Prof.dr. B. Mons

Kennis is als liefde …
men wordt van het delen niet minder

Universiteit Leiden

Bij ons leer je de wereld kennen

Kennis is als liefde …
men wordt van het delen niet minder

Oratie uitgesproken door

Prof.dr. B. Mons

bij de aanvaarding van het ambt van bijzonder hoogleraar in de Biosemantiek

aan de Universiteit Leiden

vanwege het Netherlands BioInformatics Centre

op maandag 25 februari 2013.

Universiteit
Leiden

Prof.dr. B. Mons

Meneer de Rector Magnificus, leden van het bestuur van de Stichting 'Netherlands Bioinformatics Centre', leden van het curatorium voor de bijzondere leerstoel in de biosemantiek,

Het is een grote eer voor mij om vandaag aan u allen het nieuwe vakgebied uit te leggen dat we *Biosemantiek* hebben gedoopt. Het instellen van een bijzondere leerstoel voor deze embryonale discipline is een teken van visie van zowel NBIC als het LUMC.

Deze rede gaat over de 'post-print' periode. Daarom heeft u ook geen geprint boekje voor u. Wel heeft u allen bij de ingang een kaartje gekregen met een QRcode erop. Deze leidt naar de volledige tekst van mijn voordracht.

Ik sta hier eerlijk gezegd met enige gêne, omdat deze voordracht ten diepste gaat over de allermodernste, internationale communicatietechnieken, terwijl ik hier in een 438-jarige traditie sta en daardoor voor het eerst in mijn carrière een lezing moet houden zonder visuele ondersteuning en dan ook nog in het Nederlands. Maar we gaan het proberen. Als ik hier om mij heen kijk, en besef hoeveel lieve familie, goede vrienden en collega's ik heb, en ook dat er nog veel meer mensen hadden kunnen zitten als ik dit betoog niet in het Nederlands had hoeven houden, voel ik mij een bijzonder bevoorrecht mens. Zonder veel van deze mensen, maar vooral de steun en de tolerantie van Joke en mijn gezin, had ik hier natuurlijk niet gestaan. Ik denk dat we best mogen stellen dat ik hier vandaag (eindelijk) een beetje een succesverhaal mag vertellen, en er zijn heel wat mensen in het gehoor die zelf wel weten welke bijdrage ze aan de totstandkoming daarvan hebben geleverd.

Mijn ex-malariacollega's hoeven zich ook niet te vervelen. Ik word in mijn nieuwe bioinformatica wereld nog regelmatig uitgelachen om mijn uitgekauwde malariavoorbeelden. Toch ga ik die ook **nu** weer gebruiken. Het is allemaal al ingewikkeld genoeg!

Biosemantiek brengt ons over een drempel

Wij bevinden ons op (in feite al over) de drempel van een fundamenteel veranderde manier van wetenschap bedrijven, aangeduid als 'enhanced Science' of kortweg eScience. De biologie is momenteel de meest data-intensieve wetenschap. **Biosemantiek** is een onderdeel van eScience. Ik hoop u vandaag een indruk te kunnen geven van de enorme complexiteit waarmee biologen tegenwoordig worstelen. Niet omdat de biologie zelf zoveel ingewikkelder is dan vroeger, maar simpelweg omdat wij laag na laag doordringen in een complexiteit die er altijd al was maar die ons nu, bijna letterlijk, de wetenschappelijke keel dichtknijpt.

Big Data dringt lezen naar een andere positie

Door de zogenaamde 'high-throughput' technieken in de levenswetenschappen is er een crisis in wetenschappelijke communicatie binnengeslopen die ons nu met volle kracht treft: *een onhandelbare hoeveelheid relevante data.*
Veel van de eminente geleerden in deze zaal zijn, net als ik, opgevoed met het mantra dat '*men alles gelezen dient te hebben wat relevant is, alvorens men met een experiment begint*'. Als een begeleider dit nu, in de Big Data era, tegen studenten zou zeggen, zou dit waarschijnlijk tot grote hilariteit leiden.
Lezen is nu veel meer een kwestie geworden van achteraf controleren of een door de data opgehoeste hypothese waar zou kunnen zijn. Als we de methoden die wij op data loslaten niet snel aanpassen zal dit de ontwikkeling van de wetenschap ernstig gaan vertragen.
Twee getallen om u de ernst van de situatie te schetsen: Tijdens de 45 minuten die deze oratie duurt, sterven niet alleen zo'n 150 kinderen aan malaria, maar worden ook wereldwijd 65 nieuwe biomedische artikelen gepubliceerd. Aan die artikelen zijn vaak ook nog Gigabytes aan 'ondersteunende data' gelinkt. De wereldwijde verbeteringen in de kwaliteit van leven zijn vrijwel allemaal direct gerelateerd aan doorbraken in de levenswetenschappen. Toch is het altijd al een strijd geweest om wetenschappelijke inzichten te vertalen naar maatschappelijke toepassingen, met name in de biomedische

wetenschappen. Naast noodzakelijke maar vertragende regelgeving zou ik het feit dat er na decaden van onderzoek nog steeds geen afdoende middelen zijn tegen veel ziekten zoals bijvoorbeeld malaria onder andere willen toeschrijven aan een *systematische onderschatting van de complexiteit van de biologie*. Nu er wereldwijd miljoenen wetenschappers zijn met ieder hun eigen 'doorbraakjes' zou je misschien mogen verwachten dat ook de maatschappelijke vertaling van biologische inzichten in een stroomversnelling zou komen. In sommige technische vakgebieden lijkt dit wel te kloppen, maar in de levenswetenschappen lijkt eerder het tegendeel waar. De tijd tussen het ontdekken van een mogelijk geneesmiddel en het op de markt brengen ervan is bijvoorbeeld inmiddels juist gestegen tot 15 jaar en de gemiddelde kosten tot 800 miljoen dollar. Het is dus niet zo dat beter weten automatisch leidt tot beter eten, betere gezondheidszorg en betere medicijnen. Het **Biosemantiek** vakgebied kan voor een enorme verbetering zorgen in de communicatie binnen de wetenschap en ook van de wetenschap naar de maatschappij en weer terug.
Om in de enorme data hooiberg die wij creëren de spelden te blijven vinden, moeten echter niet alleen technisch, maar ook sociaal nog heel wat hobbels worden genomen.

Korte historie
Aangezien de Biosemantiek een nieuw vakgebied is, dat niet zonder slag of stoot tot stand is gekomen, wil ik beginnen met een korte historische schets. Hierbij zullen ook enkele namen van collega's worden genoemd, maar dat scheelt dan later weer in mijn dankwoord.

Laat ik beginnen met het adresseren van een hardnekkig punt van kritiek op mij, om daar iets uit te leren. Niet alleen mijn collega's door de jaren heen, maar ook mijn belangrijkste manager sinds bijna vijf en dertig jaar hebben mij regelmatig voorgehouden: '*je zwabbert alle kanten op, er is geen touw aan vast te knopen*'
Mijn antwoord is de laatste jaren: 'dat is mijn academische zeilreis'. Ik neem u even mee op die reis, omdat deze rede ook iets van een maatschappelijke verantwoording is, met name ook naar alle mensen die mijn 'grillige gedrag' al die jaren hebben verdragen.

Eerst ging ik biologie studeren in plaats van 'gewoon' dokter, dierenarts, rechter of dominee te worden. Weliswaar sta ik vandaag alsnog op de preekstoel, maar dit is wellicht ook weer direct de laatste preek. Dat ik bioloog werd was nog tot daar aan toe, maar toen koos ik ook nog voor 'malaria' in plaats van voor een van de ziekten waarmee je *wel* beroemd kon worden. Daarna werd ik tijdelijk bureaucraat, notabene in het verstikkende Brussel, terwijl een van de stellingen in mijn proefschrift luidde: 'Ambtenaar is een Te naar Ambt'. Vervolgens richtte ik opeens een bedrijf op terwijl ik inmiddels bij NWO werkte! *En daarna nog een*! Eduard en René hebben wat met me te stellen gehad in die tijd! Na veel vallen en opstaan kwam ik via Rotterdam weer terug op mijn oorspronkelijke academische nest: Leiden.
Het startschot van deze zeilreis werd gegeven door Hugo van der Kaay, de eerste wending rond de boei kwam met Marc de Bruycker (Brussel), de tweede rond Eduard Klasen (NWO), de derde rond Johan van der Lei (Rotterdam), de vierde rond Gert Jan van Ommen en nu, midden in de eindsprint, met mijn zeilmaatje Ruben Kok en onze NBIC- en LUMC-bemanning, heb ik uiteindelijk het gevoel dat we dicht bij het doel zijn gekomen.
De regressielijn tussen al deze keerpunten is dat ik altijd heb geloofd in het '**delen van kennis**', niet alleen als bevoorrechte wetenschappers onderling, maar ook met collega's in ontwikkelingslanden. Vandaar ook mijn titel: liefde en kennis hebben gemeen dat zij vermenigvuldigen als je ze deelt. Nog sneller dan broden en vissen… er zijn nogal wat manden data over op het moment, maar daar kom ik later op terug.

De eerste jaren van de Biosemantiek groep waren niet gemakkelijk. Er was, en is, een diep gewortelde scepsis onder biologen over de waarde van textmining, computer reasoning en bioinformatica. Bij verschillende artikelen die wij volgens

het traditionele systeem wilden publiceren schreven reviewers vele varianten op de opmerking: "*ik kan niet accepteren dat computers slimmer zijn dan ik*". Nu is dat ook allerminst wat wij suggereren. Wel kan de computer veel beter dan wij omgaan met de complexiteit en het volume van de huidige wetenschappelijke kennis zoals die is neergelegd in miljoenen publicaties en duizenden databases. Er ligt een schat aan verborgen kennis in deze goudmijn die momenteel ongebruikt blijft. Wanneer je zo'n boodschap hebt die ingaat tegen de *status quo* dan heb je, met name ook in de wetenschap, 'de wind tegen'. Als je geen boot met een sterke geldmotor hebt, maar 'met een zeilboot' een doel moet bereiken dat 'tegen de wind in ligt' dan moet je laveren… Dat 'scherp aan de wind varen' lijkt voor niet-zeilers op 'zwabberen'. Als iemand vanuit een helikopter momentopnames maakt van een zeilboot die 'tegen de wind in vaart' dan zal de conclusie zijn dat die nooit op koers ligt. Slechts op de schaarse momenten dat de zeilboot 'door de wind' gaat ligt die een ondeelbaar moment op koers. Ik heb mijn hele carrière scherp aan de wind gevaren en kan heel goed begrijpen dat dit voor mijn naaste omgeving vaak verwarrend is geweest. Aan de andere kant, *ik sta tenslotte maar een keer op de preekstoel*, als u altijd op koers ligt, vaart u dan niet teveel met de wind mee?…

Het uiteindelijke doel komt nu heel dichtbij

Natuurlijk ben ik ook wel eens van de koers afgeraakt, maar uiteindelijk weet ik intuïtief al heel lang waar we naar toe moeten.

Alle informatie zo open mogelijk beschikbaar en begrijpelijk voor iedereen, inclusief computers

Natuurlijk was dit doel nog ver weg en veel minder duidelijk toen ik nog 'aan malaria werkte', maar het delen van kennis met onze collega's uit ontwikkelingslanden was ook toen al een grote drijfveer. Toen ik bij de Europese Commissie een systeem had opgezet om door taalbarrières heen vergelijkbare projecten en goede onderzoekspartners in ontwikkelingslanden te vinden, kwam gaandeweg de mogelijkheid van een universele computertaal in zicht die redeneren met behulp van computers mogelijk maakte. Het bloed kruipt waar het niet gaan kan: ik wilde dit ook in de biologie toepassen!
In het jaar 2000 kwam het verzoek van Johan van der Lei, om een groepje te starten op het Erasmus Medical Centre om onze technieken toe te passen op de (bio)medische informatica. Daar ligt in feite de oorsprong van de eerste Biosemantiek groep met mijn collega's van het eerste uur: Erik van Mulligen en Jan Kors.

Eerst moest er gegraven worden en een fundament gelegd. Gedurende bijna zeven magere jaren beschreven onze artikelen voornamelijk onze methodes om uit traditionele geschreven tekst en databases weer informatie te reconstrueren waar de computer *in elk geval iets van snapt*. Pas daarna (in 2007) kwam het eerste 'bedreigende' artikel waarin wij aantoonden (met Martijn Shuemie voorop) dat we het koppelen van functies aan eiwitten sneller met de computer konden doen dan met menselijke experts. Inmiddels was ik door Gert-Jan van Ommen, in goed overleg met Johan, naar Leiden 'gehaald' om ook hier, een Biosemantiek groep op te zetten. Sindsdien is ruwweg de helft van onze, vaak gezamenlijke, publicaties gericht op het genereren van biologische hypotheses met behulp van de computer. In 2009 kwam de volgende doorbraak: een artikel waarin we aantoonden (met Herman voorop) dat we ook daadwerkelijk *nog onbekende* eiwit-eiwit interacties konden voorspellen, die later in het laboratorium konden worden bevestigd. Denk echter niet dat hiermee de scepsis van de baan was. Elk artikel waarin wij weer nieuwe verbanden voorspellen ontmoet altijd wel weer minstens één reviewer die het '*gewoon niet wil geloven*'.
Toch doen wij in feite niet veel anders dan wat bijvoorbeeld in grootschalige DNA-analyse gebeurt, namelijk het leggen van statistische verbanden tussen begrippen die nog niet eerder expliciet met elkaar waren verbonden. Ik heb echt bewondering voor mijn huidige collega's waaronder Peter Bram 't Hoen, Erik Schultes en Marco Roos en Rob Hooft, dat zij meesturen op de laverende zeilboot met de naam Biosemantiek op de boeg.

Gedurende deze eerste tien jaar rijpte langzaam het idee van een dynamische begrippenruimte ofwel een 'Concept Web' dat zowel voor mensen als voor computers begrijpelijk zou zijn. Een heel abstract begrip voor de meesten van u, en ik ga zo meteen een poging doen het uit te leggen, maar eerst even de zeilreis afmaken. In 2009, kwam er op initiatief van NBIC het moment (achteraf gezien een keerpunt) waarop een aantal 'opinieleiders' samenkwam in New York en we de Concept Web Alliance (CWA) oprichtten. Het CWA werd een denktank die probeerde een antwoord te vinden op de enorme uitdagingen die gepaard gaan met de steeds groter wordende datasets in de levenswetenschappen. Uit deze Alliantie is de nieuwe methodologie naar voren gekomen die nu wereldwijd wordt aangeduid met de term **Nanopublicatie** (*u dacht al, waar blijft ie?*). Inmiddels worden nanopublicaties toegepast in een groot internationaal farmaceutisch project (Open PHACTS) met meer dan 30 bedrijven en publieke partners, en worden er al nieuwe bedrijven op gebaseerd. Over een maand is er een bijeenkomst in Amsterdam met meer dan 30 organisaties die zich zullen buigen over de noodzaak van het oprichten van een 'vereniging' van publieke en private partners die nanopublicatie als een belangrijke ontwikkeling voor de toekomst van de levenswetenschappen zien: de **Nanorepubliek**.

Maar nu nog de zilverruggen in de wetenschap om krijgen
Het systeem dat we vandaag de dag nog steeds gebruiken om te communiceren in de wetenschap is sinds de uitvinding van de boekdrukkunst rond 1450 niet meer wezenlijk veranderd. Nu moet dat toch echt gebeuren, anders lopen we krakend vast. Hopelijk is er, net als over de boekdrukkunst over 600 jaar nog steeds een debat gaande over wie nu precies nanopublicaties heeft bedacht. We zullen daar nooit uitkomen, want zoals bijna elke doorbraak zijn ook nanopublicaties een resultaat van multidisciplinair samenwerken. Achteraf zegt dan iedereen: '*eigenlijk logisch en simpel; dat we dat nou niet eerder hadden bedacht*'.

De tijd waarin 'Web Publishing' niet meer was dan dode PDF's op het Internet zetten lijkt met alle 'interactieve formats' nu langzaam tot een einde te komen. Toch voorspel ik u met klem dat zelfs het zogenaamde 'artikel van de toekomst', een tekst die mismaakt is tot een kerstboom van hyperlinks, *de oplossing niet gaat bieden*. Het probleem is namelijk dat men een *totaal verouderde eenheid van communicatie*, namelijk het klassieke wetenschappelijke artikel, blijft gebruiken als basis voor wetenschappelijke uitwisseling en beoordeling. Het compleet achterhaalde universitaire systeem dat wetenschappers nog steeds beoordeelt op het aantal gepubliceerde artikelen, vermenigvuldigd met hoeveel mensen deze artikelen citeren is zo langzamerhand een van de grootste vertragende factoren voor eScience aan het worden.

In de biologie voelen we nu al de pijnlijke consequenties van de steeds groter wordende kloof tussen onze mogelijkheden om massale data te genereren en ons vermogen om daar vervolgens iets nuttigs over te zeggen. Nu het gemiddelde experiment zoveel data oplevert dat die niet meer zonder hulp van de computer kunnen worden begrepen, is er een interessante maar heel gevaarlijke paradox ontstaan: de moderne datasets zijn '*in principe*' zeer waardevol; ze voeden het proces van hypothesevorming nu heel direct en de datasets kunnen '*in principe*' hergebruikt worden door een hele generatie wetenschappers om er weer nieuwe kennis uit te halen. *Het paradoxale is helaas* dat er nog steeds geen enkele wetenschappelijke beloning staat op het goed beschrijven, het behandelen, laat staan op het openlijk delen van data. Iedereen die iets met Big Data doet kent de term 'Data Kerkhoven'.

In het 'Egosysteem' dat de wetenschap ten diepste is, blijft het verleidelijk om data liever te koesteren als brandstof op weg naar de Nobelprijs dan om die te delen. Toch zijn naar mijn stellige overtuiging juist ook de bestuurders van universiteiten en de financiers van onderzoek hier echt 'mede schuldig'. Het 'eindproduct' van wetenschap lijkt nog steeds *een artikel* waarvan je al blij bent als 100 mensen het lezen en 20 het

citeren… Data kun je al helemaal niet formeel publiceren en citeren, laat staan goed toegankelijk maken, laat staan hergebruiken. Een wat slappe poging tot een tussenoplossing was om enorme datasets die 'het artikel ondersteunen' aan het artikel te linken. Klinkt mooi, maar je kunt er niet veel mee: *weer wordt dus de fout gemaakt om niet van het artikel als kapstok te willen afstappen.* De eerste klassieke tijdschriften stoppen zelfs al met het accepteren van deze datasets. Ten eerste omdat de referenten er vaak niet aan toe komen bij de beoordeling van het artikel. Ten tweede omdat de meeste datasets op de spreekwoordelijke 'laptop van de AIO' staan en na een paar jaar dus niet meer te vinden zijn. Ten derde omdat die datasets door bioinformatici in de meest exotische, ongetwijfeld 'briljante', maar onleesbare formats zijn opgeslagen.

Kortom, er moet snel iets gebeuren om 'data driven science' ook echt de drijver te laten zijn van wetenschap en te voorkomen dat datasets overal ongestructureerd ronddrijven en wij er tenslotte tussen verdrinken.

De technische oplossingen zijn er!
eScience, door Jacob de Vlieg gedefinieerd als 'wetenschap die niet zonder computers kan' heeft dus ook een *aan eScience aangepaste manier van communiceren* nodig. Wat moet er gebeuren? Om te beginnen moeten we er eindelijk aan geloven dat *computers onze belangrijkste analisten buiten het laboratorium zijn geworden.* Laten we ze dus ook toegang tot onze resultaten geven en dan niet slechts in bloemrijke mensentaal, doorspekt met stijlfiguren, synoniemen en vakjargon: dat is namelijk de ergste nachtmerrie voor computers. Het is dan ook niet toevallig dat ik al weer sinds meer dan tien jaar met mijn taalkundige broer Albert samenwerk.

Een belangrijke taak voor eScience methodologen is in feite de constructie van zogenaamde '***Social Machines***'. Dat zijn Web-omgevingen waarin de mens en de computer in een constante 'brein-machine interactie' naadloos kunnen samenwerken. In een wetenschappelijk veld waar niet alleen al meer dan 22 miljoen artikelen zijn gepubliceerd, maar waar er dus elke 40 seconde 1 bij komt en waar de datasets vaak honderden miljoenen nieuwe associaties bevatten is het natuurlijk niet eenvoudig om de dagelijkse input in zo'n Social Machine effectief te verwerken. Toch zal het moeten, en wel snel! Als een gevolg van het achterblijven van adequate uitwisseling van deze data blijft op dit moment *het overgrote deel* van nieuw gepubliceerde associaties in datasets volledig buiten beeld. Als we niet snel ingrijpen moeten we de mythische Tantaluskwelling accepteren, waarbij we zelfs het meest laaghangende fruit in de bomen boven ons niet kunnen bereiken en ook niet meer uit het grote stuwmeer kunnen drinken van bestaande kennis waarin we nieuw verworven inzichten moeten spiegelen.

In de Social Machine fase van de wetenschap moeten dagelijks nieuw binnenkomende, massieve data- en informatiestromen onmiddellijk worden geanalyseerd door computers in de context van alle data die we al hadden. Als de wetenschappelijke 'Social Machines' eenmaal effectief beginnen te draaien, zal de *manier waarop wij wetenschap bedrijven fundamenteel veranderen.* Zonder Biosemantiek kunnen biologische Social Machines niet functioneren, maar er zijn meerdere disciplines nodig om ze te maken en te laten draaien: van pure computerwetenschap om de prestaties en de schaalbaarheid van het systeem omhoog te brengen, via het bouwen van wereldwijd geaccepteerde en biologie-omspannende terminologiesystemen tot het maken van gebruikersvriendelijke Smartphone applicaties voor annotatie door miljoenen experts rond de hele wereld, *inclusief in ontwikkelingslanden.*
Mede op basis van dit vergezicht heb ik zo'n twaalf jaar geleden besloten het malaria-onderzoek vaarwel te zeggen en mij op dit fundamentele probleem te gaan richten. Voor mijn malaria-collega's in de zaal: mijn persoonlijke overtuiging is dat ik met Biosemantiek een grotere bijdrage aan malaria zal kunnen

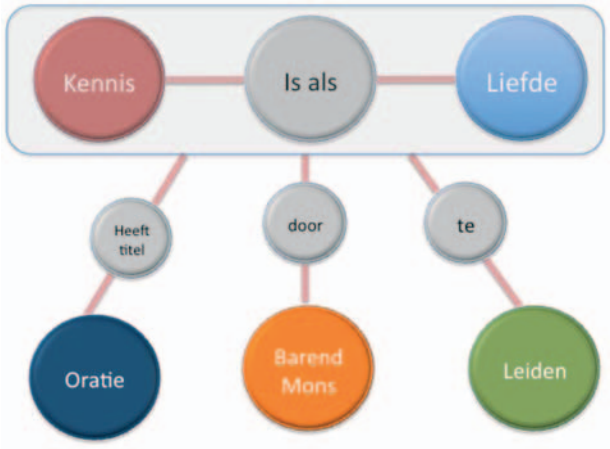
leveren dan wanneer ik was gebleven (dat laatste zal Chris Janse zeker beamen).

Nu naar nanopublicaties

Voorlopig zal het nog wel zo blijven dat *mensen* de uiteindelijke beslissing nemen of ze een bepaalde bewering als 'waar' accepteren. Toch zal er een snel groeiende bijdrage zijn van het genereren van hypotheses met behulp van computers: met name verbanden die nu toe diep verborgen zaten in de ondoorzichtige brij van alles wat we al gepubliceerd hadden. Om wat Biosemantiek doet toch zoveel mogelijk in 'biologische termen' uit te drukken noem ik alles wat we al eens een keer 'beweerd' hebben (oftewel expliciet gemaakt hebben in artikelen of databases) het '**Explicitome**'. Uiteindelijk is dat, net als het Genome, het Proteome en het Metabolome, gewoon een extra aanvullende informatiebron waarin we patronen moeten zien te ontdekken in wat we gisteren nog als chaos beleefden.

Het Explicitome van de levenswetenschappen bevat op dit moment ruw geschat 10^{14} betekenisvolle beweringen. Er is geen computersysteem (en zeker geen hersenpan) dat een dergelijke hoeveelheid informatie kan verwerken. Om de enorme data- en informatiestromen het hoofd te kunnen bieden moet dus allereerst een intelligente vorm van *datareductie* plaatsvinden zonder verlies van essentiële informatie. Gelukkig zijn veel van de 10^{14} beweringen citaties van eerdere beweringen, oftewel 'letterlijke herhalingen van een eerdere bewering'. Als het even kan zouden we zelfs de eindeloze herhalingen in het Explicitome moeten gebruiken om het wetenschappelijke betrouwbaarheidsgehalte van elke kernbewering te schatten. De vorm van data- en informatiemodelering die dit mogelijk maakt hebben we dus 'nanopublicatie' gedoopt (waarvoor dank aan Jan Velterop). Nanopublicaties vormen de noodzakelijke brandstof, of zo u wilt, het substraat voor Social Machines.

Een nanopublicatie is in feite de kleinst denkbare, betekenisvolle publicatie, namelijk een bewering met de bijbehorende 'provenance'. Provenance is de informatie over '*waar deze nanopublicatie vandaan komt*'. Net als een klassiek artikel heeft een nanopublicatie auteurs, een unieke referentie, een publicatietijdstip enzovoort. Ook kan er 'context' aan een bewering meegegeven worden. Bijvoorbeeld: deze reactie loopt alleen bij een bepaalde temperatuur. Op het kaartje dat u zojuist hebt ontvangen vindt u een 'artistieke impressie' van hoe dit er ongeveer uitziet.



Nu kan dus elke bewering in het Explicitome uniek worden herkend, maar de computer kan ook identieke beweringen herkennen die uit verschillende bronnen komen. Het Explicitome bevat bijvoorbeeld tienduizenden beweringen die uit artikelen, blogs en databases zijn verzameld en die in feite allemaal zeggen: '*malaria wordt overgebracht door muskieten*'. We kunnen nu deze 'provenance' informatie gebruiken om een zogenaamde 'kernbewering' te construeren die in het geval: *malaria wordt overgebracht door muskieten* een heel hoog betrouwbaarheidsgehalte heeft. Dit is ook direct de eerste vorm van datareductie, met in feite *informatiewinst.*
Morgen is het Explicitome van 10^{14} nanopublicaties alweer met een paar honderdduizend, en op piekdagen met miljoenen, beweringen gegroeid. Als we nu even teruggaan naar onze voorbeeldbewering '*malaria wordt overgebracht door muskieten*'

dan zien we dat die eigenlijk drie begrippen bevat, namelijk *Malaria*, *Transmissie* en *Muskieten*. Als ik nu dezelfde bewering in het Engels stel: *malaria is transmitted by mosquitoes*' of in het Frans: *le paludisme est transmis par des moustiques*, dan zou de computer direct moeten kunnen zien dat deze uitspraken in feite allemaal hetzelfde beweren. Daarvoor moet elk uniek begrip een uniek nummer krijgen. Op die manier worden alle varianten van de bewering '*malaria wordt overgebracht door muskieten*' omgezet in een zogenaamde 'triple' van drie unieke getallen die altijd hetzelfde zijn. Triples zeggen dus eigenlijk: 'A is op deze manier verbonden met B'. Soms bestaat de kernbewering van een nanopublicatie overigens uit meerdere triples. Nanopublicaties zijn dus heel goed door de computer te interpreteren, maar ze kunnen ook op elk gewenst moment aan mensen worden getoond en zelfs in hun eigen taal.

Nu kan het feit dat ik de bewering '*malaria wordt overgebracht door muskieten*' in deze rede, in Leiden, op 25 februari om precies 16.4x voor de zoveelste keer heb gedaan ook in een set van triples worden weergegeven. Het begrip 'Barend Mons' (zoals die hier voor u staat) is namelijk uniek en heeft dus een uniek nummer. Ook de datum is uniek en… Leiden is een unieke stad. Dus is de kernbewering: *malaria wordt overgebracht door muskieten*', gekoppeld met de 'provenance' dat die bewering door Barend Mons in Leiden op een bepaald tijdstip is gedaan. Dat is dus een *nieuwe* nanopublicatie. De kernbewering zelf is echter *allesbehalve nieuw*. Computers kunnen heel goed met cijfers omgaan en detecteren dus onmiddellijk dat dit de zoveelste keer is dat deze bewering in het systeem voorkomt.

Stel dat deze voordracht onmiddellijk door de computer zou worden geanalyseerd, dan zou door wat ik tot nu toe heb gezegd al 9 keer een nanopublicatie met dezelfde kernbewering zijn gedetecteerd met als enige verschil het tijdstip waarop die werd uitgesproken. De Social Machine zou die echter allemaal 'negeren'. Het feit dat ik vandaag nog maar weer eens in het Nederlands 7 maal heb gezegd dat malaria door muskieten wordt verspreid, zou natuurlijk nergens iets moeten veranderen in het 'begrippen netwerk' rond malaria. *Dit is cruciaal*. Onze voorbeeldbewering is namelijk zo 'waar' als iets in de wetenschap ooit kan worden. Als er op dit zelfde moment echter ergens in de wereld iemand in het Hindi zou bloggen dat malaria ook kan worden overgebracht door *zandvliegen* dan zouden er eigenlijk in deze zaal minimaal twee Smartphones moeten gaan trillen omdat dat groot nieuws zou zijn (maar hoogstwaarschijnlijk niet waar). Als Hugo van der Kaay en Chris Janse hun Smartphones voelden trillen zouden zij onmiddellijk hun ongeloof over deze bewering moeten kunnen intypen. Het is dus van het grootste belang voor een Social Machine om te 'weten' of een bewering nieuw is, maar ook of het herhalen (citeren zeg maar) van de bewering de waarschijnlijkheid ervan nog verhoogt of verlaagt. Een Social Machine moet dus een enorm aantal bestaande en nieuwe beweringen per seconde verwerken en beslissen of het nut heeft om mensen op de hoogte te stellen van nieuwe of veranderde inzichten.

Zo krijgt dus elke 'kernbewering' een dynamisch uitgerekend 'gewicht' dat in feite het 'wetenschappelijke betrouwbaarheidsgehalte' weergeeft. Deze factor kan gebaseerd zijn op 1 tot en met vele duizenden nanopublicaties met dezelfde kernbewering. Door soms honderdduizenden nanopublicaties met een identieke kernbewering terug te rengen naar 1 kern-nanopublicatie met slechts een gewicht eraan hebben we een enorme eerste *datareductie* stap gerealiseerd. Onze schatting is dat deze 'eerste ZIP' de druk op de Social Machine al met een factor duizend verlaagt. Met andere woorden, de Social Machine die wij voor de biomedische wetenschappen aan het bouwen zijn hoeft dus niet door 10^{14} nanopublicaties heen te rekenen maar 'slechts' door zo'n 100 miljard kernbeweringen. Frank van Harmelen kan u vertellen dat dit met de huidige redeneertechnieken en flink wat computergeweld nu al een behapbaar aantal is. In de praktijk hoeven we echter door nog veel minder beweringen heen te rekenen, omdat eigenlijk alleen de delen van het begrippennetwerk die logisch beïnvloed worden door

de vandaag binnenkomende kernbeweringen nog hoeven te worden doorgerekend. Aangezien ik hier alleen maar ouwe koek sta te vertellen komt de Social Machine door deze oratie dus niet eens uit de slaapstand.

De Knowlet

Toch heeft de Biosemantiek groep een nog veel grotere datareductie weten te bewerkstelligen. Vergelijk het (alhoewel deze vergelijking mank gaat) met een 'ZIP file van een Zip file'. Stelt u zich dit als volgt voor: we stellen ons de kernbeweringen die als 'onderwerp' (eerste nummertje zeg maar) '*malaria*' hebben even voor als een stokje met drie bolletjes. Dus bijvoorbeeld: **malaria** wordt overgebracht door muskieten, maar ook: '**malaria** doodt 3 miljoen mensen per jaar', '**malaria** komt voor in Afrika' en '**malaria** wordt veroorzaakt door parasieten van het geslacht *Plasmodium*'. Dan kunnen we alle stokjes in het systeem met als eerste bolletje '**malaria**' clusteren. Zo'n cluster, met in het centrum dus in dit geval het begrip 'malaria' noemen we een **Knowlet**. De Knowlet bevat dus eigenlijk alle kernbeweringen ooit gedaan over het begrip '**malaria**' die het systeem kent. Een Knowlet bevat typisch honderden tot duizenden kernbeweringen.

Nu een paar ontnuchterende, maar toch ook heel geruststellende cijfers: er zijn 'slechts' zo'n 3 miljoen relevante begrippen in de levenswetenschappen, waarover voldoende is gepubliceerd om er een zinvolle Knowlet van te maken. Met andere woorden, we hebben maar een paar miljoen Knowlets.

En wat heb je daar dan aan?

De belangrijkste missie van de Biosemantiek groep op dit moment is om met die Knowlets voorspellingen te doen over de verbanden tussen begrippen die nog niet eerder zijn gevonden en gepubliceerd. Ik ga proberen dat proces uit te leggen, waarbij ik de wiskundige bewerkingen die hier onder de motorkap voor nodig zijn maar even laat voor wat ze zijn.

Neem de Knowlet van 'malaria' met duizenden kernbeweringen. Deze Knowlet wordt aangetrokken of afgestoten door alle andere Knowlets in de ruimte. Nu bestaat er een andere Knowlet, die van het begrip 'Tegafur', een geneesmiddel dat oorspronkelijk op de markt is gebracht voor kanker, ook met zo'n duizend kernbeweringen. De bolletjes aan het eind van elk stokje zijn ook hier weer begrippen. Gemeenschappelijke begrippen tussen de Knowlets van malaria en Tegafur zorgen dat die elkaar aantrekken. Zo kunnen dus twee begrippen (in dit geval 'malaria' en 'Tegafur') op een bepaald moment steeds dichter bij elkaar komen in de ruimte van het Concept Web, doordat ze meer begrippen gemeenschappelijk hebben. Net zoals twee mensen die elkaar nog niet als vriend hebben op Facebook, door steeds meer gemeenschappelijke vrienden aan elkaar worden gekoppeld.

Op een gegeven moment komen de Knowlets van malaria en Tegafur zo dicht bij elkaar dat er letterlijk een belletje gaat rinkelen en de Smartphones van mensen die op zoek zijn naar nieuwe malariageneesmiddelen gaan trillen. Er verschijnt vervolgens een nanopublicatie op het scherm die Tegafur suggereert als mogelijk anti-malariamiddel. *Let wel, dit was tot op dat moment nog nergens expliciet gesuggereerd.* Er is ook nog geen laboratoriumexperiment aan te pas gekomen.

Dit is geen verzonnen voorbeeld, maar het komt uit een van onze artikelen in 2008. Helaas bleek het molecuul Tegafur te groot te zijn om de rode bloedcellen binnen te komen waarin de malariaparasiet zich schuilhoudt. Ook dat kan ik weer als nanopublicatie publiceren zodat niemand deze op zich logische aanname opnieuw doet en een hoop tijd en geld in een kansloos experiment gaat steken. Als we in 2008 al de grootte van elk molecuul en die van de poriën van een rode bloedcel hadden opgenomen in de Knowlet, dan zou het systeem de suggestie niet eens hebben gedaan.

Natuurlijk is het statistisch associëren van Knowlets niet een vervanger voor andere vormen van computer reasoning. Hypothetische verbanden die uit associatief redeneren komen, kunnen altijd in meer detail worden beredeneerd in een veel kleiner afgebakend deel van het Explicitome.

eScience heeft behoeft aan verschillende lagen van redeneren. Het is eigenlijk vergelijkbaar met het ontdekken van een afwijkend groeipatroon in een korenveld vanuit een helikopter, dat je nooit zou zien als je er middenin liep. Op grond van die observatie vanuit de lucht rijst het vermoeden dat daar de ruïnes van een oud Romeins fort liggen. Om dat echt aan te tonen zul je vervolgens moeten landen (met beide benen op de grond), een schep pakken en gaan graven. Als je de eerste stenen vindt, dien je vervolgens nog experimenten te doen om aan te tonen dat het hier inderdaad een gebouw uit de Romeinse tijd betreft.

Dit is allemaal vandaag of in de zeer nabije toekomst al operationeel. Maar ik wil niet eindigen zonder u een blik in de iets verdere toekomst van de Biosmantiek te gunnen. Ik ben bang dat wat ik nu nog ga zeggen makkelijker te accepteren zal zijn voor de leken in de zaal dan voor de meeste collega's in de flanken. Ik weet dat zij gelukkig te beschaafd zijn om op te staan en weg te lopen zoals echte kerkgangers die het niet met de dominee eens zijn soms doen, maar kijkt u vooral af en toe goed naar hun wenkbrauwen.

'Wetenschap' is eigenlijk altijd al een min of meer elitaire zaak geweest. Toen de boekdrukkunst werd uitgevonden waren er al wetenschappers die voorspelden dat dit een ramp zou veroorzaken omdat 'jan en alleman' zich met hun heilige vakgebied zou gaan bemoeien. Toen het Internet ontstond en nog erger, Wikipedia en Social Networks, wemelde het van de waarschuwingen dat dit de integriteit van de wetenschap zou bedreigen. Totdat Wikipedia onherroepelijk alle klassieke encyclopedieën simpelweg achter zich liet en nu zelfs topwetenschappers toch 'eerst even snel in Wikipedia kijken' als ze met een nieuw begrip worden geconfronteerd. Anders dan nog maar vijf jaar geleden zijn er nu Appstores, Twitter en citeerbare nanopublicaties. Ik ben er diep van overtuigd dat deze drie elementen een volgende revolutie teweeg zullen brengen in de manier waarop wij collectief wetenschap bedrijven. Ik sta hier dan ook als een man met hele hoge verwachtingen voor de komende jaren,

Binnenkort zullen dagelijks miljoenen nieuwe nanopublicaties *bijna real-time* naar miljoenen Smartphones van wetenschappers, patiënten en andere geïnteresseerde leken worden gestuurd ter informatie en commentaar. Wetenschappers zullen alleen nanopublicaties krijgen die 'nieuw voor hen zijn' omdat hun Smartphone ook hun eigen Knowlet bevat (alles wat zij over bijvoorbeeld hun favoriete gen al weten). Zo zullen dus nanopublicaties uit patiëntenblogs, datasets en andere bronnen van informatie die vandaag de dag grotendeels buiten het radarscherm van de geleerden vallen onmiddellijk op de telefoon of de tablet van de wetenschapper verschijnen en becommentarieerd worden en soms weer tot nieuwe hypotheses leiden. Dat is de Social Machine in volle actie.

Misschien nog wel belangrijker is dat we van elke genanopubliceerde bewering weten, door wie, wanneer en in welke context deze bewering is gedaan. Daardoor zijn beweringen van patiënten, artsen en wetenschappers duidelijk te onderscheiden in het Explicitome en kunnen ze dus ook letterlijk 'op waarde kunnen worden geschat'. Daardoor kunnen nanopublicaties op hun beurt net zo belangrijk worden voor de carrière van wetenschappers als artikelen vandaag. Hierdoor is het eindelijk technisch mogelijk om het hopeloos verouderde en zelfs verlammende systeem van de Journal Impact Factor aan te passen voor eScience.
Denk niet dat dit verre toekomstmuziek is. Later dit jaar zal de eerste Smartphone app namens Nature Genetics worden gelanceerd, en deze wordt in samenwerking tussen het LUMC en een Nederlands spin off bedrijf gemaakt!

Zoals eerder beloofd: het goede nieuws voor klassieke wetenschappelijke uitgevers is dat het artikel als *beschrijving van de details van een experiment* en de *retoriek rond het bereiken van de conclusies* die uiteindelijk als nanopublicaties het licht zien altijd van belang zal blijven als referentiepunt voor mensen in het Social Machine tijdperk. Volgens het principe: *'waarom zou ik deze losse nanopublicatie eigenlijk*

geloven?'. Ik voorspel dat wij in de komende jaren volledig nieuwe en onderhoudbare businessmodellen zullen zien ontstaan rond deze *mind-machine interaction* in eScience.

Vandaar ook weer mijn titel: Kennis is als Liefde….*men wordt van het delen niet minder.*

Tenslotte mijn meest controversiële bewering tot nu toe: het is onvermijdelijk dat zowel biologische als computationele technieken op zeer korte termijn zoveel hypothetische nanopublicaties zullen produceren dat die onmogelijk allemaal experimenteel getest kunnen worden. In feite worden onze recente artikelen soms in eerste instantie op die gronden afgewezen. Dit is echter een conservatieve en elitaire drogreden: *Zelfs wij, biologen zullen moeten accepteren dat grote hoeveelheden door de computer geformuleerde hypotheses noodgedwongen als 'voorlopige waarheid' zullen moeten functioneren zonder eerst in het laboratorium gevalideerd te worden.* Tenzij we natuurlijk naast datakerkhoven nu ook hypothesekerkhoven willen accepteren.

Als we zo doorgaan met data produceren dan zullen we die *of* heel snel weer moeten weggooien, *of* Nederland moet binnen enkele jaren haar volledige bruto nationaal product aan databeheer uitgeven. U begrijpt allemaal dat dit niet gaat gebeuren, al was het alleen maar omdat dataproduceren ook geld kost, hoewel, slechts een fractie van databeheer en -analyse.
Het moge echter uit mijn betoog wel duidelijk zijn dat de schrijnende onderwaardering voor bioinformatica en goed data-rentmeesterschap een grote bedreiging vormt voor de volgende fase van de moderne wetenschap. Het is dan ook vijf voor twaalf als het gaat om maatregelen die de transitie naar 'eScience' mogelijk maken. Het Dutch TechCentre voor de Life Sciences met daarbinnen het programma voor het Data Integration and Stewardship Centre DiSC, is een cruciale ontwikkeling. Dit federatieve initiatief zal niet alleen binnen Nederland onontbeerlijk zijn, maar ook voor internationale kennisinfrastructuren zoals ELIXIR, BBMRI, EATRIS, EuroBioImaging, ISBE en alle grote internationale projecten in bijvoorbeeld het Innovative Medicines Initiative.

Nederland wil een '*top* kenniseconomie zijn, we noemen alles '*top*sector' en '*top* instituut', maar zonder een *top*-beleid op het gebied van linked open datascience en zonder de duizend dataspecialisten op te leiden die we anders in 2018 al tekort zullen hebben zal dit *niet* gaan lukken.

Kortom: indien we als Nederlandse wetenschap als een topzeiler 'aan de wind' willen blijven varen, dan zal effectief omgaan met 'het nieuwe goud', Big Data, een centrale rol in moderne eScience moeten vervullen. In een wereld waar dit jaar het aantal mobiele apparaten groter zal zijn dan het aantal mensen is *mobile science* niet meer weg te denken, hoe eng het ook klinkt.

Ik heb gezegd.

Prof.dr. B. Mons (Den Haag 1957)

2013	Bijzonder Hoogleraar Biosemantiek, Leiden University Medical Center (LUMC)
2012 - heden	eScience Integrator for the Life Sciences, Netherlands eScience Centre
2010 - heden	Wetenschappelijk Directeur Netherlands Bioinformatics Centre
2009	Oprichter en voorzitter Concept Web Alliance.
2003 - 2013	Universitair Hoofd Docent afdeling Humane Genetica LUMC
2000 - 2003	Universitair Hoofddocent Erasmus Medical Center Rotterdam
2005 - 2008	Mede Oprichter Knewco. Inc.
2000 - 2005	Mede oprichter Collexis BV (geacquisiteerd door Elsevier)
1996 - 2000	Senior Adviseur NWO Medische Wetenschappen en WOTRO
1993 - 1996	Seconded National Expert to the Europese Commissie INCO-DC programma
1990 - 1993	Hoofd van de Malaria Groep TNO Primaten Centrum.
1984 - 1993	Hoofd van de Malaria Research Group, afdeling Parasitologie, LUMC
1986	Promotie Faculteit der Wiskunde en Natuurwetenschappen
1982	Doctoraal Biologie (cum laude).
1978-1982	Leraar Biologie (part-time) middelbare school (eerste graads onderwijsbevoegdheid)

Biosemantiek is een nieuwe discipline die zich richt op twee uitdagingen in de moderne levenswetenschappen. Ten eerst het verwerken van de steeds grotere datastromen die door de zogenaamde High-Throughput Technologie worden geproduceerd. Deze data moeten niet alleen door mensen maar ook door computers worden begrepen. Ten tweede genereert de Biosemantiek met behulp van de computer hypotheses die gebaseerd zijn op deze enorme databestanden. Door een combinatie van tekst en data mining, computer reasoning en vergelijkingen met alles wat al eerder gepubliceerd is in grote wereldwijde databases. Zowel bestaande als nieuwe vindingen worden gepubliceerd in een zowel voor de computer als voor mensen leesbaar format, de zogenaamde nanopublicaties. Het is aangetoond dat op deze manier voorspellingen kunnen worden gedaan van bijvoorbeeld eiwit-eiwit interacties en genen die betrokken zijn bij erfelijke ziekten.

Zie ook:
www.biosemantics.org
www.nanopub.org

Universiteit Leiden