Naar een Antroponymisch Woordenboek van de historische Nederlanden (tot 1225)

1. De aanloop tot het project

1.1. De aanleiding

In 1963 formuleerde de grote filoloog Maurits Gysseling (1919-1997) drie doelstellingen voor de historische neerlandistiek :

«*Naast de publicatie van het Toponymisch Woordenboek van de historische Nederlanden ontbreken nog twee andere belangrijke materiaalverzamelingen voor de studie van de vroege geschiedenis van het Nederlands : ten eerste de publicatie van alle Nederlandse teksten tot 1300; ten tweede de publicatie van een persoonsnamenboek als pendant van het toponymisch woordenboek.*»[1]

Gysseling had de neerlandistiek al drie jaar eerder, in 1960, zijn *Toponymisch Woordenboek* als eerste van deze drie standaardwerken geschonken. Het tweede, het *Corpus Middelnederlandse Teksten* verscheen tussen 1977 en 1987. Beide werken gelden sindsdien als monumenten van de neerlandistiek.

Gysselings derde onderzoeksdesideratum, het persoonsnamenboek, heeft hij echter niet gerealiseerd. Het was met zijn eigen woorden «*een zware taak* [...], *waarvan de voltooiing zeker niet voor morgen is.*»[2] Gysseling had voor dat persoonsnamenboek nochtans in de jaren 1940-50, tegelijk met zijn werk voor het Toponymisch Woordenboek, al het nodige archief- en excerpeerwerk verricht. Het vele malen omvangrijkere en meer heterogene persoonsnamenmateriaal heeft hij echter niet meer kunnen ordenen, laat staan tot een woordenboek bewerken. Een paar weken vóór zijn overlijden, in 1997, heeft Gysseling in zijn wils-

[1] *Proceedings of the 8th International Congress of Onomastic Sciences* (1963), Amsterdam-Den Haag, 1966 : 220-223 (vertaling uit het Duits).
[2] Ibidem.

beschikking dan de wens uitgedrukt dat de tweede auteur van dit artikel het materiaal verder zou ontsluiten. Andere bezigheden verhinderden voorlopig dat de 'executeur-testamentair' meteen de hand aan de ploeg kon slaan. Pas toen de voltooiing van het digitale *Toponymisch Woordenboek* (www.wulfila.be/tw) in zicht kwam en ervaring was opgedaan met databanken, rijpte het plan om nu ook Gysselings persoonsnamenmateriaal te bewerken tot een *Antroponymisch Woordenboek*. Een eerste aanvraag tot financiële steun bij het FWO werd in 2006 afgewezen. Een tweede poging, in 2007, had wel succes, nadat blijkbaar was ingezien dat het project geen zoveelste *data crunching* beoogde, maar de samenstelling van een onderzoeksinstrument van grote duurzaamheid.[3] Desondanks duurde het nog bijna anderhalf jaar vóór een onderzoeksmedewerker kon worden gevonden en het project kon starten. Het contrasteert wel opvallend met de situatie van nog geen twintig jaar geleden toen voor een gelijkaardig digitaliseringsproject moeiteloos medewerkers konden worden gevonden. Wellicht getuigt het van een kenniscrisis in de geesteswetenschappen dat steeds minder jonge onderzoekers beschikbaar zijn met voldoende onderlegdheid in uiteenlopende gebieden als mediëvistiek, paleografie, Latijn en de digitale letteren.

In wat volgt, wordt ingegaan op de doelstellingen van het project en de mogelijkheden die het zal bieden voor vernieuwend onderzoek van de vroege geschiedenis van het Nederlands en de Lage Landen in het algemeen. In het tweede deel wordt de methodologie voorgesteld die bij de digitalisering van Gysselings antroponymisch materiaal wordt aangewend.

1.2. De voorgeschiedenis van het project en zijn internationale context

De vraag naar een antroponymisch woordenboek van het «Oudduits» als onderzoeksdesideratum (inclusief het Oudnederlands) dateert reeds van 1846. De grondlegger van de Germaan-

[3] Copromotores van het project zijn G. De Schutter (UA en KANTL), Th. de Hemptinne en R. Van Deyck (beiden UGent en Einhard-Instituut).

se taalkunde, Jakob Grimm (1785-1863), liet het toen als onderzoeksproject door de Berlijnse Academie uitschrijven. Het project werd uitgevoerd door een schare internationale medewerkers (voor België was dat o.m. de Duitse Antwerpenaar Heinrich Pottmeyer) en gecoördineerd door Ernst Förstemann (1822-1906). Het resulteerde in de publicatie van het eerste deel van het *Altdeutsches Namenbuch,* de persoonsnamen, in 1856. In het uiteindelijk driedelige werk, waarvan de tweede en derde band de plaatsnamen tot 1200 bevatten, waren volgens de toenmalige opvattingen ook de Nederlanden opgenomen. Förstemann publiceerde zelf nog een aanzienlijk verbeterde herwerking van het deel persoonsnamen in 1900, terwijl H. Jellinghaus een (niet-geslaagde) herwerking van de delen met de plaatsnamen bezorgde in 1913 en 1916. Het woordenboek werd nadien nooit meer uitgegeven. Nochtans bestond daaraan grote behoefte zoals een anastatische herdruk van het driedelige werk in 1966 en de *Ergänzungsband* van Henning Kaufmann in 1968 bewezen.

Omdat een opvolger van Förstemanns *Namenbuch* beschouwd werd als «*ein dringendes, eigentlich völlig unabweisbares Desiderat*» (Haubrichs 1997 : 194), werd in de jaren 1960 met de steun van de Deutsche Forschungsgemeinschaft een grootschalig project opgezet, de zogenaamde «Neue Förstemann». Er werden in de diverse Duitse deelstaten, in Oostenrijk, Zwitserland, Luxemburg, Frankrijk (Elzas), Nederland en België centra opgericht, die gecoördineerd werden vanuit Freiburg (zie daarover Boesch 1971). Het project moest echter bij het begin van de jaren 1980 jammerlijk worden afgebroken, onder meer door een gebrek aan coördinatie en homogeniteit (voor de historiek zie o.a. Sonderegger 1997 en Geuenich 2002 : 83 vv).

Niet lang nadien ontstonden, maar nu los van elkaar, in diverse landen opnieuw initiatieven voor de samenstelling van historische antroponymische woordenboeken. Het waren niet langer taal- of naamkundigen die daarbij het voortouw namen, maar onderzoekers met een ruime mediëvistische achtergrond. De louter naamkundige benadering werd inhoudelijk verruimd naar thema's als prosopografie, groepsbinding en etniciteit tijdens de Vroege Middeleeuwen. Diverse nationale onderzoeksinstellingen

stelden ruime financiële middelen ter beschikking. In Oostenrijk kwam met de steun van het Fonds zur Förderung der wissenschaftlichen Forschung van 1973 tot 1987 de *Thesaurus Palaeogermanicus* tot stand (o.l.v. H. Reichert). In Duitsland financierde de Deutsche Forschungsgemeinschaft het project *Nomen et Gens* (o.l.v. D. Geuenich, W. Haubrichs e.a.). In Frankrijk startte met steun van het CNRS het project *Genèse médiévale de l'anthroponymie moderne*, waaraan verscheidene onderzoeksmedewerkers verbonden zijn (o.l.v. P. Beck en M. Bourin). Het verst gevorderd is het imposante Engelse project *Prosopography of Anglo-Saxon England* (*PASE)* o.l.v. J.L. Nelson, dat gefinancierd wordt door the Arts and Humanities Research Board/Council (AHRB) (www.pase.ac.uk). Het succes van dit grote databankproject is te danken aan de nauwe en goed gecoördineerde interdisciplinaire samenwerking tussen historisch-taalkundigen, mediëvisten en specialisten in *humanities computing*.

De promotoren van het Duitse project *Nomen et Gens*, D. Geuenich (Duisburg/Essen), W. Haubrichs (Saarbrücken), J. Jarnut (Paderborn) namen een tiental jaren geleden het gelukkige initiatief de afzonderlijk werkende nationale projecten in een internationale onderzoeksgemeenschap samen te brengen. Onder hun impuls werden tot hiertoe een vijftal internationale congressen georganiseerd. De onderzoeksresultaten die er werden gepresenteerd, verschenen onder meer in in drie *Ergänzungsbände* bij het prestigieuze *Reallexikon der germanischen Altertumskunde* (o.a. Geuenich, Haubrichs, Jarnut 1997, 2002, 2004).

Door de recent verkregen financiering van het FWO heeft nu ook het naamkundig onderzoek van bij ons opnieuw aansluiting kunnen vinden bij deze Europese onderzoeksgemeenschap. De tweede auteur van dit artikel gaf in 2007 te Magdeburg en in 2009 te Wenen, respectievelijk op uitnodiging van *Nomen et Gens* en van het Institut für Mittelalterforschung van de Oostenrijkse Academie, lezingen over de mogelijkheden van het antroponymisch onderzoek voor de prosopografie van de Oudheid en de vroege Middeleeuwen in de Lage Landen (cfr Infra, sub 3).

Hij kon er bij die gelegenheden tevens het nu gestarte project voorstellen.

Een pilootversie van het Antroponymisch Woordenboek van de historische Nederlanden zou volgens de planning in de loop van 2012 opgeleverd moeten worden. Dat streefdoel is realistisch. Het project kan immers voortbouwen op het omvangrijke materiaal dat in de jaren 1940-1950 door M. Gysseling werd verzameld. Er is ondertussen ook de nodige ervaring opgedaan in het opstellen van databanken bij vorige projecten zoals de digitale editie en databank van Gysselings *Toponymisch Woordenboek* en de editie van de Gotische Bijbel. [4] Het project beschikt bovendien over de specifieke expertise van de copromotores. De Koninklijke Academie voor Nederlandse Taal- en Letterkunde (KANTL) heeft een grote deskundigheid in huis zowel op het gebied van de historische Nederlandse taalkunde als op het gebied van het elektronisch editeren in het Centrum voor Teksteditie en Bronnenstudie (CTB). Voor de mediëvistische aspecten is er de inbreng van de Vakgroep Middeleeuwse Geschiedenis en het Einhard-Instituut van de Universiteit Gent, dat interdisciplinaire deskundigheid op het gebied van de historische en taalkundige mediëvistiek van de Lage Landen verenigt.

1.3. Het materiaal

M. Gysseling had een exhaustief persoonsnamenboek voor ogen van de (historische) Lage Landen en de aangrenzende gebiedsdelen tijdens de Vroege Middeleeuwen. In principe bevat zijn materiaal alle persoonsvermeldingen van vóór 1100 en alle persoonsnamen van vóór 1226 uit de archieven van België, Nederland, Luxemburg, de Franse departementen Nord en Pas-de-Calais, en de voormalige Rijnprovincie. [5] Hij nam alle attestaties op die voorkomen in originelen tot 1225 en alle attestaties in niet als origineel bewaarde teksten van vóór 1100. De

[4] De databank is voorlopig ondergebracht op http ://www.wulfila.be, waar ook de elektronische editie van de Gotische bijbel te vinden is (ed. Tom J. De Herdt).

[5] Een groot deel van noordwestelijk Duitsland, dat in taalkundig opzicht nauw verwant is met het Nederlands, is aldus ondervertegenwoordigd. Het was een van de voornaamste punten van kritiek die door recensenten op het Toponymisch Woordenboek van Gysseling zijn geleverd (Van Loon 2008 : 9).

afbakening van zijn materiaal is dus grotendeels identiek aan die van zijn *Toponymisch Woordenboek*, al zijn er enkele afwijkingen. Voor de periode 1100-1225 werden van zeer frequente voornamen, die vanaf dan in zo'n massale hoeveelheid in de bronnen voorkomen dat ze meer dan de helft van de hele voornaamgeving voor zich gaan opeisen (namen als Johannes, Gerardus, Walterus, Willelmus ...), enkel nog die opgenomen die in originelen voorkomen. Gysseling zag zich die beperking opgelegd door de omvang van het materiaal, dat vele keren omvangrijker is dan dat van het *Toponymisch Woordenboek*.

Het meesterschap van Gysseling staat borg voor de paleografische en filologische accuratesse van de transcripties en voor de homogene kwaliteit van het materiaal. Het woordenboek zal zich daardoor onderscheiden van vergelijkbare werken die precies in die opzichten tekortschieten (zie daarover o.m. Geuenich e.a. 1997 : V). Dat geldt in de eerste plaats uiteraard voor de nu totaal verouderde werken van Förstemann of Searle, maar ook voor het recentere woordenboek van M.-Th. Morlet (1971-85). Dit werk is bij gebrek aan enig ander compilatiewerk nuttig bij een eerste terreinverkenning, maar is onvolledig en niet gebaseerd op betrouwbare uitgaven.

Gysseling heeft het derde standaardwerk dat hij zich voor ogen had gesteld, weliswaar nooit gerealiseerd, maar uit de nalatenschap blijkt dat hij er ooit aan is begonnen. Bij zijn archiefbezoeken in de jaren veertig en vijftig transcribeerde hij alle plaats- en persoonsnamen en ook andere randinformatie op velletjes papier in A5-formaat (Afbeelding 1). Op grond van deze velletjes werden de steekkaarten gemaakt voor het *Toponymisch Woordenboek*.[6] Op dezelfde manier maakte hij ook een begin van een steekkaartenbestand voor zijn antroponymisch woordenboek. Die steekkaarten zijn van tweeërlei soort. De met de hand geschreven steekkaarten bevatten meestal excerpten uit bestaande uitgaven waarvoor collationering met het origineel

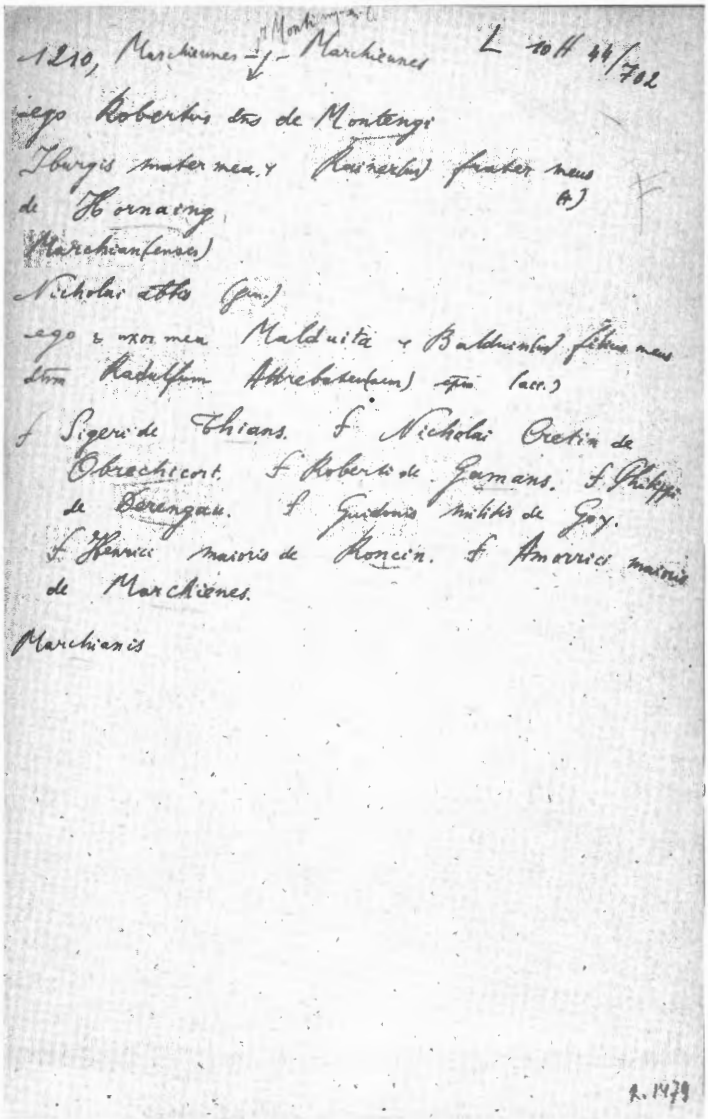
[6] Gysselings handschriftelijk materiaal werd in 1997 door zijn weduwe overgemaakt aan L. Van Durme en J. Van Loon. Het grootste deel ervan werd in de archieven van de Koninklijke Academie in Gent gedeponeerd. L. Van Durme heeft het materiaal geordend en een getypte inventaris gemaakt. Het is mede aan zijn inzet te danken dat grote delen uit de nalatenschap van M. Gysseling niet veloren zijn gegaan.

niet mogelijk was of door Gysseling niet nodig werd geacht. Het betreft edities zoals de *Diplomata Belgica*, die door hemzelf in 1950 samen met Koch waren uitgegeven, of de *Werdener Urbare* door Kötzschke. Voor de namen uit de Oudheid en de Vroege Middeleeuwen ging Gysseling voort op de grote klassieke edities of op publicaties van inscripties zoals Brigitte en Hartmut Galsterers *Römische Steinschriften* (1975). Naast de geschreven fiches zijn er een paar duizenden getypte steekkaarten. Ze zijn recenter dan de eerste en zijn meestal, maar niet altijd, gebaseerd op de handgeschreven A5-velletjes. Het gehele steekkaartenbestand was alfabetisch gerangschikt, met voorafgaand aan elke groep identieke namen een lemmasteekkaart waarop de genormaliseerde naamvorm, zijn etymologie en eventueel nog andere gegevens (Afbeelding 3). In totaal zijn zo'n zevenduizend steekkaarten gemaakt of althans teruggevonden. Dat is erg weinig als men weet dat het totale materiaal voor het Antroponymisch Woordenboek op zijn minst zo'n 200.000 attestaties bevat (zie infra).

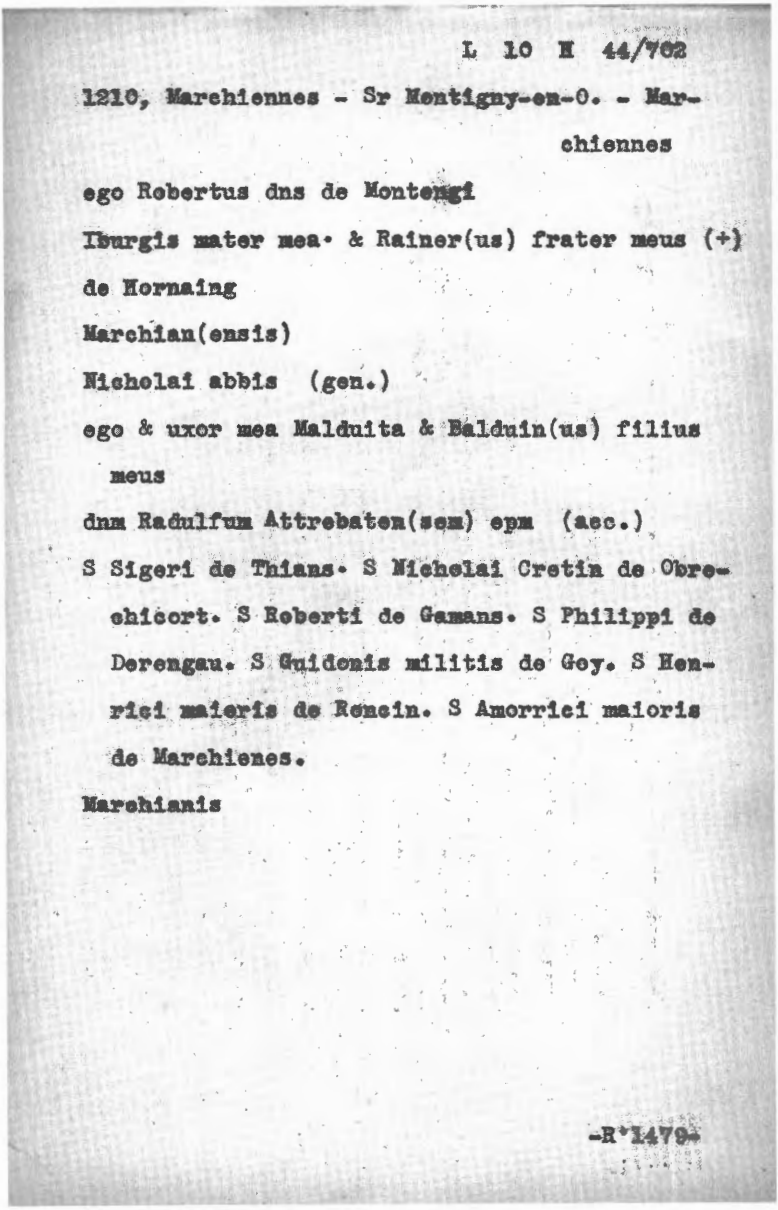
De onoverzienbare hoeveelheid materiaal betekende bij de planning van het project een schier onoverkomelijke hinderpaal. Gysselings transcripten zouden immers eerst nog manueel van de handgeschreven excerpten op moderne informatiedragers moeten worden overgebracht. Intussentijd waren echter aanwijzingen opgedoken dat Gysselings transcripten ooit moesten zijn overgetikt. Wijlen Cl. Marynissen maakt in het voorwoord van zijn proefschrift allusie op een kopie die in Leuven van het toponymisch en antroponymisch materiaal-Gysseling is gemaakt (Marynissen 1986 : 10). De zoektocht naar dit gekopieerde materiaal is een verhaal op zich want ei zo na had de tijd alle sporen ernaar uitgewist.[7] Bij schriftelijke navraag bij em. prof.

[7] Het is niet de eerste keer dat belangrijke naamkundige collecties verloren gaan. In de jaren 1970 werd als gevolg van een administratieve vergissing op het Ministerie van Onderwijs de volledige voorraad oplagerestanten van het Toponymisch Woordenboek door een archiefvernietigingsdienst opgeruimd. Iets gelijkaardigs overkwam het zogeheten Corpus Molemans-Thiry. Deze verzameling historische naamattestaties, die bijna heel Vlaanderen besloeg en waaraan jarenlang was gewerkt, was opgeslagen op magneetbanden. Als gevolg van interne reorganisaties en verhuizingen gingen die reddeloos verloren. We vermelden deze *petite histoire* omdat ze nog maar eens laat zien welke desastreuze gevolgen hervormingswoede en afbreken van tradities kunnen hebben.

K. Roelandts (jaargang 1919) bevestigde deze dat M. Gysseling in de jaren 1960-70 zijn handgeschreven materiaal in bruikleen had afgestaan aan hemzelf en professor Henri Draye, met wie hij via de Academie goed bevriend was, en dat het aan het toenmalige Instituut voor Naamkunde in Leuven was overgetypt. Bij een eerste navraag, in 2006, bleek echter niemand uit de omgeving van M. Gysseling in de Letterenfaculteit te Leuven of in de Koninklijke Academie te Gent weet te hebben van dergelijke transcripten, laat staan van de plaats waar ze zich bevonden. Een zoektocht in de bergruimtes van het Erasmusgebouw in Leuven in 2006 leverde geen resultaat op. Een jaar later echter, toen alle hoop al was opgegeven om het materiaal ooit terug te vinden, leverde een toevallige ontmoeting met A. De Koninck, voormalige rechterhand van wijlen prof. Draye en nog net niet geprepensioneerd, wel succes op. Deze welhaast laatste getuige wist zich nog goed te herinneren dat hij Gysselings materiaal bij de verhuizing van het statige herenhuis in de Blijde-Inkomstraat naar het Erasmusgebouw, nu bijna veertig jaar geleden, eigenhandig had verpakt en het in de kelderruimtes van het gebouw had gedeponeerd. Op zijn aangeven kwam van onder stapels rommel en stof van veertig jaar een grote doos te voorschijn met daarin de verloren gewaande typoscripten. Ofschoon het niet volstrekt zeker was dat hiermee het volledige persoonsnaamkundige materiaal van M. Gysseling was teruggevonden, was de hoeveelheid materiaal zo groot dat het de grondslag kon vormen voor een antroponymisch woordenboek.



AFBEELDING 1 – *Voorbeeld van een transcript in het handschrift van M. Gysseling*



L 10 H 44/702

1210, Marchiennes - Sr Montigny-en-O. - Marchiennes

ego Robertus dns de Montengi

Iburgis mater mea· & Rainer(us) frater meus (+)

de Hornaing

Marchian(ensis)

Nicholai abbis (gen.)

ego & uxor mea Malduita & Balduin(us) filius meus

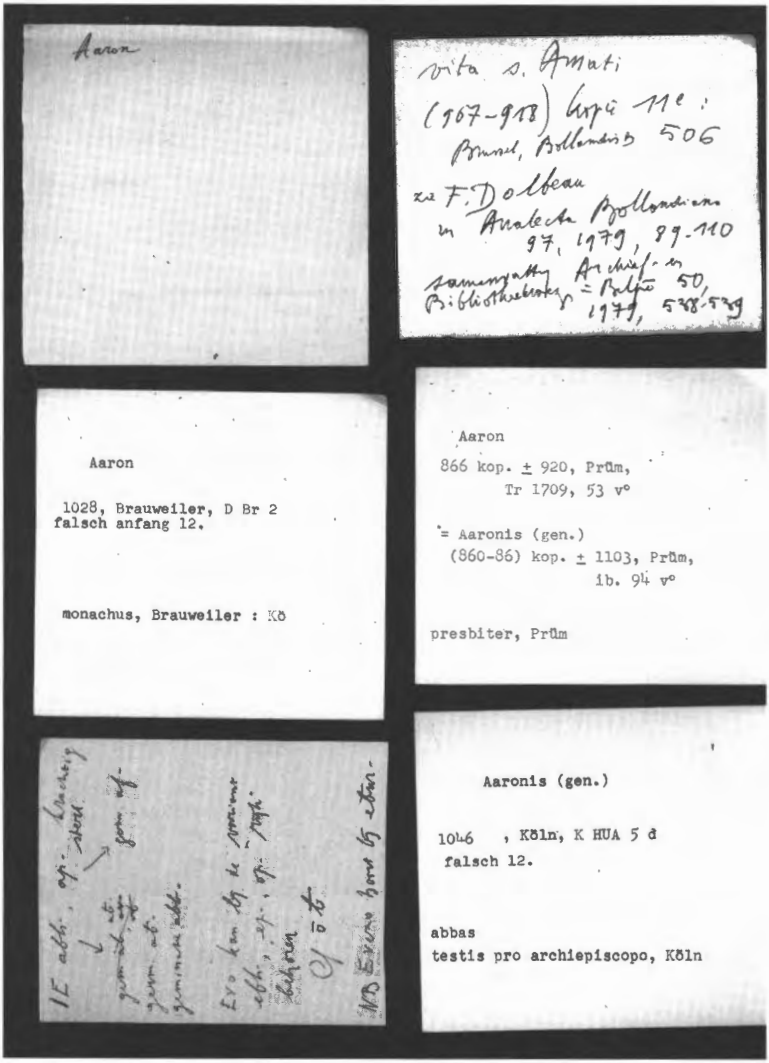
dnm Radulfum Attrebaten(sem) epm (acc.)

S Sigeri de Thians· S Nicholai Cretin de Obrechicort. S Roberti de Gamans. S Philippi de Derengau. S Guidonis militis de Goy. S Henrici maioris de Roncin. S Amorrici maioris de Marchienes.

Marchianis

-R°1479-

AFBEELDING 2 – *Voorbeeld van een typoscript uit de jaren 1960*



AFBEELDING 3 – *Voorbeeld van een lemmafiche in het handschrift van M. Gysseling*

De begin 2007 teruggevonden typoscripten werden in de zomermaanden van 2007 op elektronische drager ingelezen.[8] De weergavekwaliteit van deze digitale facsimile's was, ondanks de sterk wisselende inktcontrasten van de getypte velletjes, erg bevredigend : bij een eerste proef bleek al dat de facsimile's vrijwel foutloos door courante OCR-software (*Optical Character Recognition*) werden gelezen.

Dat neemt niet weg dat de gedigitaliseerde output nog gecollationeerd zal moeten worden met de typoscripten en met de handgeschreven transcripten van Gysseling. Controlesteekproeven wijzen uit dat de typoscripten van zeer betrouwbare kwaliteit zijn. Uit randnotities in potlood blijkt trouwens dat ze na het overtypen al met Gysselings transcripten zijn gecollationeerd. Wel is te zien dat de revisors typisch paleografische afkortingstekens in Gysselings handschrift (bovenliggende streepjes, doorhalingen e.d.) niet hebben herkend. Deze tekortkomingen zijn echter gemakkelijk automatisch recht te zetten via een herkenningsprogramma. Deze en andere collationeringen zullen echter niet meteen worden uitgevoerd maar pas nadat de hele databank is afgewerkt, met behulp van een *user interface* die toelaat efficiënt en eenduidig aanvullingen en correcties door te voeren.

Zoals al vermeld, is nog niet gecontroleerd of Gysselings transcripten allemaal zijn overgetypt. Waar dat niet gebeurd zou zijn, moeten ze nog manueel worden ingevoerd. Dat laatste zal zeker nodig zijn voor de met de hand geschreven steekkaarten (cfr. supra), die Gysselings excerpten uit bestaande edities bevatten. Daarentegen zullen Gysselings lemmasteekkaarten met genormaliseerde persoonsnaamvormen en etymologieën niet of toch niet volledig worden overgenomen. Bij de huidige stand van zaken lijkt het aangewezen daarvoor het lemmasysteem over te nemen dat door de onderzoekers van *Nomen et Gens* is ontwikkeld.

[8] De afstelling van de apparatuur op de wisselende contrastkwaliteit van het typoscript gebeurde door Tom J. De Herdt, die we hierbij van harte danken. Een aantal specimens zijn te vinden op www.wulfila.ba/aw.

1.5. De doelstelling : een digitale, interactieve en online uitgave

Van bij de aanvang was het de bedoeling van de promotores van het project om het *Antroponymisch Woordenboek* in digitale vorm uit te geven en als databank beschikbaar te stellen via een *web interface*, die zowel een snel occasioneel naslaan als complexe zoekopdrachten moet mogelijk maken. De grote meerwaarde van zo'n editie is ten overvloede gebleken bij het recent gedigitaliseerde *Toponymisch Woordenboek*. Sinds zijn vrijgave op het internet in 2007 is dat al frequenter geraadpleegd dan het gedrukte werk uit 1960. De mogelijkheden tot het doorzoeken van het materiaal zijn immers exponentieel verhoogd tegenover de beperkte zoekmogelijkheden van het gedrukte woordenboek. De digitale versie maakt immers talloze indexen mogelijk, die voor corpora nu eenmaal het aangewezen ontsluitingsmiddel zijn.

Een bijkomend voordeel van een digitale editie is dat eventuele lees- of scanfouten of nog ontbrekende gegevens niet onmiddellijk een catastrofe hoeven te betekenen. Het materiaal kan als het ware van dag tot dag *up-to-date* worden gehouden en in principe *on-the-fly* verbeterd en aangevuld worden zonder dat jaren gewacht hoeft te worden op een herwerkte druk met alle daarmee gepaard gaande kosten. De digitalisering laat met andere woorden toe het *Antroponymisch Woordenboek* dynamisch te beheren. Nog meer dan bij Gysselings twee andere standaardwerken het geval is, zullen immers voortdurend nieuwe gegevens opduiken : nieuwe attestaties, correcties van lezingen, van localiseringen of dateringen, betere etymologische voorstellen, persoonsidentificaties enz. Een webversie biedt bovendien de mogelijkheid bepaalde gebruikers rechten te verlenen om zelf verbeteringen en toevoegingen aan de databank aan te brengen. Om de integriteit van de gegevens te garanderen, zullen zulke schrijfrechten op de databank vanzelfsprekend enkel aan geauthentificeerde gebruikers worden toegekend en zullen contribuanten hun emendaties en aanvullingen telkens wetenschappelijk moeten legitimeren (cfr. infra, sub 2.4).

Een webversie biedt ten slotte de mogelijkheid het digitale woordenboek op termijn te verbinden met andere digitale corpo-

ra, zoals historische lexica, archiefinventarissen en regesten, verhalende teksten, oorkondenedities en facsimile's e.d., en het te integreren in een groot portaal van historische bronnen.[9] Die doelstelling zal binnen het nu lopende project al ten dele worden gerealiseerd door het *Antroponymisch Woordenboek* te koppelen aan het al gedigitaliseerde *Toponymisch Woordenboek*. De lemma's en velden van beide zijn namelijk voor een groot deel identiek.

Ook andere, meer populaire toepassingen door de creatieve gebruiker zelf zijn denkbaar, zoals *mash ups* met een steeds groeiend aantal *online* databanken en *data stores*, zoals recombinatie met geografische en cartografische gegevens uit bijvoorbeeld *Google Earth* en *Google Maps*. De gebruiker kan dan zelf gedetailleerde taalkundige kaarten op maat laten maken door de data aan te passen, weg te laten of af te stemmen op zijn eigen behoeften. Zo zal het mogelijk zijn het programma geografische kaarten te laten tekenen met aanduiding van alle locaties waarin een bepaalde persoonsnaam werd geattesteerd binnen een bepaald gebied, of de verspreiding van een naam diachroon te traceren. De implementatie van dergelijke *mash ups* is betrekkelijk eenvoudig, ofschoon niet goedkoop en arbeidsintensief. Sommige uitgevers van databanken stellen nu al gedetailleerde *application programming interfaces* (APIs) ter beschikking, waardoor twee of meerdere webapplicaties met elkaar kunnen communiceren en hun gegevensbanken kunnen worden samengevoegd volgens de specifieke wensen van een gebruiker.

Hoewel een digitale editie van het Antroponymisch Woordenboek tal van voordelen biedt en dus de voorkeur verdient, hoeft een gedrukte uitgave niet meteen te worden uitgesloten. Die heeft immers het voordeel dat het werk niet verdwijnt met de samenstellers ervan, dat het niet teloor gaat in de dreigende digitale *black out* van onleesbare bestandsformaten, en dat toekomstige onderzoekers eenduidig naar een en dezelfde versie kunnen refereren. Veel hangt ervan af in welke mate het digitale medium in de toekomst zijn eigen vluchtigheid zal weten te overwinnen

[9] Zoals aan de databanken Thesaurus Diplomaticus Belgicus (Koninklijke Commissie voor Geschiedenis), en Narrative Sources (UGent en KCGB).

en een even duurzame, eenduidige en niet-manipuleerbare informatiebron kan worden als het gedrukte woord.

2. Technische Problematiek

De digitale ontsluiting van Gysselings materiaal is een meervoudige taak die bestaat uit (1) de digitale invoer en tekstherkenning als zodanig, (2) de verwerking en semantische analyse van de gegevens, (3) hun structurering en opslag volgens een intelligent schema, en, ten slotte, (4) het ontwerp van een platform dat de interactie met de gebruiker regelt. Het digitaliseren en collationeren van het reeds overgetypte materiaal is arbeidsintensief, maar technisch gezien relatief eenvoudig. Aanzienlijk meer tijd en deskundigheid dan de digitalisering en tekstherkenning, vergen het structureren en verrijken van de data, en vooral, de technische implementatie van een methodiek die dergelijke verrijking efficiënt, principieel en op permanente basis moet mogelijk maken.

2.1. Digitalisering

Mochten Gysselings handgeschreven velletjes niet als typoscript bewaard zijn gebleven, zou de digitalisering bijzonder kostenintensief zijn geworden. Vanuit informatica-technisch standpunt zijn echter ook de typoscripten louter grafische data (d.w.z. niet-digitale tekstgegevens), die op de een of andere wijze als platte tekstgegevens moeten herkend worden, om naderhand semantisch verrijkt te worden. Zoals hierboven al vermeld, werden alle typoscript-fiches in de zomer van 2007 ingescand als digitale afbeeldingen, die dienen als verdere *input* voor de tekstherkenningssoftware (*optical character recognition* of OCR). De kwaliteit van de afbeeldingen is zeer behoorlijk voor een optimaal resultaat tijdens de OCR; bovendien zullen deze scans later ook gebruikt worden als facsimile's, aangeboden aan de gebruikers van de *web interface*, om controle van de digitale transcriptie te vergemakkelijken. Er werd veel zorg besteed aan het uittesten en *fine tunen* van de instellingen van de soft-

ware [10] om leesfouten tijdens de OCR zoveel mogelijk te beperken en de *output* af te stemmen op de *post-processing*. Ondertussen is ongeveer een vierde van de gescande beeldbestanden verwerkt en omgezet naar tekstbestanden, geformatteerd volgens eenvoudige *html-markup*, die verdere verwerking toelaat.

```
<p>L 10 H 44/702</p>
<p></p>
<p>1210, Marchiennes - Sr Montigny-en-O. - Mar-</p>
<p>chiennes</p>
<p>ego Robertus dns de Montengi</p>
<p>Iburgis mater mea- & Rainer(us) frater meus (+)</p>
<p>de Hornaing</p>
<p>Marchian(ensis)</p>
<p>Nicholai abbis (gen.)</p>
<p>ego & uxor mea Malduita & Balduin(us) filius</p>
<p>meus</p>
<p>dnm Radulfum Attrebaten(sem) epm (acc.)</p>
<p>S Sigeri de Thians. S Nicholai Cretin de Obre-</p>
<p>chicort. S Roberti de Gamans. S Philippi de</p>
<p>Derengau. S Guidonis militis de Goy. S Hen-</p>
<p>rici maioris de Roncin. S Amorrici maioris</p>
<p>de Marchienes.</p>
<p>Marchianis</p>
<p></p>
<p>-R.1479-</p>
```

AFBEELDING 4 – *Voorbeeld van de door de OCR geleverde html-output van Afb. 2*

Omdat het tekstmateriaal zeer hybride is en zowel Oud-Franse, Latijnse als Nederlandse teksten in ongeregelde spelling bevat, en omdat het bovendien voor een groot deel bestaat uit unieke persoonsnamen, is het bijzonder moeilijk de tekstherkenningssoftware regels op te leggen die een beslissing tussen twee of meer alternatieve lezingen mogelijk maakt op basis van een voorafgekende woordenschat. Onvermijdelijk worden er dus leesfouten gemaakt tijdens de OCR; de tolerantie die we daarbij aanvaarden bedraagt – op basis van steekproeven – echter niet

[10] We gebruiken Abby FineReader v.9.0 Professional onder Windows XP.

meer dan circa 8% foutmarge. Maar ook het bronmateriaal, m.n. de typoscript fiches, is in principe zelf niet vrij van fouten : het gaat immers om een kopie (typ- en interpretatiefouten van Gysselings handschrift) van een kopie (Gysselings paleografische transcriptie van het archiefmateriaal). Het materiaal is evenwel te omvangrijk om manueel, letter per letter, op transcriptie-, typ- en leesfouten gecontroleerd te worden, en ook dan bestaat het risico op nieuwe fouten tijdens de verbetering t.g.v. menselijke onachtzaamheid. Om al die redenen gaan we ervan uit dat het uiteindelijke tekstmateriaal, zoals het zal worden gepubliceerd en aangeboden aan de gebruiker, een foutmarge zal bevatten van om en bij de 10%. Om deze fouten gaandeweg uit te zuiveren, lijkt het aangewezen een beroep te doen op de hedendaagse praktijk van webgebaseerd gegevensbeheer, m.n. *crowdsourcing*.[11] Deze bestaat erin, de gebruikers van het Woordenboek – die zelf alle belang hebben bij accurate en betrouwbare gegevens – de gelegenheid te geven het digitale tekstmateriaal te vergelijken met zowel de facsimile *scans* van de typoscripten, als met Gysselings handschrift, en hen in staat te stellen fouten ad hoc te verbeteren. Bovendien kan in een later stadium vergelijking met de diverse oorspronkelijke archiefbronnen worden toegevoegd, wanneer die eveneens als digitale facsimile's zouden worden ontsloten door hun bewaarinstellingen (zie noot 9). Daarom worden Gysselings scrupuleus opgetekende herkomstgegevens nu reeds mee in het platform geïntegreerd, zodat de koppeling met dergelijke digitale archieven later eenvoudig gelegd kan worden. Correcties en emendaties die gebruikers zouden toevoegen op basis van vergelijking met de facsimile's of op basis van eigen onderzoek, zullen de integriteit van de gegevensbank ten goede komen, en dat in de mate waarin dat wenselijk en dringend is : fouten worden immers pas vastgesteld (en verbeterd) op het ogenblik dat het specifieke onderdeel wordt opgevraagd. Op die wijze werkt de academische onder-

[11] De term werd bedacht door Jeff Howe in een beroemd geworden bijdrage aan het internetmagazine *Wired* uit 2006. Bedoeld wordt het uitbesteden van grote, arbeidsintensieve taken aan een groep mensen door middel van een open oproep, waar dergelijke taken voordien typisch werden uitbesteed. Omvangrijke opdrachten worden dus ge-*outsourced* aan de *crowd*, in plaats van aan één enkele (dure) subcontractor.

zoeksgemeenschap op permanente basis mee aan de verbetering en aanvulling van het *Antroponymisch Woordenboek* en houdt de juistheid en actualiteit ervan gelijke tred met de mate waarin het wordt geraadpleegd en dat het dus relevant is. Deze aanpak stelt uiteraard bijkomende eisen aan de architectuur van zowel de databank als van de applicatielogica. Hoe dan ook levert zij het meest efficiënte compromis tussen kosten en baten.

2.2. Semantische analyse

De *output* van de OCR – zoals boven reeds vermeld – bestaat uit platte tekstbestanden die in eenvoudige *html-markup* worden gestructureerd. [12] Deze structuur geeft slechts een interpretatie weer van de oorspronkelijke grafische indeling van de typoscripten. Meer bepaald, voegt de OCR aan elke tekstregel in het typoscript een aanduiding of *tag* (label) toe die een opeenvolging van lettertekens (*character string*) aanduidt als een regel in het origineel (cfr Afb. 4). Deze OCR-*output* zal verder dienen als *input* voor een proces dat de tekstbestanden moet interpreteren en verder ontleden naar kleinere stukjes informatie die naar de databank kunnen worden weggeschreven en die als zodanig de kleinste gegevensbestanddelen zullen leveren voor raadplegingen van het Woordenboek. Die analyse is veel complexer dan de tekstherkenning, omdat ze semantiek moet toevoegen aan de onsemantische plattetekst-invoer. Complexe zoekopdrachten en indexaties bij bevraging van de databank kunnen immers pas gebeuren, indien bepaalde elementen in de tekst (ambten en beroepen, data, locaties, namen, personen, sigla, enz.) expliciet worden gemarkeerd en in overeenkomstige velden of *database*-tabellen worden opgeslagen, die bovendien onderling gerelateerd zijn. Deze platte tekst moet met andere woorden worden gestructureerd d.m.v. *markup* of *tagging*. Dat gebeurt door het

[12] Html (*HyperText Markup Language*), is de taal waarin documenten als webpagina's op het internet worden gepubliceerd. Met *markup* worden de metagegevens bedoeld die aan tekstdocumenten worden toegevoegd en de structuur en vorm ervan bepalen. Html is een zodanig universeel geworden taal voor de beschrijving van tekstdocumenten dat ze ook wordt gebruikt voor de uniforme uitwisseling ervan. Precies omdat html echter zo universeel en voor algemeen gebruik werd geconcipieerd, schiet de taal tekort voor meer complexe documenten. Om die reden werd xml ontworpen (cfr infra).

toekennen van *tags* (*labels* van *entities*) aan *strings*, op basis van een voorgedefinieerd schema, waarin wordt vastgelegd welke elementen (*tags*) beschikbaar zijn, wanneer, in welke mate en volgorde, en wat ze mogen bevatten. Dergelijk schema is eigenlijk een afgeleide of variant van het *data model* dat ook voor de databank zal worden gebruikt (cfr infra, sub 3.3). Door de grote omvang van het materiaal kan deze semantische *markup* echter onmogelijk handmatig gebeuren. Binnen het bestek van het project is het immers onbegonnen werk om de meer dan 10.000 fiches handmatig te *taggen* met bijvoorbeeld een *xml editor*.[13] Het risico op fouten zou daarbij bovendien te groot worden ten gevolge van menselijk falen of door het gebrek aan consistentie bij interpretatie.

Een groot aantal van de gewenste semantische elementen of types (categorieën) kan evenwel formeel worden geïdentificeerd op basis van positie en/of vorm. Zo delen bijvoorbeeld persoons- en plaatsnamen een gelijkaardige spelling, waardoor verschillende varianten semi-automatisch kunnen worden geclusterd en gekoppeld aan een gemeenschappelijk lemma. Zodra een substantieel gedeelte van de variante spellingen op die wijze werd gelemmatiseerd, is de software in staat voor het resterend gedeelte van het tekstmateriaal instanties van deze lemmata te herkennen en toe te kennen aan de *identifier* (uniek volgnummer) van de overeenkomstige genormaliseerde vorm in de databank. Elementen die bijvoorbeeld als naam bevestigd worden, zullen verderop automatisch herkend worden. Dergelijke methode wordt in softwareontwerp *Named Entity Recognition* (NER)[14] genoemd, op zich een onderzoeksdomein van *Natural*

[13] Voor occasionele xml-markup en vooral voor het ontwerp van de xml-schema's gebruiken we <oXygen/> XML Editor v.10 onder Mac OS X v.10.5.7 Leopard in de Java Virtual Machine-runtime.

[14] Zie o.m. Nadeau 2007 (diss. Ottawa), die een dergelijke methode beschrijft. Aan de universiteit van Karslruhe heeft men op basis van NER een xml-editor ontworpen die gebruikers tijdens de markup assisteert d.m.v. semi-automatische toekenning, deze *GoldenGATE*-editor is vooral uitgerust voor documenten met biologisch tekstmateriaal (http ://idaho.ipd.uni-karlsruhe.de/GoldenGATE/). Voor onze toepassing komt mogelijk het GATE-framework in aanmerking (http ://gate.ac.uk/), met een eigen editor, ontwikkeld door de Natural Language Processing onderzoeksgroep te Sheffield, die een meer «general purpose» en taalkundig gerichte aanpak heeft. Andere dergelijke NER-editoren (openbron en commercieel) staan opgelijst op http ://en.wikipedia.org/wiki/Named_entity_recognition#NER_Software.

Language Processing (NLP), die ook binnen de computerlinguïstiek wordt bestudeerd. Vooral in de biomedische wetenschappen wordt hiermee uitgebreid geëxperimenteerd, duidelijk vanwege het grote toepassingsgebied bij het herkennen en *taggen* van unieke namen van species en moleculestructuren. In grote lijnen komt het erop neer dat de analyse-software lijsten of indices bijhoudt van alle waarden die reeds voor een bepaalde entiteit (*entity*, *tag*) werden ingevoerd en gekend zijn. Deze lijsten worden vervolgens door de software geconsulteerd telkens een *string* van lettertekens moet worden herkend en toegewezen aan een genormaliseerd lemma. De opzoeklijsten (registers van genormaliseerde vormen) kunnen handmatig worden opgesteld op grond van een representatief aantal fiches (een steekproef uit het materiaal), en gaandeweg uitgebreid door de software zelf. In dergelijke lijsten kan eveneens worden bijgehouden hoe vaak een bepaalde waarde voor een bepaalde entiteit voorkomt (frequentie). Het systeem is dan in staat om zelf woorden of combinaties van woorden (*strings*) te herkennen door ze op te zoeken in de lijst, te herkennen en na te gaan van welke entiteiten de *string* een instantie kan zijn en hoe vaak het er wordt aangetroffen. Vervolgens kan het dan een beslissing nemen voor definitieve toewijzing op basis van een afweging van de frequenties (probabiliteitsberekening). Op dit ogenblik wordt een dergelijk op NER-principes gebaseerd algoritme bedacht voor de *post-processing* van de OCR-*output* en de semi-automatische of geassisteerde toekenning van klassen. Tegelijk wordt ook een *interface* ontworpen via dewelke de gebruiker de automatische toekenning kan controleren en de software assisteren of corrigeren.

Moeilijker te automatiseren is persoonsidentificatie – die nochtans onontbeerlijk is voor biografisch, genealogisch of prosopografisch onderzoek – of identificatie van de naamdrager met zijn functie, ambt of ambacht (getuige, mancipium, enz.) of zijn relatie tot andere personen-naamdragers (zoon van, enz.). Een specifiek probleem bij dit antroponymisch onderzoek is onder meer de koppeling van gelijkende attestaties enerzijds aan één dan wel meer lemmata (vooral problematisch bij ondoorzichtige namen), anderzijds aan één dan wel meer historische personen.

Dergelijke semantische verrijking moet noodzakelijkerwijs nog steeds uitgevoerd worden door menselijke tussenkomst. Door de speciaal daartoe ontworpen *user interface* zal dit echter gelukkig snel, efficiënt en gebruiksvriendelijk kunnen gebeuren, en uitbreidbaar zijn naar de uiteindelijke webapplicatie, waardoor elke gebruiker zijn bijdrage aan de semantische en historische verrijking van het materiaal zal kunnen leveren.

2.3. Datamodellering en -opslag

Om efficiënte toegang en zoekopdrachten mogelijk te maken, dienen de gegevens geïndexeerd en beheerd te worden in een *database management system* (DBMS). Dergelijk gegevensbeheer vereist een specifiek *data model*, of een schema waarin alle wenselijke en mogelijke gegevenseenheden zijn gedefinieerd en hun onderlinge relaties worden uitgedrukt. Er moet voor het project met andere woorden een specifieke semantische structuur worden ontworpen, waarin klassen of categorieën van *instanties* of *tokens* voorgedefinieerd worden en restricties worden opgelegd aan de mogelijke relaties of doorsnedes van de gegevensverzameling. Zo zijn er velden voor attestaties, het bewaardepot en fonds, de bron (origineel, kopie van kopie van kopie, enz.), de gebruikte edities, de context (goederenlijst, inscriptie, munt, oorkonde, vita, enz.), en velden met een genormaliseerde datering (met het oog op numerieke bewerking) en lokalisatie (taalgebied, land, arrondissement of provincie, gemeente). Ook velden voor etymologie en grammaticale informatie (bindmorfemen, casus, genus) zijn van belang, naast, ten slotte, een algemeen veld voor de invoer van overig commentaar. Het datamodel legt daarnaast op dat een attestatie bijvoorbeeld een datum en een plaats kan of zelfs moet hebben, en dat ze betuigd moet zijn in een document van een bepaald type of context. Door deze verweven gegevensstructuur wordt het dan uiteindelijk mogelijk dat gebruikers snel en volgens eigen parameters, de databank kunnen bevragen naar specifieke doorsnedes zoals «alle persoonsnamen met een variante van de eindvorm -bald, voorkomend in oorkonden, tijdens de twaalfde eeuw, in het graafschap Vlaanderen».

Het spreekt voor zich dat het ontwerp van een flexibel datamodel dat zulke specialistische en principieel onvoorzienbare opzoekingen moet mogelijk maken, geen sinecure is. Hoe meer *fine grained* en modulairder het datamodel is, des te complexer kunnen de opzoekingen worden, en des te waardevoller het digitale *Antroponymisch Woordenboek* als instrument voor allerlei gespecialiseerde studies. Buitendien moet het conceptuele datamodel ook worden geïmplementeerd in een logische pendant, waarbij de technische performantie en onderhoudbaarheid van het systeem centraal staan. Het is dus van groot belang dat een software-omgeving wordt gebruikt die geëigend is voor de doelstellingen van de onderneming.

Typologisch bevinden de gegevens zich op het kruispunt tussen gestructureerde data en semi-gestructureerde data. In het eerste geval kan een beroep gedaan worden op het traditionele *entity-relationship* (ER) datamodel, waarvoor verschillende *relational database management*-systemen (RDBMS) bestaan, zoals het commerciële *MS Access* of de meer courante openbronsoftware *MySQL*. De relaties tussen attestaties, lemmata en personen, laten zich efficiënt coderen in een relationeel systeem (met automatische controle op referentiële integriteit); ook de hoger opgesomde velden en hun satellietinformatie (bronnen, geografie, enz.) passen goed in een traditionele, genormaliseerde tabelstructuur. In het tweede geval kan een beroep worden gedaan op allerhande methodes voor semantische en structurele *markup* van tekstgegevens. De meest gebruikte *document markup*-taal is de *extensible markup language* (xml). Zoals de *hypertext markup language* (html) – die met name voor algemeen representationeel gebruik werd ontworpen – is xml een web-standaard die in velerlei *dialecten* beschikbaar is, telkens met zeer specifieke doeleinden. In het bijzonder wordt binnen de *digital humanities* veel gebruik gemaakt van *TEI*, een xml-dialect dat door het gelijknamige *Text Encoding Initiative* werd ontworpen voor onder andere digitale edities van literaire, taalkundige en historische documenten.[15] Met dergelijke *document markup* kan het ruwe materiaal (de gedigitaliseerde typoscript-fiches) seman-

[15] http ://www.tei-c.org/

tisch verrijkt, opgeslagen en doorzoekbaar gemaakt worden. Geciteerde namen worden dan in hun tekstuele context gemarkeerd, voorzien van commentaarvelden (vrije tekst met minimale logische *markup*, bijvoorbeeld voor verwijzingen), de etymologische structuur kan worden gecodeerd (met operatoren om onzekerheidsgraden en alternatieven aan te geven), en tenslotte kan ook de synthese van alle beschikbare informatie in de vorm van een ingang in het uiteindelijke woordenboek worden gegenereerd, bijvoorbeeld volgens de TEI-richtlijnen voor «Print Dictionaries» (P5/9). [16]

De grootste uitdaging bestaat er echter in beide systemen met elkaar te integreren. Zo lijkt met deze tweeledige aanpak voor het gegevensbeheer dubbel werk alvast onvermijdelijk : het datamodel moet immers (minstens gedeeltelijk overlappend) zowel in de *relationele database* worden geïmplementeerd, als in een (op TEI gebaseerd) xml-schema, dat voor de structurering van de *tekstdocumenten* moet dienen. Veelbelovend in dit verband zijn relationele databanken die xml ondersteunen, of, beter nog, volwaardige (*native*) xml-databanken, die vaak reeds als openbronsoftware beschikbaar zijn. Dit gebied is volop in beweging en het heeft weinig zin om op dit ogenblik concrete technologie te noemen. Het nadeel van zowel xml-gebaseerde als traditionele relationele databanken is dat zodra het datamodel vastgelegd is, het later slechts zeer moeizaam kan worden bijgesteld of veranderd. Dit zou problematisch kunnen blijken wanneer conceptuele fouten worden gemaakt bij het ontwerp van het schema, of wanneer uit het gebruik blijkt dat bijkomende velden en relaties (opzoekmogelijkheden) wenselijk zijn. Belangrijk is ook dat ernaar wordt gestreefd om ook de eindgebruiker toegang en schrijfrechten te verlenen die hem in staat stellen het materiaal te bewerken en daardoor te verbeteren, zoals al meermaals werd aangegeven. Vooral wanneer de gebruiker bijkomende gegevensvelden zou willen toevoegen (zoals biografische, geografische of bijzondere historische data), is een rigide datamodel hoogst onpraktisch. Daarom is een flexibel datamodel meer dan wenselijk, dat immers kan groeien en evolueren volgens de vraag

[16] http ://www.tei-c.org/release/doc/tei-p5-doc/en/html/DI.html

van de onderzoeksgemeenschap en waarin alle wijzigingen bovendien ook worden geversioneerd, onontbeerlijk voor de veiligheid en integriteit van de gegevens (cfr infra, sub 3.4). Zulk veeleisend systeem valt nauwelijks elegant te implementeren in een hybride xml-compatibele relationele databank.

Op dit ogenblik wordt het *Resource Description Framework* (rdf)[17] door het World Wide Web Consortium (W3C) aanbevolen als toekomstige standaard voor het semantische web. Veeleer dan een specifieke implementatie in één of andere programmeer- of markup-taal, is rdf een conceptueel model voor generische gegevensopslag en -beheer. Zoals bij xml het geval is, kunnen geïsoleerde tekstgegevens of documenten met rdf gestructureerd worden; zo bestaat er ook een xml-notatie voor rdf. Maar daarnaast kunnen er met rdf tegelijk ook ettelijke honderdduizenden tot miljoenen drie- of meerledige attestaties in gigantische *data warehouses* of *triple* en *tuple stores* worden opgeslagen met een verbluffende performantie op het gebied van *querying* (uitvoeren van zoekopdrachten). Nu al worden belangrijke gegevensbanken gepubliceerd in rdf-formaat, waardoor onderlinge koppeling zeer efficiënt tot stand kan worden gebracht. *Mash ups* van de gegevens in het *Antroponymisch Woordenboek* met andere gegevensbanken worden dan technisch bijzonder eenvoudig. Maar veel relevanter nog, met het oog op de flexibiliteit en uitbreidbaarheid van het datamodel en de facilitering van gebruikersbijdragen (*user input*), is de elegantie van een op rdf-gebaseerd systeem. In plaats van of aanvullend bij een *database management system* kan met rdf immers ook een *content management system* (cms) worden geïmplementeerd, waardoor gebruikers bijzonder eenvoudig aanvullingen kunnen doen, correcties invoeren, of zelf zeer specifieke *queries* opstellen om de gegevens te bevragen.

Op dit ogenblik is nog niet geheel duidelijk welke omgeving zich het beste zal lenen voor de structuur en implementatie van het datamodel; rdf lijkt alvast een goede kandidaat, hoewel een meer traditionele aanpak met een relationele databank uit pragmatisch oogpunt nog het meest realistisch blijft. Er zal in ieder

[17] http ://www.w3.org/rdf/

geval over gewaakt worden dat de gegevens steeds beschikbaar blijven in een duurzaam, transparant en goed gedocumenteerd formaat en dat de applicatie zoveel mogelijk gebruik maakt van (web)standaarden als css, rdf, (x)html, xml en xsl.

2.4. Ontsluiting via een webapplicatie

Nadat (1) het omvangrijke materiaal op de fiches uit Gysselings nalatenschap volledig is gedigitaliseerd, (2) het verrijkt werd met allerlei semantische en historische metadata, en (3) al deze gegevens op een adequate wijze zijn gestructureerd en opgeslagen in een performant systeem, kan het eindelijk ook ontsloten worden via een programma waarmee de gebruiker opzoekingen kan verrichten, bewerkingen uitvoeren en combinaties maken met andere gegevensbronnen. Dit programma regelt dus de interactie met de gebruiker, behandelt de vragen aan de databank en waakt over de integriteit en de veiligheid van de gegevens. In de huidige, web-georiënteerde wereld van academisch uitgeven, is het een evidentie dat dergelijk programma web-gebaseerd zal zijn en dus niet als *stand-alone* desktopapplicatie zal worden aangeboden, noch op CD-rom of DVD, maar als webapplicatie. Bovenop de interne applicatielogica (*back-end*) wordt een externe *user interface* gebouwd (*front-end*). Via de *back-end* kunnen de gegevens in het digitale *Antroponymisch Woordenboek* worden gecombineerd met gegevens uit andere digitale naslagwerken, databanken, geografische of cartografische applicaties of digitale archieven. Via de *user interface* krijgt de individuele gebruiker op een gebruiksvriendelijke en praktijkgerichte wijze toegang tot alle faciliteiten.

Omdat de kwaliteit van het woordenboek in belangrijke mate mee bepaald zal worden door de individuele bijdrage van gebruikers (*user input*), is het van belang om, tot slot, nog even in te gaan op de veiligheidsprocedures die daarbij zullen worden gehanteerd. Gebruikers die de gegevens wensen te corrigeren of aan te vullen, of gebruikers die specifieke velden voor nieuwe types van gegevens aan het datamodel wensen toe te voegen, zullen zich moeten authentificeren via een meervoudige login-procedure. Alle ingrepen op de originele data kunnen daardoor

worden gelogged en zijn traceerbaar tot een unieke gebruikersidentificatie. De systeembeheerder kan dan op elk ogenblik foutieve of kwaadwillige wijzigingen ongedaan maken. Bovendien kunnen andere gebruikers mede toezien op de integriteit van de correcties door ze ten allen tijde te vergelijken met voorgaande versies. Gedetailleerde versionering is dus van groot belang. Ofschoon het *Antroponymisch Woordenboek* via het internet aan elke geïnteresseerde vrij zal worden aangeboden, zullen *user IDs* met schrijfrechten op de databank, om dezelfde redenen van veiligheid en integriteit bovendien ook maar alleen worden toegekend aan gekende of traceerbare personen met voldoende academische merites in het vakgebied.

De online toegang tot de gegevensbank blijft tijdens de duur van het project geïntegreerd in de onderzoekssite http://www.wulfila.be/aw (Universiteit Antwerpen). Om de continuïteit van het webadres te garanderen wordt het beheer van de moederversie na afloop van het project toevertrouwd aan de Universiteit Antwerpen (CNTS, Taal- en Spraaktechnologie) en/of de Koninklijke Academie voor Nederlandse Taal- en Letterkunde (Centrum Teksteditie en Bronnenstudie), die als copromotor van het project optreedt.

3. Onderzoeksperspectieven op basis van het *Antroponymisch Woordenboek*

Het is eigen aan nieuwe wetenschapsinstrumenten dat onmogelijk voorzien kan worden welke onderzoeksresultaten er op termijn uit voort zullen komen. Dat basisinzicht, dat maar al te vaak over het hoofd wordt gezien bij toekenning van onderzoeksgelden, geldt ook voor het *Antroponymisch Woordenboek*. De waarde ervan als duurzaam naslagwerk is echter evident. Zulks kan alleen al worden afgeleid uit de weerklank die de twee vorige standaardwerken van Maurits Gysseling hebben gevonden. Het zogeheten Corpus-Gysseling (1977-87) – de korte benaming wijst er al op – is in weinige jaren een begrip geworden in de neerlandistiek. Het heeft sinds zijn verschijnen, ook in digitale vorm, een stroom aan artikelbijdragen doen verschijnen, was de

basis voor een tiental proefschriften aan Belgische en Nederlandse universiteiten en vormde de grondslag voor het vierdelige *Vroegmiddelnederlands Woordenboek* (Leiden, 2003). Het belang van Gysselings andere standaardwerk, het *Toponymisch Woordenboek* uit 1960, blijkt uit het grote aantal omvangrijke recensies die er in de jaren 1960-1964 van verschenen (synthese in Van Loon 2008). Doordat het echter slechts in gedrukte vorm voorhanden was, vormde het op zichzelf nooit het voorwerp van een monografie of een proefschrift, al biedt het daartoe ongeziene, nog nauwelijks geëxploreerde mogelijkheden. Dat laatste bleek nadat het woordenboek in een databank was omgezet en in alle mogelijke richtingen doorzocht kon worden. Op grond daarvan kon een monografie worden verwezenlijkt over de Europese stadsgeschiedenis (Van Loon 2000).

Het lijdt niet de minste twijfel dat ook het *Antroponymisch Woordenboek* een belangrijk basisinstrument wordt voor toekomstig onderzoek, zeker wanneer het elektronisch ontsloten zal zijn en nog meer nadat het gekoppeld zal worden aan andere gegevensbanken. Het materiaal aan voor- en toenamen, volkennamen, sacrale namen, naambestanddelen, morfemen en formantia vormt immers – samen met de gegevens van het *Toponymisch Woordenboek* – vrijwel het enige taalmateriaal waarover we beschikken om het oudste Nederlands (tot circa 1200) te bestuderen. Het werk is door zijn omvang en de periode die het bestrijkt, evenzeer voorbestemd een basisreferentiewerk te worden voor andere disciplines zoals de oudgermanistiek (Oudsaksisch, Oudfries, Oudhoogduits), de historische romanistiek (Medio-Latijn, het noordelijke Gallo-Romaans) en de medievistiek in het algemeen.

Een van de eerste concrete realisaties van het project wordt een klein antroponymisch lexicon dat zal worden afgeleid uit een (nog niet beschikbare) etymologische databank die in de jaren 2005-2007 bij het *Toponymisch Woordenboek* werd opgebouwd. Daaruit zullen die toponiemen worden gelicht die persoonsnamen bevatten. Het betreft namen zoals Bavo in plaatsnamen als Bevekom, Bevingen, Bavikhove, of Wilmar in plaatsnamen als Wilmarsdonk en Wommersom, en etnoniemen zoals Geldumni

in Jodoigne-Geldenaken. Dit topoantroponymisch materiaal valt chronologisch (van de Oudheid tot 1100 c.q. 1225) en geografisch (van Keltisch tot Fries) volledig samen met dat van het *Antroponymisch Woordenboek*. De homogeniteit van beide corpora wordt bovendien gegarandeerd doordat ze berusten op het onovertroffen excerptiewerk van één en dezelfde auteur, m.n. Gysseling. Het belang van dergelijke «topo-antroponymische» lexica is al eerder onderstreept door een aantal buitenlandse onderzoekers (onder meer Greule 1997 en 2002 voor Duitsland, Insley 2002 voor Engeland, Eichler 2002 voor het Slavische taalgebied), maar ze zijn tot nu toe buiten Morlet 1985 nog nooit gerealiseerd. Zo'n lexicon voor de historische Nederlanden is niet alleen van belang als aanvulling op het persoonsnamenmateriaal in het *Antroponymisch Woordenboek*. In mediëvistisch opzicht zal het nieuwe inzichten opleveren over de meest duistere maar tegelijk zo cruciale periode van de West-Europese geschiedenis (5de tot de 10de eeuw), over het verloop van de taalgrens, de germanisering van Noord-Gallië en de zgn. Frankische landname, de kersteningsperiode enz. De honderden (vooral Germaanse) persoonsnamen die in het plaatsnamenmateriaal zijn gefossiliseerd, zijn immers die van de Frankische (of andere Germaanse) grondheren die blijkens hun naam aan de grondslag lagen van de vele nederzettingsnamen op *-gem, -sele, -ingen, -court, -ignies, -ville, -y, -tun, -wic*, enz. in de Nederlanden en Noord-Frankrijk. De waarde van een topoantroponymisch lexicon zal vooral blijken zodra het digitaal met het grote *Antroponymisch Woordenboek* wordt verbonden. De namen van de stichters van deze nederzettingen kunnen dan immers worden vergeleken en mogelijk zelfs geïdentificeerd met de namen van daadwerkelijk overgeleverde schenkers, getuigen en oorkonders in vroegmiddeleeuwse teksten. Het instrument kan daardoor o.m. de internationale actieradius van sommige vroegmiddeleeuwse optimaten in Noord-West-Europa (specifieke namen van Franken, Friezen, Gallo-Romanen, Goten, enz.) zichtbaar maken of het aandeel van de Nederlanden in de frankisering van Noord-Gallië. Het woordenboek wordt aldus ook een basiswerk voor de prosopografie van de Vroege Middeleeuwen in de Neder-

landen. Het knoopt daarin historisch aan bij een onderzoekstraditie van belangrijke mediëvisten als K. Schmid, Tellenbach, Wenskus en Werner, of bij naslagwerken als de imposante *Prosopography of the Later Roman Empire* (3 vols.; Cambridge: 1971, 1980, 1992).

De nieuwe inzichten inzake etymologie en identificatie van afzonderlijke namen en personen kunnen ook resulteren in een reeks losse etymologische en historische namenstudies, zowel met betrekking tot het Oudnederlands als het Oudfrans, Oudhoogduits, Oudfries, enz. Behalve lexicale kwesties kunnen die studies nieuwe gegevens aan het licht brengen in verband met historische fonologie, morfologie en dialectgeografie van het Oudnederlands en het Vroegromaans (bijvoorbeeld, de *r*-metathese, de evolutie van de diftong /eu/, Romaanse casusmorfemen, assibilatie, pari- en imparisyllabische nomina, genitiefmorfemen). De buitengewone paleografische kwaliteit van het materiaal maakt het bovendien uitermate geschikt voor de studie van vroegmiddeleeuwse scripta.

Uiteraard zal het *Antroponymisch Woordenboek* het aangewezen basiswerk worden voor de louter naamkundige studie van de Vroege Middeleeuwen. Het zal een synthese mogelijk maken van de persoonsnaamgeving in de Lage Landen van de Oudheid tot circa 1200. Zo'n werk, dat de zeer lange periode bestrijkt die aan de opkomst van onze huidige familienamen voorafging, is nog onbestaande. Andere naamkundige verschijnselen die kunnen worden onderzocht, zijn modenamen, lokaal beperkte namen, het opkomen en verdwijnen van bepaalde toenaamtypes (zoals de Latijnse trianomina of de eerste toenamen), telkens met hun historische motivering.

Dat de genoemde onderzoeksmogelijkheden geen fictie zijn, is al aangetoond in een aantal recente studies. Zo konden twee volkennamen in Caesars *De Bello Gallico*, die al sedert de tijd van de humanisten de geleerden voor grote raadsels plaatsten, de etnoniemen Geldumni en Caerosi, worden ontcijferd. De naam van de Geldumni, die met Ambiorix deelnamen aan de opstand tegen Caesar, is te identificeren met de plaatsnaam Geldenaken/Jodoigne (Van Loon en Wouters 1994). De ontcijfering van de

naam Caerosi, die de oervorm is van het woord *heer* en mogelijk voortleeft in de plaatsnaam Heers (Belgisch Limburg) of Heer (Maastricht), levert het wel definitieve bewijs dat de zuidelijke Nederlanden al in de late prehistorie gegermaniseerd waren en werpt een nieuw licht op de beruchte namenzin van Tacitus (Van Loon 2006, 2009). In een recente studie (Van Loon 2009), waarin de *Vita Landoaldi* uit de 10de eeuw als historische bron wordt gerehabiliteerd, kon op grond van de namen van 7de-eeuwse heiligen worden aangetoond dat omstreeks 650 Zuid-Franse of Italiaanse religieuzen in Zuid-Limburg als zendelingen werkzaam moeten zijn geweest.

Het *Antroponymisch Woordenboek* zelf zou in een eerste versie opgeleverd moeten worden tegen eind 2012. Met het project hopen de initiatiefnemers te voorkomen dat een filologisch-taalkundige en historisch-taalkundige onderzoekstraditie waarin Vlaanderen en Nederland in het verleden een vooraanstaande plaats bekleedden, over enkele jaren volledig zal verdwenen zijn. Bij voortzetting van de huidige trend zullen er immers over afzienbare tijd in Vlaanderen en Nederland geen deskundigen meer zijn die weerwerk kunnen bieden aan de vaak Becanistische naamsverklaringen die de laatste jaren tot zelfs in zogenaamde wetenschappelijke tijdschriften opnieuw de kop kunnen opsteken. Hopelijk kan de online-uitgave van het *Antroponymisch Woordenboek* aankomende generaties taalkundigen, historisch-taalkundigen en historici er opnieuw toe aanzetten om onze rijke historisch-taalkundige en naamkundige onderzoekstraditie voort te zetten. De immense inspanningen die onderzoekers van vorige generaties op dit gebied hebben geleverd – waaronder Maurits Gysseling beslist niet de minste was – zijn dan niet verloren, maar kunnen zich als waardevol erfgoed bewijzen voor de toekomst.

Bibliografie

Boesch, Bruno. Zur Gestaltung des neuen Förstemann. *Beiträge zur Namenforschung* NF 6 (1971) : 307 vv.

EICHLER, Ernst. Zum Stand der slavistischen Personennamenforschung. In Geuenich e.a. 2002 : 195-203.

FÖRSTEMANN, Ernst. Altdeutsches Namenbuch. 1. Band : *Personennamen*. Nordhausen 1856 (Bonn 1900[2]); 2. Band : *Ortsnamen* (zwei Lieferungen).Nordhausen 1872 (Bonn 1913-1916 (hg. von H. Jellinghaus).

GEUENICH, Dieter, Wolfgang HAUBRICHS & Jörg JARNUT (uitg.). *Nomen et Gens. Zur historischen Aussagekraft frühmittelalterlicher Personennamen* (Ergänzungsband 16 zum *Reallexikon der Germanischen Altertumskunde*). Berlijn-New York 1997.

GEUENICH, Dieter, Wolfgang HAUBRICHS & Jörg JARNUT (uitg.). *Person und Name* (Ergänzungsband 32 zum *Reallexikon der Germanischen Altertumskunde*). Berlijn-New York 2002.

GREULE, Albert. Personennamen in Ortsnamen. In : Geuenich e.a. 1997 : 242-258.

GREULE, Albert. Personennamen in Ortsnamenbüchern. Plädoyer für ein Namenbuch der toponymischen Personennamen. In : Geuenich e.a. 2002 : 305-309.

INSLEY, John. The Study of Old English Personal Names and Anthroponymic Lexica. In Geuenich e.a. 2002 : 148-176.

MARYNISSEN, Clem. *Hypokoristische suffixen in Oudnederlandse persoonsnamen*. Gent 1986.

MORLET, M.-Th. *Les noms de personne sur le territoire de l'ancienne Gaule du 6[e] au 12[e] siècle* (3 dln.). Parijs (CNRS) 1971-85.

NADEAU, David. Semi-Supervised Named Entity Recognition. Learning to Recognize 100 Entity Types with Little Supervision. Diss. Ottawa 2007.

SEARLE, William George. *Onomasticon Anglo-Saxonicum*. Cambridge 1897 (herdruk Hildesheim 1969).

SONDEREGGER, Stefan. Prinzipien germanischer Personennamengebung. In :Geuenich, Dieter, Wolfgang Haubrichs & Jörg Jarnut (uitg.) 1997 : 1-29.

VAN LOON, Jozef en Annelies WOUTERS. De stamnaam Geldumni. *Belgisch Tijdschrift voor Filologie en Geschiedenis* 74 (1994) : 25-34.

VAN LOON, Jozef. *Maurits Gysselings Toponymisch Woordenboek. Receptie, aanvullingen en correcties*. Brussel (KCTD) 2008.

VAN LOON, Jozef. Een etymologie en haar historische implicaties : de stamnaam Caerosi (Caesar, BG 2.4). *Verslagen en Mededelingen KANTL* 116 (2006) : 375-400.

Wouter SOUDAN en Jozef VAN LOON
(FWO en Universiteit Antwerpen)