Dies Natalis 2009

Web & wetenschap op weg naar science 2.0

Het Web
als wetenschaps-
versneller

vrije Universiteit *amsterdam*

Rede uitgesproken ter gelegenheid van
de 129ste stichtingsdag van de Vrije Universiteit
op dinsdag 20 oktober 2009

Prof.dr. Frank van Harmelen

connect & reflect
Jaarthema VU 2009-2010: netwerken

Dies Natalis 2009
Web & wetenschap:
op weg naar science 2.0

Het Web als wetenschapsversneller

Het web zal in de nabije toekomst een veel diepgaander invloed op de wetenschap hebben dan we op het moment om ons heen zien

Web en Wetenschap: op weg naar science 2.0

Met Sir Tim Berners-Lee in ons midden ligt het voor de hand om ook in de Dies lezing te reflecteren op het World Wide Web. En omdat we de geboortedag vieren van een academische instelling ligt het voor de hand om te reflecteren op de rol van het WWW voor de wetenschap.

Het is moeilijk te overschatten hoe transformatief het Web geweest is in allerlei sectoren van onze maatschappij. Ik voorspel mijn studenten dat als we over 50 jaar terugdenken aan de jaren '90, we ons nog maar twee dingen zullen herinneren: de hereniging van West- en Oost-Europa en de opkomst van het World Wide Web. Al die andere zaken zijn we al lang vergeten. Denkt u nog wel eens aan de Golf Oorlog? Aan Dolly het gekloonde schaap? Aan de landing op Mars? Maar het Web en het Internet in bredere zin, verandert nog elke dag uw leven. Het verandert uw privé leven (mijn kinderen communiceren met hun grootouders in Schotland via het Internet), het verandert ons commerciële leven (we zien zelden nog een bankfiliaal van binnen), het verandert ons culturele leven (we luisteren muziek en bekijken films online, we gaan steeds meer digitale boeken lezen).

En natuurlijk heeft het Web ons wetenschappelijke leven veranderd. Als wetenschappers bloggen en hyperlinken we dat het een lieve lust is. Maar dat is allemaal oude koek. Ik ga beweren dat het Web in de nabije toekomst een veel diepgaander invloed op de wetenschap zal hebben dan we op het moment om ons heen zien.

ad Schotenij opinionem.

Primum, in Algebra falsum verum ex falsis non elicitur, sed omnino nihil ut praeter hypothesin argumentationem. Numerus v.g. quem quaero ut per 2 multiplicatus dabit 6. erit 6. ergo 6 per 2 si dividatur oritur quaesitus numerus. Nihil autem hic falsi fingitur. Nam signum quod pono, scribitur loco vocabuli (numerus) sed non supponitur illud signum esse numerum quaesitum.

In regula falsi non aliter. Ponamus in regula falsi duplicis positionis, quid hic falsi fingitur, nisi [illegible] pro numero quaesito, qui non sit verus [illegible] illum qui non sit esse quaesitum numerum. Hic contingere videtur, nisi hoc [illegible] ubi quos ita multiplicamus et divisores quaesito praescribitur. Deinde [illegible] interim certum per regulam numerum [illegible] autem numeros ut ut eos pro numero quaesito [illegible] quidem magnitudinem signa in algebra, diuerso modo dicam.

Ad syllogismum primum respondeo, videlicet ex falsis vera directe elici. mihi sic videtur. Ex falsis vera elici quando positis falsis aliquibus hypothesi tanquam si vera esset, aliquid vere esse, ita demonstratur, ut persuadeat lectori. In demonstratione istud verum esse licet hypothesin deprehendat falsam. Nam si quae talis hypothesi utebatur. Quod omnis figura curvilinea in data parte cava est, si basin et altitudinem cum rectangulo aequalem habet, sit ejus rectanguli subsesquialtera

Terug naar vroeger

De beste manier om te voorspellen is achteruitkijken. Voor de niet-wetenschappers in het publiek: dat is wat wetenschappers bedoelen als ze de technische term “extrapoleren” gebruiken.

Op de afbeelding hiernaast ziet u een brief die Christiaan Huygens schreef in 1652 aan zijn vroegere leermeester Frans van Schooten, indertijd Hoogleraar Nederduytsche Mathematique te Leiden. Huygens betwist in deze brief onder andere een aantal bevindingen van Descartes. Hij moppert dat Descartes “vaak zonder bewijs zaken pleegde te beweeren”.

En dit was slechts een van de vele briefwisselingen van Huygens. Hij was een toegewijde briefschrijver die onder andere correspondeerde met de grote wiskundige Mercenne, met Antoni van Leeuwenhoek en met de grote Rene Descartes. Dat waren nog eens tijden... In zijn vruchtbare leven schreef Huygens in totaal maar liefst bijna 3.000 brieven.

Die brieven deden er vaak weken over voordat ze bezorgd werden en het antwoord liet dan ook vaak op zich wachten. In vergelijking met de huidige wetenschappelijke briefwisseling is dat een ongelooflijk laag tempo. Wanneer schreef u voor het laatst een brief aan een collega? Een maand geleden? Een jaar geleden? Maar, wanneer schreef u voor het laatst een email aan een collega? Het zou me verbazen als uw antwoord is ”een dag geleden“. De meerderheid van de antwoorden zal luiden: “een uur geleden”.

Hetzelfde geldt voor andere aspecten van onze wetenschappelijke arbeid. De frequentie waarmee we wetenschappelijke artikelen en boeken publiceren, lezen en citeren is enorm toegenomen. Wanneer heeft u voor het laatst een wetenschappelijk artikel gelezen? Als u een actieve wetenschapper bent, en velen van ons zijn dat, dan zal uw antwoord variëren van “een uur geleden” tot “een dag geleden”, op zijn hoogst “een week geleden”. Uw antwoord zal niet zijn “een maand geleden”. Echter, als ik u vraag: wanneer was u voor het laatst in de bibliotheek om een artikel op te zoeken, dan zal het antwoord van velen van u anders zijn. De belangrijkste dienst die onze bibliotheek mij als wetenschapper verleent is de toegang tot de talloze online verzamelingen die via het Web beschikbaar zijn.

Het Web werkt als een enorme communicatie-versneller voor wetenschappers. En dat zou ons ook niet moeten verbazen. Wat we nu het WWW noemen is immers ooit begonnen als een manier om wetenschappers (in CERN) eenvoudiger informatie te laten uitwisselen. Om mijn natuurkunde collega’s in de Faculteit wat te plagen zou ik kunnen zeggen dat het Web veruit de meest praktische spin-off is van al die jaren elementaire deeltjesonderzoek in CERN, maar dat zou flauw zijn.

Het is duidelijk, de onderlinge communicatie van wetenschappers is drastisch veranderd: we sturen e-mails in plaats van brieven, we downloaden publicaties in plaats van dat we ze op papier lezen. Maar feitelijk zijn dat oppervlakkige veranderingen: het medium is weliswaar veranderd, en daarmee de snelheid, maar feitelijk doen we nog steeds hetzelfde. Is dat het enige dat het Web ons als wetenschappers te bieden heeft?

Het zal Mendeleev ongetwijfeld maanden gekost hebben om sommige van zijn bronnen te vinden en te analyseren.

Meer dan slechts snelle communicatie?

Ik beweer inderdaad dat dit niet alles is wat het Web de wetenschap te bieden heeft. Ik zal u laten zien dat het Web een veel verdergaande invloed zal hebben op de wetenschap dan alleen de manier hoe we onderling alledaags communiceren.

Eén van die veranderingen, namelijk hoe we omgaan met onze experimentele data, is al aan het gebeuren. Daar zal ik eerst iets van laten zien.

De tweede verandering betreft de hele notie van wat een wetenschappelijke publicatie is of zou moeten zijn. Zover zijn we nog lang niet, en hiervoor ga ik dus de toekomst voorspellen, altijd gevaarlijk...

Data Driven Science

De traditie vereist dat de wetenschappelijke methode altijd hypothesegedreven is: eerst voorspelt een wetenschapper het goede antwoord, daarna verzamelt ze experimentele data om haar voorspelling te testen.

Echter, in de nieuwe data-gedreven benadering beginnen we juist met het verzamelen van data en gaan dan kijken wat die data ons vertelt. Deze benadering wordt doenlijk wanneer experimenten enorme hoeveelheden data produceren, genoeg data om meer hypotheses te testen dan welke wetenschapper ook ooit zou kunnen bedenken.

In principe is zulke data-gestuurde wetenschap helemaal geen nieuw fenomeen. Veel klassieke resultaten in de wetenschap zijn voortgekomen uit een analyse van beschikbare gegevens in de open literatuur. Een klassiek voorbeeld is Mendeleev's periodiek systeem der elementen. In zijn Faraday lezing voor de Fellows van de Chemical Society in 1889 in het Theatre of the Royal Institution vertelde Mendeleev het als volgt:

> *"Mijn periodiek systeem der elementen was dus een direct resultaat van een grote verzameling feiten en generalisaties die zich tegen 1870 opgestapeld had. Het systeem is de belichaming van al die data in een min of meer systematisch vorm."*

De data waarop Mendeleev zijn periodiek systeem baseerde waren in het geheel niet systematisch vergaard maar stonden in de chemische literatuur verspreid, gepubliceerd in verschillende talen en met gebruik van verschillende symbolische notaties. Het zal Mendeleev ongetwijfeld maanden gekost hebben om sommige van die bronnen te vinden en te analyseren. Moderne *data-driven science* doet dat anders en sneller: resultaten worden opgeslagen in grote databases en via het Web beschikbaar gesteld aan collega wetenschappers.


HELPDESK: 86882
Applications Places System
File Edit View History Bookmarks Tools Help
http://localhost:8080/search?
Java Eclipse python Lucene Navigator WindGURU: Netherl...
Blogo.nl
AmCAT Navigator - Index Se...
Report
User List
Ontology
Help »
Admin
Full User List
Site Status
Logs
Recent Projects
VN project
Media
All Media
Date
All Dates
Action
Show Summary
Show Table
Show Graph
Show Cluster Map
Save as CSV
Submit
Total number of articles: 8255
4526
mensen* (6211)
545
verhagen (446)
279
handel* (1327)
738
Downloads
Done
han2009.zip
Aqua Data Studio ...
Nel Ruigrok | fijn is ...

Collega's in de levenswetenschappen vertellen me dat er in dat brede vakgebied zo ongeveer elke maand een nieuwe database met wetenschappelijke gegevens online bij komt. Niet alleen is de hoeveelheid data niet meer bij te houden, het is zelfs niet meer bij te houden waar de data beschikbaar is!

Dat levert natuurlijk een nieuw probleem op: wie gaat al die data analyseren. Tot u spreekt een Informaticus, dus u kunt het antwoord raden. Natuurlijk gaat de computer dat doen.

Peter Murray Rust van het Department of Chemistry in Cambridge zegt het als volgt:

"Wat we daarom nodig hebben zijn computersystemen die automatisch de chemische literatuur kunnen lezen, die de data kunnen verzamelen, betekenis kunnen toekennen, en die het mogelijk maken om wetenschappelijke hypotheses te kunnen toetsen. Het zou mogelijk moeten zijn om met behulp van zo'n systeem patronen of ongebruikelijke observaties te detecteren en daaruit nieuwe hypotheses te kunnen vormen. Onze stelling is dat de huidige wetenschappelijke literatuur een vracht aan nieuwe resultaten bevat, als het maar mogelijk zou zijn om die literatuur op zo'n manier te analyseren."

Maar in weerwil van Murray Rust's verzuchting wordt het overgrote deel van alle wetenschap nog steeds gepubliceerd in gefragmenteerde vorm en niet in de vorm van de systematische dataverzamelingen zoals we die vinden bij bijvoorbeeld astronomie, deeltjesfysica of (delen van) genomics. Als voorbeeld schat Murray Rust dat er jaarlijks 2 miljoen chemische verbindingen gepubliceerd worden zonder enige machinaal toegankelijke semantiek. Als gevolg daarvan vereist het enorme menselijke inspanning om iets nuttigs met al die data te doen en staat de computer buiten spel.

En dit is niet bedoeld om met de beschuldigende vinger naar de chemici te wijzen: hetzelfde geldt voor vrijwel alle takken van wetenschap, met slechts enkele gunstige uitzonderingen.

Katy Börner van het Cyberinfrastructure for Network Science Center in Indiana zegt het als volgt:

"De verzamelde menselijke kennis is opgeslagen in een exponentieel groeiend aantal artikelen, boeken, e-mails en allerlei andere formaten. Mens noch machine kan deze informatie verwerken in deze enorme hoeveelheid en in deze vorm. Het gevolg is dat veel kennis opnieuw wordt uitgevonden, dat onderzoek wordt gedupliceerd tussen wetenschappelijke disciplines, of dat de kennis simpelweg na een korte tijd alweer verloren gaat".

De moderne *data-gestuurde* benadering is momenteel het best te zien in de levenswetenschappen, waar standaarden zijn ontwikkeld voor datarepresentatie en waar uitgevers zelfs eisen dat auteurs hun data ter beschikking stellen. De levenswetenschappen zijn op het moment het beste voorbeeld van data-gestuurde wetenschap, en een hoop biowetenschappers doen een hoop goede wetenschap zonder dat ze ooit nog in een laboratorium hoeven te komen. Zoals gezegd lopen ook grote delen van de astronomie en de deeltjesfysica hierin voorop.

Ik heb tot nu toe slechts voorbeelden uit de natuurwetenschappen genoemd, maar er beginnen ook steeds meer gammawetenschappen te komen die sterk *data-driven* zijn.

Een goed voorbeeld hiervan is de samenwerking binnen het Netwerk Instituut, één van de interfacultaire onderzoeksinstituten van de VU, waar informatici samenwerken met onder andere sociale wetenschappers. Die intensieve samenwerking over een aantal jaren maakt ook binnen de gammawetenschappen een sterk data-gedreven benadering mogelijk. Laat ik u kort twee voorbeelden schetsen:

Prof. Jan Kleinnijenhuis beoefent Communicatie Wetenschap. Hij bestudeert onder andere hoe de media rapporteren over bijvoorbeeld de politieke actualiteit en hoe politici op hun beurt weer reageren op die rapportages in de media. Een elementaire dataset voor hem zijn natuurlijk de mediaberichten zelf. Decennia lang huurden onderzoekers zoals Kleinnijenhuis kleine legertjes mensen in om kranten artikelen te *annoteren*: precies aangeven in welk artikel of zelfs in welke zin van welk artikel, wat gezegd wordt door welke politicus over welke andere politicus. U kunt zich voorstellen dat dat een enorm monnikenwerk is. Het is duur om die annotators te betalen, waardoor data schaars is. En die annotators zijn ook maar mensen, dus de data is ook nog vaak onbetrouwbaar. Met behulp van technieken die mijn promovendus Wouter van Atteveld heeft ontwikkeld lukt het nu om die krantenannotatie door de computer te laten doen: grootschalige syntactische analyse van krantenartikelen. Vele duizenden artikelen worden "gelezen" door een supercomputer cluster, wat leidt tot veel grootschaliger datasets waarop Kleinnijenhuis en zijn collega's hun analyses kunnen doen. Zo kunnen ze bijvoorbeeld heel aardig laten zien dat wat politici tijdens de verkiezingscampagne over elkaar zeggen, een voorspeller is van de uitkomst van de kabinetsformatie. Kunnen we misschien die langdradige kabinetsformatie de volgende keer overslaan. Ik zeg nadrukkelijk: Kleinnijenhuis én zijn collega's. Want niet alleen is het belangrijk om zulke grote datasets te maken. Het is bij data-driven science minstens even belangrijk dat zulke datasets *via het Web* gedeeld worden met anderen. En met behulp van moderne technieken die voor een groot deel hier aan de VU ontwikkeld zijn, begint ook dat te lukken.

Een tweede voorbeeld, ook in de Faculteit der Sociale Wetenschappen, is het werk van prof. Peter van den Besselaar. Hij bestudeert een eigenaardige diersoort. Hij bestudeert namelijk u, dames en heren wetenschappers. Prof. van de Besselaar is geïnteresseerd in het gedrag van wetenschappers en dan voornamelijk hun groepsgedrag. Of zou ik moeten zeggen hun kuddegedrag? Hij bestudeert hoe vakgebieden zich afsplitsen of juist meer met elkaar verweven raken, hoe schoolvorming binnen een vakgebied terug te vinden is in de literatuur, etcetera. Tot op heden was daar eigenlijk maar één soort dataset voor, namelijk de databases met publicatie-gegevens: wie publiceert er samen met wie, en wie citeert wie. Maar die dataset heeft zeer sterke beperkingen. Ten eerste loopt die dataset altijd een paar jaar achter de werkelijkheid aan: als informatici veel gaan samenwerken met biologen dan markeert dat wellicht het ontstaan van zoiets als Bio-informatica, maar de eerste publicaties daarover en zeker de eerste tijdschriftpublicaties, verschijnen pas vele jaren later. Daarnaast is die dataset natuurlijk ook heel beperkt. Wetenschappers doen veel meer dan alleen maar publiceren: ze gaan naar congressen, ze zitten in beoordelingscommissies, ze verschijnen in de media, ze bloggen, etcetera. Al die activiteiten waren vroeger moeilijk te traceren, maar tegenwoordig laten al die activiteiten sporen na... op het Web! Dat betekent dus dat we ook hier weer het Web kunnen gebruiken als een belangrijke databron die een wetenschap, in dit geval de wetenschapsdynamica, veel meer data-driven maakt dan voorheen. De samenwerking met collega Van Den Besselaar is pas recent begonnen dus ik kan u nog geen resultaten melden, maar we hebben beiden hoge verwachtingen.

We hebben nu uitgebreid gezien dat het Web niet alleen de rol speelt van publicatie medium voor data, maar dat het ook de wellicht nog veel belangrijker rol speelt van *observatorium*: het is de plek waar data verkregen wordt. In het bijzonder voor de sociale wetenschappers beloofd het Web te worden wat de telescoop is voor de astronomen. Zoals u hebt gezien, lopen de VU als universiteit en het Netwerk Instituut in het bijzonder voorop bij deze ontwikkelingen.

Na al deze voorbeelden zal het duidelijk zijn wat de enorm toegenomen betekenis is van grote en wereldwijd gedeelde datasets voor zeer uiteenlopende takken van wetenschap. Het zal ook duidelijk zijn dat

het Web een cruciale rol gaat spelen en in een aantal takken al speelt. Als publicatiemedium voor het uitwisselen van data en als *observatorium*, als middel om datasets te verkrijgen.
Het zou dan ook duidelijk moeten zijn dat het publiceren en onderhouden van een dataset een even belangrijke bijdrage in de wetenschap zou moeten zijn als een wetenschappelijke publicatie. En net zoals belangrijke publicaties veel geciteerd worden, worden belangrijke datasets veel gebruikt. Download-statistieken zouden dus helemaal niet moeten misstaan op een wetenschappelijk CV. Dat zou een belangrijke verandering zijn in de cultuur van wetenschappelijk publiceren.

Dat brengt ons bij het laatste onderwerp waar het Web een diepgaande invloed zal hebben op de wetenschap: de toekomst van de wetenschappelijke publicatie.

De toekomst van wetenschappelijke publicaties

De belangrijkste vorm van communicatie in de wetenschap is nog steeds de wetenschappelijke publicatie.

Meestal is dat een tijdschrift artikel; in snelle en jonge disciplines zoals de Informatica is dat vaak een conferentie artikel, in de meer eerbiedwaardige humaniora is dat vaak een boek. Die publicaties zijn de kern van onze communicatie. En behalve de kern van wetenschappelijke communicatie zijn ze ook de munteenheid waarin wetenschappers afgerekend worden: hoe meer publicaties, hoe beter. En vooral: hoe vaker je publicaties geciteerd worden door anderen, hoe beter. Zou die belangrijkste valuta van de wetenschap onveranderd blijven onder invloed van het Web?

We bekijken een sterk vereenvoudigd beeld van de klassieke productiecyclus van de gemiddelde wetenschappelijke publicatie. Die begint voor de wetenschapper met (1) het zoeken naar relevante literatuur, gevolgd door (2) de literatuur bestuderen en interpreteren. Dat leidt vervolgens tot (3) het opstellen van hypothesen en (4) het ontwerpen van een experiment om die hypothesen te toetsen. Als het experiment voldoende data heeft opgeleverd kan de wetenschapper (5) haar hypothese toetsen en (6) de resultaten publiceren in een fraai artikel.

Feitelijk gaat deze cyclus terug tot zelfs de tijd van Huygens. Ook Huygens las de literatuur (bijvoorbeeld de boeken van Blaise Pascal over diens waarschijnlijkheidsleer of de werken van Descartes). Hij verrichtte experimenten (bijvoorbeeld ter bestudering van de slingerbeweging, of met zijn lensopstellingen) hij deed observaties (bijvoorbeeld aan de manen van Saturnus) en hij publiceerde zijn resultaten. Al was dat niet in wetenschappelijke tijdschriften, maar in de vorm van brieven aan collega wetenschappers en in een aantal boeken.

Van al deze stappen wordt eigenlijk maar een klein deel echt goed ondersteund door de computer. Natuurlijk, het zoeken van literatuur en het publiceren van de resultaten gaan beter en sneller vanwege het Web. We hebben het hergebruik van data via het Web al besproken, maar alle andere stappen zijn nog steeds even handmatig als bij Huygens zonder het Web: literatuur lezen en interpreteren, hypothesevorming, experimenteel ontwerp, experimenten uitvoeren, de publicatie schrijven. Kan dat niet beter? Moet dat echt nog allemaal op feitelijk dezelfde wijze als Huygens dat deed in de 17e eeuw? Dat kan zeker beter.

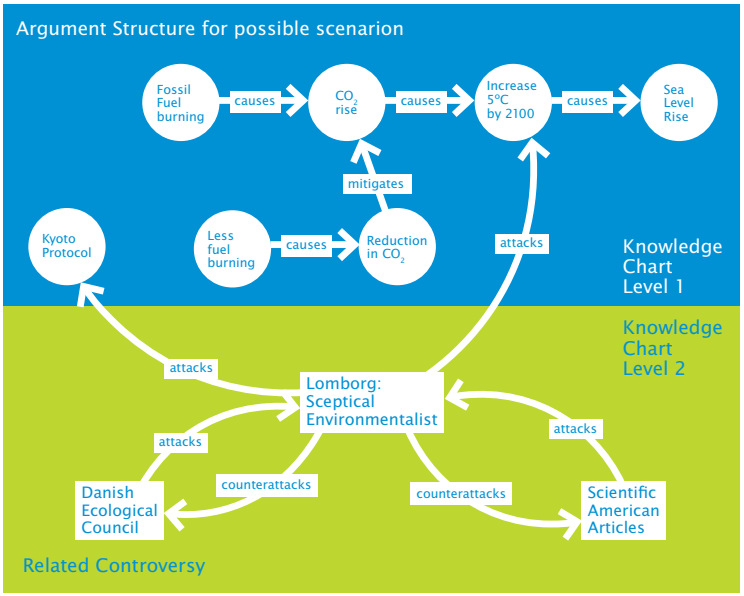
In de Informatica bestaat een deelgebied onder de naam *Information Extraction*. Doel van dat vakgebied is het automatisch extraheren van gestructureerde informatie uit ongestructureerde bronnen. Bijvoorbeeld: uit een k rantenartikel achterhalen welke politicus wat heeft gezegd over welke andere politicus. Of uit een weblog halen wie er welke mening heeft over welk onderwerp. Of... uit een wetenschappelijk artikel achterhalen wie er welke claim maakt op basis van welke data.

Maar is het niet eigenaardig dat een computer dit moet achterhalen door middel van complexe *information extraction*. Als de informatie er "uitgetrokken" moet worden dan was het blijkbaar eerst goed verstopt. En vanuit het oogpunt van de computer is dat ook inderdaad zo. Een gewaardeerde collega in mijn afdeling beschrijft een wetenschappelijk artikel altijd als "een staatsbegrafenis voor je resultaten". Je hebt prachtige gestructureerde data verkregen op grond waarvan je heldere conclusies trekt en vervolgens schrijf je al die prachtige gestructureerde gegevens op in proza (Engels, Nederlands, Chinees) dat alleen begrepen kan worden door tussenkomst van een menselijke lezer, namelijk een collega wetenschapper. Die vervolgens op zijn computer gestructureerde experimenten gaat opzetten, op wellicht dezelfde datasets (op het Web) als die beschreven werden in het oorspronkelijke artikel.

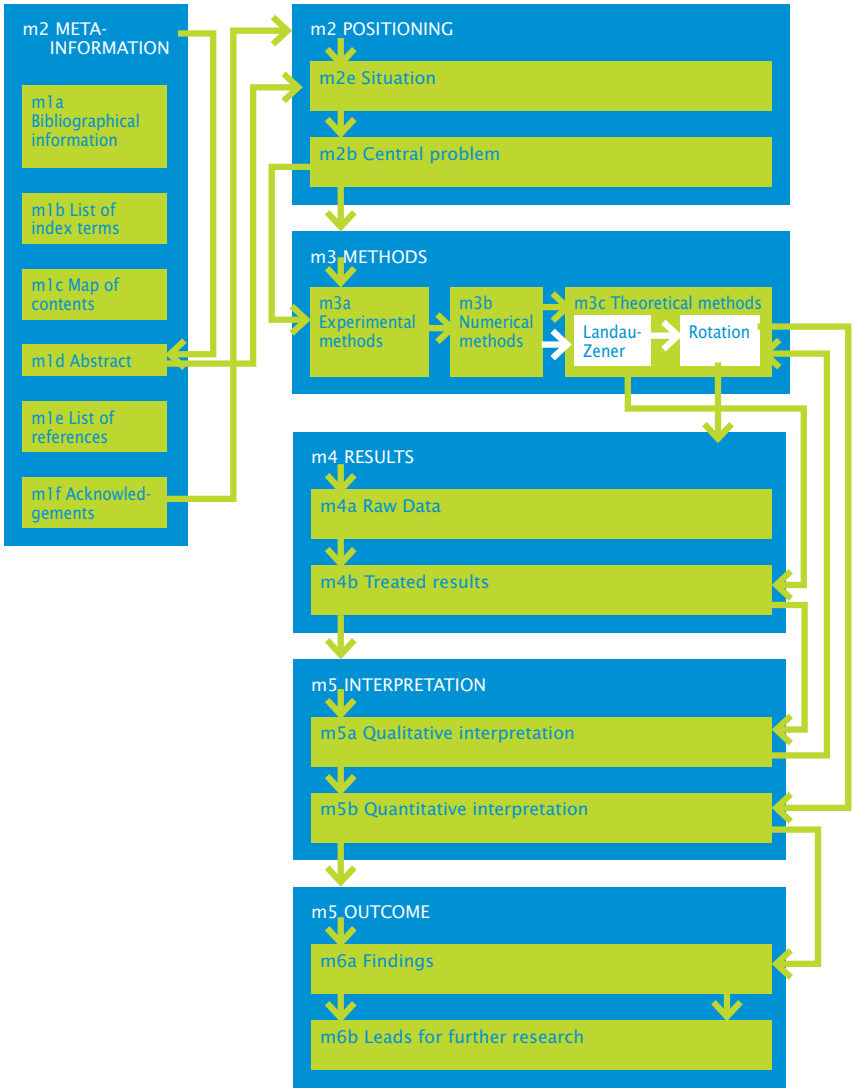
Waarom slechts die staatsbegrafenis in proza? Waarom de wetenschappelijke gegevens, de hypothese, de conclusie, de argumenten niet ook opgeschreven in een expliciete, gestructureerde vorm die toegankelijk is voor directe computer analyse en behandeling? Want voor alle duidelijkheid: of u uw artikel nou in het Nederlands of het Engels schrijft, voor een computer is het natuurlijk allemaal Chinees. De conclusies van een artikel zijn nu zo goed verstopt in proza dat mijn computer me er niet op kan wijzen dat twee artikelen strijdige conclusies bevatten. Zou het niet prettig zijn als mijn computer me daar op zou kunnen wijzen? En als mijn computer me dan een voorstel zou doen voor een "cruciaal experiment" dat zou uitwijzen wie er gelijk had?

Dat is allemaal minder sciencefiction dan u wellicht zult denken. Er bestaan inmiddels experimentele formaten om wetenschappelijke artikelen op het Web te publiceren zodat de structuur en argumenten van een wetenschappelijk artikel niet alleen toegankelijk zijn voor collega wetenschappers (door het op te schrijven in Engels, Nederlands of Chinees), maar die het ook toegankelijk maken voor die andere belangrijke speler in de wetenschap: de computer.

Dit werk gaat al terug tot de late jaren '90. Werk van Moens in Edinburgh, werk van Buckingham-Shum aan de Open Universiteit van Engeland, maar ook werk hier in Amsterdam door Joost Kircz en anderen.

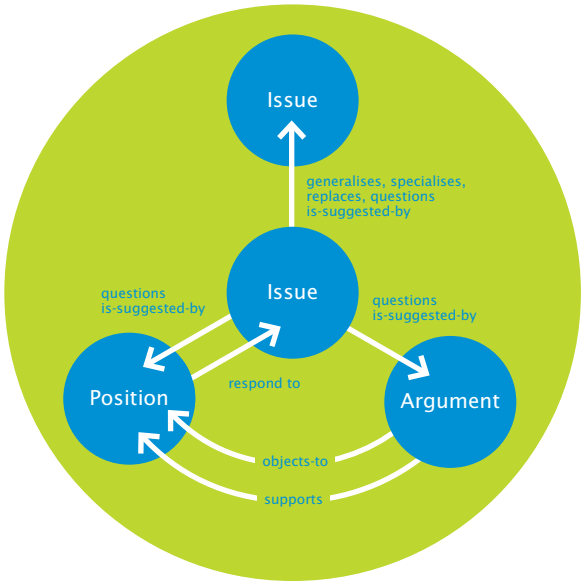


Hier ziet u een vereenvoudigde grafische weergave van een deel van het debat over klimaatverandering inclusief twijfels over het Kyoto verdrag, de rol van de Deense klimaatveranderingscepticus Lomborg en de structuur van zijn argumenten.



Een ander figuur, dit keer van Kircz, geeft een beeld van de standaard onderdelen van een wetenschappelijk artikel en hun samenhang: initiële positionering, probleemstelling, uiteenzetting van methode, beschrijving van resultaten, interpretatie en conclusies. De veelheid aan pijltjes en lijntjes geeft aan dat er veel variatie in deze structuur mogelijk is, maar ook dat deze structuren traceerbaar en veelal canoniek zijn.

De figuur hieronder geeft een eenvoudig argumentatiemodel weer met de typische relaties die we vaak terugvinden in een wetenschappelijk debat, als het ware "onder de huid" van een wetenschappelijk artikel.



Door dit soort structuren expliciet te publiceren en ze niet langer te verstoppen in Nederlands, Engels of Chinees, legt een wetenschapper dus niet alleen de resultaten van haar wetenschappelijk werk uit aan haar collega's, maar ook aan de computers van haar collega's. Hiermee wordt die meest kostbare van alle wetenschappelijke valuta, het wetenschappelijk artikel zelf, toegankelijk voor computers. Daarmee wordt het in principe mogelijk om (delen van) die andere stappen (artikel lezen en interpreteren, hypothesevorming en experiment ontwerp, experiment uitvoeren, publiceren) met ondersteuning van de computer uit te voeren.

Daarnaast wordt ook *modulair* publiceren mogelijk: waarom zouden wetenschappers steeds weer dezelfde hypotheses uiteenzetten of dezelfde basisaannames moeten formuleren? Als een wetenschappelijk artikel eenmaal is opgebouwd uit componenten, zoals in het diagram van Kircz, dan kan een artikel heel goed bestaan uit een netwerk van delen uit andere online artikelen (bijvoorbeeld de aannames die ik deel met ander werk, of juist de hypothese uit ander werk die ik probeer te verwerpen), in combinatie met delen die ik zelf aan dat netwerk van argumenten toevoeg.

Overigens zal het duidelijk zijn dat zo'n ecosysteem van modulair publiceren alleen zal gedijen in een cultuur met open toegang tot wetenschappelijke publicaties. Het is dan ook toepasselijk dat de komende week de eerste wereldwijde "Open Access Week" plaats vindt, een activiteit waar ook de VU van harte aan meedoet. Het resultaat van zulke modulaire, Open Access en machinaal toegankelijke publicaties is dat de structuur van het wereldwijde wetenschappelijke discours expliciet beschikbaar komt, leesbaar voor zowel mensen als machines. Met gebruik en hergebruik van feiten en bevindingen, en het online aan elkaar verbinden van resultaten op verwachte, maar ook op onverwachte manieren. Daarmee wordt het Web een nog veel grotere wetenschapsversneller dan het nu al is.


Huygens zou die wereld
niet meer herkennen, maar
ik verheug me er op.

Ten slotte

Dit alles overziend zou het wel eens kunnen zijn dat het klassieke wetenschappelijk artikel als munteenheid zijn langste tijd heeft gehad. Immers, het wordt hoog tijd dat we andere wetenschappelijke producten ook gaan waarderen als primaire bijdragen: het produceren en beschikbaar maken en houden van datasets bijvoorbeeld, met misschien wel download statistieken als graadmeter om het belang van de bijdrage te meten. Wetenschappelijke publicaties zullen drastisch van vorm gaan veranderen: van monolithische blokken tekst (op papier of in PDF) tot modulaire structuren die bestaan uit delen geschreven door anderen en daaraan toegevoegd onze eigen bijdragen in het Wereld Wijd Web van online wetenschappelijke argumentaties.

We moeten met elkaar maar eens goed gaan nadenken wat dit voor consequenties zal hebben voor de inrichting van ons werk en onze wetenschappelijke wereld. Huygens zou die wereld niet meer herkennen, maar ik verheug me er op. Het jaarthema van de VU: “Connect & Reflect” had niet beter gekozen kunnen worden.

Curriculum Vitae

Frank van Harmelen (1960) studeerde Wiskunde en Informatica aan de Universiteit van Amsterdam, en promoveerde in de Kunstmatige Intelligentie aan de Universiteit van Edinburgh (1989). Hij werkt sinds 1995 aan de VU, waar hij sinds 2002 is aangesteld als Hoogleraar Kennisrepresentatie en Redeneren.

Van Harmelen is vooral geïnteresseerd in het gebruik van klassieke kennis-formalismen (zoals de logica) in grootschalige en open informatie-omgevingen zoals het Web. Hij schreef het eerste leerboek over dit onderwerp (inmiddels verschenen in 5 talen), hij droeg bij aan de internationale standaard computertaal op dit gebied, en hij is een van de editors van het vuistdikke standaard werk “Handbook of Knowledge Representation”. Van Harmelen is een van de meest geciteerde Nederlandse computerwetenschappers.

Colofon

Ontwerp
Rudie Jaspers
Dienst Marketing & Communicatie

Fotografie
Yvonne Compier en Riechelle van der Valk
Dienst Marketing & Communicatie

Druk
Drukkerij Papyrus bv. te Diemen

Uitgave
Vrije Universiteit Amsterdam
20459/11 oktober 2009

Bloemen
De bloemdecoraties zijn ontworpen en verzorgd door
de Hortus Botanicus van de Vrije Universiteit