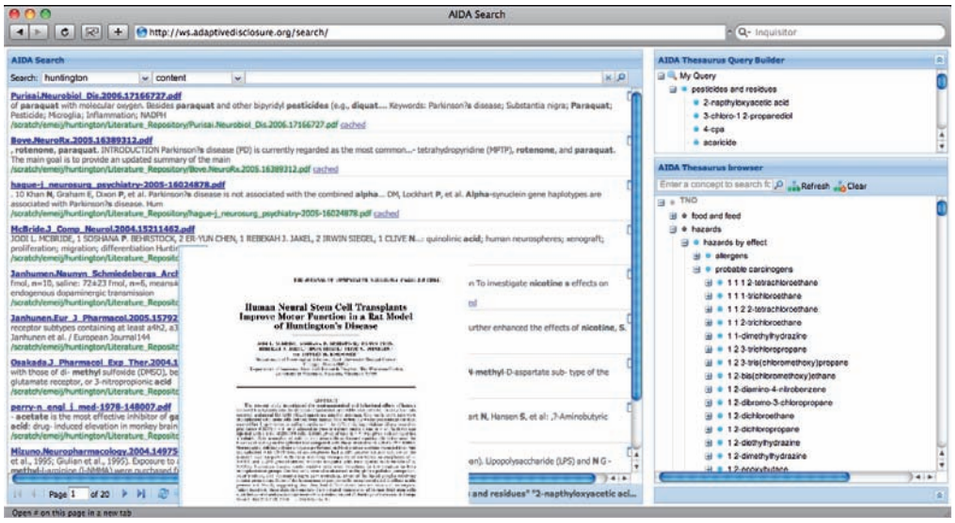
De AIDA toolbox: een gecombineerde aanpak voor het beheren van kennis

M. Scott Marshall[1], Marco Roos[1], Edgar Meij[1], Sophia Katrenko[1], Willem Robert van Hage[2], and Pieter Adriaans[1]

[1]Institute for Informatics, University of Amsterdam, Kruislaan 403, 1098 SJ, Amsterdam, The Netherlands
[2] Vrije Universiteit Amsterdam, de Boelelaan 1081a, 1081HV Amsterdam, The Netherlands

In een computationele netwerk omgeving zoals het grid is een overvloed aan zeer uiteenlopende soorten bronnen aanwezig. Denk bijvoorbeeld aan tijdschrift artikelen, beelden, massa spectrometrie data, R scripts voor statistiek, web services, workflows of spreadsheets. Deze overvloed kan een grote belemmering vormen. Hoe moet een gebruiker de juiste bronnen vinden voor een voorliggend probleem? Vele factoren maken het matchen van de benodigdheden en gebruikerswensen aan wat de bronnen kunnen leveren en de regels ten aanzien van hun gebruik een complex probleem. Het probleem doet zich voor op verschillende niveaus. Eindgebruikers willen het benodigde vinden in hun eigen domein. Applicatie en middelware ontwikkelaars moeten services en data kunnen vinden, bij voorkeur geautomatiseerd zodat veranderingen in aanwezigheid en toegankelijkheid kunnen worden opgevangen. Dit probleem beperkt zich niet tot grids; ook het Web en allerlei dataopslag toepassingen hebben er mee te maken. Ook voor 'enhanced science' (e-science) is het beheren van heterogene bronnen een belangrijke uitdaging.



figuur 1: AIDA screenshot

Het Virtueel Laboratorium voor e-Science (VL-e) is een project met academische en industriële partners waarvoor de toepassing van e-science in meerdere wetenschappelijke onderzoeksdomeinen een centraal thema is. 'Adaptive Information Disclosure' (AID) is een subprogramma in VL-e. De AID groep is een multidisciplinaire samenwerking van experts in 'information retrieval', 'machine learning' en het semantische web – een krachtige combinatie van technologieën voor het extraheren en opslaan van kennis, en daarmee het vinden van heterogene bronnen. In het kader van VL-e heeft AID samengewerkt met leden van het 'Food Consortium' om toepassingen te bouwen gebaseerd op haar sleuteltechnologieën en web services. Om het gebruik van metadata en kennis te ondersteunen voor verschillende toepassingen heeft AID web services gebouwd die als generieke componenten kunnen worden ingezet in toepassingen in specifieke domeinen, zoals hier voor voedsel onderzoek. De AID web services en toepassingen daaromheen vormen samen de AIDA gereedschapskist[1] ('AID Application toolkit'). De AIDA gereedschapskist is gericht op groepen kenniswerkers die gezamenlijk grote verzamelingen documenten over verspreide locaties willen kunnen doorzoeken, annoteren, interpreteren en verrijken. Met AIDA gereedschap kunnen verschillende taken worden uitgevoerd, zoals het leren van nieuwe patronen voor kennis extractie, gespecialiseerde zoekopdrachten op verzamelingen en het opslaan van kennis (zie fig. 1). De meeste componenten zijn beschikbaar als web service en zijn 'open source' onder een Apache licentie.

De AIDA gereedschapskist en haar toepassingen zijn ontwikkeld in samenwerking met partners uit verscheidene toepassingsgebieden en lossen een aantal use-cases op. De AIDA gereedschapskist is bijvoorbeeld gebruikt voor bioinformatische, medische beeldbewerking en voedsel informatica use-cases. Zelfs als het om samenwerking tussen sterk gerelateerde disciplines (zoals machine learning, information retrieval en het semantische web) gaat, is multidisciplinaire samenwerking is een aanzienlijke uitdaging. De verschillende terminologieën

1 http://www.adaptivedisclosure.org/aida

en aanpakken leiden tot een intensief proces van definiëren en verifiëren van doelen en intenties. Natuurlijk is dit proces niet wezenlijk verschillend van waar iedere software engineer mee te maken krijgt. Er is echter nog geen algemeen geldend raamwerk waarin kennis-gebaseerde representaties worden gedefinieerd, ondanks de uitgebreide onderzoekshistorie in filosofie, logica en kunstmatige intelligentie. Zo zou bijvoorbeeld een kennis 'engineer' de term *metadata* kunnen gebruiken om te refereren aan zowel een semantische als een syntactische representatie, terwijl een database 'engineer' tabelstructuur en syntactische informatie als metadata beschouwt. Een ander gebied dat wrijvingen kan opleveren is de afstand tussen 'middleware' ontwikkelaars en de (eind)gebruikers in een bepaald toepassingsgebied. De gebruikers vinden het vaak lastig om te begrijpen wat bepaalde software doet en de ontwikkelaars weten veelal niet voldoende van het toepassingsgebied om het gebruik van hun applicaties in voldoende mate toe te lichten. De enige manier om deze ruimte te vullen is om ofwel de ontwikkelaars uitgebreid kennis te laten nemen van het toepassingsgebied, of om de gebruikers te verdiepen in de ontwikkelomgeving. Deze stap is noodzakelijk voor een praktische en succesvolle samenwerking. Desondanks blijft het een uitdaging om de problemen en taken van het toepassingsgebied te koppelen aan de mogelijkheden van een nieuwe technologie.

Een belangrijke technologische hoeksteen van AIDA is het semantische web. De semantische 'stack' van het World Wide Web Consortium (W3C) is gecreëerd om kennis uitwisseling tussen computers mogelijk te maken. Deze stack is opgebouwd met steeds specifieker wordende vereisten. Aan de basis staat XML, wat een data uitwisseling faciliteert door een representatie, schema en syntax voor te schrijven. Na XML komt RDF (Resource Description Framework), waarmee uitdrukkingen in de vorm van subject-predicaat-object uitgedrukt kunnen worden. De modelleertaal RDF-Schema (RDF-S) vormt de basis voor hiërarchieën en redeneren aan de hand van subsumpties. OWL (Web Ontology Language) breidt de basale klasse definities van RDF uit om redeneren en modelleren aan de hand van descriptie logica mogelijk te maken. Tenslotte sluit een laag met regels de semantische stack af. De stack vormt een praktische basis voor het opslaan, ontsluiten en uitwisselen van kennis, ondanks dat deze niet alle vormen van kennis kan ondersteunen. Een groot aantal open source en vrijelijk beschikbare implementaties geeft een indicatie van de ruime inzetbaarheid en toepasbaarheid. Verdere W3C standaarden zijn ook beschikbaar, waaronder SKOS (Simple Knowledge Organization System) voor vocabulaires en FOAF (Friend of a Friend) voor sociale netwerken. Terwijl er meer en meer gebruikers deze Linked Open Data principles[2] volgen en gebruiken, worden er steeds meer bronnen en data gekoppeld en ontstaat het semantische web.

Een aantal technologieën worden gebruikt in de AIDA web services, ieder met specifieke toepassingen en voordelen. Sesame RDF repositories faciliteren bijvoorbeeld het opslaan en terugvinden van kennis, gebruikmakend van semantische web technologieën. Lucene zorgt voor het indexeren en terugvinden van documenten aan de hand van hun inhoud, wat er voor zorgt dat we door gehele document collecties kunnen zoeken vanuit applicaties. Niet alleen kleine, toepassingsspecifieke document collecties kunnen worden doorzocht, maar bijvoorbeeld ook grote, algemeen beschikbare collecties zoals MedLine[3]. De machine learning technieken die zijn geïmplementeerd in AIDA geven applicatie ontwikkelaars de mogelijkheid om entiteitherkenning in teksten uit te voeren, waarmee bijvoorbeeld eiwitten kunnen worden geïdentificeerd. Tevens bestaat de mogelijkheid relaties tussen entiteiten, zoals "eiwitA heeft een interactie met eiwitB", te herkennen en op te slaan. De voornaamste horde die bij het trainen van deze statistische modellen genomen moet worden is het annoteren van training data en vooralsnog biedt AIDA slechts beperkte ondersteuning in dit gebied. In de toekomst zullen dergelijke modellen wellicht beschikbaar worden gemaakt, alsmede een uitbreiding waarmee patroonherkenning in visuele data kan worden gedaan.

Alhoewel iedere op zich staande technologie interessant en nuttig is op zichzelf, is een combinatie het meest krachtig en kan de meeste voordelen bieden. In een 'text mining' toepassing bijvoorbeeld, kan de data waarop gemined wordt specifiek worden toegespitst door gebruik te maken van information retrieval technieken en methodologieën zoals query expansie (het al dan niet automatisch toevoegen van query termen om de *recall* te verhogen). In het geval van een gepersonaliseerde zoek toepassing kan de RDF repository web service worden aangewend om alternatieve formuleringen van termen te vinden. Hiermee kunnen conceptgebaseerde zoekopdrachten worden uitgevoerd, alsmede nieuwe concept-gerelateerde annotatie suggesties worden verkregen.

De AIDA applicaties kunnen, onder meer, uitgevoerd worden middels drie 'interfaces': Taverna (workflows), een web-browser applicatie en een Java applicatie genaamd de VBrowser[4] die grid bronnen toegankelijk maakt. Met Taverna hebben we een workflow gemaakt die hypotheses genereert in een biologische context middels 'information extraction'. Dit heeft geleid tot discussies en oplossingen aangaande computationele experimenten, zoals de keuze van een kennis representatie die het hergebruik en de herkomst van kennis kan opslaan, alsmede de ondersteuning voor semantische data types in workflows. De AIDA web-browser applicatie geeft toegang tot op maat gemaakte (Lucene[5]) indexen en thesauri, waaronder de SKOS vertaling van een voedsel ontologie die is ontwikkeld door het Food Consortium. Dezelfde combinatie van op maat gemaakte indexen en thesauri is terug te vinden in de VBrowser, waarmee bronnen op het grid op een semantische wijze kunnen worden geannoteerd en ontsloten. De VBrowser wordt bijvoorbeeld gebruikt door medische beeldbewerkingsexperts om experimenten ten aanzien van functionele MRI op het grid te ondersteunen. De VBrowser is uitgebreid om deze afbeeldingen op een semantische wijze te annoteren en terug te vinden.

2 http://linkeddata.org/

3 http://www.nlm.nih.gov/pubs/factsheets/medline.html

4 http://www.vl-e.nl/vbrowser/

5 http://lucene.apache.org/