DK-CLARIN:
METADATA FOR RESSOURCER

DK-CLARIN WP 5.2

Version 2.0 2011-03-04

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

NB: Afsnit 3.6 er ikke ajourført

Versionshistorie

Dato	Version	Logtekst	Ansvarlig	Status
9/09-2009	0.8	Dokumentet sendt til kommentering i WP3,WP4 og WP5.1	Lene Offersgaard	Første samlede version
29/09-2009	1.0	Dokumentet afleveret T21. Input fra WP3 mangler. Dokumentet skal måske justeres når det er lagt fast hvordan rettigheder håndteres i DK-CLARIN.	Lene Offersgaard	Afleveret T21 Input fra WP3 mangler.
1/2-2011	1.1	Tilføjet info ang. WP3. Korrektur: DateAccepted og HasVersion udgår	Lene Offersgaard	Korrektur
4/3/2011	2.0	Opdateret efter test af deposit og samarbejde med ressource-leverandører	Lene Offersgaard	Opdateret

INDHOLD

1 INTRO

Formålet med dette dokument er at specificere metadata for dataressourcer. Der redegøres for generelle metadata, ressourcespecifikke metadata, samt i hvilket omfang disse metadata er obligatoriske. Fastlæggelsen af værdier for de enkelte metadata-elementer vil ske i andre dokumenter, se http://www.cst.dk/dk-clarin/?q=WP5.2_task5_metadata for yderligere dokumentation.

Modtagergruppen for dette dokument er alle projektdeltagere. Dokumentet benytter betegnelsen 'bruger' både for den person som bidrager med ressourcer til DK-CLARIN og den som søger information om ressourcer vha. metadata.

Dette dokument bygger videre på "DK-CLARIN Metadata for ressourcer – forslag". DK-CLARIN har til hensigt at følge EU-projektet CLARIN's indstillinger og anbefalinger i det omfang det er muligt. Da EU-projektet CLARIN har arbejdet med deres specifikationsdokumenter i samme periode som DK-CLARIN afslutter sin specifikationsfase, har det være begrænset i hvilket omgang der har været mulighed for at følge ændringer og forbedringer i EU-projektet CLARIN's anbefalinger som ikke har været lagt fast tidligt i projektet. Nærværende arbejde har haft som basis at inddrage den seneste version af EU-projektet CLARIN-dokumenterne "Metadata Infrastructure for Language Resources and Technology 2009-02-04 Version 5" og listen "Data Categories"[1], samt "Standardisation Action Plan for Clarin(draft)" udleveret fra CLARIN. På nuværende tidspunkt (februar 2011) kan der henvises til **www.isocat.org** og http://www.clarin.eu/cmdi for information om EU-projektet CLARIN.

1.1 HVAD ER 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 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. med EU-projektet CLARIN blive brugt engelsk navngivning for de enkelte metadataoplysninger. Hvilket sprog der anvendes når metadataoplysningerne præsenteres for brugeren fastlægges af brugerinterfacet og behøver naturligvis ikke være begrænset til engelsk navngivning.

1.2 SAMARBEJDE MED EU-PROJEKTET CLARIN ANGÅENDE METADATA

[1] http://www.clarin.eu/view_datcats

DK-CLARIN ønsker så vidt muligt at kunne sameksistere med de valg der træffes i EU-projektet CLARIN. DK-CLARIN vil derfor også tilstræbe at følge forslag fra EU-projektet CLARIN i det omfang de foreligger og passer til DK-CLARIN's forhold.

I EU-projektet CLARIN har man foreslået en række generelle metadatainformationer som vil kunne specificeres til alle ressourcer, og derudover vil der være mulighed for at 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 der for visse ressourcer skal specificeres nogle særlige metadata, fx ønsker man for en lydoptagelse at registrere oplysninger om en talers bopæl, mens dette for et skrevet korpus normalt ikke er relevant.

I DK-CLARIN vil vi gerne vægte brug af standarder højt. I det indledende arbejder har vi haft mest fokus på DC Dublin Core Metadata Terms og OLAC (Open Language Archives Community) for generelle metadata og TEI Text Encoding Initiative for tekster. EU-projektet CLARIN har til WP2-arbejdsmødet i Oxford feb. 2009 skitseret en dobbelt strategi: henholdsvis en meget fleksibel arkitektur for metadata-håndtering generelt i EU-projektet CLARIN og en liste med konkrete forslag til metadata.

DK-CLARIN har i videst muligt omfang fulgt den del af CLARIN-strategien, som arbejder på at fastlægge en liste af metadata, og dette har WP5 i samarbejde med indholdsarbejdspakkerne specificeret. Sådan at de enkelte arbejdspakker kan følge disse specifikationer i deres dataindsamling og dataregistrering, samtidig med at WP5 implementerer håndtering af metadata. I samarbejde med de enkelte arbejdspakker fremlægges DK-CLARIN's skabelon for metadata, der primært vil blive benyttet for de nye ressourcer, der opbygges som en del af DK-CLARIN. Vi har i vores udgangspunkt taget hensyn til EU-projektet CLARINs liste af metadata som nu foreligger på den åbne del af EU-projektet CLARIN's hjemmeside http://www.clarin.eu/view_datcats [2].

For eksisterende ressourcer, der bliver indlagt i DK-CLARIN repositoriet, anføres senere i dokumentet hvordan de håndteres. Det er meget vigtigt at der ikke stilles uoverstigelige krav til disse eksisterende ressourcer, når de skal tilføjes til repositoriet, da det kan virke som en barriere for brugeren når denne overvejer at tilføje data. DK-CLARIN vil derfor også kun specificere en mindre mængde metadata-elementer som obligatoriske. EU-projektet CLARIN havde ikke ultimo 2009 opdelt deres metadataelementer i obligatoriske og optionelle metadata, men den opdeling ønsker vi i DK-CLARIN for at vi kan være sikre på hvilke informationer, der som minimum er til stede for en ressource.

EU-projektet CLARIN har også planer om at skulle implementere en helt fleksibel metadatastruktur-håndtering, sådan at alle potentielle CLARIN-brugere selv kan vælge hvilke metadata der er relevante for den enkelte ressource. Denne fuldstændig fleksible løsning forsøger DK-CLARIN ikke at implementere på nuværende tidspunkt, men hvis der senere er grund til at tilslutte sig, vil muligheden blive overvejet.

1.3 GENERELT OM STANDARDER I EU-PROJEKTET CLARIN

EU-projektet CLARIN arbejder med at fastlægge hvilken række af standarder som man generelt vil anbefale øvrige CLARIN-partnere og andre interesserede at overholde, og at gøre sig kompatible med.

[2] Herfra henvises nu til: **www.isocat.org**

Den nuværende version af dokumentet kan ses i "Standardisation Action Plan for Clarin[3]". Disse standarder omfatter bl.a. XML, ISO 10646 (Unicode), ISO 3166 (Country codes), ISO 639-3 (Language codes), ISO 15924 (Codes for names of scripts for written languages).

2 GENERELLE METADATA

I dette afsnit fokuseres på generelle metadata, mens metadata der er aktuelle for særlige ressourcetyper behandles i afsnit 3.

I det følgende fokuseres på hvilke metadataoplysninger, der generelt skal håndteres af repositoriet, og dette gøres med fokus på Dublin Core Standarden(DC). Det er nemlig fastlagt at repositoriet skal kunne udveksle metadata vha. OAI-PMH-protokollen[4] og denne protokol stiller krav om at der kan udveksles metadata i Dublin Core. Der kan i OAI-PMH-protokollen også udveksles metadata jf. andre standarder, men data i Dublin Core-format er obligatorisk.

2.1 GENERELLE DK-CLARIN METADATA

2.1.1 TIDLIGERE FORSLAG FOR GENERELLE DK-CLARIN METADATA

DK-CLARINs tidligere forslag til generelle metadata, hvor optionelle metadata angives med *. DC (Dublin Core) 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. Dette forslag er i det følgende udvidet jf. tilbagemeldinger fra partnere. Metadata for ressource-rettigheder specificeres ikke i denne rapport, da de håndteres af licens-arbejdsgruppen, og specificeres i grænsefladen til infrastrukturen.

2.1.2 GENERELLE OBLIGATORISKE DK-CLARIN METADATA

DK-CLARIN fastholder at de metadata-elementer der er opført som obligatoriske i det oprindelige forslag fortsat skal være obligatoriske. I tilbagemeldingerne på det tidligere forslag ønskedes kontaktinformation tilføjet metadata. Kontaktdata for udgiveren tilføjes ikke direkte i metadata for den enkelte ressource, til håndtering af dette kan man implementere en særlig udgiverdatabase, hvor rettelser i kontaktdata kan udføres uden at metadata-oplysningerne for de enkelte tekstressourcer

[3] http://www.clarin.eu/clarin-members/members-documents/standardisation-action-plan-for-clarin

[4] Open Archives Initiative Protocol for Metadata Harvesting http://www.openarchives.org/pmh

skal ændres. Selvom det kan anses for vigtigt at brugerne af DK-CLARIN kan få adgang til disse kontaktoplysninger ved søgning og inspektion af metadata, så er det ikke udgiverdatabasen ikke en facilitet der er inkluderet i implementeringen af DK-CLARIN.

Dette leder hen til en ny obligatorisk liste som nu indeholder otte elementer som skal oplyses ved tilføjelse af data i repository. Listen beskrives på engelsk for at kunne udveksles med EU-projektet CLARIN. Yderligere fem metadata-elementer udfyldes og vedligeholdes automatisk af repositoriet.

Alle ressourcer vil inde i repositoriet have mindst 14 metadata-oplysninger. Disse metadata-oplysninger er alle søgbare. De enkelte ressourcetyper kan have flere metadata-oplysninger som er søgbare.

Metadata	DC equivalent[5]	Legal values	Description
Type	dc.type	See section 2.2.1	Resource type
Title	dc.title	Literal value. Annotations title value starts with ”Annotation:”[6]	Name of the metadata resource representing the linguistic resources.
Language	dc.language	ID (ISO 639) of the languages included in the resource	ID (ISO 639) of the languages included in the resource or the languages of input resources for a tool
Format	dc.format	http://www.iana.org/assignments/media-types and other values	open vocabulary describing the medium used to exchange the resource
PublicationDate	dcterms.issued	dcterms:W3CDTF, eg. YYYY-MM-DD or YYYY-YYYY	date of the publication of the resource
Publisher	dc.publisher	List of publishers/ rightsholders/copyright owners	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
Rights	dc.rights	“Link”/Reference to licence agreement	Information about rights held in and over the resource.
AccessRights	dcterms.accessRights	This information has to be added in interface when depositing resources as the depositor needs to accept the conditions.	Information about who can access the resource or an indication of its security status.
Description	dc.description	Literal value.	Information describing the resource such as manuals, papers etc. Attribute xml:lang specify language if language is not English

[5] Prefikset ”dc:” henviser til namespace http://purl.org/dc/elements/1.1/, og prefiket ”dcterms:” henviser til namespace http://purl.org/dc/terms/

[6] Der skelnes ved eksport ikke mellem Tekstannotationer og Mediaannotationer her.

ADDED BY REPOSITORY			
PID	dc.identifier	dcterms:URI	An unambiguous reference to the resource within a given context. Unique value internally in repository. Added by repository. dc:identifiers will use a PID as an URI encoding scheme
DateSubmitted	dcterms.dateSubmitted	dcterms:W3CDTF, eg. YYYY-MM-DD or YYYY-YYYY	Date when the metadata description and resource are added to repository
DateModified[7]	dcterms.modified	dcterms:W3CDTF, eg. YYYY-MM-DD or YYYY-YYYY	The date of the last update.
ContentProvider	dc.mediator	clarin.dk user identification	The user/institution making the resource available (Content Provider).
DKCLARIN:type	Internal used type specification used by repository	DKCLARIN defined list	Used internally to describe resourcetype.

2.1.3 GENERELLE OPTIONELLE DK-CLARIN METADATA

DK-CLARIN opfordrer til at der specificeres optionelle metadata når en ressource tilføjes, når det er muligt. Særligt bør 'Creator' udfyldes, men det er ikke et krav.

'Creator' rummer plads til at specificere fx forfatternavn eller servicenavn, så man i metadata kan se hvem eller hvad, der har skabt ressourcen.[8]

'ConformsTo' bruges til at give information om standarder som ressourcen følger. En sådan standard kan være specificeret ved et XML Schema eller ved en reference til dokumentation for formatet.

Metadata	DC equivalent[9]	Legal values	Description
Creator	dc.creator	Literal value	A person, an organization, or a service.
CreationDate	dc.date	dcterms:W3CDTF, eg. YYYY-MM-DD or YYYY-YYYY	Date when the resource are processed by Content Provider

[7] DateModified will only be available when the facility to update resources is implemented.

[8] Det skal her bemærkes at selvom denne metadata-oplysning i repositoriet betegnes 'Creator', så er det ikke på nuværende tidspunkt fastlagt at netop denne betegnelse benyttes i web-grænsefladen. Grænsefladen vil sandsynligvis benytte danske navne i den danske grænseflade.

[9] Prefikset "dc:" henviser til namespace http://purl.org/dc/elements/1.1/, og prefiket "dcterms:" henviser til namespace http://purl.org/dc/terms/

SourceTitle	dc.source	Literal value	A related resource from which the described resource is derived. Eg. info about the situation, web-address or the book where the source of the digital resource is found.
ConformsTo	dcterms.conformsTo	Literal value. DK-CLARIN-interface uses a list of often used standards.	An established standard to which the described resource conforms. The repository has a list of formats, including xml schemas.
InfoAbout	dc.relation	InfoAbout:PID or InfoAabout:url	A related resource or a url that contains information about the resource. Eg. documentation, description, original metadata
Subject	dc.subject	DK-CLARIN do not use common subject domain classification system.	Subject domain for resource
ContributorResponsible	dc.contributor	Literal Value eg. project name	A short name or abbreviation of the project that lead to the creation of the resource or tool/service
DKCLARIN: ContentprovidersId		Id used by content provider to identify resource locally or in private storage. Can be used in relation files when describing relations between resources before ingest.	String of characters and numbers
ADDED BY REPOSITORY			
IsVersionOf[10]	dcterms.isVersionOf	PID values.	A related resource of which the described resource is a version

2.1.4 RELATIONER MELLEM RESSOURCER

Relationerne mellem ressourcerne skal angive hvilke forhold der gælder mellem forskellige ressourcer. Hvis infrastrukturen fx skal kunne vise hvilke annotationer der er tilknyttet en bestemt tekstfil, så skal denne information tilføres infrastrukturen. Til dette benytter vi relationer, og disse genbruges i delvist fra Dublin Core standarden dcterms[11]. Dcterms beskriver et antal relationer, bl.a. 'hasFormat' hvor det er fastlagt at denne relation benyttes for at angive at en anden ressource grundlæggende er den samme ressource men kan findes i et andet format i en anden ressource. Et

[10] IsVersionOf will only be available when the facility to update resources is implemented.

[11] DCMI Metadata Terms: http://dublincore.org/documents/dcmi-terms

eksempel kunne være to mediaannoteringer som egentlig indeholder den samme annotering, men forskellen er at der er benyttet to forskellige formater[12].

Relationerne håndteres af infrastrukturen, og kan importeres selvstændigt som relationer i en relationsfil. Enkelte relationer[13] kan udtrykkes eksplicit i metadata for visse ressource. Disse muligheder er beskrevet af de enkelte arbejdspakker for de enkelte ressourcer. Informationerne om relationer lagres i et særligt segment af et objekts metadata benævnt "relations".

Der er mulighed for at specificere at en ressource har relationer til andre ressourcer. Det gælder bl.a. for relationerne 'HasFormat', 'References' og 'Requires', hvor der er mulighed for at specificere relationer til andre ressourcer i infrastrukturen ved hjælp af PID-værdier eller ved brug af ContentProvider'sId(CPsID)-værdier.

'HasFormat' kan fx bruges til at relatere to mediaannoteringer som beskriver den samme annotering, men blot er i forskelligt format.

'References' kan fx bruges når output-ressourcen fra et værktøj vil beskrive hvilke input-ressourcer der har været brugt til at producere ressourcen. Det er muligt at angive mere end en ressource som 'References'.

'Requires' kan fx bruges når en ressource har referencer som refererer ind i en anden ressource. For aligneringsannoteringer af to parallelle tekster vil det være relevant at anføre to referencer i 'Requires', hvor hver reference refererer til en af de parallelle tekster. For STO-DanNet-ressourcen(WP4.2.2) vil her kunne angive at både STO- og DanNet-ressourcen er 'required' for ressourcen.

Der er brug for yderligere relationer, nogle af disse fokuserer på specifikke ressourcetyper. Alle relationer er anført i nedenstående tabel.

I det omfang en relation har en omvendt relation, fx Required og isRequriedBy vil begge disse blive oprettet ved import af den ene relation. Visse relationer angives på samme måde for den omvendte relation, dette gælder fx for HasFormat og ParallelText. Mens en enkelt relation ikke har en omvendt relation, nemlig InfoAbout.

Relation	Invers relation	Description
HasFormat	hasFormat	Equivalent to dcterms.hasFormat. A related resource that is substantially the same as the pre-existing described resource, but in another format.
InfoAbout	N/A	Can be specified in metadata for lexicon.

[12] Der kan være tilfælde hvor der ved konvertering mellem to formater sker informationstab og man derfor gerne vil have muligheden for at have ressourcen liggende i to forskellige formater.

[13] AnnotationOf og InfoAbout

References	isReferencedBy	Equivalent to dcterms.references. A related resource that is referenced, cited, or otherwise pointed to by the described resource.
isReferencedBy	References	Equivalent to dcterms.isReferencedBy. A related resource that references, cites, or otherwise points to the described resource.
Requires	isRequiredBy	Equivalent to dcterms.requires. A related resource that is required by the described resource to support its function, delivery, or coherence.
isRequiredBy	Requires	Equivalent to dcterms.isRequiredBy. A related resource that requires the described resource to support its function, delivery, or coherence.
hasDependent	isDependentOf	A related annotation resource that is dependent of this annotation.
isDependentOf	hasDependent	A related annotation resource, being dependent of this annotation.
AnnotationOf	hasAnnotation	A related resource for which this is an annotation. Can be specified in metadata for annotation.
hasAnnotation	AnnotationOf	A related resource that is an annotation of this resource.
ParallelText	ParallelText	A text which is parallel to the resource. EU texts often have parallel texts, where information about the language of the original is not known.
TranslationOf	hasTranslation	A relation to the original text.
hasTranslation	TranslationOf	A related resource that is a translation of the resource.
Relationer for samlinger	**Invers relation**	**Only relevant for collections.**
hasPart	isPartOf	A related resource that is included either physically or logically in the described resource. Equivalent to dcterms.hasPart
isPartOf	hasPart	A related resource in which the described resource is physically or logically included.

2.2 YDERLIGERE SPECIFIKATION AF GENERELLE METADATA

De generelle metadata er primært oplysninger som ejeren af ressourcen skal opgive ved tilføjelse af en ressource til repositoriet. I dette afsnit angives uddybende oplysninger om tilladte værdier for de enkelte metadata. Visse værdier bl.a. PID-værdien tildeles af infrastrukturen og dette beskrives ikke nærmere her.

2.2.1 TYPE OG DKCLARIN:TYPE

Internt i infrastrukturen benyttes en mere finkornet ressource-typebetegnelse, end der er mulig ved at udtrykke i DC-standarden. Disse typebetegnelser giver også information om hvilken metadata-specifikation en ressource overholder. Typebetegnelse angives som metadata-oplysningen: DKCLARIN:type.

Ved høstning af data fra clarin.dk til andre infrastrukturer vil dc.type blive angivet. Herunder angives den række af ressourcetyper, der er defineret i "WP5.1 Specifikation af Teknisk infrastruktur:

Ressourcetyper og –formater", suppleret i kolonnen til venstre med de betegnelser som bruges når en ressources type bliver høstet. I kolonne nr.2 angiver således den måde ressourcetypen angives på ved eksport vha. fx OAI-PMH-protokollen.

Tekst: http://purl.org/dc/dcmitype/Text

Fx <dc:type xsi:type="dcterms:DCMIType">Text</cd:type>

Lyd: dcterms http://purl.org/dc/dcmitype/Sound

Video: dcterms http://purl.org/dc/dcmitype/MovingImage

Billede: dcterms http://purl.org/dc/dcmitype/Image

Leksikon: olac:linguistic-type http://www.language-archives.org/REC/type.html#lexicon

Data: dcterms http://purl.org/dc/dcmitype/Dataset

TextAnnotation: dcterms http://purl.org/dc/dcmitype/Dataset.

Annotationens "title" starter med "Annotation:"

MediaAnnotation: dcterms http://purl.org/dc/dcmitype/Dataset.

Annotationens "title" starter med "Annotation:"

Tekstsamling: dcterms http://purl.org/dc/dcmitype/Collection.
Fx <dc:type xsi:type="dcterms:DCMIType">Collection</cd:type>
<dc:type xsi:type="dcterms:DCMIType">Text</cd:type>

Lydsamling: dcterms http://purl.org/dc/dcmitype/Collection.

Fx <dc:type xsi:type="dcterms:DCMIType">Collection</cd:type>
<dc:type xsi:type="dcterms:DCMIType">Sound</cd:type>

Mediasamling: dcterms http://purl.org/dc/dcmitype/Collection.

Billedsamling: dcterms http://purl.org/dc/dcmitype/Collection.

Datasamling: dcterms http://purl.org/dc/dcmitype/Collection.

Annotationssamling: dcterms http://purl.org/dc/dcmitype/Collection.

Værktøj, eksternt: dcterms http://purl.org/dc/dcmitype/Software

Værktøj, internt: dcterms http://purl.org/dc/dcmitype/Service

Som nævnt ovenfor bruges de anførte DC-typer i kolonne nr 2 ved eksport. Bemærk at der ved eksport vha. OAI-PMH-protokollen for annotationer er brug for at angive at ressourcen er en annotation vha. 'title'-feltet.

2.2.2 TITLE

Internt i DK-CLARIN indeholder dette felt fri tekst.

Da DC ikke har en datatype der specifikt omhandler annoteringer fastlægges det at annotationens titel ved udveksling med andre repositorier foranstilles med teksten "Annotation:" for at kunne skelne mellem Data og Annotationer i metadata.

2.2.3 *LANGUAGE*

De fleste tekster og ressourcer i DK-CLARIN vil være på dansk og kun 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 opfordrer til at angive indholdssprog for den enkelte ressource. Hvis det fx er en lydfil hvor flere sprog tales, kan der derfor specificeres flere sprog for ressourcen.

En vejledning til language codes kan ses på: http://www.w3.org/International/articles/language-tags/Overview.en.php. Denne anbefaling foreslår brug af korteste kode der er tilstrækkelig. For WP2's TEI-tekstressourcer angives 'language codes' således med kortest mulige kode, fx en, da. I IMDI filer er der en tradition for at bruge koder med tre karakterer, derfor er WP3's language-codes i IMDI filer angives som language som: ISO639-2:eng, eller ISO639-2:dan

2.2.4 *FORMAT*

Specifikation af værdier kan ses på http://www.iana.org/assignments/media-types. Der kan dog være enkelte ressourcetyper i DK-CLARIN som ikke er detaljeret beskrevet på denne liste, i de tilfælde må man vælge den bedste passende værdi.

2.2.5 *PUBLICATIONDATE, DATESUBMITTED, DATEMODIFIED, DATE*

Datoer bør alle steder registreres jf. W3CDTF, fx YYYY-MM-DD eller YYYY-YYYY. Hvis andre datoformater tillades i brugerinterfacet bør datoerne konverteres til W3CDTF-formatet ved tilføjelse til repositoriet. Yderligere info kan ses på http://www.w3.org/TR/NOTE-datetime

2.2.6 *PUBLISHER, RIGHTS OG ACCESSRIGHTS*

I metadata anføres informationer om 'Publisher'. Det forventes at det er 'Publisher', som har rettighederne for anvendelse af ressourcen[14]. Rettighederne for anvendelsen af ressourcen specificeres af 'ContentProvider' ved deponering af ressourcen, hvor det skal bekræftes hvilken licens ressourcen kan deponeres under.

Håndtering af licenser vil ikke blive specificeret nærmere her.

I feltet 'Rights' vil der for nogle ressourcer være mulighed for at angive et dokument der angiver mere specifikke rettighedsbetingelser.

2.2.7 *PID*

En PID (persistent identifier) som gør det muligt for brugere at identificere en ressource, og dermed gør det muligt at fremfinde netop denne ressource i fremtiden.

2.2.8 *CONTENTPROVIDER*

'ContentProvider' bruges til information om hvem der har tilføjet ressourcen. 'ContentProvider' specificeres ved hjælp af WAYF-identiteten for den bruger der tilføjer ressourcen i repositoriet.

---

[14] I DK-CLARIN vil 'Publisher' derfor indeholde samme information som andre kunne specificere i 'RightsHolder'

2.3 MINIMALT KRAV TIL METADATA

For at kunne udveksle og aflevere data til DK-CLARIN vil vi som minimum kræve at man skal kunne levere de specificerede generelle obligatoriske metadata. Dette skal sikre at der på den ene side er en fast mængde metadata man altid kan finde for en ressource, samtidig med at byrden for at specificere metadata for en ressource er minimal. Når man ønsker at finde ressourcer på baggrund af en søgning i metadata vil det ofte være en fordel hvis der er specificeret flere metadata end de obligatoriske, og grænsefladen til aflevering af data bør derfor tilskynde brugeren til at specificere så mange metadata som muligt/relevant for en ressource.

Er ressourcen en samling af ressourcer, skal der specificeres minimale metadata både for samlingen og for hver enhed i samlingen.

3 RESSOURCESPECIFIKKE METADATA

De generelle metadata kan suppleres af ressourcespecifikke metadata. Disse ressourcespecifikke metadata specificeres for alle de ressourcetyper DK-CLARIN har defineret som ressourcetyper repositoriet kan arkivere.

For tekster er metadata-formatet for nye tekstressourcer, der indsamles i WP2 ,beskrevet i TEI-P5 i dokumentet "Metadata for corpus texts" http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf. Formatet kaldes i det følgende TEIP5DK-CLARIN. I afsnit 3.1 anføres bl.a. konvertering mellem ovennævnte generelle metadata og TEIP5DKCLARIN-metadata.

For lydoptagelser, video og annotationer af disse mediaressourcer specificeres metadata vha. formatet IMDI. Disse specifikationer kan læses i rapporten "DK-CLARIN Metadata for WP3 ressourcer" http://intern.dkclarin.dk/files/DKCLARIN_MetadataWP3_Mar11ver1.0.pdf.

For billeder er metadata-specifikationen endnu ikke gennemført, men der arbejdes med et forslag. Se evt. afsnit 3.2.

For leksika specificeres metadata i rapporten "DK-CLARIN Metadata for WP4 ressourcer" http://intern.dkclarin.dk/files/DKCLARIN_MetadataWP4final.doc

Ressourcer af typen 'data' kan kun arkiveres i og downloades fra repositoriet. Der specificeres ingen særlige metadata for ressourcetypen 'data'.

Metadata for tekstannoteringer specificeres i afsnit 3.3.

Metadata for homogene og heterogene samlinger specificeres i afsnit 3.4

Metadata for værktøjer og services specificeres i afsnit 3.5.

Oversigt over de enkelte ressourcetypers metadata i flere detaljer end der ses i dette dokument vil på sigt kunne findes på hjemmesiden http://www.cst.dk/dk-clarin/?q=WP5.2_task5_metadata. Her vil også hentes en execl-fil med en samlet tabeloversigt over metadata i DK-CLARIN.

3.1 METADATA FOR TEKSTER

Specifikation af metadata for tekster der bearbejdes i WP2 foreligger og er baseret på TEIP5 og kaldes her TEIP5DKCLARIN. Specifikationen er defineret i "Metadata for corpus texts", se http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf.

De tekstspecifikke metadata er således fastlagt som de metadata, der er defineret i nævnte dokument og som ikke er indeholdt i afsnit 2 om generelle metadata.

Herunder angives hvordan visse metadata i TEIP5DKCLARIN header-elementer konverteres til de generelle metadata specificeret i afsnit 2. Den fulde beskrivelse kan findes på http://www.cst.dk/dk-clarin/?q=WP5.2_task5_metadata.

Metadata Repository	Metadata TEIP5DKCLARIN	Notes
Type	<teiHeader type="text">	
Title	<fileDesc><titleStmt><title> SamplingDeclaration textTitle</title>	
Language	<profileDesc ><langUsage> <language ident="languageID"> languageCharacterisation</language>	
Format	text/xml	Specified as text/xml when <teiHeader type="text">
PublicationDate	<profileDesc><creation><date when="YYYY-MM-DD"/>	
Publisher	<sourceDesc> <biblStruct> <monogr> <imprint> <publisher>#VALUE</publisher>	
Description	<noteStmt><note> note</note>	
Creator	<sourceDesc><biblStruct><analytic><author ><name ref="#personID></name>	
Date	<revisionDesc> <change when="#DATE">	
SourceTitle	<sourceDesc><biblStruct><analytic><title level="a">textTitle</title>	
ConformsTo	TEIP5DKCLARIN Schema	
InfoAbout		To be specified as a relation
Subject	<textDesc><domain>tdDomain</domain> </textDesc>	
ContributorResponsible	<fileDesc><titleStmt> <sponsor>DK-CLARIN</sponsor>	
DKCLARIN:ContentProvidersId	/teiHeader/fileDesc/publicationStmt/idno@ type!=file and externalUri	Can be used when specifying relations

3.2 METADATA FOR BILLEDER

I projektet inddrages forskellige billeder, bl.a. billeder af scannede tekster, og billeder fra Nationalmuseet.

Billeder af scannede tekster kan deponeres som en del af en tekstressource, og behandles i dette tilfælde uden selvstændige metadata.

For billeder fra Nationalmuseet er metadata-specifikationen endnu ikke afsluttet[15].

Følgende metadata er obligatoriske for billeder ud over de generelle obligatoriske metadata:

Metadata	Legal values	Description
Creator	Literal value	For image resources this contains information about the photographer or the creator of the image.
Object	Literal value	Information about the object of the image. Eg. a scanned text, an object or a location

3.3 METADATA FOR TEKSTANNOTERINGER

De enkelte annoteringsressourcer optræder som selvstændige ressourcer. Ressourcerne tilknyttes de ressourcer de er annoteringer af ved hjælp af relationen 'AnnotationOf'.

For tekstannoteringer gælder så vidt muligt de samme specifikationer som for tekster. Dog vil nogle af metadata-elementerne specificeret for tekst enten være irrelevante eller omdefinerede for annotationer.

De vigtigste annotationsspecifikke metadataelementer er beskrevet under tekster for metadata-elementet AppInfo og Application. Her angiver attributten fx subtype hvilket værktøj der har udført annotationen. Subtypes kan bl.a. være segmenter, s-splitter,regularizer,lemmatizer, pos-tagger, morph-tagger, term-tagger.

Metadata-elementer som SourceTitle, CreationDateCertainty, Subject og Editor er ikke anvendbare for annotationer. Metadata elementer fra TEIP5: /teiHeader/profileDesc/particDesc/person, /teiHeader/profileDesc/textDesc/channel og /teiHeader/fileDesc/sourceDesc/biblStruct/monogr/imprint/biblScope benyttes heller ikke af annotationer.

3.4 METADATA FOR SAMLINGER

DK-CLARIN skelner mellem homogene samlinger og heterogene samlinger. En homogen samling er en samling af samme type af ressourcer, fx en tekstsamling eller en billedsamling. Den generelle metadataoplysning 'PublicationDate' angiver for en samlingen den dag samlingen er dannet.

[15] Evt kan flere informationer ang. metadata for billeder ses i "Guidelines for handling image metadata http://www.metadataworkinggroup.org/pdf/mwg_guidance.pdf.

Rettigheder til at tilgå hele samlingen bestemmes af den ressource i samlingen som har de mest restriktive rettigheder.

Samlinger har kun en særlig metadata-oplysning, nemlig listen af PID'er til de ressourcer der er i samlingen.

Metadata	Legal values	Description	Obligatory
ContentRessources	List of PIDs	List of PID's describing all resources in the collection	Yes

Bemærk at der i DKCLARIN:type specificeres hvilken type samling der er tale om. WP3's IMDI-filer kan også beskrive en samling.

3.5 METADATA FOR VÆRKTØJER OG SERVICES (TO BE UPDATED)

Metadata-registreringen af værktøjer skal overordnet håndtere to forskellige måder værktøjer kan være tilgængelige på. Hvis værktøjet ikke kan aktiveres internt fra infrastrukturen defineres værktøjet som eksternt, hvis værktøjet er integreret vil det blive omtalt som en service. Metadata skal derfor indeholde information om værktøjet er eksternt eller er en service. Dette angives i feltet 'Integration', hvor services får en af værdierne 'REST' eller 'SOAP'.

Værktøjer og services inddeles desuden i undertyper i forhold til deres funktion, jf. DK-CLARIN_værktøjer-typer-if_19marts09.doc. Denne opdeling foretages som en hjælp til brugeren, så det er nemmere at finde de værktøjer der kan komme i betragtning til en ønsket opgave. Denne klassifikation angives i 'SubType'.

Metadata skal være informative for brugeren og indholdet i visse felter skal være søgbart. Fx skal felterne 'SubType' og 'Integration' være søgbart, mens de fleste andre felter fx 'Task' og 'Ortography' ikke er søgbare.

Værktøjer og services har mulighed for nedennævnte typespecifikke metadata.

Metadata	Legal values	Description	Obligatory
Integration	REST/SOAP/No	Specifies if the tool is integrated in the infrastructure.	Yes
SubType	List of types of tools	Specified subtype of tool, eg.lemmatiser, tokeniser, search tool	Yes
Dialect		Names of the dialects that the tool supports. (suggested by CLARIN)	
ResourceFormat		Specification of the input format.	Yes
ResourceFormatOutput		Specification of the output format.	
ExecutionLocation	"Integrated"/url to tool	Identification of the location where the tool/service is being executed. (suggested by CLARIN).	

OperatingSystem		Identification of the operating systems and its exact specification that is required to execute the tool/service. N/A for services. (suggested by CLARIN)	
RunningEnvironment		Specification of the running environment that is required to execute the tool/service. (suggested by CLARIN). N/A for services.	
CharacterEncoding		The character encodings accepted by the tool/service.	
CharacterEncodingOutput		The character encodings of the output produce by the tool/service.	
Tagset	PID/url to tagset-documentation	Specifies the tag set used in the annotation of the resource or used by the tool/service or it contains a URL that points to the information about the tag set. (suggested by CLARIN)	
Task	Literal value	The major task carried out in the resource or a typical task description of the tool/service. (suggested by CLARIN)	Yes
Orthography	Literal value	Description of special orthography that the tool can handle, eg. Danish text from 1872 to 1948.	

Eksempel på headerinformation specificeret for det integrerede værktøj CSTlemma:

Metadata Categories	Values for CST's lemmatizer	Notes
Type	Service	
Title	CSTlemma	
Language	Danish, British English, German	
Format	N/A	The tool is not to be exchanged.
PublicationDate	2009-10-30	
Publisher	Københavns Universitet	
Rights		"Link"/Reference to licence agreement.

		To be specified at ingest
Context	Non-commercial use allowed	
Creator	Bart Jongejan	
Description	REST-based Service Rule based lemmatizer for inflected languages. In addition to rules, the lemmatizer can employ a full form-lemma list.	
Source	N/A	
HasFormat	http://cst.dk/online/lemmatiser	
BasedOn	N/A	
HasVersion	N/A	
Requires	N/A	
ConformsTo	TEIP5DKCLARIN Schema	To be specified at ingest
InfoAbout	http://cst.dk/online/lemmatiser	
Subject	General Language	
Integration	REST	
Subtype	Lemmatizer	
Dialect	N/A	
ResourceFormat	TEIP5DKCLARIN-Baseformat	Input format
ResourceFormatOutput	TEIP5DKCLARIN-Baseformat	
ExecutionLocation	http://dkclarin.dk/tools/cstlemma	
OperatingSystem	N/A	N/A for a service
RunningEnvironment	N/A	N/A for a service
CharacterEncoding	UTF8	
CharacterEncodingOutput	UTF8	

Tagset	http://dkclarin.dk/resources/annotation/CSTtagset	
Task	1) lemmatization of plain text 2) lemmatization of PoS-annotated text	
Orthography	1950-present	

4 OPSUMMERING

Dokumentet indeholder specifikation af generelle og typespecifikke metadata i det omfang de enkelte arbejdspakker har specificeret typespecifikke metadata. Der henvises til andre rapporter i det omfang specifikationerne er dokumenteret der. Afsnittene ang. værktøjer og billeder er endnu ikke afsluttet.

5 REFERENCER

DK-CLARIN: specifikation af teknisk infrastruktur, 04.05.2010: Rapport

EU-projektet CLARIN's foreslåede datakategorier 2009: http://www.clarin.eu/view_datcats

EU-projektet CLARIN's 'Resources overview': http://www.clarin.eu/view_resources

Dublin Core Metadata Element Set, Version 1.1: http://dublincore.org/documents/dces/

DCMI Metadata Terms: http://dublincore.org/documents/dcmi-terms

IMDI: ISLE Metadata Initiative: http://www.mpi.nl/IMDI og
http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf

ISOcat http://www.isocat.org

WP2's TEIP5DKCLARIN metadata-beskrivelse: "Metadata for corpus texts" Jørg Asmussen,
http://korpus.dsl.dk/clarin/corpus-doc/text-header.pdf

DK-CLARIN: Metadata for WP3 ressourcer, 2011, version 1.0, se
http://intern.dkclarin.dk/files/DKCLARIN_MetadataWP3_Mar11ver1.0.pdf

"DK-CLARIN Metadata for WP4 ressourcer"
http://intern.dkclarin.dk/files/DKCLARIN_MetadataWP4final.doc

"Specifikation af teknisk infrastruktur: Værktøjstyper og -interfaces" Lene Offersgaard & Bart Jongejan, DK-CLARIN_værktøjer-typer-if_19marts09.doc