Textbank workflow

Møde med Jakob & Sussi 19/5-2009 kl. 13:30 i DSL's lille mødelokale

1. Status vedrørende CTB-tekstformat (JA)
2. Demo af header-generator (JH)
3. Demo af CTB-kigger (JA)
4. Workflowet

   a) Indkommende tekstmateriale skal **formatkonverteres**, før det kan lægges i tekstbanken. Tekstmateriale kan komme i to prototypiske varianter:

      i. Tekst uden meta-data => Teksten skal konverteres til CTB-tekstformatet, der tilføjes en header-skabelon til manuel udfyldning (som kan ske i CTB-kiggeren). Dette kræver **tekstkonvertere**, der i første omgang udvikles som specifikke ad-hoc køkkenbordsprogrammer.

      ii. Tekst med meta-data (fx InfoMedia) => Teksten skal konverteres til CTB-tekstformatet, eksisterende meta-data konverteres til CTB-headerformatet (som evt. skal suppleres manuelt). Dette kræver desuden **metadatakonvertere**, som laves på samme grimme måde som tekstkonverterne.

   b) Tekstmaterialet skal lægges ind i eXist. Dette kræver måske en **collection-grundstruktur** i db'en, fx alle WP 2.1's tekster for sig, alle WP 2.2's tekster for sig e.l. Alternativt kan man lægge dem hulter til bulter i én collection, idet kiggeren tillader at filtrere dem ud, man vil se nærmere på. Dette kræver endvidere en **indlæggelses-procedure**. Måske det bedst gøres med eXist-clienten og i første omgang som en centralt styret proces.

   c) I CTB-kiggeren skal det være muligt at oprette **profiler**, der gør det muligt kun at se de tekster, der matcher dem.

   d) De samme profiler anvendes til at udtrække de tekster, der skal underkastes en særlig behandling, fx **tilretning af headeroplysninger** eller

en særlig **opmærkning af teksten**. Headertilretninger foregår manuelt i (fx) CTB-kiggeren. Opmærkningskørsler foregår ved at udtrække de relevante tekster, udføre opmærkningen på dem og lægge dem tilbage. Enten processeres de én efter én eller som batch.

e) De samme profiler anvendes til at **udtrække** de tekster, der tilsammen skal udgøre et korpus og skal gøres søgbare via (fx) CQP

5. Spørgsmål

   a) Skal vi satse på en udbygning af CTB-kiggeren i stedet for eller som supplement til oXygen?

   b) Hvordan indgår header-generatoren bedst i workflowet?