DK-CLARIN:
METADATA FOR WP3 RESSOURCER

DK-CLARIN WP 3 og WP 5.2

Version 1.0 2011-03-04

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

Michael Barner-Rasmussen, KU, mbr@hum.ku.dk

Frans Gregersen, KU, fg@hum.ku.dk

Johannes Wagner, SDU, jwa@sitkom.sdu.dk

Peter Juel Henriksen, CBS, pjh.isv@cbs.dk

Costanza Navarretta, KU, costanza@hum.ku.dk

Versionshistorie

Dato	Version	Logtekst	Ansvarlig	Status
13/1 2010	0.5	Samlet information til fælles dokument	Lene Offersgaard	Korrektur
2/2 2010	0.6	Opdateret efter review i WP3	Lene Offersgaard	Korrektur
9/6 2010	0.7	Opdateret efter kommentarer fra WP3 og andre	Lene Offersgaard	Korrektur
22/12 2010	0.9	Opdateret baseret på erfaring med produktion af IMDI-data	Lene Offersgaard	Justeringer
4/3/2011	1.0	Opdateret med afsnit om validator, samt øvrige ændringer.	Lene Offersgaard	Til godkendelse af partnere.

INDHOLD

1 INTRO

Formålet med dette dokument er at specificere metadata for WP3's ressourcer der består af lyd- og videoressourcer, der i det følgende betegnes media-ressourcer, samt annoteringer af disse, der betegnes mediaannoteringer.

Dokumentets seneste version kan findes på http://intern.dkclarin.dk/?q=WP5.2_task5_metadata. Dokumentet baserer sig på det generelle WP5.2 dokument "Metadata for ressourcer" som også kan findes på http:// intern.dkclarin.dk/?q=WP5.2_task5_metadata og aftaler truffet i WP3.

WP3 har besluttet at tage udgangspunkt i IMDI-standarden[1] ved specifikation af metadata. Det har i arbejdet vist sig at WP3 ønsker en del metadataelementer, der afviger fra allerede fastlagte del af IMDI-specifikationen, men dette kan godt udtrykkes i overensstemmelse med IMDI-standarden.

WP3 ønsker desuden et fåtal af obligatoriske metadataelementer, hvilket IMDI-standarden også kan honorere. Ud fra dette og at WP3.1 fandt det relevant at konvertere metadata for deres ressourcer til IMDI, samt at IMDI forventes at kunne konverteres forholdsvis nemt til CMDI (ClarinMetadataInfra-structure-format) har WP3 besluttet at anvende IMDI-formatet. IMDI-formatet giver også mulighed for udvidelser bestående af nogle WP3-definerede elementer. WP3's selvdefinerede elementer er nævnt i dette dokument.

Arbejdet i WP3 har et stykke tid afventet at arbejdet i EU-projektet CLARIN kom så langt frem som muligt og metadata-arbejdet med CMDI og ISOcat-projektet[2] findes nu i en propotype. Men vi tog i sommeren 2010 den beslutning ikke at vente yderligere på resultater fra CLARIN. og afleveringerne vil derfor foregå i IMDI-formatet. Det bør senere være muligt at konvertere IMDI-metadata til CMDI-format, men dette specificeres ikke i nærværende rapport, og dette arbejde forventes ikke udført som en del af DK-CLARIN projektet. Før en evt. konvertering til CMDI vil det desuden være hensigtsmæssigt at disse WP3-definerede elementer også defineres i ISOcat.

Metadata afleveres således i IMDI-formatet, som følger WP3's specifikation og IMDI-standarden. Repositoriet skaber ud fra disse et lille sæt metadata i Dublin Core (DC) format. Disse bruges af infrastrukturen sammen med IMDI-metadata, hvis metadata høstes fra repositoriet vha. OAI-server.

Sammen med metadata i IMDI-format afleveres de tilhørende ressourcer. Der er referencer i metadata-filen til de enkelte ressourcer og evt. supplerende dokumenter, der afleveres.

I defineringen af WP3's metadata tages der udgangspunkt i IMDI-skemaet, der er defineret som xsi:schemaLocation=http://www.mpi.nl/IMDI/Schema/IMDI ./IMDI_3.0.xsd. Der implementeres desuden en validator som kan validere i forhold til WP3's specifikation af aflevering af data. Denne validering vil også blive benyttet ved deponering af media-ressourcer i fremtiden.

---

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

[2] http://www.isocat.org

Man kan foretage en grov evaluering ved at benytte IMDI-skemaet, men dette vil kun checke en del af specifikationen af WP3's format, og kan derfor ikke alene bruges til at validere audio- og video-afleveringerne inden eller ved deponering.

Dokumentet indeholder først information om generelle metadata for alle ressourcer i DK-CLARIN ressourcer, hvor det er udtrykt hvordan IMDI-angivelserne, der benyttes af WP3, konverteres til de generelle metadata. Dernæst redegøres for de specielle WP3-metadata for henholdsvis mediaressourcer og annoteringsressourcer. Transskriptioner af mediaressourcer anses i denne sammenhæng for at være en særlig slags annoteringer. Endelig beskrives validatoren, der benyttes ved validering af media-ressourcer ved deponering.

Teksten i tabellerne i resten af dokumentet er på engelsk, da disse tabeller så lettere kan indgå i den endelige dokumentation og udveksles med CLARIN.

1.1 PID

DK-CLARIN repositoriet vil tildele alle ressourcer en PID. PID'en skal gøre det muligt for brugere at identificere en ressource, og dermed gør det muligt at fremfinde netop denne ressource senere. PID-værdier kan også bruges til at angive relationer mellem ressourcer. PID-værdien er ikke eksplicit nævnt i metadata tabellerne da den tildeles af repositoriet ved import af ressourcen eller ved import af ressourcens metadata.

Det vil også være muligt at opbevare en Content Provider ID i repositoriet som kan bruges til at referere tilbage til en tidligere lokal ID, som er benyttet før ressourcen blev en del af repositoriet. For WP3 ressourcer vil Content Provider ID blive tildelt ressourcens filnavn/url.

1.2 MINIMALT KRAV TIL METADATA

For video, lyd og mediaannoteringsressourcer kan man nøjes med at specificere de obligatoriske metadata-elementer. Det er anført i tabellerne om de enkelte elementer er obligatoriske.

1.3 STRUKTUR AF IMDI-METADATA

IMDI-metadata udtrykker metadata for en samling af ressourcer. Denne samling kaldes i IMDI en session. Ud over de oplysninger, der gælder for hele sessionen, kan der specificeres metadata, som tilhører de enkelte ressourcer i filen.

I DK-CLARIN har vi brug for at kunne tilknytte metadata både til samlingen og til de enkelte ressourcer, fx den enkelte video. De kommende skemaer skal derfor læses sammenhængende, sådan at metadata i IMDI's generelle session-del også gælder for de indeholdte ressourcer.

2 METADATA-ELEMENTER FOR WP3

I dette afsnit fokuseres først på generelle metadata for både mediafiler og annoteringsfiler, mens metadata der er aktuelle for de enkelte ressourcetyper behandles bagefter.

Grå felter i tabellerne angiver generelle metadata elementer for alle ressourcer i DK-CLARIN.

2.1 GENERELLE METADATA FOR ALLE WP3 RESSOURCER

Metadata WP3	Equivalent element[3]	Description WP3	Legal values	WP3 Obligatory
Type	dc.type & IMDI:Resources.MediaFile.Type / IMDI:Resources.WrittenResource.Type	Resource type **To be derived at ingest from session as one session contains a number of resources.**	audio,video,Annotation, Document, Unspecified	Yes
Title	dc.title & IMDI:Session.Title	Name of the metadata resource. Specify the title so it makes it possible to distinct the resource from other resources.	Literal value.	Yes
Language	dc.language & IMDI:Content.Languages.Language	ID (ISO 639) of the languages included in the resource	ID (ISO 639) of the languages included in the resource[4]	Yes
Format	dc.format & IMDI:Resources.MediaFile.Format / IMDI:Resources.WrittenResource.Format	Open vocabulary describing the medium used to exchange the resource	http://www.mpi.nl/IMDI/Schema/MediaFile-Format.xml For written resources: MIMEtypes	Yes
PublicationDate	dcterms.issued & IMDI:Session.Date	Date when the resource are processed by Content Provider	dcterms:W3CDTF, eg. YYYY-MM-DD or YYYY-YYYY	Yes
Publisher	dc.publisher IMDI:Resources[MediaFile,WrittenResource].Access.Publisher	One or more persons or organizations owning or managing rights over the resource. Name of the person or organisation to be contacted for accessing the language resource / for copyright issues	DK-CLARIN defined list of publishers/ rightsholders	Yes
Rights	dcterms.rights IMDI:Resources[MediaFile,WrittenResource].Access.	**To be specified, when depositing the resources in the infrastructure** Information about rights held in and over the resource.		No. Best not to specify this in metadata!

[3] Prefikset "dc:" henviser til Dublin Core namespace http://purl.org/dc/elements/1.1/. Prefikset "dcterms:" henviser til Dublin Core namespace http://purl.org/dc/terms. Prefixet "IMDI" henviser til http://www.mpi.nl/IMDI/

[4] En vejledning til language codes kan ses på: http://www.w3.org/International/articles/language-tags/Overview.en.php. WP3's language-codes i IMDI filer angives lige nu language som: ISO639-2:eng, eller ISO639-2:dan I modsætning til TEI-tekstressourcer, hvor language codes angives med kortest mulige kode, fx en, da.

	Description			
~~AccessRights~~[5]	dcterms.accessRights IMDI:Resources[MediaFile,WrittenResource].Access.Availability	**To be specified, when depositing the resources in the infrastructure** Information about who can access the resource or an indication of its security status. DK-CLARIN vocabulary is used describing the license conditions.	public access, academic access, restricted access	Will **not** be read from IMDI-file.
Contributor	dc.contributor IMDI:Session.Project.Name	Literal Value eg. project name	A short name or abbreviation of the project that lead to the creation of the resource or tool/service	Yes
Description	dc.description IMDI:session.Description IMDI:session.[MediaFile, WrittenResource].Description	Information describing the session or the resource. Attribute xml:lang specifies language if language is not English	Literal value.	Yes
Creator	dc.creator	A person, an organization, or a service.	Literal value	No
CreationDate	dc.date & IMDI:session.[MediaFile, WrittenResource].Date	Recording Date of the (first version of the) resource	dcterms: W3CDTF, eg. YYYY-MM-DD or YYYY-YYYY	No
ConformsTo	dcterms.conformsTo	The repository has a list of formats, including xml schemas.	*DK-CLARIN-repository supplies this from type*	Repository adds
InfoAbout[6]	IMDI:session.Resources.WrittenResource.Type=”Document”	InfoAbout:PID or InfoAabout:url	A related resource or a url that contains information about the resource. Eg. documentation, description.	No
Subject	dc.subject		N/A for WP3 resources	No
InteractionType	WP3 definition: IMDI:Session.Content.Keys.Key.InteractionTy	See section 3.1.4 and http://dgcsspublish.hum.ku.dk/InteractionTypes.xml	Unknown, Unspecified, Single person interview,	Yes

[5] AccessRights are specified by depositor as part of ingest process, information in ingested IMDI-file is not used. Caused by the legal aspects, the depositor will have to accept the license type for the resource as an act, not just as something written in a file.

[6] Some WP3 IMDI-files will use “document” for files containing a number of TextGrid annotations. This leads to the situation that at ingest, information in InfoAbout can only be created for WrittenResource=”document”, if and only if WrittenResource.Format equals text/TextGrid or text/x-chat.

	pe		Two-person interview , Group interview, Conversation, Focussed discussion, Focus group, Monologue, Experiment, Constructed, Computer, Phonecall, Telechat, Meeting, Work, Medical, Classroom, Tutorial, Private, Family, Sports, Religious, Legal, Face to face, Institutional, Other	
NumberOfParticipants[7]	WP3 definition: IMDI:Session.Content.Keys.Key.NumberOfParticipants	Number of participants	1,2,3,4,5,more,audience	Yes
SocialContext	IMDI:Content . Communication Context.Social Context	Indicates the social context the event took place in. See section 3.1.5	Unknown Unspecified Family Private Public Controlled Environment	No
For each actor *:[8]				
ActorCode	IMDI:Session.Actor.Code	Short unique code to identify the person participating in the session. Comments: Mostly the code is used in the transcription and annotations to identify parts belonging to this specific Actor.	String	Yes
FamilySocialRole	IMDI.Session.Actor.SocialRole	See section 3.1.6	String	Yes
ActorRole	IMDI.Session.Actor.Role		Interviewer Informant	No

[7] The validator only implements numbers as legal values.

[8] Metadata about participants states for audio and video resources information about people participating in the recording. WP3 has chosen not to use the actor/participant section for annotations metadata, but the IMDI standard allows using actors for annotations.

ActorGender	IMDI:Session.Actors.Sex	The sex of the person participating in the session. When the data about the sex of the Actor is lost or simply not recorded, the sex 'Unknown' should be selected. In case of an artificial Actor (a computer) 'Undefined' should be selected.	Unknown, Male, Female, Undefined	Yes
AgeAtRecording	IMDI:Session.Actors.Keys.Key.Age	The age of the person participating in the session.	The age is encoded as years;months.days from Codes for the Human Analysis of Transcripts [AGECHAT]. Regular expression: Unknown|Unspecified|[0-9]+(;[0-1]?[0-9](\.[0-3]?[0-9])?)?	No
BirthDate	IMDI:Session.Actors.*.BirthDate	Year or date of birth.	The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format.	No
BirthPlace	WP3 definition: IMDI:Session.Actors.Actor.Keys.BirthPlace	Adress, region (postal code), administrative unit (institutional address or place) with country as an obligatory feature	String	No
SocialClass	WP3 definition: IMDI:Session.Actors.Actor.Keys.SocialClass	Se afsnit 3.1.8	Upper Class, Middle Class, Working Class, Lower Working Class, Outside the work force	No
Anomalies	WP3 definition: IMDI:Session.Actors.Actor.Keys.Anomalies	Any diagnosed or by the informant acknowledge speech deficit or pathology may be noted here ranging from aphasia to stuttering.	String	No
Citizenship	WP3 definition: IMDI:Session.Actors.Actor.Keys.Citizenship	Actors National Citizenship. SDU suggested this extension	ISO 3166-1-alpha-2 code	No

2.2 METADATA KUN FOR AUDIOFILER[9]

Metadata WP3	Equivalent element	Description WP3	Legal values	WP3 Obligatory
NoOfChannels	WP3 definition: IMDI:Session.Re	Number of Channels	Number	No

[9] Metadata I dette afsnit kan også udtrykkes som en del af filens 'Description', dette benyttes bl.a. af WP3.1

	sources.Media File.Keys.Key.NoOfChannels			
SampleRate	WP3 definition: IMDI:Session.Resources.Media File.Keys.Key.	Sample rate given in kHz, e.g. 44.1 kHz	Number	No
SampleDepth	WP3 definition: IMDI:Session.Resources.Media File.Keys.Key.SampleDepth	e.g. 16 bit	String	No

2.3 METADATA KUN FOR VIDEOFILER

Metadata WP3	Equivalent element	Description WP3	Legal values	WP3 Obligatory
Compressor	WP3 definition: IMDI:Session.Resources.Media File.Keys.Key.Compressor	e.g. Sorensen, Cinepak	String	No
Video Sound Compression	WP3 definition: IMDI:Session.Resources.Media File.Keys.Key.VideoSoundCompression		String	No
Video Sound Sample Rate	WP3 definition: IMDI:Session.Resources.Media File.Keys.Key.VideoSoundSampleRate		String	No

2.4 METADATA FOR BÅDE VIDEOFILER OG LYDFILER

Metadata WP3	Equivalent element	Description WP3	Legal values	WP3 Obligatory
Strict Speaker Separation	WP3 definition: IMDI:Session.Resources.Media File.Keys.Key.StrictSpeakerSeparation	Tells if it is possible to strictly separate the speakers.	Yes/No	No
Recording equipment used	IMDI:Media File. Recording Conditions	Description of the technical conditions under which the media file was recorded. Comments: Used to describe the equipment used for the recording (e.g. microphone type, amplifier type etc.). This element is not constrained and covers prose text. Nevertheless, short typical descriptions are recommended.	String	No

Recording Quality	IMDI:Media File. Quality	A numeric indication of the quality of the media file.	It is suggested to describe the quality of the recordings with help of a number between 1 and 5 where 1 stands for low and 5 for high quality. Eg.1: "very noisy/reverberant" 4: "very quiet/low-reverberant" 5: ="non-echoic"	No
Room Layout	WP3 definition: IMDI:Session.Resources.MediaFile.Keys.Key.RoomLayout	Alment-sproglig karakteristik hvor fx denne slags udtryk kan bruges: "køleskab brummer", "gadestøj", "fjernsynsstøj i baggrund"	String	No
Microphone type	WP3 definition: IMDI:Session.Resources.MediaFile.Keys.Key.MicrophoneType	Microphone type	String	No
Microphone position	WP3 definition: IMDI:Session.Resources.MediaFile.Keys.Key.MicrophonePosition	Focus on "distance to primary sound source"	String	No
Microphone target[10]	WP3 definition: IMDI:Session.Resources.MediaFile.Keys.Key.MicrophoneTarget	e.g. "speaker NN", "room"	String	No

2.5 METADATA FOR MEDIAANNOTERINGER

Metadata WP3	Equivalent element	Description WP3	Legal values	WP3 Obligatory
AnnotationLevel [11]	IMDI:WrittenResource.Keys.Key.AnnotationLevel	See section 3.1.12	Comma separated open list of: Unknown, Unspecified, Gesture, Orthography, Phonetic, Phonology, Morphology, Morphosyntax,	Yes

[10] For MediaFiles: The content for MicrophoneTarget can also be a comma separated list as recordings can have more channels and a channel dedicated to specific speakers microphones.

[11] For WrittenResources: The content for AnnotationLevel, AnnotationMode and Tagset for 'WrittenResources' can be comma separated lists, and not just a single value while a Praat TextGrid-fil can contain a number of annotationlayers: each having a -Level, -Mode and Tagset value. The validator checks if the number of values in the lists fits together.

			Syntax, Semantics, Pragmatics, Transcription, Typology	
AnnotationSublevel	WP3 definition: IMDI:WrittenResource.Keys.Key.AnnotationSublevel	See section 3.1.13	Open list, see section 3.1.13	No
AnnotationFormat	IMDI:WrittenResource. Keys.Key.AnnotationFormat	Specifies the annotation format, since often mime type will not be sufficient	String	No
AnnotationMode	WP3 definition: IMDI:WrittenResource.Keys.Key.AnnotationMode		automatic, manual	No
Proof Reading	WP3 definition IMDI:WrittenResource.Keys.Key.ProofReading	Has annotation been proof read	Yes/No	No
Proof Reading Date	WP3 definition: IMDI:WrittenResource.Keys.Key.ProofReading Date	Date for latest proof reading	The date is encoded according to a profile of [ISO8601] as described in [W3CDTF] and follows the YYYY-MM-DD format.	No
Proof Reading ID	WP3 definition: IMDI:WrittenResource.Keys.Key.ProofReading ID	ID for proof reader	String	No
Tagset	WP3 definition: IMDI:WrittenResource.Keys.Key.Tagset	Url to tagset-documentation	Specifies the tag set used in the annotation of the ressource	No

3 YDERLIGERE SPECIFIKATION AF METADATA

I dette afsnit angives uddybende oplysninger om tilladte værdier for de metadata, hvor værdierne ikke allerede er beskrevet i tabellen herover eller hvor der er brug for uddybende kommentarer.

3.1.1 TITLE

Internt i DK-CLARIN indeholder dette felt fri tekst.

Da Dublin Core standarden ikke har en datatype, der specifikt omhandler annoteringer fastlægges det at annotationens titel ved udveksling med andre repositorier kan foranstilles med teksten 'Annotation' for at kunne skelne mellem ressourcetyperne 'Data' og 'Annotationer' i metadata.

3.1.2 FORMAT

Specifikation af værdier kan ses på http://www.iana.org/assignments/media-types

IMDI skemaet benytter følgende typisk følgende værdier for WrittenResources: http://www.mpi.nl/IMDI/Schema/WrittenResource-Format.xml og for MediaResources http://www.mpi.nl/IMDI/Schema/MediaFile-Format.xml, men skemaet validerer feltet som "open vocabulary".

Projektet har besluttet at benytte følgende formater.

For lydressourcer: wav, flac (NB: mp3 is not a recommended format)

For videoressourcer: mov, mp4, wm, avi

3.1.3 *PUBLISHER, RIGHTS OG ACCESSRIGHTS*

I metadata anføres informationer om 'Publisher'. Det forventes at det er 'Publisher', som ejer rettigheden for anvendelse af ressourcen[12]. Ved deponering i infrastrukturen skal den der deponerer vælge licens og rettigheder for anvendelse i clain.dk. Det er således hvad der oplyses ved deponering der gælder for ressourcen. Der bliver mulighed for at angive 'public', 'academic' og 'restricted' access.

3.1.4 *INTERACTIONTYPE*

InteractionType kan angives som en af følgende værdier:

- Unknown
- Unspecified
- Single person interview (definition: One person is the only interviewee although others may accidentally be present. One (other) person is the interviewer.)
- Two-person interview (definition: Two persons are the only interviewees although others may accidentally be present. One (other) person is the interviewer)
- Group interview (definition: more than two persons are the only interviewees although others may accidentally be present. One (other) person is the interviewer)
- Conversation (definition: None of the participants are ratified as interviewer or interviewee, i.e. the conversation is ideally symmetrical in terms of rights and obligations)
- Focussed discussion (as the above, but with a specified theme)
- Focus group (as the above but there is a facilitator present)
- Monologue (definition: One person talks to the microphone or to an imagined but not present audience)
- Experiment (definition: The setting is defined as one of elicitation of some sort of linguistic behaviour; the experimenter may but need not be present)
- Constructed,
- Computer,
- Phonecall,
- Telechat, meeting,
- Work, medical,
- Classroom,
- Tutorial,
- Private,
- Family,
- Sports,
- Religious,

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

- Legal,
- Face_to_face
- Institutional: Definition: The recorded event is an event which regularly occurs as part of the normal life in the institution where it is recorded. Examples: counselling, lectures, educational interaction in a kindergarten
- Other

3.1.5 SocialContext

This category is kept as IMDI suggests. It is not obligatory; information can also be expressed in InteractionType.

This element is referred to as 'IMDI:Content . CommunicationContext . Social Context' and has a closed vocabulary in the IMDI standard.

- Family: Definition: The access to the communication event is restricted to relatives.
- Private: Definition: The access to the communication event is restricted to specific individuals of the social environment. Examples : Friends, colleagues, professionals etc.
- Public: Definition: The access to the communication event is allowed to whoever, in a free or in a regulated manner.
- Controlled Environment: Definition: The access to the communication event undergoes the agreement to elicit a linguistic behaviour.
- Mass Media: Definition: Data drawn from mass media, e.g. TV, Radio.

3.1.6 Actor.ActorRole

WP3 benytter IMDI's Family Social Role som udgangspunkt. Listen er en åben liste. IMDI listen er allerede nu udvidet med 'Friend', men yderligere udvidelser vil komme.

IMDI:Actor . Family Social Role

- Unknown
- Unspecified
- Mother
- Father
- Child
- Husband
- Sibling
- Boss
- Partner
- Student
- Teacher
- Shaman/Priest
- Mayor
- Doctor

Extension:

- Friend

3.1.7 *ACTOR.BIRTHPLACE*

Geografisk information i form af en adresse, en egn eller en administrativ enhed, dvs. address, region (postal code), administrative unit (institutional address or place) with country as an obligatory feature. Åbent felt.

3.1.8 *ACTOR.SOCIALCLASS*

- Upper Class (reference to definition)
- Middle Class (reference to definition)
- Working Class (reference to definition)
- Lower Working Class (reference to definition)
- Outside the work force
- Unknown

3.1.9 *ACTOR.ANOMALIES*

Definition: Any diagnosed or by the informant acknowledge speech deficit or pathology may be noted here ranging from aphasia to stuttering.

Feltet indeholder fri tekst.

3.1.10 *ACTOR.LANGUAGES*

Language codes anvendes. IMDI har tradition for at anvende tre-bogstavs sprogkoder, se også s.6 fodnote 4.

3.1.11 *ACTOR.CITIZENSHIP*

ISO 3166-1-alpha-2 code anvendes.

3.1.12 *ANNOTATIONLEVEL*

Der benyttes følgende lukkede liste:

Unknown, Unspecified, Gesture, Orthography, Phonetic, Phonology, Morphology, Morphosyntax, Syntax, Semantics, Pragmatics, Transcription, Typology

3.1.13 *ANNOTATIONSUBLEVEL*

For hvert af elementerne i AnnotationLevel kan der specificeres en attribut, hvis man har brug for at angive en undertype til AnnotationLevel. Åben liste.

For Gesture: Open list: “Facial expression”, “Hand gesture”, “Body posture”, "Head movement" og "Gaze".

4 VALIDATOR

DGCSS har implementeret en validator der vil blive benyttet til at validere IMDI-filer , der afleveres både fra WP3 og fra andre leverandører.

Validatoren validerer en IMDI-fil, givet en liste med tilhørende filnavne, både om IMDI-filen overholder IMDI-skemaet, om den refererer til de filer der er nævnt i fillisten, og om den overholder den række af betingelser for obligatoriske metadata som er udtrykt i tabellerne i afsnit 2. Disse betingelser er udtrykt i en regelfil, som kan ses i næste afsnit.

Et eksempel på fillisten kan ses her:

```
<?xml version="1.0" encoding="UTF-8"?>
<validator>
   <imdi>http://dkclarin.dk/testdata/WP3/WP31_feb2011/SamtaleBank/Radio/natteravn.imdi</imdi>
   <filelist>
      <file>http://talkbank.org/media/CABank/SamtaleBank/Radio/natteravn/anita.mp3</file>
      <file>http://talkbank.org/data-orig/CABank/SamtaleBank/Radio/natteravn/anita.cha</file>
   </filelist>
</validator>
```

Regelfilen for validatoren, indeholder regler der udtrykkes på følgende måde:

```
<Rule>
      <KeyBinding>
            <ResourceType>All</ResourceType>
            <XPath>/Session/MDGroup/Content/Keys</XPath>
            <RequiredKey>InteractionType</RequiredKey>
            <Schema>http://DGCSSPublish.hum.ku.dk/InteractionTypes.xml</Schema>
      </KeyBinding>
</Rule>
```

5 METADATA FOR HOMOGENE SAMLINGER

En homogen samling er en samling af samme type af ressourcer, fx en lydsamling eller en vidoesamling. Den generelle metadataoplysning 'PublicationDate' angiver for en samlingen den dag samlingen er dannet. Disse samlingers metadata kan også udtrykkes som en IMDI-session, som beskrevet ovenfor.

Rettigheder til at tilgå hele samlingen bestemmes af den ressource i samlingen som har de mest restriktive rettigheder. En sådan samling har mulighed for at de samme metadata som beskrevet ovenfor, hvis den udtrykkes som en IMDI-session.

6 REFERENCER

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

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

DK-CLARIN WP5.2 Metadata for ressourcer, 2011, version 2.0: Rapport

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