Stand van zaken DCS eind december 2012

Personele bezetting

De DCS had tm december 2012 de gewenste bezetting. In oktober werd duidelijk dat Maaske Treurniet (documentalist) per 1 januari 2013 ander werk had gevonden. Momenteel loopt een sollicitatieronde. De verwachting is dat het DCS-team per 1 februari met een vervanger voor Maaske is gecompleteerd.

DBD

Zie voortgangsverslag, opgenomen in curatierapport op Google Docs: "DBD curation report" dat te vinden is op http://www.clarin.nl/group/415 > Status of resources.

Dit moet nog worden gedaan:

- Component Size in imdi is in CMDI: cmdi-totalsize geworden met daarin de elementen Number en SizeUnit. De velden moeten nog automatisch worden gevuld met gegevens.
- Keys uit Backus-Bilingual verplaatsen. Het gaat om de key genaamd "Turkish proficiency" "Kurdish Proficiency" en "proficiency".
- De warnings die het XSLT-script genereerd mbt keys controleren cq verbeteren
- Metadata moet gekoppeld worden aan de extra transcripties (van CHILDES) en aan de extra .wav files die zijn aangemaakt.
- de pdf-files die op groepsniveau stonden in .imdi: daar moet nu ook naartoe gelinkt worden vanuit de cmdi files van de individuele sessies. Daarna kan de data naar het MPI,
- Het probleem met de ResourceProxy moet nog door het MPI worden opgelost.
- de documentatiefile (curatiereport op Google Drive) van DBD moet meegenomen worden in de uiteindelijke database

CLARIN-datacentrum: MPI

Verwachte levering aan CLARIN-datacentrum: eind februari 2013

IPNV

Zie voortgangsverslag, opgenomen in curatierapport op Google Docs: "IPNV curation report" dat te vinden is op http://www.clarin.nl/group/415 > Status of resources.

In september 2012 zijn we gestart met de curatie van de resterende IPNV interviews. Zo'n 250 interviews in deze collectie zijn reeds gecureerd in het kader van het INTER-VIEWs-project (ook CLARIN-NL). Blijven er nog zo'n 850 te doen door de DCS. De IPR-rechten zijn afgezekerd in een convenant tussen het Veteranen Instituut en DANS. DANS Heeft de audiofiles in beheer en bovendien (de metadata van) een groot aantal interviews al via in EASY beschikbaar gesteld.

Het curatiewerk omvat CMDI metadata files maken en die vervolgens harvestable maken. Dat laatste is een taak van DANS evenals het genereren van de PIDs.

Er is ruim een maand voor nodig geweest om het Veteranen Instituut duidelijk te maken dat deze curatie past in het convenant en in de verwerkingslijn van het eerdere INTER-VIEWs project. Dat had alles te maken met de tijdelijke vervanging van een leidinggevende daar. Toen we alles voor elkaar dachten te hebben, kwam de vorige weer terug zodat we alles opnieuw moesten uitleggen. Daarna zijn we gedurende de maand november in de weer geweest om van de IT-beheerder van de collectie Become-IT (Leo Kodde) de benodigde metadata te verkrijgen. Deze meende dat hiervoor een aparte toestemming van het veteraneninstituut nodig was.
Stand van zaken eind december is dat de scripts voor de conversie gereed staan. De conversie zelf zal in januari worden gedaan.

CLARIN-datacentrum: DANS

Verwachte levering aan CLARIN-datacentrum: februari 2013

Gelders Dialectwoordenboek (WGD)

Zie voortgangsverslag, opgenomen in curatierapport op Google Docs: "WGD Curation Report" dat te vinden is op http://www.clarin.nl/group/415 > Status of resources.

Charlotte Giesbers die werkt aan de Gelderse dialect woordenboeken heeft contact opgenomen met de DCS om de mogelijkheden voor curatie ervan te verkennen.

Voor Gelderland bestaan er 3 dialectwoordenboeken: 1 voor het Rivierengebied, 1 voor de Veluwe en 1 voor de Achterhoek en Liemers.

-De data voor het Rivierengebied is ontvangen. In totaal zijn het 19.000 records.

-De data voor de Veluwe is afgerond en wordt in februari verwacht, samen met een plaatselijk woordenboek voor het Groesbeeks dat in hetzelfde formaat wordt opgeleverd.

-De data voor de Achterhoek en Liemers is nog niet digitaal beschikbaar. Charlotte werkt daar nog aan.

De data zijn de resultaten van de vragenlijsten, de data bestaan uit:

- lemma
- dialectwoord
- betekenis
- plaats
- kloekecode
- beschrijvingen

We verwachten de curatie van de data van het Rivierengebied in februari 2013, en voor de Veluwe en het Groesbeeks in maart 2013 te kunnen afronden.

Er is een afspraak met Folkert de Vriend gemaakt dat hij deze data zal cureren en wel op dezelfde manier als hij dat voor CLARIN (COAVA-project) reeds voor het Brabants (WBD) en Limburgs (WLD) woordenboek gedaan heeft. Folkert zal voor zijn inspanningen een tegemoetkoming krijgen.

CLARIN-datacentrum: Meertens

Verwachte levering aan CLARIN-datacentrum: maart 2013

Juridisch Corpus

Verzoek van Arjan van Hessen, data van Martha Komter en Wilbert Spooren. Dit corpus bestaat uit 4 onderdelen:

(1) 31 geluidsopnames van strafzittingen (meervoudige kamer, in 3 arrondissementen), begin jaren '90.

(2) 20 geluidsopnames van politieverhoren (bureaus Nieuwmarkt en Warmoesstraat, Amsterdam), rond de eeuwwisseling.

(3) 14 geluidsopnames van politieverhoren (roofteam in de Bijlmer), 2007-2009.

(4) 5 video-opnames van zittingen van dezelfde zaken (jeugdzaken), 2007-2008.

Alle opnames zijn digitaal opgeslagen.

Geconcludeerd werd (juni 2012):

- IPR van het materiaal is niet geregeld, het mag aan niemand ter beschikking worden gesteld. Wellicht dat dit later nog eens gebeurt, maar daar is nu echt niets van te zeggen
- Het is nu zinloos als de DCS er desondanks toch (veel) werk insteekt om het te cureren: oplijning, anonimiseren van tekst en audio, metadata, PIDs
- Curatie daarom voorlopig achterwege laten.
- Jammer dat dit zoveel tijd heeft gekost

Verzoek Nicoline van der Sijs:

Het omzetten van ca. 480.000 microfiches van de vragenlijsten van het Meertens Instituut naar jpg/pdf. De eerste vraag is of het omzetten van de 480.000 microfiches naar jpg/pdf kan vallen onder de DCS: er zijn al veel kosten en moeite gedaan om de data van papier om te zetten naar een digitaal medium, maar inmiddels voldoet het medium microfiche niet meer aan de eisen die we momenteel aan data stellen. De tweede vraag is of het scannen van de papieren antwoorden wellicht ook onder de DCS kan vallen.

Besloten is dat het Meertens deze kosten zelf zou moeten betalen. Wel wil de DCS bemiddelen bij het inzetten van Breed voor het scannen/digitaliseren van de gegevens.

LESLLA

Zie voortgangsverslag, opgenomen in curatierapport op Google Docs: "LESLLA Curation Report" dat te vinden is op http://www.clarin.nl/group/415 > Status of resources.

In nauw overleg met Ineke van de Craats wordt LESLLA gecureerd. Hier is in september 2012 aan begonnen. De bedoeling is dat de data in ieder geval veilig gesteld worden (door overdracht aan de DCS) en dat er enkel een inspanning geleverd wordt om het corpus middels de metadata te ontsluiten

Het copiëren van de data van een harde schijf was onverwacht tijdrovend maar is uiteindelijk via de DVD-backup in orde gekomen. Er is een metadataprofiel opgesteld (dat is afgeleid uit dat van de DBD).

LESLLA-files waren in praat-collectionfiles opgeslagen. Deze zijn uiteengerafeld in wav files en (praat-)transcriptiefiles. De laatste zijn deels geconverteerd naar EAF-files (ELAN). De rest moet nog. Dit zal gebeuren als er een vervanger voor Maaske is gevonden.

Te doen:
- Metadata omzetten naar CMDI via .CSV.
- Batch-extractie van collection naar losse .textgrid en .wav,

CLARIN-datacentrum: MPI

Verwachte levering aan CLARIN-datacentrum: maart 2013

Representatie DCS

De DCS is gepresenteerd op de Dag voor de Fonetiek op 13 december 2012 door middel van een presentatie van Maaske Treurniet:
http://www.fon.hum.uva.nl/FonetischeVereniging/DvdFonetiek/DagvdFonetiek2012abstracts.html

De DCS zal tevens met een poster en een mondelinge presentatie aanwezig zijn op de International Data Curation Conference 2013 in Amsterdam (14-16 januari 2013). Zie ook http://www.dcc.ac.uk/events/idcc13/programme en http://www.dcc.ac.uk/events/idcc13/posters-and-demonstrations. Er is een paper geschreven dat gepubliceerd zal worden in het International Journal for Data Curation van de DCC.

DCS Curatieplannen 2013

Naast het afronden van de curatie van bovengenoemde dataverzamelingen, zijn de werkzaamheden van de DCS vooralsnog gericht op curatie van de Gelderse dialectwoordenboeken, diverse werken van het Centrum voor Parlementaire Geschiedenis en data van deelprojecten van het Traces of Contact project.

Boeken van het Centrum voor Parlementaire Geschiedenis (CPG)

In samenwerking met het CPG worden diverse belangrijke publicaties van het CPG gecureerd. Het gaat om de volgende werken:

1 Jaarboeken van de Nederlandse parlementaire geschiedenis (periode 1999-2009): 11 delen, totaal ca. 2200 pagina's (140 bijdragen) Serie boeken (6 delen) over de Nederlandse kabinetten (in de periode 1945-1956):

- *Het kabinet-Schermerhorn-Drees (1945-1946), 756 pagina's*
- *Het kabinet-Beel (1946-1948), 6 volumes, 4422 pagina's*
- *Het kabinet-Drees -Van Schaik (1948-1951), 3 volumes, 2585 pagina's*
- *Het kabinet-Drees II (1951-1952), 817 pagina's*
- *Het kabinet Drees III (1952-1956), 632 pagina's*
- *Het kabinet Drees IV en het kabinet-Beel (1956-1959), 371 pagina's*

3. Biografieën:

- *Prof.dr. G.M.J. Veldkamp, Herinneringen 1952-1967. Le carnaval des animaux politiques, 229 pagina's*
- *Architect van onderwijsvernieuwing. Denken en daden van Gerrit Bolkestein 1871-1956, 279 pagina's*

De pdf's van deze werken zijn te zien op http://www.ru.nl/cpg/onderzoek/online-cpg/

Curatie omvat het cureren van de text en de metadata.Met de uitgevers is een regeling getroffen v.w.b. de IPR en de werken mogen openbaar gemaakt worden. De boeken staan al online in pdf-formaat bij het CPG en kunnen nu in het kader van CLARIN-NL breder toegankelijk gemaakt worden.
Beoogd CLARIN Centrum: DANS
Betrokken partnersDe curatie wordt uitgevoerd door de DCS in samenwerking met het CPG en DANS. Voor de inzet van UCTO en NERD wordt Martin Reynaert (Tilburg University) geraadpleegd (Maarten van Gompel is uitvoerende voor de DCS)..

Data van het Traces of Contacts project

Het betreft hier data die voortkomen uit deelprojecten van het Traces of Contacts project (Muysken; zie: http://www.ru.nl/linc/projects/erc-traces-contact/) Het gaat hierbij om

- Multilingual Netherlands (betrokken onderzoekers: Linda van Meel, Pablo Irizarri van Suchtelen, Suzanne Aalberse, Margot van den Berg
- Multilingual processing (betrokken onderzoeker: Gerrit-Jan Kootstra)
- Suriname
- South America

Deze deelprojecten zijn recentelijk afgerond of staan op het punt afgerond te worden. Dit is dan ook een optimaal moment om de curatie ter hand te nemen.

De werkzaamheden/acties die door de DCS uitgevoerd zullen worden zijn de volgende:
- het regelen van de IPR; er zal een MPI-licentie worden opgesteld die …; de licenties moet vervolgens ondertekend worden door proefpersonen. Daarnaast moet er een gebruikerslicentie worden opgesteld waarin is vastgelegd onder welke voorwarden en voor welke doeleinden de data gebruikt mogen worden (vertaling van dataleverantielicentie naar gebruikerslicentie)
- het DBD-metadataprofiel moet worden aangepast zodat de aanvullende metadata die de nieuwe data met zich meebrengen kunnen worden geaccommodeerd.
-alle data en aanvullende documentatie/informatie moet worden verzameld en overgedragen aan de DCS. De DCS onderhoudt hierover contact met de betrokken onderzoekers.
- voor elk van de datacollecties wordt een curatieplan geschreven.

Overige

Nadat de DCS eerder was geattendeerd op het LULC dat over een reeks van resources zou beschikken die wellicht voor curatie door de DCS in aanmerking zouden komen, heeft de DCS bij herhaling gepoogd deze te achterhalen. Aanvankelijk door contact te leggen met Marjan Klahmer. Nadat duidelijk was geworden dat met haar curatiewerkzaamheden in een door CLARIN in Call 3 gehononeerd project in de curatie van alle data van haar en directe collega's (o.a. Mous) voorzien was, hebben we vervolgens op haar suggestie geprobeerd contact te leggen met Ton van Haaften. Dit heeft niet tot resultaat geleid.

De DCS bekijkt met Onno Crasborn de mogelijkheden voor curatie van de transcripten die horen bij het videomateriaal dat reeds CLARIN-conform werd gearchiveerd. Het gaat hierbij om rijke transcripten op papier bij de dataset van Heleen Bos die gericht was op morfosyntactisch onderzoek naar gebarentaal en de dataset van Beppie van den Bogaerde en Anne Baker met kindertaalopnames. Deze twee dataverzamelingen met Nederlandse Gebarentaal vormen samen het enige materiaal buiten het Corpus NGT dat algemeen beschikbaar is voor andere onderzoekers, en daarmee ook het oudst beschikbare materiaal. Eerdere opnames zijn ofwel verloren gegaan ofwel niet gedocumenteerd. De dataset van Bos omvat in totaal 20 uur video. Voor 10 uren hiervan bestaat een handmatige transcriptie (ca. 2000 transcriptiebladen). Voor de dataset van Van den Bogaerde & Baker (80 video) is voor ca. 20 uren een transcript (ca. 4000

transcriptiebladen) beschikbaar. Doel van de curatie zou zijn om de glos- en vertalingtiers (Bos) resp. de glos- en spraaktiers (Van den Bogaerde & Baker) te converteren naar EAF-files.

Verder overweegt de DCS curatie van de Database Streng n.a.v. het in Call 4 ingediende, maar niet-gehonoreerde LION voorstel (CLARIN-NL-12-017. Het IAP suggereerde hier dat het curatiedeel van het project wellicht door de DCS zou kunnen worden uitgevoerd. De DCS legt hiervoor contact met Ton van Kalmthout en het Huygens ING.

De DCS zal daarnaast nagaan welke van de eerder geïnventariseerde kandidaten concrete mogelijkheden bieden voor curatie op korte termijn.