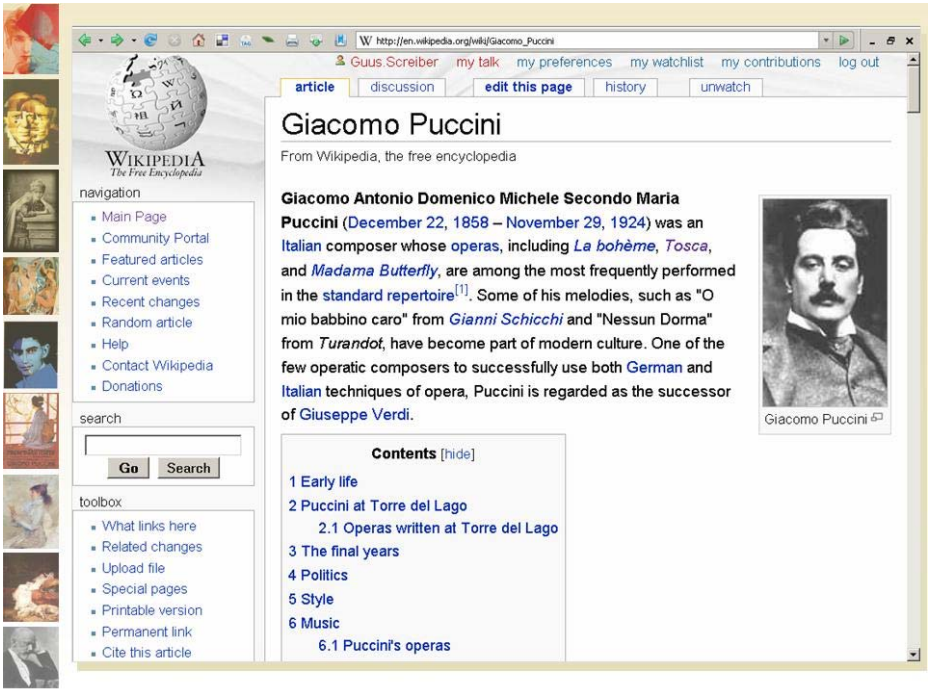
Oratie 24 november 2006: "Over grijpbaarheid en grenzen van kennis"

Rede uitgesproken door dr. A. Th. Schreiber bij de aanvaarding van het ambt van hoogleraar Intelligente Informatiesystemen aan de Vrije Universiteit Amsterdam op vrijdag 24 november 2006.

INLEIDING

Mijnheer de rector magnificus, dames en heren,

Het Web is in rap tempo het medium geworden voor informatie- en kennisuitwisseling. "Googelen" is sinds kort zelfs een goedgekeurd woord in de Nederlandse taal. Als ik heden ten dage iets wil weten over, zeg, een opera van Puccini, dan pak ik niet meer een boek uit mijn boekenkast, maar kijk in Wikipedia; de informatie die ik daar vind is uitgebreider en meer up-to-date dan de informatie in mijn opera-encyclopedie. Toch levert niet elke zoekvraag zomaar het gewenste resultaat op. Als je op zoek bent naar plaatjes waarop Parijs staat afgebeeld is dat lastig. We krijgen misschien wel wat plaatjes over Parijs, als we er tenminste aan gedacht hebben om "Paris (en)" of zo u wilt "Paris (fr)" in te typen, maar we zien ook afbeeldingen van Paris Hilton, van hotels met Paris in de naam, enzovoorts. Een afbeelding van Montmartre zal er niet bij zitten, terwijl toch iedereen weet dat dit een wijk van Parijs is.



Puccini in Wikipedia

De vraag die ons als knowledge engineers bezighoudt is óf en hoeverre wij dit proces van informatie-ontsluiting kunnen verbeteren door het gebruik van expliciete kennis over webobjecten. Met webobjecten, in het jargon "resources" genoemd, bedoel ik tekstfragmenten, plaatjes, kortom alles waar men naar kan verwijzen met een weblink. Zou het niet mooi zijn als de zoekmachine zou weten dat een bepaald plaatje over de stad Parijs gaat (en niet over een andere betekenis van Parijs), en dat de zoekmachine ook kennis heeft over de wijken van Parijs. Om dit doel te bereiken dienen wij te beschikken over metadata, dat wil zeggen data over de webobjecten, die zodanig gerepresenteerd zijn dat een computerprogramma afleidingen kan maken over de betekenis van het object. Deze toekomstvisie is voor het eerst geformuleerd door sir Tim Berners-Lee, de bedenker van het World-Wide Web zoals wij dat nu kennen, en door hem het Semantische Web genoemd. In dit uur wil ik met u nagaan in hoeverre dit toekomstbeeld een realistische scenario is. In hoeverre zijn wij in staat kennis over

webobjecten formeel te beschrijven, met andere woorden: te "grijpen", en wat zijn de grenzen die wij ons daarbij moeten stellen?

PRINCIPES VOOR WEB-GEBASEERD KNOWLEDGE ENGINEERING

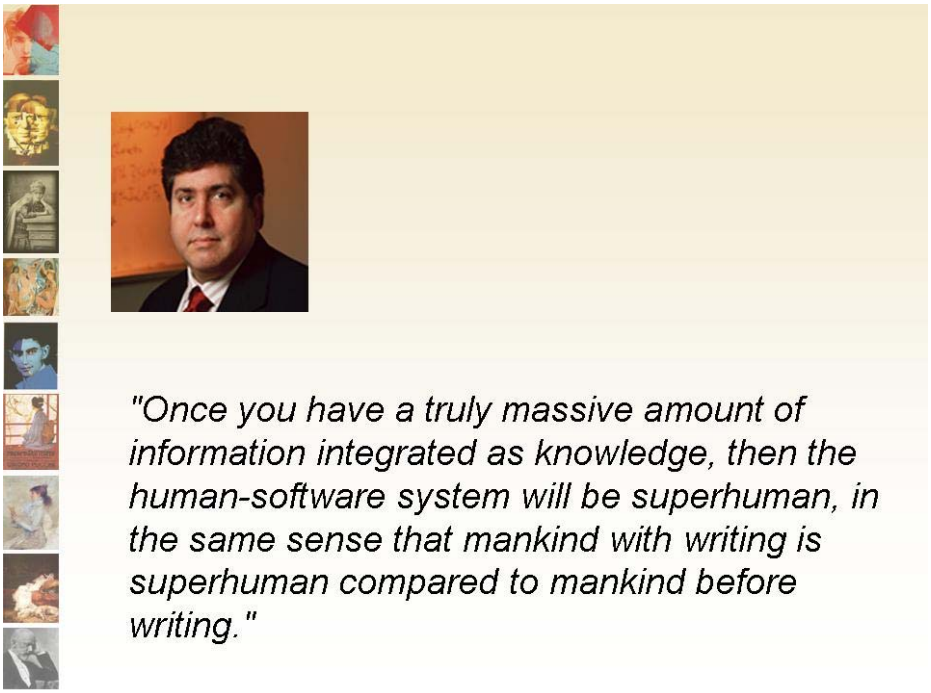
De vraag naar formalisering van kennis is natuurlijk niet pas opgekomen met het Web. Zij is zo oud als de Griekse filosofen. Zelf heb ik mij lange tijd beziggehouden met formalisering van kennis voor zogenaamde kennissystemen. Dit zijn computerprogramma's die problemen op kunnen lossen, zoals het bepalen of een hypotheekaanvraag moet worden toegekend of het maken van een ontwerp van een liftinstallatie. Ik wil u hier niet vermoeien met een lange historische verhandeling over knowledge engineering, een term waarvoor ik nooit een adequate Nederlandse vertaling heb kunnen bedenken. Ik wil volstaan met het formuleren van een viertal principes voor het formaliseren van kennis op een schaal zoals dat nodig is voor het Web.

Griekse filosofen

Het eerste principe zou ik het *bescheidenheidsprincipe* willen noemen. Kennis is beschikbaar in vele domeinen, waarin mensen reeds lang, vaak decennia of zelfs eeuwen, zijn bezig geweest om expertise op te bouwen. Informatici hebben de betweterige neiging om al gauw te zeggen, dat de kennis in een bepaald domein incorrect gerepresenteerd is. Om een voorbeeld te noemen: in het cultureel-erfgoeddomein (waaraan ik al mijn voorbeelden in deze voordracht zal ontlenen) is de Art & Architecture Thesaurus, kortweg AAT, ontwikkeld, die 120.000 onderling gerelateerde termen omvat. Op deze kennisbron kan men kritiek hebben. Natuurlijk is de AAT niet volledig, natuurlijk hadden de begrippen die erin voorkomen met meer betekenis (semantiek in het jargon) beschreven kunnen worden. Men kan ook betogen dat de AAT hierarchie verkeerd opgebouwd is en dat men niet had moeten vaste houden aan de enkele-boom structuur. Maar het gaat veel te ver om daarom deze bron

plompverloren af te schrijven als onbruikbaar. Ik moet hier een korte excursie maken naar het begrip "ontologie". Deze oude filosofische term wordt tegenwoordig gebruikt om te spreken over een geformaliseerde verzameling begrippen, die ons in staat stellen kennis met elkaar te delen. De populariteit van ontologieën sinds eind jaren negentig heeft alles te maken met de toenemende behoefte aan gemeenschappelijke definities van begrippen in een explosief-groeiende gedistribueerde informatieruimte, lees: het Web. Hel simpel gezegd is een ontologie een verzameling afspraken over wat we met een bepaalde term bedoelen. Informatici leggen in hun onderzoek veel nadruk op de mate van formalisering van een ontologie, d.w.z. de noodzaak voor het gebruik van een formeel-logische taal. Voor uitwisselbaarheid is een machine-interpreteerbare representatie een minimum voorwaarde en een taal zoals de webstandaard RDF is voor dit doel zeer geschikt. Voor al het meerdere zou ik willen zeggen: *ja graag, maar niet tot elke prijs!* Er bestaat een nogal kinderachtige neiging om de kwaliteit van een ontologie af te meten aan het aantal logische symbolen. Dit is een denkfout, die gebaseerd is op een beperkt wereldbeeld en een gebrek aan respect voor de resultaten van andere disciplines. Kwaliteit van een ontologie is slechts af te meten aan het gebruik als instrument om informatie te delen. Ik zou zelfs de stelling willen poneren dat er op dit moment een invers verband is tussen de mate van formalisering van een ontologie en haar nut in de praktijk. Dat heeft echter ook alles te maken met het feit dat te veel onderzoekers in hun artikelen niet gebruik maken van bestaande vocabulaires, maar hun eigen speelgoedvoorbeeld ontwikkelen. Ik voorzie dat op dit punt een radicale omslag nodig is in het denken binnen ons vakgebied. Als wij een Semantisch Web willen realiseren moeten wij niet op de stoel van domeinexperts gaan zitten, maar ons bescheiden opstellen en voortbouwen op de veelheid aan bestaande kennisbronnen.

Het tweede principe is het *schaalprincipe*. Kort samengevat: "denk groot!". Voor dit principe wil ik teruggrijpen op een rede die Doug Lenat in 1987 bij de IJCAI in Milaan hield met als titel "On the Thresholds of Knowledge". Deze rede verwekte destijds nogal wat opschudding. Feitelijk vertelde hij het aanwezige wetenschappelijke publiek dat het voor de realisering van intelligente computerprogramma's noodzakelijk is om grote hoeveelheden alledaagse kennis, zoals opgeslagen in encyclopedieën, te formaliseren, d.w.z. in een voor computers leesbaar formaat te representeren. Ik was zelf niet op de conferentie aanwezig, maar herinner mij goed hoe mijn collega's bij terugkomst verhalen vertelden over Lenat, die helemaal gek geworden was. Hij had toen al een enigszins gemankeerde reputatie. Zijn ego was legendarisch en Amerikanen hadden zelfs de zogenaamde "Lenat" eenheid geïntroduceerd, die een maat zou zijn voor wat ik, vrij vertaald, het "kwakzalver"-niveau zou willen noemen, en die voor normale stervelingen, net zoals bij de Faraday eenheid, alleen in micro-Lenats gemeten kan worden. . Echter, na bijna 20 jaar terugkijkend kan men niet anders dan vaststellen, dat Lenat gelijk had. Als onze computertoepassingen Parijs als plaats moeten kunnen onderscheiden van Paris Hilton en moeten weten dat Montmartre onderdeel is van Parijs, dan is geformilseerde encyclopedische kennis precies hetgeen nodig is. Het goede nieuws is: dit soort kennis is overvloedig in semi-formele vorm aanwezig. In veel domeinen zijn kennisbronnen in de vorm van vocabulaires, thesauri, geografische databases en wat dies meer zei beschikbaar. Hier ligt een kans voor ons informatici om deze rijke kennisbronnen op een uitwisselbare manier ter beschikking te stellen, zolang we maar niet teveel zeuren over de beperkte semantiek, zie het eerste principe. Voor een Semantisch Web is het noodzakelijk dat wij zo veel mogelijk van deze kennisbronnen bij elkaar brengen en zorgen dat deze gezamenlijk, met een lelijk woord "interoperabel", gebruikt kunnen worden. Als eerbetoon aan Lenat is de titel van deze oratie een parafrase op die van de genoemde voordracht.



Doug Lenat

Het derde principe wil ik het *verrijkingsprincipe* noemen. Dit principe houdt direct verband met de zojuist genoemde veelheid aan kennisbronnen, waar wij in de praktijk mee te maken hebben. De wereld is veel te divers om met één ontologie te kunnen worden beschreven, althans niet in de voorzienbare toekomst. Dit betekent een breuk met de filosofische traditie, waarin "ontologie" de unieke leer was van datgene wat bestaat. Ook heden ten dage zijn er nog veel informatici die deze Heilige Graal najagen. Vanuit het enge systeemdenken tolereren zij slechts één manier waarop de wereld gemodelleerd kan worden, en is het "fout" als er meerdere interpretaties of, nog erger, inconsistenties bestaan. Dit wordt mede gevoed door de logisch-formele scholing, waarin logische consistentie en compleetheid als absolute waarheden worden onderwezen. Echter, het is onrealistisch om te veronderstellen dat dit in een globale informatieruimte zoals het Web het geval kan zijn. Als wij werkelijk kennisrepresentatie op globale schaal willen realiseren, dan dienen wij om te kunnen gaan met kennisbronnen, die slechts gedeeltelijk overlappen en onderling inconsistent kunnen zijn. Als wij dit *idee fixe* van unificatie loslaten kunnen wij ons richten op een beperkter maar realistischer doel, namelijk het verrijken van bestaande bronnen met additionele kennis. Daarbij moet men in de eerste plaats denken aan het identificeren van een beperkte verzameling verbanden of links tussen kennisbronnen, bijvoorbeeld tussen begrippen uit de eerder genoemde AAT aan de ene kant en begrippen uit de lexicale database WordNet aan de andere kant. Om die reden dient het onderzoek dat aangeduid wordt als *ontology alignment* een van onze eerste onderzoeksprioriteiten te zijn. De verbanden, die wij hierbij vinden, zijn altijd partieel, maar kunnen desondanks het gezamenlijk gebruik van kennisbronnen kwalitatief sterk verbeteren. Een andere verrijkingsmogelijkheid is het expliciteren van impliciete kennis. Dit kan men op verschillende manieren doen. Eén methode is de zogenaamde semantische analyse van bestaande relaties in kennisbronnen, zoals bijvoorbeeld met de OntoClean methode van Guarino en Welty of met de methodes zoals voorgesteld door Rector. Ook kan men met natuurlijk-taal verwerkingstechnieken bijbehorende tekstbronnen, zoals bijvoorbeeld scope notes van thesauri, analyseren en begrippen uit de tekst extraheren. In het geval van visuele bronnen is het soms mogelijk om met beeldanalyse begrippen in stilstaande

beelden of video te herkennen. De aldus verkregen kennis komt niet in plaats van de bestaande kennis; zij voegt kennis toe en leidt tot een verrijking van het het gebruik van de bron.

De Heilige Graal

Als vierde en laatste principe noem ik het *patroonprincipe*. Zoals gezegd zie ik weinig mogelijkheden voor unieke representaties van kennis, met uitzondering van een selecte verzameling standaardbegrippen voor bepaalde meeteenheden, zoals tijd. Dat betekent niet dat wij bij het modelleren van kennis volledig aan ons lot zijn overgelaten. In de knowledge engineering zijn in de loop der jaren een grote verzameling patronen opgebouwd voor het modelleren van kennis. Voorbeelden daarvan zijn te vinden in de publicaties van de Semantic Web Best Practices groep van W3C, de organisatie verantwoordelijk voor de webstandaarden; ik verwijs u naar de patronen voor deel-geheel relaties en de patronen voor waardebereiken. Patronen veranderen het knowledge-engineering proces van een kunst in een methodologisch ondersteunde discipline. In het CommonKADS onderzoek hebben wij eerder goede ervaringen opgedaan met patronen van kennisintensieve taken, zoals beoordeling en diagnose. Ik verwacht overigens, dat dit soort taak-gebaseerde patronen weer heel actueel zullen worden in het onderzoeksgebied van automatische webdiensten, de zogenaamde "web services", maar dit punt valt buiten mijn huidige betoog. Voor de beschrijving in webformaat van vocabulaires en thesauri ben ik momenteel met andere onderzoekers in W3C bezig de SKOS specificatie te standaardiseren. SKOS kan men het beste opvatten als een verzameling patronen voor een interoperable specificatie van dit soort kennisbronnen. SKOS kan een belangrijk brug vormen waarover bronnen uit het bibliotheek- en archiefwezen voor het Semantisch Web ter beschikking komen.

Patronen

Samenvattend, als wij gepaste bescheidenheid in acht nemen ligt in veel domeinen kennis voor het grijpen; daarbij dienen wij in eerste instantie te denken aan formalisering op grote schaal van bestaande, relatief eenvoudige maar uitgebreide kennisbronnen. Deze grenzen kunnen wij verleggen door het ontwikkelen van verrijkingstechnieken. Bij de representatie van deze kennis dienen modelleerpatronen een centrale rol te vervullen.

KENNISREPRESENTATIETALEN OP HET WEB

Alvorens verder te gaan met een onderzoeksprogramma gebaseerd op bovenstaande principes, wil ik kort de aandacht vestigen op een paar punten aangaande kennisrepresentatietalen voor het Web. Mijn collega Frank van Harmelen heeft hier recentelijk een aantal zeer behartenswaardige opmerkingen over gemaakt, zoals bijvoorbeeld over denkfouten met betrekking tot beslisbaarheid. Ik zal daar vanuit mijn specifieke achtergrond een aantal observaties aan toevoegen.

De webontologietaal OWL beschouw ik als een belangrijke stap voorwaarts. OWL bevat een aantal uiterst nuttige instrumenten voor het representeren van semantiek van web resources, met name de mogelijkheden voor het definiëren van de logische betekenis van relaties, alsook de "sameAs" constructie voor het disambigueren van identiteit. Ik ben zelf een fel pleitbezorger geweest van de mogelijkheid tot metamodelleren binnen OWL. Bijna elke gedistribueerde toepassing vereist dit; het is absoluut onrealistisch om te veronderstellen dat iedereen de klasse/instantie scheidslijn op dezelfde manier trekt. Het Web is immers geen ideale wereld: mensen modelleren de wereld vanuit hun eigen perspectief en met die riemen zullen we moeten roeien, zie alweer principe 1. Deze eis voor metamodelleren heeft geleid tot felle debatten met de hardcore logici, die actief zijn in dit veld, en die de doem van Russel's paradox over niet-gelovigen uitspraken. Het destijds bereikte compromis van één taal met twee verschillende interpretaties, hoe bizar dit ook moge lijken, vond ik derhalve een goede tussenweg. De recent door Motik en

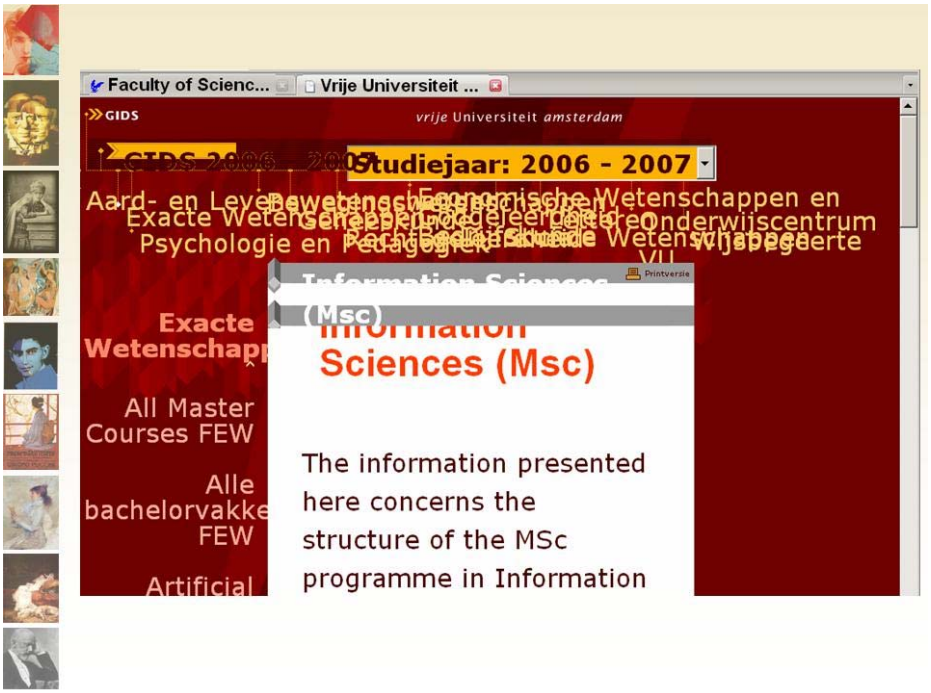
anderen voorgestelde oplossingen voor metamodelleren binnen het keurslijf van beschrijvingslogica maakt in de toekomst mogelijk zelfs deze tussenoplossing overbodig. Daarbij dient overigens de relatie met RDF in stand gehouden te worden.

OWL ontwikkelaars

De aanwezigheid van een taal als OWL leidt echter ook tot nieuwe valkuilen. Sommige onderzoekers menen dat, als zij iets in OWL opschrijven, het automatisch een ontologie wordt. Uit mijn eerder definitie zal u wel duidelijk zijn, dat dit niet het geval is: een specificatie wordt pas een ontologie als deze binnen een gemeenschap wordt gebruikt als gezamenlijk vocabulaire. Nog kwalijker is de omgekeerde redenering: als het niet in OWL is opgeschreven, dan is het geen ontologie. Met deze argumentatie heb ik een onderzoeker de stelling zien verdedigen dat zijn eigen, idiosyncratische specificatie wel een ontologie was en WordNet, het veelgebruikte lexicale netwerk van 100.000 concepten, onderling verbonden met 17 soorten relaties, niet. Een absurde denkfout die helaas nog wijdverbreid is.

Voor ik verderga wil ik even stilstaan bij de rol van W3C, de organisatie die ik al een aantal keren noemde en die opgericht is en geleid wordt door Tim Berners-Lee. W3C, voluit het World-Wide Web Consortium, is een organisatie waarvan bijna alle grote informatietechnologiebedrijven lid zijn. Binnen W3C komen zij samen tot afspraken, die ervoor zorgen dat onze web browsers webpagina's op dezelfde manier interpreteren, qua inhoud, qua vormgeving, enzovoorts. Toen mij als co-voorzitter van de OWL werkgroep voor het eerst een kijkje in de keuken van W3C werd gegund, wist ik eerst niet wat ik van deze organisatie moest denken. Het leek een soort kerk, met een goeroe aan het hoofd, omringd door apostelen, die strenger in de leer zijn dan de meester. Het interne proces lijkt een jaren-zeventig-achtige anarchistische bureaucratie. Ik heb echter mijn mening snel bijgesteld. Het is een organisatie die volledig naar buiten gericht is, nauwlettend de belangen van iedereen, vooraleerst de gewone webgebruiker, in de gaten houdt. Het is feitelijk één van de weinige tegenwichten tegen monopolistische bedrijven, met name één in het bijzonder, waarvan ik u de naam niet hoef te noemen. Ik

zou graag willen dat de Europese Unie de W3C richtlijnen ten aanzien van toegankelijkheid tot wetgeving verheft, zoals in de Verenigde Staten al aan het gebeuren is. Zelfs mijn eigen universiteit is niet bij machte een website aan te bieden, waarop iemand zoals ik met beperkt gezichtsvermogen in staat is zoiets elementairs als een studiegids te raadplegen. En het is toch o zo eenvoudig. Ik heb veel geleerd van W3C en ook van het sociale proces dat hoort bij het opstellen van standaarden. Het was alle stress meer dan waard. En nog even over die ontoegankelijke websites: volgend jaar wil ik met groepjes studenten websites gaan keuren en gele en rode kaarten gaan uitdelen aan overtreders. U hoort nog van ons.



VU studiegids in grote fonts

HET CULTUURWEB

Goed, over naar het onderzoeksprogramma. In mijn opinie is de hypothese, die ten grondslag ligt aan het Semantisch Web, namelijk dat informatie-uitwisseling op het web ondersteund en verbeterd kan worden door middel van expliciete achtergrondkennis, zeker waard verder onderzocht te worden. Ik noem het bewust een hypothese, omdat wij als onderzoekers op dit vakgebied nog relatief weinig evidentie hebben verzameld ter ondersteuning van deze hypothese, d.w.z. nog nauwelijks delen van een Semantisch Web gebouwd hebben, waar gewone webgebruikers hun voordeel mee kunnen doen. Dat wil niet zeggen dat er geen veelbelovende voorbeelden zijn, integendeel. Even dicht bij huis blijvend denk ik aan het DOPE systeem, ontwikkeld door Frank van Harmelen, Heiner Stuckenschmidt en collega;s, dat achtergrondkennis gebruikt voor het vinden en visualieren van medfisch-wetenschappelijke informatie, de Flink applicatie van Peter Mika voor semantisch ondersteunde sociale netwerken, en onze eigen E-Culture toepassing. Ik geloof dat met name in domeinen waarin veel specifieke kennis aanwezig is, de kansen voor toegevoegde waarde van een Semantisch Web aanzienlijk zijn. Daarbij moet in eerste instantie vooral gedacht worden aan domeinen, waarin de informatie én de kennisbronnen publiek toegankelijk zijn.

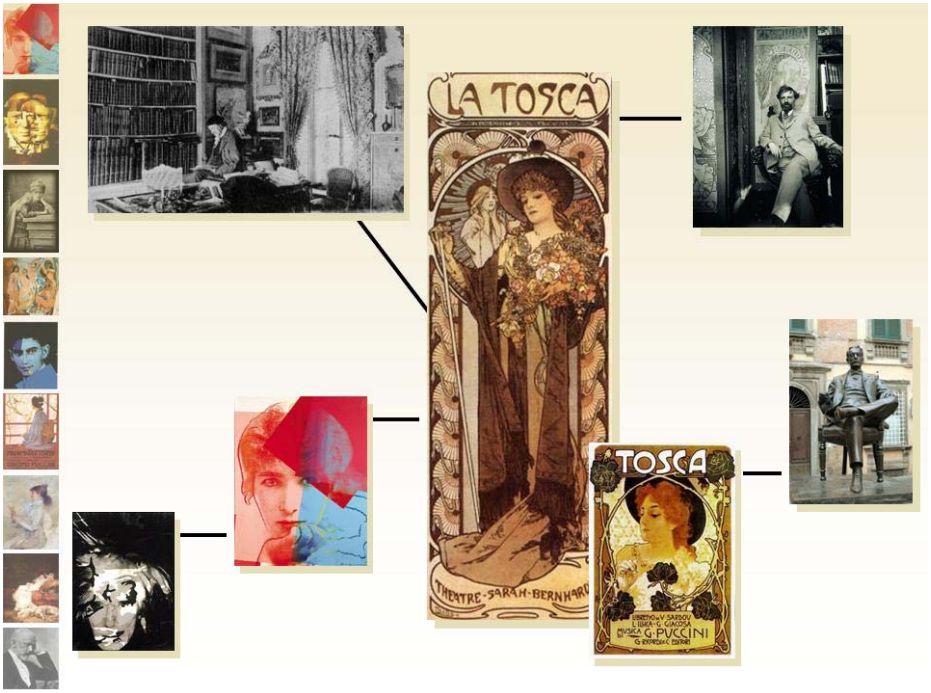
Ik wil mij de komende jaren concentreren op het cultureel-erfgoeddomein als toepassingsgebied, zoals dat de afgelopen jaren ook al in toenemende mate het geval was. Deze keuze is niet toevallig: in dit domein is een rijkgeschakeerde verzameling kennisbronnen aanwezig, er is veel publieke informatie, en er is een groeiende behoefte van zowel erfgoedinstellingen als gebruikers om deze informatie breed toegankelijk te maken. In Nederland is het CATCH programma van NWO een belangrijke drijvende kracht en ook op Europees niveau zijn er veelbelovende initiatieven.

Het is mijn expliciete doel om de komende jaren een concrete aanzet te geven tot wat ik een semantisch Cultuurweb zou willen noemen. Dit Cultuurweb heeft een tweeledige doelstelling: het dient aan te tonen dat een Semantisch Web, althans in bepaalde domeinen, daadwerkelijk gerealiseerd kan worden, én het geeft ons als onderzoekers een platform voor verder onderzoek. Voordat ik inga op de noodzakelijk technische infrastructuur en de daaraan gerelateerde onderzoeksvragen, wil ik twee voorbeelden geven van wat ik met zo'n Cultuurweb zou willen kunnen doen.

Mattise, Derain, Moreau en Salomé

Stel ik zoek naar de schilder Matisse. Mijn zoekmachine zou nu onderscheid moeten kunnen maken tussen een schilderij gemaakt door Matisse, zoals het schilderij van de "dame met hoed", dat destijds zoveel opschudding verwekte, en een door een collega gemaakt schilderij, waarop Matisse staat afgebeeld. Ik wil tevens werken van gerelateerde schilders kunnen vinden, zoals het schilderij van collega-Fauvist Derain midden-boven, en het schilderij van zijn leermeester Moreau midden-onder. Op dit laatste schilderij staat Salomé afgebeeld; een Cultuurweb moet mij kunnen verwijzen naar andere cultuuruitingen betreffende Salomé, zoals de opera van Richard Strauss. De opera en het dame met hoed stammen beiden uit het jaar 1905; ook dat tijdsverband wil ik kunnen zien. Met andere woorden, een Cultuurweb moet niet alleen relevante resultaten opleveren, maar ook aan kunnen geven *om welke reden* de gevonden objecten relevant zijn.

Een tweede voorbeeld: een zoektocht naar Tosca zou mij moeten kunnen brengen bij een afbeelding van de schrijver van het originele toneelstuk, te weten Victorien Sardou, bij de componist Puccini, die op basis hiervan de gelijknamige opera componeerde, bij Sarah Bernhardt, de beroemde actrice die de rol op het toneel bracht, en bij Andy Warhol die een portret van Sarah Bernhardt maakte.



Tosca, Sardou, Puccini, Bernhardt en Warhol

Wat is er nu nodig om zo'n semantisch Cultuurweb op te bouwen? De architectuur moet *low-tech* zijn, daarmee bedoel ik dat deze met relatief eenvoudige webtechnologie gebouwd kan worden. De infrastructuur, zoals wij die in het MultimediaN E-Culture project hebben opgebouwd met collega's van de UvA en het CWI, zie ik als een goede basis. Deze architectuur is volledig gestoeld op het gebruik van open webstandaarden zoals XML, RDF/OWL, SVG en AJAX. Wel zal de hardware-kant aanzienlijk versterkt dienen te worden. Vorige week, na publicatie van een artikel over ons systeem in de NRC Next, crashte de E-Culture server onmiddellijk. De hardware backbone moet berekend zijn op 100+ collecties om van een echt Cultuurweb te kunnen spreken. Dit zal de nodige schaalbaarheidsproblemen opleveren, een kolfje overigens naar de hand van mijn waarde college Jan Wielemaker. Wat betreft de RDF/OWL representatie van erfgoedthesauri, de kennisbronnen dus, verwacht ik geen grote problemen. Door het werk van Mark van Assem en het werk van de nieuwe W3C Semantic Web Deployment groep hebben we dit proces methodologisch goed onder de knie; de representatiepatronen voor thesauriconversie zijn redelijk uitgekristalliseerd. De situatie ligt ingewikkelder voor het metadata-conversieproces, waarbij bestaande semi-gestructureerde metadata van collecties afgebeeld moeten worden op concepten uit kennisbronnen. Dit gebeurt in het E-Culture project nog op een ad-hoc basis; professionalisering is hier dringend gewenst. Ik ben van plan hiervoor samenwerking te zoeken met professionele partners met ervaring in informatie-extractie. Daarbij denk ik bijvoorbeeld aan de ontwikkelaars van de KIM workbench, die hebben laten zien dat zij dit probleem goed onder de knie te hebben. Ook de technieken, die in het kader van het CHOICE project door Luit Gazendam en Veronique Malaisé ontwikkeld worden, kunnen hierbij helpen. Op den duur moet dit leiden tot een methodologische aanpak voor dit

specifieke conversieprobleem, zodat bedrijven dit als service kunnen aanbieden voor cultuurinstellingen, die hun collectie binnen het Cultuurweb willen brengen.

nrc·next

16 Nov 2006

▶ webvoorkeur

Kunstweb

Onderzoekers van de UvA en de VU hebben, in samenwerking met Digitaal Erfgoed Nederland (DEN) en het Instituut Collectie Nederland (ICN), een programma ontwikkeld dat het mogelijk maakt kunstcollecties 'intelligenter' te doorzoeken met zoek- en informatievragen. Dat maakt het

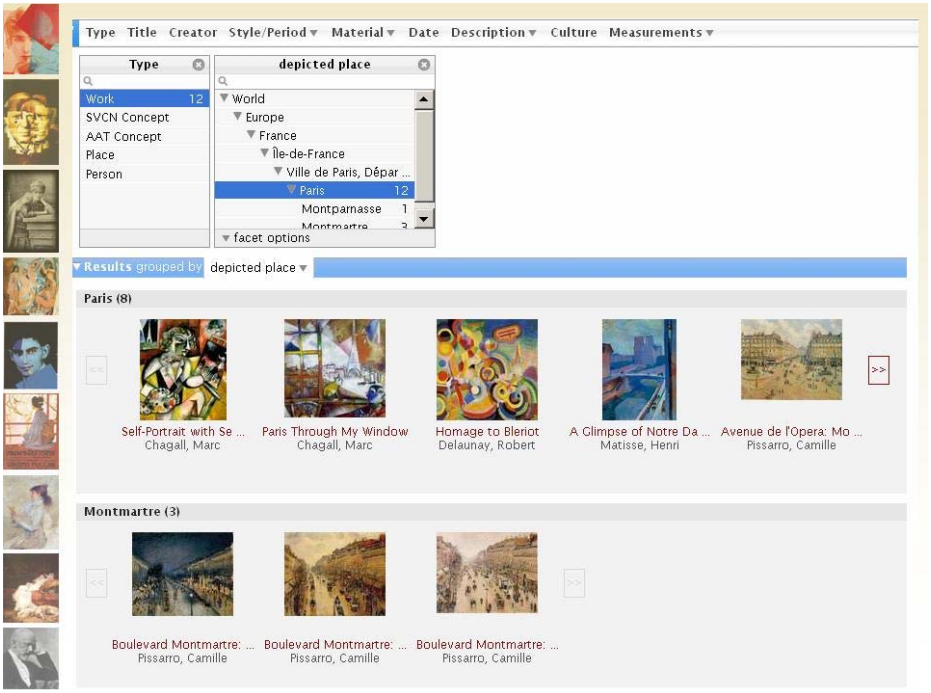
Zelf zoeken: e-culture.multimedian.nl/demo/search of pluk: sms 56647 naar 7585

Artikel over E-Culture project

Wat betreft de zoektechnieken zet ik voorlopig in op de *basic search* en de *facet browser*, zoals die binnen E-Culture door Jacco van Ossenbruggen, Michiel Hildebrand en anderen ontwikkeld zijn, gecombineerd met een tijd- en plaats-gebaseerde visualisatie van de zoekresultaten. Kortom, de basis van het Cultuurweb kan in mijn opinie een professioneel opgezette versie van de prijswinnende E-Culture demonstrator zijn. Het is werk, veel werk, maar vergt eigenlijk alleen technologie die wij nu reeds goed beheersen. Zoals Tom Gruber, de *godfather* van het gebruik van ontologieën in de hedendaagse informatica, tijdens de laatste internationale Semantic Web conferentie opmerkte: de Semantic-Web gemeenschap moet, naast verder onderzoek, gewoon beginnen daadwerkelijk zo'n Web te realiseren, anders verliest het de slag van andere Web ontwikkelingen, zoals Web 2.0. De technologie is er immers klaar voor. Het is ook een middel om cultuurinstellingen over de brug te krijgen met hun collectiedata, iets wat wij nu al voorzichtig zien gebeuren. De barrières zijn overigens met name van sociale aard en niet zozeer technologisch. In dit acceptatieproces is een centrale rol weggelegd voor evaluatiestudies van zowel de functionaliteit en als van de presentatiemogelijkheden. Voor deze studies worden op dit moment de fundamenten gelegd door Alia Amin en Lora Aroyo. Evaluatiestudies zijn een schromelijk onderschat aspect van de informatica, en verdienen zowel in het onderzoek alsook in het onderwijs veel meer aandacht dan zij momenteel krijgen.

Natuurlijk is hiermee onderzoeksmatig de kous niet af. Een Cultuurweb kan, zoals gezegd, een platform vormen voor verdere exploratie. Allereerst gaat dan de gedachte uit naar *alignment* van de kennisbronnen. Door het identificeren van dergelijke kruisverbanden zal de zoekfunctionaliteit aanmerkelijk kunnen worden verbeterd. Een voorbeeld zijn de relaties tussen kunstenaars en stijlen, zoals die door Victor de Boer door middel van analyse van kunsthistorische teksten zijn gevonden. Ik denk ook aan de

technieken ontwikkeld door Willem van Hage binnen het VL-e project en aan samenwerking met projecten zoals STITCH, waar dit onderwerp bovenaan de onderzoeksagenda staat.



Zoeken naar plaatjes van Parijs

In een Cultuurweb hebben wij te maken met veelal visueel materiaal. Een uitdaging waar ik met Arnold Smeulders, Marcel Worring en collega's al enige jaren aan werk is het overbruggen van de afstand tussen beeldnalysetechnieken en semantiek. Wij begonnen ooit met het automatisch herkennen van kleur op foto's van apen en dat bleek geen sinecure. De beeldanalyse gaat echter met sprongen vooruit en ik verwacht dan ook veel van de methoden en technieken uit deze hoek. Het proefschrift dat Laura Hollink vorige week verdedigde biedt hiervoor concrete aanknopingspunten. Een andere belangrijke bron is tekstueel materiaal: artikelen, webpagina's en dergelijke over culturele onderwerpen. Voor een Cultuurweb zal het noodzakelijk zijn de samenwerking met onderzoekers op het gebied van natuurlijke-taal verwerking te intensiveren. Om die reden ben ik ook blij met het nieuwe MuNCH project, een multi-disciplinair project waarin beeldanalyse, tekstanalyse en semantiek samengebracht worden voor het zoeken in de archieven van TV programma's beheerd door het Instituut voor Beeld & Geluid.

Wat betreft de interface kunnen wij in het multi-culturele Europa niet volstaan met Engels als communicatietaal. Voor een Cultuurweb zijn daarom lexicale bronnen in andere talen een must, zoals bijvoorbeeld een Nederlandse versie van WordNet waar Maarten de Rijke en collega's momenteel aan werken.

Het resulterende Cultuurweb moet niet alleen een rijkgeschakeerde, semantisch geïndexeerde virtuele collectie worden; het moet ook nieuwe zoekparadigma's introduceren: zeg maar, vragen die je niet aan Google kunt stellen. Zo werken wij binnen MultimediaN aan het vinden van verbanden tussen twee door de gebruiker opgegeven termen. Bijvoorbeeld: vertel mij hoe Van Gogh en Gauguin aan elkaar gerelateerd zijn. Ik verwacht dat dit voor professionele gebruikers, zoals samenstellers van

tentoonstellingen, een waardevolle zoekmogelijkheid toevoegt. Ook moet het naast de beschikbare reguliere collectiedata mogelijk worden voor gebruikers om zelf beelden en annotaties toe te voegen. Dit stelt wel hoge eisen aan de interface. Een vraag hierbij is hoe een gebruiker efficiënt een relevante indexeringsterm kan selecteren, liefst net zo snel en gemakkelijk als momenteel zogenaamde "tags" bij populaire fotosites als Flickr toegevoegd kunnen worden. Uiteindelijk wordt immers het succes van een Cultuurweb bepaald door wat tegenwoordig de "lange staart:" van het Web genoemd wordt: betrokkenheid van de vele anonieme gebruikers *out there*.

Van Gogh en Gauguin

In het erfgoeddomein is er één kennisbron, die voor zover wij hebben kunnen nagaan echt ontbreekt en waar wij enorm vel plezier van zouden kunnen hebben: een historische thesaurus. Een dergelijke kennisbron stelt ons in staat om een cultuurobject in de historische context te presenteren, zoals men bijvoorbeeld vaak wel ziet in biografiën. Onze nieuwe promovenda Anna Tordai heeft zich op dit probleem geworpen. Als haar eerste haalbaarheidsstudie positief uitvalt willen wij in samenwerking met historici (alweer principe 1) proberen een dergelijke kennisbron te bouwen op basis van de SKOS patronen.

De lijst van genoemde onderzoeksthema's is puur indicatief en zeker niet volledig; ik wil hier volstaan met op te merken dat voortgang binnen dit onderzoeksveld alleen mogelijk als men multi-disciplinair denkt: een mono-benaderingen is simpelweg onrealistisch. Voor mijzelf sprekend: ontologieën alleen zijn niet zaligmakend.

ONDERZOEKSFINANCIERING

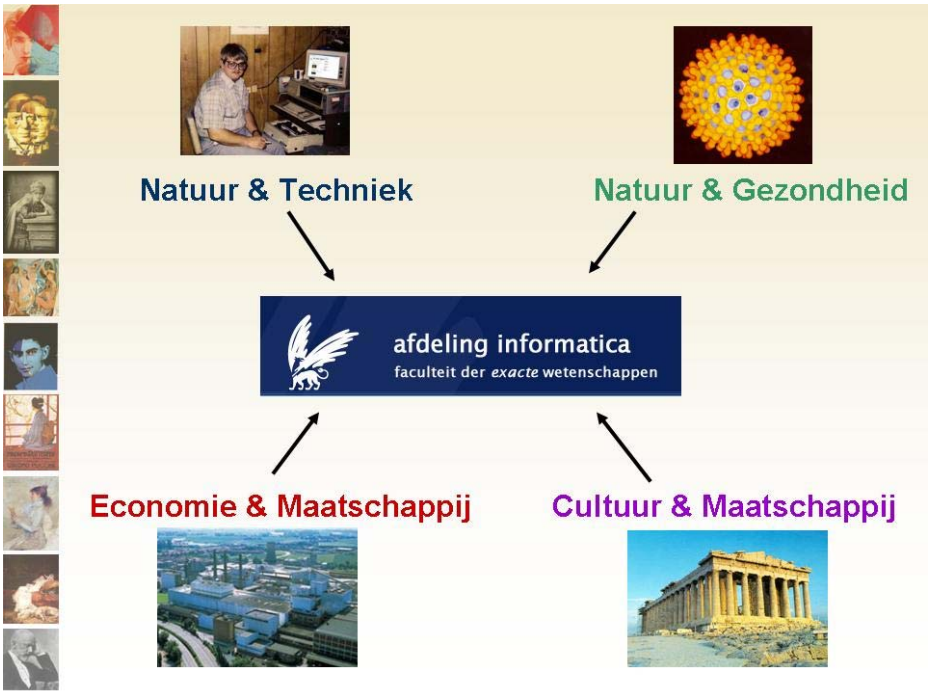
Voor het Cultuurweb is uiteraard geld nodig, en dat brengt mij op het thema van onderzoeksfinanciering. Universitair onderzoek wordt steeds meer vormgegeven volgens het Amerikaanse model, dat wil zeggen. de universiteit financiert een kleine *core*

van senior onderzoekers, soms een enkele promovendus; de rest moet gefinancierd worden uit externe bronnen. Laat ik voorop stellen dat ik prima kan leven met dit model. Het legt wel een zware druk op de senior onderzoekers, maar houd je ook scherp. Tot enige jaren geleden waren wij vrijwel volledig aangewezen op de Europese fondsen voor extra onderzoeksgeld, via programma's als ESPRIT en IST. Een voorbeeld is Knowledge Web, het Network of Excellence waar ik wetenschappelijk coördinator van mag zijn en waar speerpuntonderzoek gebeurd op het gebied van het Semantische Web. Ondersteuning van de EU heeft er mede voor gezorgd dat Europa op dit gebied voorloopt op de Verenigde Staten. De laatste jaren zijn tot mijn vreugde ook de nationale programma's meer in beeld gekomen. De gelden voor de kenniseconomie hebben geleid tot verniewende projecten. Ik noemde al het MultimediaN project, dat onder deze regeling gefinancierd wordt; daarnaast participeren wij ook met meer dan verwacht resultaat onder leiding van Pieter Adriaans in het *Advaced Information Disclosure* project binnen het VL-e programma. De thematische NWO programma's dienen in dit kader tevens genoemd te worden, zie het CATCH programma. SenterNovem doet ook een duit in het zakje, getuige onder meer het aan het Cultuurweb gerelateerde RNA project onder leiding van Hans Nederbragt. Anders dan wel eens gedacht wordt is het onderzoeksklimaat in Nederland eigenlijk best goed. Ik wil drie kanttekeningen plaatsen. Ten eerste, subsidiegevers moeten reeële verwachtingen hebben ten aanzien van te bereiken resultaten. Ik bespeur een jachtige sfeer, waarbij resultaten eigenlijk gisteren verwacht worden; het zal wel iets met de tijdsgeest te maken. Dit leidt tot projectvoorstellen, waarin aanvragers meer beloven dan zij waar kunnen maken, een ongezonde situatie. Ten tweede, in dit polderland lijkt men er nog steeds naar te streven om de koek gelijkmatig te verdelen. Zo werd in de Open Competitie van NWO onderzoekers afgeraden om een voorstel in te dienen als zij al in thematische programma's van NWO participeren. Ik vind dit ongewenst en slecht voor de ontwikkeling van het Nederlandse onderzoek. Ten derde, bij de uitkomsten van onderzoek kijkt men zowel in Nederland al binnen de EUvooral naar het directe economische belang. Dat vind ik nogal kortzichtig. Zoiets als een Cultuurweb heeft geen direct economisch belang, maar kan wel meehelpen de kwaliteit van leven te verbeteren. De economische voordelen zijn meer indirect: diensten, *merchandising* e.d. Voor alle succesvolle webbedrijven geldt trouwens, dat het primaire proces (zoeken, foto's en video oplsaan) gratis is en dat geld wordt verdiend met aanvullende diensten. Overigens is cultuur sowieso een stimulans voor de economie, zeker in een stad als Amsterdam met haar creative industrie. Los van deze kanttekeningen hoop ik natuurlijk wel dat een volgend kabinet beried zal zijn om te blijven investeren in onderzoek.

ONDERWIJS

Laat ik afsluiten met een aantal observaties over het onderwijs. De informatica-opleidingen in Nederland hebben al jaren te maken met een lage instroom; de VU is hierop geen uitzondering. Gedeeltelijk is dit een maatschappelijk fenomeen; gedeeltelijk moeten wij ook de schuld bij onszelf zoeken. Wij dienen te beseffen dat een kleine minderheid van de studenten die instroomt klassieke Informatica studenten zijn; de meerderheid komt om multi-disciplinaire opleidingen zoals Informatiekunde en Kunstmatige Intelligentie te volgen. Informatica is steeds minder een pure bèta-wetenschap en dient ook niet als zodanig naar buiten toe geprofileerd te worden. Ik ben blij dat wij op de onlangs gehouden bezinningsdag met elkaar hebben vastgesteld dat een dergelijk cultuuromslag noodzakelijk is. En het wordt ook tijd dat onderwijsvisitatiecommissies dit feit onder ogen gaan zien. De kritiek die ik vorige maand hoorde van de visitatiecommissie Informatica, namelijk dat er te weinig aandacht is voor wiskunde vond ik van een hoog absurditeitsgehalte. Kritiek was zeker op zijn plaats geweest, maar dan toch vooral op het feit dat onderwerpen als *information*

retrieval en webtechnologie nog niet of maar mondjesmaat aan bod komen. Ik zie het geschetste Cultuurweb niet alleen als een goede mogelijkheid voor onze studenten om interessante afstudeerstages te doen; ook kunnen wij dit als één van de kapstokken gebruiken om methoden en technieken te identificeren waar het in ons vakgebied echt om draait. Een dergelijke heroriëntatie maakt onze opleidingen voor een veel grotere groep scholieren een aantrekkelijk keuze. Ik zou willen pleiten voor een pakket van vier opleidingen, één voor elk van de VWO profielen. Het zal wel een zware dobber worden om de vakken zodanig om te vormen dat deze in dergelijke brede curricula passen. Overigens, ik heb harde woorden gesproken over sommige beoefenaars van de logica, maar dat laat onverlet dat ik de logica zelf een noodzakelijk basisvak vind, dat een plaats moet krijgen in alle curricula.



Informatica en de VWO profielen

DANKWOORD

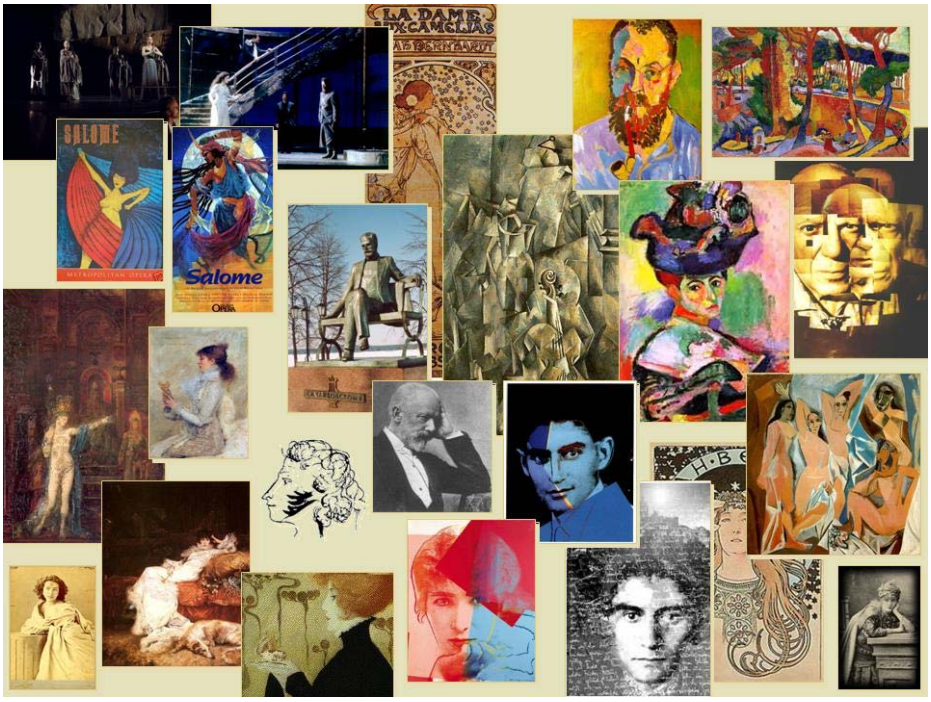
Voor ik eindig wil ik graag mijn dank uitspreken, in de eerste plaats aan de faculteit Exacte Wetenschappen en aan de Afdeling Informatica voor het in mij gestelde vertrouwen. Van dit vertrouwen gaf u materieel al blijk door mij reeds kort na mijn aanstelling bij het management van de afdeling te betrekken, hetgeen door de latere reorganisatie overigens een *mixed blessing* was. Ik kan echter oprecht zeggen, dat ik mij van het begin af aan welkom heb gevoeld op de VU.

Bepalend voor mijn ontwikkeling in de wetenschap is geweest dat ik in 1986 terecht kwam bij Bob Wielinga. Ik heb geen idee waarom hij mij uitnodigde voor een gesprek, want bij mijn sollicitatie was ik zo vreselijk naïef, dat ik zelfs vergat een publicatielijst mee te sturen. Bob beschouw ik als mijn leermeester, in vele opzichten: in het analytisch denken, in het wetenschappelijk debat en in de o-zo-moeilijke taak van het begeleiden van promovendi. Ik bewaar hele goede herinneringen aan mijn oud-collega's van de SWI groep aan de Universiteit van Amsterdam, met wie ik 17 jaar heb samengewerkt, dank daarvoor, Jan, Jacobijn, Robert, Saskia en alle anderen.

De hoop dat ik op de VU een bloeiende samenwerking met collega Frank van Harmelen zou kunnen opbouwen is volledig uitgekomen; het is heel bevredigend om na ruim drie jaar de interne cohesie én de externe uitstraling te zien. Hans Akkermans haalde mij over naar de VU te komen; het doet mij genoegen met hem en Jaap Gordijn te werken aan verdere profilering en verbetering van de Informatiekunde opleiding. Ik heb ontzettend veel plezier in de samenwerking met Lora, Borys, Laura, Veronique, Mark, Willem, Luit en Anna. Hetzelfde geldt voor de samenwerking met de andere medewerkers van onze virtuele Semantic Web groep en met alle collega's in de eerder genoemde onderzoeksprojecten. Jullie allemaal maken het dat ik eigenlijk elke dag met plezier naar mijn werk ga. Als ik zo de gesprekken om mij heen in de trein hoor, is dat echt iets om zuinig op te zijn

Wilma, Niels en Judith, jullie waren er voor mij op belangrijke momenten, en zonder jullie was ik hier nooit gekomen. Ik dank en gedenk hier ook mijn ouders André en Fien en mijn stiefmoeder Mia. Mijn veel te vroeg overleden moeder prentte mij al vroeg in dat professor best een aardig beroep is. Zij had gelijk: het is een privilege om in de wetenschap te mogen werken.

Ik heb gezegd.



Dankbetuiging: De figuren in dit document zijn een selectie uit de slides, die als visuele achtegrond voor de presentatie dienden. Ik ben Lora Aroyo zeer erkentelijk voor de hulp bij de vervaardiging van de slides.