

Verslag 1^e CODE-Congres: *Data, tools en output*

Datum: 21 maart 2025

Plaats: Vrij Universiteit Amsterdam

Verslag: Dr. mr. D.A. Sirks; Dr. M.A.J. Vermeer en drs. R. Haverman

1. Dr. **Erika Kuijpers**: *Het kroniekencorpus*

Dr. Erika Kuijpers (VU Amsterdam) presenteerde het onlangs afgeronde project 'Chronicling novelty. New knowledge in the Netherlands 1500-1850', dat zij samen met Judith Pollman (Universiteit Leiden) tussen 2018 en 2023 heeft geleid op basis van een NWO Vrije Competitie-beurs. Aan het [project](#) waren twee doctorandi verbonden, Alie Lassche en Theo Dekker, die beiden in 2024 hun proefschrift verdedigd hebben. Het project is geënt op het idee van de revolutie in informatievoorziening en wetenschap die in de vroegmoderne tijd plaats vond. Kuijpers en Pollman stelden de vraag hoe nieuwe kennis zich verspreidde naar en gereciperd werd door de gewone bevolking. Om deze vraag te beantwoorden, hebben zij een corpus van kronieken samengesteld. Enerzijds zijn kronieken een specifiek historiografisch genre, anderzijds waren zij de ideale plaats voor het noteren en bijhouden van observaties uit de eigen tijd en omgeving. Veel vroegmoderne kronieken vervulden zodoende ook een dubbelfunctie als kennisverzameling.

In totaal verzamelde het project maar liefst 204 kronieken in 308 delen uit het hedendaagse Nederland en Noord-België. Geselecteerd werden alle kronieken die geschreven waren op eigen initiatief, niet in opdracht van vorstenhoven of kloosters. Deze handschriften werden gescand via Vele Handen en DBNL, en middels Transkribus ontsloten. Vrijwilligers werden ingezet om modellen te trainen. Voor het doctoraatsonderzoek van Lassche hebben de vrijwilligers aanvullend op detailniveau annotatie toegevoegd aan tekstpassages waarin observaties. Dit bleek een zeer tijdrovende bezigheid: binnen de duur van het project is uiteindelijk ongeveer tweederde van de tekst doorgenomen en geannoteerd.

De onderzoekers hebben een dataset van deze kronieken gecreëerd. Zij konden enkele observaties doen, zoals de aanmerkelijke toename van historiografische activiteit in tijden van crisis. Deze observaties hebben ze gepresenteerd in de publicatie. Het materiaal wordt in de toekomst aan de bewaarders van de handschriften worden verstrekt; de metadata, documentatie en datasets zijn online geplaatst.

2. Prof. dr. **Matthias van Rossum**, *Koloniale archieven, rechtszaken en sociale geschiedenis*

Prof. dr. Matthias van Rossum (IISG & RU) hield een lezing getiteld 'Koloniale archieven, rechtszaken en sociale geschiedenis – enkele mogelijkheden van indexen, datasets en digitale infrastructuur.' Voor bijvoorbeeld het [project](#) GLOBALISE maakt hij gebruik van de criminele procesdossiers van de VOC om de sociale geschiedenis van onderbelichte groepen op wereldschaal te bestuderen. Die dossiers waren al gescand beschikbaar gesteld door het Nationaal Archief, maar de ontsluiting was nog niet verder dan op inventarisniveau. Door te indexeren op basis van bijvoorbeeld persoonsnamen, aanklachten en locaties kunnen de archieven digitaal doorzoekbaar gemaakt worden. In een tweede cluster van data-projecten poogt Van Rossum kennis met betrekking tot de slavenhandel van de VOC in Azië bloot te leggen. Daarbij kunnen bijvoorbeeld de individuele juridische documenten die zich in de procesdossiers bevinden (zoals transportakten, oftewel bewijzen van eigendoms-overdracht) inzicht bieden in de grotere lijnen van de Aziatische slavenhandel: wie wordt aan wie verkocht, en hoe oud zijn deze mensen? Op die manier is het zelfs mogelijk om individuen te volgen door de tijd heen. Doel van de projecten is niet alleen het aanleggen van een uitgebreide dataverzameling, maar ook het opzetten van een digitale infrastructuur waarin de procesdossiers doorzoekbaar zijn, aangevuld met semantische taaltechnologie en historische referentiedata.

3. Prof. dr. **Stephan Dusil**, *De consilia van de juridische faculteit in Tübingen*

Prof. dr. Stephan Dusil deed verslag van zijn onderzoeksproject naar de *consilia* van professoren aan de rechtenfaculteit van de universiteit van Tübingen. Tot het jaar 1879 schreven deze professoren veelvuldig juridische adviezen op verzoek van individuen of rechtbanken. De oudste adviezen zijn bewaard als expedities; vanaf de zeventiende eeuw hield de faculteit zelf registers met afschriften. Daarnaast is een deel in druk uitgegeven. De duizenden adviezen beslaan alle velden van recht, waardoor het een zeer waardevolle seriële bron is voor een veelheid aan disciplines, zoals receptiegeschiedenis, regionale geschiedenis, sociale geschiedenis en filologie.

Dusil heeft als casus het register genomen dat het zomerdekanaat van 1808 beslaat. Dit heeft hij met een groep vrijwilligers getranscribeerd en ondergebracht in een omvangrijke XML-structuur. Het huidige stadium van het project omvat het taggen van verschillende typen entiteiten, zoals persoons- en plaatsnamen, en het produceren van regesten met de meest essentiële informatie.

Op basis van praktische gronden heeft Dusil momenteel gekozen om zich te richten op het publiceren van de data, zonder deze eerst te corrigeren. Middels deze 'vuile' data zou de bron in ieder geval geopend zijn voor onderzoek. Aan deze keuze kleven

evenwel voor- en nadelen. Het gebruik van *Large Language Models*, in casu ChatGPT, heeft tot enigszins succesvolle resultaten geleid, maar niet geheel zonder problemen. Op het moment is Dusil daarom aan het overwegen wat de beste keuzes zijn voor de volgende stadia van het project.

4. Dr. **Rik Hoekstra**, *De webapplicatie Goetgevonden*

De bijdrage van dr. Rik Hoekstra betreft het [project](#) Goetgevonden, waarvan de digitale omgeving in 2024 voor het publiek is opengesteld. In dit project zijn de resoluties van de Staten Generaal. Dit archief is een sleutelbron voor politieke en andere geschiedenis en is in seriële vorm overgeleverd voor bijna tweeënhalve eeuw (1576-1796). Hoewel de oudste jaren vorige eeuw zijn uitgegeven in de Grote Reeks van de RGP, is het corpus zodanig groot dat een manuele aanpak ondoenlijk is: ongeveer 650 banden met in totaal ca. 1 miljoen resoluties. Het voordeel van de seriële aard van het archief is wel dat het vrijwel volledig bestaat uit teksten met een hele duidelijke structuur, zodat gedeeltelijke automatisatie van het procedé een mogelijkheid was.

Voor dit project moest de traditionele structuur van het archief worden doorbroken. In plaats van een fysieke structuur, gedicteerd door de materiële vorm van de registers, was een logische structuur nodig. Dit hield in dat de pagina niet langer als basiseenheid diende, maar de vergadering of individuele resolutie.

Ter voorbereiding heeft het Nationaal Archief alle banden gescand. Hierna zijn ze ontsloten. Stap 1 betrof de transcriptie van de scans. Dit vond eerst plaats met de hulp van vrijwilligers via Transkribus, naderhand via herkenning door Loghi. De tweede stap omvatte het segmenteren van de teksten op basis van zittingsdagen en individuele resoluties, hetgeen geautomatiseerd kon worden door de herkenning van respectievelijk datumformuleringen en formulair taalgebruik, al dan niet met *fuzzy searching* om spellingsvarianten en tekstherkenningsfouten te omzeilen. Als derde stap werden alle entiteiten, zoals persoonsnamen, plaatsnamen en titels, herkend en geïndexeerd.

Met deze kwantitatieve, seriële gegevens bleken de onderzoekers goed in staat om een antwoord te vinden op vragen van meer kwalitatieve aard. Zo waren de gegevens geschikt om bijvoorbeeld veranderingen in de behandeling van verzoekschriften en in de toegang van individuen tot de Staten Generaal te herkennen, de invloed van gewesten op de besluitvorming te identificeren, of een relatie tussen taalstandaardisatie en een groeiende bureaucratie van de Staten te leggen.

5. Prof. dr. **Hylkje de Jong**, *Het HUF-project*

Prof. dr. Hylkje de Jong (VU Amsterdam) hield tot slot een lezing getiteld 'Het HUF-project.' Dit [project](#) beslaat een omvangrijk corpus bestaand uit de civiele

procesdossiers afkomstig van het Hof van Holland, het Hof van Utrecht en het Hof van Friesland tussen 1700 en 1811, en houdt een samenwerking met het Nationaal Archief, het Utrechts Archief en Tresoar in. Een pilotproject is al van start gegaan met civiele rechtszaken die speelden tussen 1716 en 1730 voor het Hof van Friesland. De gekozen periode is geen toeval: in 1723 is namelijk de herziene Landsordonnantie van Friesland afgekondigd, en de geselecteerde rechtszaken bieden dus inzicht in de toepassing van het recht in de periode voor en na de inwerkingtreding van deze ordonnantie. Hoewel aanvankelijk de civiele procesdossiers bewerkt zouden worden, bleek dat de techniek hier onvoldoende op aansluit. Door de vele aantekeningen in de marges leverde het automatisch transcriberen door een computermodel te grote foutmarges op. Om die reden is in plaats van die dossiers gekozen voor de sententieboeken, die zich beter lenen voor het trainen van een computermodel. Het pilotproject dient vervolgens twee doelen. Ten eerste worden op macroniveau de sententies automatisch getranscribeerd met een daarop getraind computermodel en geïndexeerd op bijvoorbeeld persoonsnamen en rechtsbronnen. Daarmee wordt een basis gelegd voor een digitale infrastructuur. Ten tweede worden op microniveau de sententies bestudeerd vanuit bepaalde juridische thema's, zoals bijvoorbeeld het delict krenking (*iniuria*). Hoe ziet de toekomst eruit? Idealiter ontsluit het HUF-project het gehele corpus, op zo een manier dat het doorzoekbaar is, en verrijkt met links naar soortgelijke zaken uit het corpus, alsmede online beschikbare rechtswetenschappelijke literatuur en primair bronnenmateriaal.

Discussie: Vier specialisten in gesprek

Na de lezingen in de ochtend, was het na de lunch tijd voor een paneldiscussie. Vier experts gingen – onder leiding van hoogleraar **Gijs van Dijck** – met elkaar in gesprek. Waar de sprekers focusten op data en output in hun digitaal historisch onderzoek, richtte de discussie zich vooral op de onderliggende techniek.

De deelnemers aan het panel waren onderzoeker **Katrien Depuydt** (Instituut voor de Nederlandse Taal), Transkribus-community director **Annemieke Romein**, KNAW-software engineer **Rutger van Koert** en projectleider **Edwin Klijn** (beiden actief voor het CABR – Centraal Archief Bijzondere Rechtspleging). Het panel bestond kortom vooral uit mensen uit het werkveld die hun ervaringen met technische hulpmiddelen deelden.

Klijn trapte af. Op de vraag van Van Dijck: “Wat zijn de mogelijkheden van data en techniek?”, antwoordde hij dat data de ‘grondstof’ vormt voor gedegen digitaal onderzoek. Volgens hem is een goede, gestandaardiseerde datastructuur daarvoor essentieel. Hij kreeg bijval van Depuydt en Van Koert. “Met technologie kun je bovendien dingen ontdekken die vroeger verborgen bleven,” voegde Van Koert toe.

Verbinden

Van Koert werkte eerder als techneut in de commerciële sector, maar richt zich nu op softwareontwikkeling voor de analyse van historische, handgeschreven documenten (zoals [Loghi](#)). Ook Romein houdt zich daarmee bezig als [Transkribus](#)-adept. “Eigenlijk ben ik een beetje lui,” grapte ze. “Ik heb geen zin om teksten handmatig door te ploegen. Elke tool die helpt bij transcriberen is voor mij dus welkom.”

Net als Loghi maakt Transkribus gebruik van *Handwritten Text Recognition* (HTR)-software om oude documenten te ontcijferen. Romein vertelde dat het nieuwste model – ‘Dutch Dean’ – is getraind op maar liefst 25 miljoen woorden. Toch is inzicht in andere tools en hun toepassingen belangrijk, benadrukte ze. “Techniek werkt verbindend. Het is mooi dat op dit congres ideeën en tools uit verschillende disciplines samenkomen.”

“Maar hoe begin je met dit soort *tech tools*?”, vroeg Van Dijck vervolgens. “Als je materiaal uploadt – zoals gescande documenten – is het belangrijk dat je begrijpt wat je hebt en bedenkt wat je eruit wilt halen,” aldus Romein. “Dan kun je de verschillende handschriften herkennen en de lay-out inrichten voor optimale tekstherkenning.”

Van inhoud naar techniek

Ook Klijn benadrukte dat je altijd het doel voor ogen moet houden bij technische keuzes. “Het CABR is een gigantisch rechtbankarchief over mensen die tijdens de Tweede Wereldoorlog onderzocht zijn op collaboratie. We bouwen [een website](#) om dit archief toegankelijk te maken voor een breed publiek. Dan is het cruciaal dat we techniek inzetten op basis van de vraag: wat moet de gebruiker kunnen?”

Volgens Klijn zijn de eindgebruikers vaak nabestaanden. “Voor hen is het belangrijk om houvast te bieden in de wirwar van gedigitaliseerde documenten. Zoektermen moeten direct leiden naar relevante stukken, waarin de resultaten zijn te vinden. Daarnaast hebben we de rechtsgang gereconstrueerd, en de computer geleerd om sleuteldocumenten – zoals besluiten en vonnissen – daaraan op te hangen.”

Van Koert was nauw betrokken bij de ontwikkeling van de CABR-website, die zo’n anderhalf jaar in beslag nam. “We hebben ervoor gekozen om documenttypen automatisch te laten herkennen via *text similarity search*. Daarmee werden gelijksoortige stukken – zoals dagvaarding – automatisch gegroepeerd en daarna geclassificeerd.”

KISS

Ook Katrien Depuydt hanteert het principe ‘bezint eer ge begint’ in haar werk. Ze is verantwoordelijk voor de bouw en ontsluiting van digitale corpora en [\(historische\) lexica](#) bij het Instituut voor de Nederlandse Taal. “Je moet je data-infrastructuur zó ontwerpen dat je er later op kunt voortbouwen,” zei ze. “Maak datasets flexibel en toekomstbestendig zodat datamodellering achteraf mogelijk blijft.”

Toen Van Dijck de panelleden vroeg naar valkuilen, sloot Depuydt hier direct op aan. “Mensen willen vaak snel resultaat, maar je moet niet te vroeg beginnen. Stel eerst de vraag: wat wil ik eigenlijk met mijn data? Als je dat niet doet, eindig je onvermijdelijk met veel nabewerking.”

Tot slot benadrukte ook Romein nog eens dat goede planning essentieel is. “*Keep it simple and straightforward,*” sloot ze af. “Ken je gebruikers, praat met ze, en test voortdurend. Sluit de output nog aan op de doelen? Een tachtigjarige die een dagboek wil digitaliseren heeft immers heel andere wensen dan een professioneel archief.”