer og værktøjer

I EU-CLARIN har man et forslag om at have en mængde generelle metadata for alle ressourcer og supplere med ekstra metadata for de enkelte typer af ressourcer. I DK-CLARIN vil vi også gøre brug af en sådan opdeling i generelle metadata og metadata der er specifikke for den enkelte ressourcetype. Dette gør det nemlig lettere at tage hensyn til at en gruppe af ressourcer har et stor behov for at specificere nogle metadata, fx ønsker man for en lydoptagelse at registrere oplysninger om en talers bopæl, mens dette for et skrevet korpus ikke er relevant. Det er dog vigtigt at denne opdeling af ressourcer i typer ikke medfører en for stor detaljeringsgrad, så det bliver svært for en bruger at få overblik over ressourcetyperne.
I DK-CLARIN vil vi gerne vægte brug af standarder højt. De to standarder der er mest i fokus i denne sammenhæng er i DC Dublin Core Metadata Terms og TEI Text Encoding Initiative. I det følgende gennemgås EU-CLARINs forslag til generelle metadata, da vi også ønsker at have et samarbejde med dette projekt, hvorefter DK-CLARINS forslag til generelle metadata præsenteres.

EU-CLARINs metadata forslag[1] indeholder følgende fælles metadata, herunder også med kommentarer om i hvilket omfang det ser ud til at der er en klar parallel til DC Element set:

- ResourceType — Ligner DC element Type
- Name — Ligner DC element Title
- Language — DC element
- Description — DC element
- Country
- Institute — Ligner DC element Creator
- Creator — DC element

[1] Se forslaget på http://www.cst.dk/dk-clarin/files/eu-clarin_ad-hoc-registry-v6_0.pdf

- Year Ligner DC element Date, men Date bruges eksplicit under flere af de enkelte grupper af data
- Format DC element
- MetadataLink
- ReferenceLink Ligner DC element Identifier

Generelle DK-CLARIN metadata

DK-CLARINs forslag til generelle metadata, hvor optionelle metadata angives med *. DC beskrivelserne er også angivet:

- **Type** (DC element set): The nature or genre of the resource.
- **Title** (DC element set): A name given to the resource
- **Language** (DC element set): A language of the resource. Repeatable.
- **Publisher** (DC element set): An entity responsible for making the resource available.
- **Date** (DC element set): A point or period of time associated with an event in the lifecycle of the resource
- **Format** (DC element set): The file format, physical medium, or dimensions of the resource.
- **Identifier** (DC element set): An unambiguous reference to the resource within a given context.
- **Description*** (DC element set): An account of the resource.
- **isVersionOf*** (refined DC element set: relation) A related resource of which the described resource is a version, edition, or adaptation.

Disse er udvalgt fra henholdsvis DC element set og relation.

Yderligere specifikation af ovenstående

Language: Her er der to muligheder med hhv. ISO 639-2(tre karakter-koder) og ISO 639-1(to-karkter-koder), hvor ISO 639-1 kan konverteres til ISO 639-2[2]. Kodelisten kan ses på

[2] From http://www.loc.gov/standards/iso639-2/langhome.html: ISO 639 provides two sets of language codes, one as a two-letter code set (639-1) and another as a three-letter code set (this part of ISO 639) for the representation of names of languages. ISO 639-1 was devised primarily for use in terminology, lexicography and linguistics. This part of ISO 639 represents all languages contained in ISO 639-1 and in addition any other language as well as language groups as they may be coded for special purposes when more specificity in coding is needed. The languages listed in ISO 639-1 are a subset of the languages listed in ISO 639-2; every language code in the two-letter code set has a corresponding language code in the alpha-3 list, but not necessarily vice versa. Both code lists are to be considered as open lists.

http://www.loc.gov/standards/iso639-2/php/code_list.php. De fleste tekster og ressourcer i DK-CLARIN vil være for dansk og få andre sprog vil være repræsenteret, så foreløbig vil der ikke være så store krav til dækningsgrad for language. Vi ser dog ingen grund til at begrænse os på dette område og for at have de bedste muligheder i fremtiden for at angive sprog benyttes ISO 639-2 (tre-karakter-koder).

Format: se http://www.iana.org/assignments/media-types

Identifier: en URI som gør det muligt for brugere at identificere en ressourceressource, og dermed gør det muligt at fremfinde denne ressource.

Øvrige metadata

Til alle ressourcer er der således knyttet nogle generelle metadata. Disse suppleres med metadata fastlagt for de enkelte typer af ressourcer, som beskriver karakteristika for netop denne ressourcetype. Nogle eksisterende ressourcer vil have yderligere individuelle metadata tilknyttet, som ikke er omfattet af hverken de generelle metadata eller af metadata for ressourcetypen. Det foreslås at disse metadata godt kan angives i metadataformatet, men de skal indordnes under den definerede struktur for 'generelle'- og 'ressourcetype'-metadata. Det foreslås at disse ekstra metadata kan gøres synlige, når man ser den enkelte ressources beskrivelse. Muligvis vil det kun være de generelle og den enkelte types metadata som er søgbare og altså ikke eventuelle ekstra-definerede metadata. Dette må afklares senere.

Det skønnes ikke realistisk at forestille sig at en DK-CLARIN-søgning kan tilfredsstille alles behov i forhold til enkelte ressourcers ønsker til søgbare metadata. Men det er vigtigt at vi lægger os fast på en fællesmængde af oplysninger som er både brugbare og realistiske at angive for ressourcerne. Vi vil derfor stille som krav at man skal kunne levere de specificerede generelle metadata for at kunne udveksle og aflevere data til DK-CLARIN.

Kategorisering i ressourcetyper

Den første kategorisering foretaget i DK-CLARIN tager udgangspunkt i 4 hovedkategorier: Skrevet sprog (Monolinvale tekstsamlinger, Multilingvale tekstsamlinger, Ord-collections), Tale, Multimodale ressourcer samt Billeder(obj), se oplæg fra stormøde 9. juni 2008

Denne kategorisering danner basis for en mere detaljeret gruppering af ressourcer, som vil blive benyttet i forbindelse med specificering af metadata for de enkelte grupper. Grupperingen foretages overordnet for at opnå så få grupper som muligt, men også for at inddele ressourcerne i grupper sådan at der er store fællestræk mellem de særlige karakteristika, anvendelser og muligheder internt i grupperne. Der kan senere ske en opsplitning eller en sammenlægning af ressourcetyper, hvis det viser sig hensigtsmæssigt.

Vi arbejder med følgende overordnede ressourcetyper:

- Basisressourcer
- Annoteringsressourcer
- Applikatoner/Værktøjsressourcer
- Services: Har en URI, inddata specificeres med URI og resultat som URI.

Basisressourcer og annoteringsressourcer

Vi foreslår at fx en tekstsamling vedbliver at være den samme ressource selvom der udføres annoteringer på tekstsamlingen. Annoteringerne anses derfor selv som værende en ressource: en annoteringsressource. Det gør det muligt at tilknytte metadata og ejerskabsforhold til annoteringsressourcen, der som en af sine metadata vil have en reference til den basisressource, den knytter sig til. Det foreslås derfor at der fastlægges et DK-CLARIN-format for annoteringsressourcer for hver af de enkelte ressourcetyper.

Basisressourcer og applikationer

Det vil i en del tilfælde gælde, at der er forskel mellem den 'rå' tekstressource (XML-fil) og en tekstressource, anvendt som til inddata til en applikation, som fx ikke kan benytte XML-filer som input. Fx vil en tekstressource fra DUDS-samlingen(KU-INSS) findes som en XML-tekst, mens den samme tekstressource, hvis den er gjort søgbar i korpus-søgeværktøjet CQP, være ændret og der er foretaget valg ang. hvilke tekstopmærkninger, der skal være søgbare i CQP.
Forslaget er at vi i princippet kun vil omtale teksten som én ressource. Men når vi beskriver en applikation, så specificerer vi hvilke ressourcer, der er brugt som inddata, angiver en konverterings- og formatterings-specifikation samt beskriver værktøjet selve applikation er baseret på. Fx kan tekstsamlingsressourcen "7 Parallelle HCA-eventyr" have en applikationsressource tilknyttet som er "CQP-søgning i '7 parallelle HCA-eventyr'". Denne applikationsressource specificeres ved metadata for tekstsamlingsressourcen, metadata for selve værktøjet CQP, samt en specifikation af hvad der er gjort ved tekstsamlingsressourcen for at indlæse den i CQP.

Basisressourcetyper

DK-CLARIN skal håndteres følgende typer af basisressourcer[3]:

---

[3] Til sammenligning listes her et uddrag af de betegnelser DC tilbyder som en del af http://dublincore.org/documents/dcmi-type-vocabulary til sammenligning:
Collection
Text
Dataset
Physical Object
Image
Moving Image
Sound
Service
Software

1. Monolingval tekstsamling ”Monolingual text collection”
2. Multilingval tekstsamling ”Multilingual text collection”
3. Leksikalsk ressource ”Lexical resource”
4. Tale korpus ”Spoken corpus”
5. Video samling ”Video Collection” [4]
6. Billedsamling(billeder af ’objekter’ med tilknyttet tekst) ”Image Collection”
7. Tekstenhed ”Text”
8. Alignering(mellem to tekster) ”Alignment”
9. Post ”Record” (del af leksikalsk ressource)
10. Taledata ”Speech segment”
11. Video ”Video”
12. Billede (billede af ’objekt’ med tilknyttet tekst) ”Image”

Bemærk at ressource-typerne 1-6 er samlinger, mens 7-12 er enheder som er mindre enheder som samlingerne består af. Hvis ressourcen består af en eller flere billedfiler af en scannet tekst, så defineres det i DK-CLARIN sammenhæng sådan at disse billeder udgør en tekstressource, ikke en billedressource.

Annoteringsressourcetyper

DK-CLARIN skal håndtere følgende typer af annoteringsressourcer:

13. Annotering af tekstsamling/tekst
14. Annotering af leksikalsk ressource
15. Annotering af taledata
16. Annotering af video
17. Annotering af billeder

Værktøjsressourcer og services

DK-CLARIN skal håndteres følgende software ressourcer:

18. Applikation/Værktøj
19. Webservice

Oversigt over ressourcetyper

Til sammenligning kan listen af ressourcetyper som EU-CLARIN benytter ses på http://www.clarin.eu/view_resources[5]. De er også anført i nedenstående tabel.

Den forventede sammenhæng mellem DK-CLARIN ressourcetyper og EU-CLARIN ressourcetyper, samt Dublin Core betegnelser er angivet i følgende tabel. I tabellen er også angivet hvilke andre

---

[4] Hermed specificeres at der med multimodale ressourcer i DK-CLARIN menes video-optagelser.

[5] Hvis man er oprettet som bruger på clarin.eu

ressourcetype-enheder der indgår i den enkelte ressourcetype, fx består en monolingval tekstsamling af tekster.

For at betegnelserne kan være forståelige internationalt, bør der defineres engelske navne for ressourcetyperne, fx dem som er angivet nedenfor[6]. Af hensyn til det internationale samarbejde vil det også være oplagt at kunne bruge engelske betegnelser i brugergrænsefladen.

Ressourcetype i DK-CLARIN	Forslag til engelsk betegnelse for DK-CLARIN ressourcetype	Indeholder ressourcer af DK-CLARIN typen	Kan klassificeres som følgende ressourcetype i EU-CLARIN	Korresponderer til Dublin Core som
Monolingval tekstsamling	Monolingual text collection	Text	Written Corpus	Collection of Text
Multilingvale tekstsamling	Multilingual text collection	Text, Alignment	Aligned Corpus	Collection of Text
Leksikalsk ressource	Lexical resource	Record	Lexicon/Knowledge Source	Collection of Dataset
Talekorpus	Spoken Corpus	Speech segment	Spoken Corpus	Collection of Sound
Videosamling	Video collections	Video	Multimodal Corpus	Collection of Dataset
Billedsamling	Image Collection	Image		Collection of Image
Tekstenhed	Text			Text
Alignering	Alignment			
Post	Record			
Taledata	Speech segment			Sound
Video	Video			
Billed	Image			Image
Annotering af tekstsamling/tekst	Annotation of text			
Annotering af leksikalsk ressource	Annotation of lexical ressource			
Annotering af taledata	Annotation of speech			
Annotering af video	Annotation of Video			
Applikation/Værktøj	Application/Tool		Application	Software

[6] Vi synes at det er oplagt at lade dem ligne ressource-navnene i EU-CLARIN i det omfang der er parallellitet, men lige nu er der ikke så stor sammenfald i opdelingen. Der er også andre muligheder fx Dublin Core's navne

Webservice	Web Service		Web Service	Service
			Treebank	
			Grammar	
			N-gram model	
			Terminological Resource	
			Other	

For de enkelte ressourcetyper er der brug for fastlæggelse af hvilke ressourcetype-metadata, der skal supplere de generelle metadata. Derudover skal der også for de enkelte typer fastlægges et opmærkningsformat for selve ressourcen, men opgaven angående opmærkningsformat behandles ikke i dette dokument.

I det følgende gives en beskrivelse for metadata for tekstsamlinger og der angives et forslag til metadata for tekstenheder. Dette forslag er stadig åbent for tilføjelser.

Metadata for tekstsamlinger

For tekstsamlinger har vi brug for metadata på to niveauer: For hele korpusset og for den enkelte tekstenhed. Da DK-CLARIN ønsker at udnytte eksisterende standarder foreslås det at der som struktur for et korpus bruges TEI's specifikation:

```
<teiCorpus>
 <teiHeader>
<!--[header information for the corpus]-->
 </teiHeader>
 <TEI>
  <teiHeader>
<!--[header information for first text]-->
  </teiHeader>
  <text>
<!--[first text in corpus]-->
  </text>
 </TEI>
 <TEI>
  <teiHeader>
<!--[header information for second text]-->
  </teiHeader>
  <text>
<!--[second text in corpus]-->
  </text>
 </TEI>
</teiCorpus>
```

Der skal defineres en teiHeader for et teiCorpus. Denne teiHeader skal indeholde metadata, som dækker de obligatoriske generelle metadata og ekstra metadata som er særligt relevante for tekstsamlinger.

Vi definerer på nuværende tidspunkt metadata for en tekstsamling som de samme som de generelle metadata, dog sådan at de to generelle metadata 'description' og 'isVersionOf' for tekstsamlinger er obligatoriske metadata-elementer. Der skal således tilknyttes en beskrivelse og en angivelse af om tekstsamlingen er en del af en anden tekstsamling for en tekstsamling.

Metadata for tekstenheder

I det følgene beskrives kort de eksisterende ressourcers metadataformater, derefter præsenteres metadataformatet for DK-CLARIN tekstenheder.

Metadata for eksisterende tekstenheder

De indberettede eksisterende tekstenheder hos projektets partnere kan grupperes i følgende header-formater: DSL's format, ADL's TEI-lite, CST's CES-format samt Paroles format. Beskrivelserne nedenfor er kun meget korte. Yderligere information kan søges i dokumenterne på http://cst.dk/dk-clarin/?q=ressourceoversigt_juni2008.

DSLs format

DSL har redegjort for deres metadata-format i "Fastlæggelse og dokumentation af headerstruktur" Jørg Asmussen, version 1.1, 20. Feb. 2008. Ikke alle korpora ligger i dette format, men DSL planlægger at håndtere dette med konvertering til og fra formatet. Her er der fem grupper af headerinformation:

- Adminitrative(adminInfo)
- Tekstinterne(tekstIntern)
- Teksteksterne om teksten (omTeksten)
- Teksteksterne om værker (omVærket)
- Sprogbrugsrelaterede (sprogbrugerInfo)

ADL i TEI-lite

Om ADLs metadata fra informationsindhentningen om ressourcer, citat Sigfried Lundberg: teiHeader metadata (http://www.tei-c.org/release/doc/tei-p5-doc/html/HD.html) supplerad med en del information indaterad direkt i databasen. Motsvarar en relativt god dublin core-record.

CES Corpus encoding initiative

CST burger i Mulinco-korpusset den minimale CES-header, med nogle få tilføjelser af bl.a. eAddress, idno, firstPubDate, samt profileDesc incl. annotation, translations og textClass.
Mulinco-headeren er ikke kompatibel med TEI-lite.

Parole-korpussets format

Parole-korpusset benytter PAROLE Corpus Encoding Standard. Denne kan jf. http://korpus.dsl.dk/paroledoc_dk.pdf afsnit 3, s9-12 parses med TEI-dtd'en (oplysningen er fra 1996).

Metadata for DK-CLARIN tekstenheder

Som metadata for den enkelte tekst foreslås det, at vi udvælger en relevant del af opmærkningerne fra TEI-lite strukturen og udbygger den med en optionel del som så vidt muligt understøtter header-information fra bl.a. ADL, DSL og Mulinco-korpusset. Vi tager dermed udgangspunkt i en anerkendt standard[7] og de enkelte ressourcer vil så skulle kunne udveksle data ved brug af denne standard.

Fra TEI header beskrivelsen (http://www.tei-c.org/release/doc/tei-p5-doc/html/HD.html) har vi følgende: "the <teiHeader> element has four principal components:

- fileDesc (file description) contains a full bibliographic description of an electronic file.
- encodingDesc (encoding description) documents the relationship between an electronic text and the source or sources from which it was derived.
- profileDesc (text-profile description) provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting.
- revisionDesc (revision description) summarizes the revision history for a file.

Of these, only the <fileDesc> element is required in all TEI headers; the others are optional."

Den minimale teiHeader er således jf. standarden:

```
<teiHeader type="text">
    <fileDesc>
            <!-- ... -->
    </fileDesc>
</teiHeader>
```

TEI-header for DK-CLARIN tekstenhed

For DK-CLARIN foreslår vi en TEI-header for tekstenheder, hvor et eksempel på en tei-header kan ses i filen tei001.xml[8], her vises for læsevenlighedens skyld kun et uddrag af filen, nemlig begyndelsen. TEI-headeren beskriver skuespillet 'Jeppe på Bierget' med referencer til det bogværk teksten er scannet fra og i denne en reference til den originale tekst fra 1722.

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader xml:lang="dan">
    <fileDesc>
```

[7] http://www.tei-c.org/

[8] Se filen tei001.xml på http://www.cst.dk/dk-clarin/?q=wp52b_metadata

```
      <titleStmt>
            <title>Jeppe paa Bierget Eller Den forvandlede Bonde</title>
            <author>Holberg, Ludvig</author>
      </titleStmt>
      <publicationStmt>
            <authority>
              <name type="organisation">Det Kongelige Bibliotek</name>
            </authority>
            <distributor>
              <name type="organisation">Det Kongelige Bibliotek</name>
              <name type="place">København, Danmark</name>
            </distributor>
            <date>2001</date>
      </publicationStmt>
      <sourceDesc>
            <bibl>
              <author>Holberg, Ludvig</author>
              <title level="a">Jeppe paa Bierget Eller Den forvandlede Bonde</title>
              <title level="m">Ludvig Holberg: Værker i tolv Bind</title>
              <editor>Billeskov Jansen, F. J.</editor>
              <publisher>Rosenkilde &amp; Bagger</publisher>
              <pubPlace>København</pubPlace>
              <idno type="isbn">??????</idno>
              <date>1969-1971</date>
              <biblScope type="vol">3</biblScope>
              <biblScope type="pp">225-277</biblScope>
              <ref target="https://rex.kb.dk/F/?func=find-
                        b&amp;local_base=KGL01&amp;find_code=SYS&amp;request=000894787"/>
              <relatedItem type="original">
                <bibl>
                  <author>Holberg, Ludvig</author>
                  <title level="a">Jeppe paa Bierget Eller Den forvandlede Bonde</title>
                  <date>1722</date>
                  <ref target="https://rex.kb.dk/F/?func=find-
                        b&amp;local_base=KGL01&amp;find_code=SYS&amp;request=002160106"/>
                </bibl>
              </relatedItem>
            </bibl>
      </sourceDesc>
    </fileDesc>
…
<teiHeader xml:lang="dan">
```

Dublin Core-header for DK-CLARIN tekstenhed

TEI-headeren kan også udtrykkes i Dublin Core standarden. Det er valgt at vise dette her, da OAI-PMH protokollen som forventes anvendt til metadata-høstning benytter Dublic Core standarden. DK-CLARIN's metadataindeks vil udbyde konverteringsscripts fra TEI-header til Dublin Core-headeren, sådan at udbyderne af tekstenheder tilbydes konverteringsmulighed fra TEI til Dublin Core for DK-CLARIN tekstenheder.

Tages der udgangspunkt i den ovennævnte TEI-header kan de generelle metadata udtrykkes i Dublin Core-headeren som følger[9]:

```
  <oai_dc:dc>
    <dc:title>Jeppe paa Bierget Eller Den forvandlede Bonde</dc:title>
```

[9]Se filen dcterms001.xml på http://www.cst.dk/dk-clarin/?q=wp52b_metadata

```
    <dc:creator>Holberg, Ludvig</dc:creator>
    <dc:publisher>Det Kongelige Bibliotek, København, Danmark</dc:publisher>
    <dc:date>2001</dc:date>
    <dc:identifier>http://clarin.dk/luddes/jeppe</dc:identifier>
    <dcterms:isVersionOf>https://rex.kb.dk/F/?func=find-
b&amp;local_base=KGL01&amp;find_code=SYS&amp;request=000894787</dcterms:isVersionO
f>
    <dc:language>dan</dc:language>
    <dc:type>comedy</dc:type>
    <dc:type>text</dc:type>
  </oai_dc:dc>
```

Se også readme-filen i vedlagte metadata.zip fil for yderligere detaljer. Her kan også ses konverteringsscripts mellem TEI og Dublin Core formaterne.

Metadata for leksikalske ressourcer

Der er endnu ikke udarbejdet noget specifik metadataforslag for leksikalske ressourcer, ud over forslaget om generelle DK-CLARIN metadata. Der indledes snarest muligt en dialog med WP4 ang. dette, hvor deltagerne i WP4.1 og WP4.2 opfordres til at komme med et forslag til metadata der er specifikke for leksikalske ressourcer, med udgangspunkt i de ovenfor generelle DK-CLARIN metadata.

Det forventes at metadata for leksikalske ressourcer kan udtrykkes med TEI-standarden.

Metadata for tale- og videoressourcer

Der er endnu ikke udarbejdet et specifikt metadataforslag for tale- og videoressourcer, ud over forslaget om generelle DK-CLARIN metadata. Der indledes snarest muligt en dialog med WP3 ang. dette, hvor deltagerne i WP3.1 og WP3.3 opfordres til at komme med et samlet forslag til metadata for video-ressourcer og annoteringer af disse. WP3.2 opfordres til at komme med et forslag for tale-ressourcer og annoteringer af disse.

Metadata for billeder og billedsamlinger

Der er endnu ikke udarbejdet et specifikt metadataforslag for billeder og billedsamlinger, ud over forslaget om generelle DK-CLARIN metadata. Denne opgave igangsættes snarest muligt i dialog med WP2.5. Deltagerne i WP2.5 opfordres til at komme med et forslag til metadata, med udgangspunkt i de specificerede DK-CLARIN generelle metadata.

Metadata for annoteringsressourcer

Det forventes at der kan tilknyttes annoteringsressourcer til alle typer af basisressourcer. Metadata for annoteringsressourcer skal derfor også fastlægges. Ud over de generelle og obligatoriske metadata, se evt. afsnittet ”Generelle DK-CLARIN metadata”, skal der angives en reference til den ressource annoteringen knytter sig til.

For tekstressourcer kunne det ske som det ses herunder, men det kunne også blot udtrykkes med en identifier(URI).

```
<teiHeader type=”annotation”>
…
  <sourceDesc>
            <bibl>
                  …
                  <relatedItem type="annotationOf">
                              <bibl>
                              <author>Holberg, Ludvig</author>
                              <title level="a">Jeppe paa Bierget Eller Den forvandlede Bonde</title>
                              <date>1722</date>
                              </bibl>
                  </relatedItem>
            </bibl>
  </sourceDesc>
…
</teiHeader>
```

Det skal specificeres yderligere hvordan man har mulighed for at angive disse referencer. Vi afventer på nuværende tidspunkt EU-CLARINs oplæg til håndtering af problemet, da vi ønsker at være så kompatible med EU-CLARIN som muligt i DK-CLARINS valg af løsning.

Vi forventer ikke i første omgang at lave særlige løsninger for annotering af leksikalske ressourcer, med mindre specifikke brugere ønsker at benytte dem og ønsker at indgå i specifikationsarbejdet i samarbejde med WP5.

Metadata for værktøjer og webservices

Det tilstræbes at der defineres et metadataformat for webservices, som også kan bruges til værktøj, blot med nogle elementer defineret som optionelle for værktøjer.
Det kunne desuden være hensigtsmæssigt at opdele værktøjer i grupper efter anvendelse. En sådan opdeling står imidlertid over for den udfordring at visse værktøjer godt kan bruges i forskellig sammenhæng og derfor ikke helt entydigt kan grupperes. Som udgangspunkt forventes det at inddele værktøjerne i hovedgrupper efter deres mest oplagte/originale anvendelse:

- Værktøj for skrevet sprog / ”Tools for written language”
- Værktøj for talt sprog / ”Tools for spoken language”
- Værktøj for videoressourcer / ”Video tools”
- Værktøj for Billedressourcer / ”Image tools”
- Værktøj til administration eller opbygning af samlinger / ”Collection tools”

Forslaget er at vi afventer konkret arbejde i WP5 eller yderligere specifikationer fra EU-CLARIN inden vi går videre her.

EU-CLARIN har på nuværende tidspunkt lavet lister af værktøjstyper i de enkelte grupperinger. Listerne kan ses i dokumentet views_on_taxonomies-v09.pdf afsnit 2.1 og 4.1. Man kan måske bruge listerne som pick-lister brugeren kan vælge en betegnelse fra, sådan at man slipper for forskellige stavemåder for samme type.

Opsummering angående typer af ressourcer

For de forskellige ressourcetyper der vedrører tekst udtrykkes metadata i TEI-standarden. Dette sker ved at der specificeres forskellige typer af <teiHeader>. Dette gælder:

- corpus
- text
- lexicon
- lexitem
- annotationText
- alignment

De ressourcer, der ikke vedrører tekst ressourcer, specificerer ikke metadata vha. TEI-standarden. Disse ressourcers header-definitioner skal dog også angive typen på ressourcen, jf. kravet angående de generelle metadata. Umiddelbart foreslås der følgende navne til typer af ressourcer:

- spokenCorpus
- videoCollection
- imageCollection
- speechSegment
- video
- image
- annotationSpeech
- annotationVideo
- annotationImage

Dette forventes endeligt fastlagt i dialog med de øvrige arbejdspakker i løbet af efteråret 2008.

Status og videre arbejde

Det nærværende dokument er den indledende specifikation ang. ressourcetyper og metadata i WP5.2. WP5 vil arbejde videre på baggrund af den opdeling af ressourcetyper der er foretaget her. Angående metadata er der fokuseret på generelle DK-CLARIN metadata og metadata for tekstsegmenter. Der udestår en dialog med WP2.5, WP3 og WP4 om specifikke metadata for de øvrige ressourcetyper.

Der blev i juni 2008 gennemført en kortlægning af hvilke ressourcer projektpartnerne på det tidspunkt allerede havde udarbejdet. En oversigt over de opnåede tilbagemeldinger kan ses på http://cst.dk/dk-clarin/?q=ressourceoversigt_juni2008. Data fra denne indsamling indtastes i EU-CLARINs ressourceoversigt i oktober 2008 på http://www.clarin.eu/view_resources.

I dialog med WP5.1 fastlægges arkitektur og de tekniske rammer for etablering af adgang til decentrale og eksisterende ressourcer fra DK-CLARINS servicelag. Og der implementeres en prototype for udvalgte ressourcer.

Inden prioriteringen af de eksisterende ressourcer med hensyn til integrering i DK-CLARIN, så gives der mulighed for at projekt-eksterne ressourceudbydere kan melde sig som interesserede i at tilbyde deres ressourcer som decentrale resurser og værktøjer.

Der udarbejdes en arbejdsplan for perioden frem til T18 i løbet af oktober.