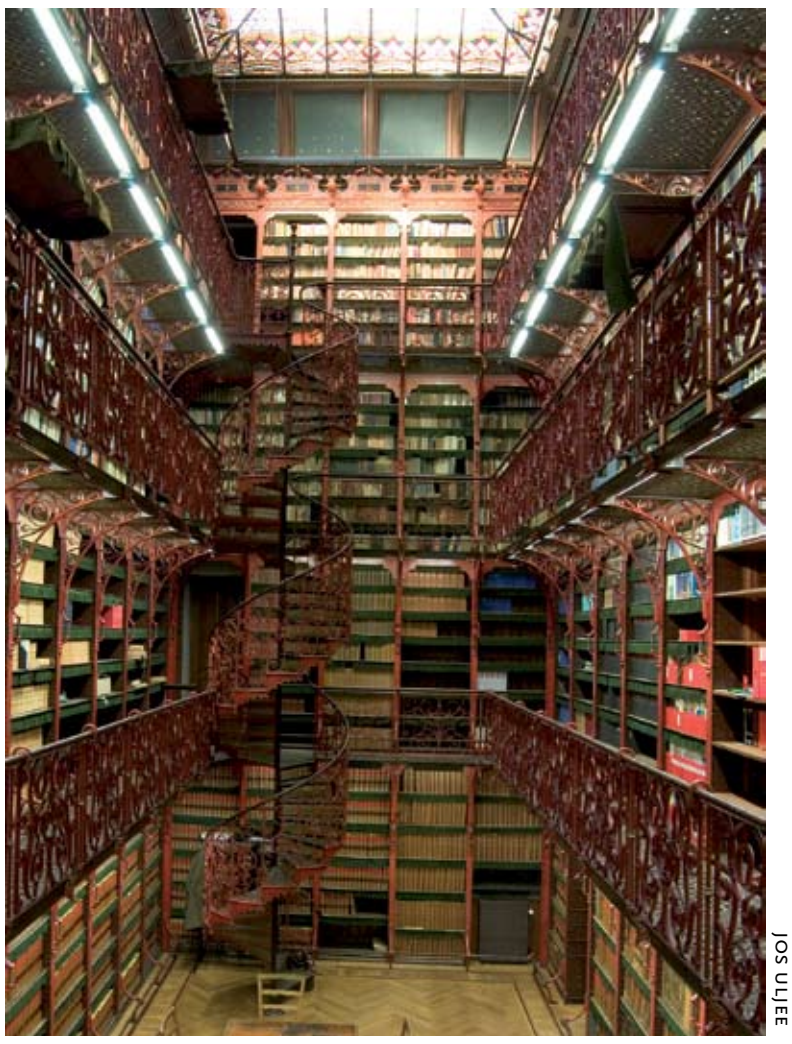
Ook registers van Handelingen ontsloten

Massadigitalisering is monnikenwerk

HUIBERT CRIJNS EN TINEKE KOSTER

Na de lancering van de website www.statengeneraaldigitaal.nl op 8 juni is het project voor de digitalisering van de Handelingen van de Staten-Generaal in volle gang. In juni waren nog alleen de jaargangen 1990-1995 in te zien en te doorzoeken. Maar in 2010 zullen de verslagen en vergaderstukken van de Eerste en de Tweede Kamer uit de hele periode van 1814 tot 1995 (2,3 miljoen pagina's) gedigitaliseerd en online zijn.



JOS ULJEE

De handelingenkamer in het gebouw van de Tweede Kamer

In het project Staten-Generaal Digitaal 1814-1995 werken de Tweede Kamer en de Koninklijke Bibliotheek samen. Doel is het conserveren en digitaal ontsluiten van de Handelingen. Behalve omvangrijk zijn de Handelingen ook complex en moeilijk toegankelijk, zelfs voor ervaren onderzoekers. Digitalisering maakt het mogelijk om via een website rechtstreeks in het materiaal te zoeken op onder andere vrije zoekterm, naam, jaartal, datum en documenttype.

Afgelopen zomer is een begin gemaakt met het ontsluiten van de registers van de Handelingen op de website. In de periode 1814-1979 zijn bij de meeste jaargangen registers verschenen. Vanaf 1979 hebben beide Kamers een geautomatiseerd zoeksysteem ontwikkeld. De registers bestaan uit alfabetische lijsten van begrippen (de zaakregisters) en personen (de persoonsregisters) met daarachter verwijzingen naar relevante pagina's in de Handelingen.

De registerpagina's worden eerst als afbeelding ingescand. Het tonen van een afbeelding alleen geeft echter geen meerwaarde boven het origineel. Het wordt pas interessant als je via het register direct door kunt linken naar de passage in de Handelingen die je zoekt. Daartoe moet het origineel of de afbeelding omgezet worden naar computerleesbare tekst. De huidige OCR-software voor automische tekenherkenning kan niet goed met historische lettertypes en taal omgaan en laat veel steken vallen, terwijl in de beknopte stijl van de registers elke letter, elk cijfer en elk leesteken telt. Volledig overtypen of handmatige correctie zal daarom voor de digitale ontsluiting noodzakelijk zijn. Bovendien is de opzet, lay-out en de methode van verwijzen niet consistent over de 165 jaar die de registers beslaan. Dat betekent dat deze informatie niet geautomatiseerd verwerkt kan worden.

Onder één lemma vallen vaak tientallen pagina's met verwijzingen, die vooral bestaan uit afkortingen gevolgd door nummers van documenten, pagina's en data. Het zijn eindeloze, bijna abstracte opsommingen van letters, Romeinse en Arabische cijfers en leestekens. Om links te kunnen creëren moeten alle verwijzingen handmatig in een database ingevoerd worden. En dat is geen domme data-invoer: de datatypisten moet afdoende geïnstrueerd worden zodat zij begrijpen waar zij welke informatie moeten invoeren. De kleinste vergissing zorgt voor een niet werkende link.

Al dit werk gaat resulteren in een aparte zoekpagina op de website www.statengeneraaldigitaal.nl. De gebruiker kan kiezen uit een lijst van alle lemma's en deze in een zoekvraag combineren met een jaar of een periode. Als resultaat krijg je een lijst met alle vermeldingen van dat lemma in de registers van de betreffende jaren. Je kunt elke vermelding aanklikken, waarna je alle verwijzingen op je scherm krijgt naar passages in de Handelingen die op dit lemma betrekking hebben. Van elke verwijzing kun je rechtstreeks naar de desbetreffende pagina doorlinken. Door heen en weer te navigeren tussen de lijst met verwijzingen en de documenten is het mogelijk om in relatief zeer korte tijd met behulp van de registers de Handelingen grondig te doorzoeken.

Aan de digitale ontsluiting van historische bronnen gaat dus een enorme hoeveelheid monnikenwerk vooraf. Het spanningsveld tussen kwaliteit (nauwkeurigheid en volledigheid) en kwantiteit (2,3 miljoen pagina's in totaal, 17 duizend pagina's registers) is hierin een voortdurend aandachtspunt.

Cohortonderzoek brengt nieuwe golf van kansen voor wetenschap

MARTIJN DE GROOT

Bijna vijfhonderd publicaties kwamen er voort uit PRIMA, een van de twee onderzoeken die vooraf gingen aan het deze zomer gelanceerde project COOL 5-18. PRIMA en het VOCL waren cohortonderzoeken voor respectievelijk het lager en voorgezet onderwijs. COOL omvat beide. De eerste gegevens worden in het nieuwe schooljaar binnengehaald. De eerste resultaten komen eind volgend jaar beschikbaar.

De indrukwekkende oogst van de voorgaande onderzoeken zal zeker een rol hebben gespeeld bij de beslissing van NWO om ook voor COOL 5-18 middelen uit te trekken, denkt dr. Lia Mulder, projectleider van het basisschoolgedeelte. 'Tal van onderzoek- en beleidsinstellingen hebben van de eerdere gegevens gebruik gemaakt en dat is ook nu weer de verwachting'.

COOL is een grootschalig cohortonderzoek: van veel mensen worden tegelijkertijd gegevens verzameld over een bepaalde tijdsspanne. In eerste instantie worden drie keer 70 duizend leerlingen van 5 tot 18 jaar in hun schoolloopbaan gevolgd door het primair-, voortgezet- en middelbaar beroepsonderwijs (mbo). In de tweede en derde ronde komen daar nog eens twintig duizend scholieren bij, waarbij speciaal wordt gelet op de cognitieve ontwikkeling (kennis en vaardigheden), de ontwikkeling van burgerschapscompetenties en de sociaal-emotionele ontwikkeling (persoonlijkheid, motivatie en betrokkenheid).

Ook tal van andere aspecten en invalshoeken kunnen in het onderzoek worden betrokken. Volgens de onderzoekers gaat het om een 'basis dataverzameling', waaraan specifieke data kunnen worden toegevoegd op aanvraag van belangstellende onderzoekers met eigen onderzoeksvragen. Deze zogenaamde moduleregeling wordt nog belangrijker dan in de eerdere onderzoeken, legt Mulder uit: 'Toen probeerden we zoveel mogelijk onderzoekvragen alvast een plek te geven in de algemene dataverzameling. Maar vaak zagen we achteraf dat de gegevens die dat had opgeleverd nèt niet beantwoordden aan de informatiebehoefte van de onderzoekers die zich meldden. Dan moest je toch weer met aanvullende vragenlijsten terug naar leerlingen of schoolleidingen, terwijl die in het eerdere onderzoek al behoorlijk belast werden. Om zeker te weten dat we alle vragen nu aan kunnen

NATIONALE BEELDBANK

zonder dat het teveel op de scholen drukt hebben we nu de basis dataverzameling versmald, met dus die moduleregeling als aanvulling'.

Een kleinere vragenlijst zou ook een kleiner budget kunnen betekenen maar daarvan blijkt geen sprake. Mulder: 'In theorie zou je daardoor minder kosten kunnen hebben, maar daar staat weer een andere verbetering ten opzichte van PRIMA en VOCL tegenover die behoorlijk arbeidsintensief is. We gaan nu ook de leerlingen volgen die uit het cohort vallen om een of andere reden, zoals verhuizen, zittenblijven of verwijzing naar het speciaal onderwijs'. Voor het eerst kunnen nu namelijk de onderwijsnummers in het onderzoek worden betrokken, waardoor het mogelijk wordt de onderwijsloopbaan te volgen ook van leerlingen die niet meer op de scholen zitten die in het onderzoek zijn betrokken.

Bij de eigen gegevensverzameling wordt ernaar gestreefd in elke nieuwe ronde steeds dezelfde leerlingen in beeld te houden. Maar na de basisschool verspreiden leerlingen zich over veel verschillende scholen en opleidingen. Om efficiënt te kunnen werken en om ook aan de betrokken scholen bruikbare resultaten te kunnen bieden zal een keus worden gemaakt uit de scholen voor voortgezet onderwijs waar de voormalige basisschoolleerlingen naar toe zijn gegaan. 'Dan gaan we door met die scholen waar een voldoende aantal leerlingen uit de vorige ronde, zeg bijvoorbeeld twintig, naar toe is gegaan. En rond die groepen nemen we dan de hele klas in de nieuwe ronde mee'.

Mulder zelf is projectleider vanuit het Instituut voor Toegepaste Sociologie (ITS) van de Radboud-universiteit, een van de deelnemers in het zogenaamde consortium dat COOL 5-18 tot een goed einde moet zien te brengen. Andere leden van dat consortium zijn het Kohnstamm Instituut dat samen met ITS de basisscholen voor zijn rekening neemt, toetsenproducent CITO en het Gronings Instituut voor Onderzoek van Onderwijs die samen het voortgezet onderwijs onder hun hoede hebben, en het CBS dat voor de onderwijsnummers zorgt. Datainstituut DANS zorgt voor de beschikbaarstelling van de resultaten.

www.cool5-18.nl