## DK-CLARIN Statusrapport

## Aflevering nr. D5.2.4a Operational version of accessibility modules

|  |  |
| --- | --- |
| Arbejdspakke, nummer og navn | 5.2, Technical Infrastructure |
| Aflevering, navn | Operational version of accessibility modules  Ressourcer, værktøjer og web-grænseflade |
| Afleveringsfrist | T42, 30/6-2011 |
| Afleveringsdato | 22/7-2011 |
| Arbejdspakkeleder | Lene Offersgaard |
| Deltagere | Lene Offersgaard (CST), Bart Jongejan (CST), Jacob Larsen (KB), Tom Juul Andersen (KB), Tomas Christensen (DSL) |
| Evt. tilhørende dokumenter på hjemmesiden | http://clarin.dk |

Afleveringen omfatter faciliteter i release 2.0 af den tekniske infrastruktur. Infrastrukturen er tilgængelig via <http://clarin.dk>. Der henvises også til Statusrapporten DK-CLARIN

D5.2.4b, hvor dokumentationen er omtalt.

**Deponering af ressourcetyper**

Version 2.0 har mulighed for deponering af følgende typer af ressourcer:

* Tekstresurser i formaterne TEIP5DKCLARIN (Det definerede TEIP5-fællesformat for WP2’s tekster), TEIP5DKCLARIN med billeder, samt TEIDKCLARIN-plain-tei (tekster udtrykt i TEIP5, med meget få formatkrav)
* Tekstannoteringer i formatet TEIDKCLARIN\_ANNOTATION
* Videoresurser, hvor metadata specificeres i en IMDI-session, jf. WP’3 beslutninger
* Audioresurser, hvor metadata specificeres i en IMDI-session, jf. WP’3 beslutninger
* Mediaannoteringer, hvor annoteringer og metadata specificeres i en IMDI-session, jf WP’3 beslutninger
* Leksikonresurser, hvor metadata specificeres i formatet TEIDKCLARIN\_LEX (Det definerede TEIP5-fællesformat for WP4’s metadata.
* Værktøjer, enten som integrerede værktøjer, der kan aktiveres i infrastrukturen eller som deponerede værktøjer, med mulighed for download. Metadata specificeres i CMD-formatet (EU-projektet CLARIN’s framework for Component MetaData Infrastructure)
* Data, hvor metadata specificeres i CMD-formatet
* Relationer. Relationer er ret beset ikke en ressourcetype, men infratrukturen giver mulighed for at man kan importere relationer mellem ressourcer efter deponering. Dog skal man på selv eje den ressource man importerer relationer til.

Der er implementeret valideringsprocedurer for metadata for alle typer, samt validering af format af indholdsdata. Herunder er der for WP3’s resurser udviklet en selvstændig validator.

Deponeringsfaciliteterne er afprøvet gennem grundig testdeponeringer af afleverede testdata.

Der er i version 2.0 fravalgt at deponere billeder som en selvstændig resursetype, de deponeres i stedet som tekster med billeder.

Tekstaligneringer deponeres foreløbig som ressourcetypen ’Data’. Tekstaligneringerne er i TEIP5-format, men da der ikke på nuværende tidspunkt er implementeret en selvstændig deponeringsmulighed for aligneringer, deponeres aligneringerne med en tilhørende CMD-metadata-beskrivelse. Dette kan på sigt ændres således at tekstaligneringer kan få sin egen deponeringsmulighed. Da tekstaligneringerne deponeres samlet i en fil for hvert aligneret korpus fra WP2.6 drejer det sig om 8 ressourcer, og det er således overskueligt at gendeponere disse senere.

Formaterne er dokumenteret gennem rapporter om metadata, valideringsskemaer og info-tekster i web-grænsefladen. Se Statusrapport D5.2.4b, for detaljer om dokumentationen. Alle resurser udviklet i WP2-4 kan således deponeres i infrastrukturen.

Der har dog ikke været ressourcer til at arbejde med at komme fra indledende kontakter til konkrete deponeringer af ressourcer som ikke umiddelbart er en del af projektet. Deponering i Clarin.dk er på den anden side tilgængelig for alle forskere i DK, hvor deres institution er tilknyttet WAYF, og der er således vide muligheder for deponering for projekt-eksterne forskere.

**Integration af værktøjer**

Der er også integreret tekstannoteringsværktøjer, som kan aktiveres direkte fra infrastrukturen.

På nuværende tidspunkt er integration af følgende værktøjer udført:

* CST-lemmatiser
* CST-tagger

Der planlægges integration af flere værktøjer i perioden.

**Web-grænseflade**

Efter aftaler med styregruppen blev licensarbejdet flyttet til WP, mens web-grænsefladeudviklingen blev flyttet til WP5.2.

Der har været nedsat en følgegruppe til grænseflade-udvikling, som har bestået af både projektdeltagere og andre forskere. Følgegruppen har gennem en møderække givet feedback til web-udviklingen. Referater kan ses på: http://intern.dkclarin.dk/?q=wp5\_web

Udviklingen af webgrænsefladen, dens browserkompatibilitet og sammensætningen til den underliggende logik er nået så langt som det har været muligt med de forhåndenværende resurser. Der er kendte fejl og problemer, bl.a. er der problemer med søgning i lydfiler og IMDI-filer, ligesom der er problemer med grænsefladen for Microsoft Internet Explorer og Apples Safari webbrowsere.

Der er implementeret basal visning af tekstressourcer, både med og uden billeder. Der er implementeret mulighed for at få sammenflettet tekstressourcer med annoteringer, givet ressourcerne overholder de aftalte formater for tekstressourcer i WP2. Der er to sammenfletningsmuligheder, enten kan der sammenflettes til et xml-format eller til CQP-format, hvorefter de sammenflettede data, er klar til import i CQP-korpusværktøjet.

**Deponeringsomfang**

Deponeringen i release 2.0 af infrastrukturen har været mulig fra 15 juni 2011. Der har i perioden siden været en del kontakt med projektets ressource-leverandører og der er oprette et supportberedskab til brugerne indtil udgangen af september 2011.

Der er 22. juli deponeret følgende ressourcer i Clarin.dk:

* TEIP5DKCLARIN: 358
* TEIP5DKCLARIN\_ANNOTATION: 2114
* TEIP5-ressourcer: 670
* IMDI-SESSION-XML: 4
* IMDI-SESSION-CHAT: 1

Denne mængde forventes at stige voldsomt efterhånden som de enkelte arbejdspakker kommer videre i deponeringsarbejdet. I den første tid har der både været behov for support og for fejl-rettelser.

**Høstning af metadata**

Clarin.dk stiller nu en OAI-PMH-service tilrådighed, som kan bruges til at høste metadata fra Clarin.dk. Metadata kan umiddelbart høstes i ’DublinCore’ og OLAC-formaterne. Efter aftale med Clarin.eu har dette projekt afprøvet høstningen, og metadata høstes til Clarin.eu fra Clarin.dk.

**Kommentarer om evt. behov for yderligere arbejde**

* Der vil sandsynligvis være behov for bedre browser-kompatibilitet
* Der bør lægges et logo på forsiden, logoet er produceret, men ikke integreret på web-siden
* Konstaterede fejl i bl.a. søgning for visse ressourcetyper og for visning af billeder bør udbedres
* Opdatering af info-sider, når brugerne kommer med tilbagemeldinger
* Når der deponeres et meget stort antal ressourcer i infrastrukturen kan der måske opstå performanceproblemer da infrastrukturen ikke er testet med et meget stort antal ressourcer