

## **Noter fra DK-CLARIN-møde 15. december 2010 kl. 9-10.30**

BM, 22.12.2010

### **Til stede:**

Tre styregruppemedlemmer: Birte Christensen-Dalsgaard, Lars Trap-Jensen, Bente Maegaard

Fem WP5-medlemmer: Anders Conrad, Lene Offersgaard, Jacob Larsen, Tomas Groth Christensen, Bart Jongejan

Baggrunden for mødet var den rapport (*Status pr 25/11-2010*) som Anders Conrad på vegne af WP5 havde indleveret til styregruppemødet 29.11.2010. Denne rapport indeholdt nogle forbehold som styregruppen gerne ville forstå bedre. Mødet var blevet indkaldt med ret kort varsel og derfor var der begrænset fremmøde fra styregruppens side.

Nedenstående gør ikke krav på at være et egentligt referat, men noter fra mødet, herunder præciseringer og beslutninger om yderligere møder mv. Noterne er nummereret, da det muligvis kan lette referencer.

1. Hele datacentret er nu klar til deponering, dvs. deponering kan nu ske i infrastrukturen er klar, backup er implementeret og testet etc. Det anslås at 11/12 af arbejdet med datacentret er gennemført. Der kan på nuværende tidspunkt ikke deponeres alle datatyper, men workflow omkring dette er klart.
2. Testdata findes for 4 af WP2's delpakker (**action:** LTJ for de to resterende), og man er parat til at modtage deponering. Det blev foreslået at WP2 laver en køreplan for deponering, møde i januar (**action:** LTJ)
3. WP4 er langt fremme. (**action:** BSP - find ud af hvad der mangler og aflever)
4. WP3: der er ikke modtaget testdata fra CBS, mens der er modtaget noget fra både KU-LANCHART og SDU. Flere testdata er således stadig ønsket for at sikre bedst muligt at alle taledata og multimodale data som ønskes deponeret, kan deponeres, dvs. for hele WP3. (**action:** FG). WP3 var ikke til stede ved dette møde, derfor arrangere møde for WP3 (**action:** BM). WP3 laver en validator der accepterer data før deponering, aftale om API findes. Validator afleveres i januar. Metadata-format er aftalt.
5. ASC foreslår en deponeringsworkshop, men der kan også være andre måder at hjælpe deltagerne i gang med at deponere (**action:** ASC)
6. Vedr. søgning blev det nævnt at man ikke vil kunne søge på kombination af tekst og annotering. CQP kunne være et værktøj der kunne bruges. Grunden til at det ikke er ligetil at søge på annotering, er at annotering lagres i en fil for sig. MarcLogic kan håndtere XML, men så skal man 'flette' filerne. WP5 vil sikre at man kan eksportere data der kan læses direkte ind i CQP hos brugeren. I et evt. opfølgningsprojekt kan man evt. implementere en CQP som del af DK-CLARIN.
7. Tools:  
Integration af tools til håndtering af tekst er undervejs, og udviklingen siden 25/11 gør at ASC nu er mere tryk ved det. Input/output for værktøjer er beskrevet; det er vanskeligt at lave en generisk løsning der kan tage alle typer af værktøjer i betragtning; derfor må der fokuseres på de værktøjer som deltagerne har indmeldt. Samle alle værktøjsudbydere snarest (**action:** BJ). Tools-modulet forstår metadata for tekst. Der arbejdes ud fra at værktøjer

køres i batch-mode således at der kan etableres en pipeline; dette kører og er demonstreret for tekst på WP5-møder. Resultatet af en kørsel med et (eller flere) tool vil tilgå brugeren som et link, og med en forespørgsel om der skal deponeres (dette sidste er under udvikling).

8. Fokus på test af tools-modulet (**action:** BJ)

9. Tools er ikke beskrevet i kravspecifikationen; men vi ønsker ikke en opdatering af selve krav-spec., hellere en tilføjelse (**action:** ASC, BJ)

10. Grunden til at ASC ikke kan følge fremdriften i tools-moduler er at BJ har givet op over for kompleksiteten i afrapporteringen. Der var enighed om at begge parter må gøre sig umage i fremtiden.

11. Visning:

**Alle** datatyper kommer til at kunne vises på en basal måde. Bredde er valgt frem for dybde, for at sikre at alle datatyper er med. Der kan derfor være specifikke forventninger for specifikke datatyper der evt. ikke vil blive opfyldt. Fx kan man ikke se en tekst og dens annotering samtidig.