School of Electronics and Computer Science

Web Intelligence
Professor Nigel Shadbolt
Why distributed and digital is good...
The work of many people…

- Harith Alani
- Steve Harris
- Nick Gibbins
- Yannis Kalfoglou
- Kieron O’Hara
- David Dupplaw
- Bo Hu
- Paul Lewis
- Srinandan Dashamapatra
- Duncan Macrae-Spencer
- Hugh Glaser

- Les Carr
- David de Roure
- Wendy Hall
- Mike Brady
- David Hawkes
- Yorick Wilks
- Enrico Motta
- Carole Goble
- Simon Cox
- Andy Keane
Drivers

- Moores Law and Powers of 10
- The WWW
- Making the Web Semantic
- Intelligence on the Web
- Research on the Semantic Web
- Futures Challenges
Faster and Smaller

- Devices are getting smaller and faster all the time
- Moore’s Law has held for 40 years
- This leads to orders of magnitude
  - Increase in power
  - Increase in memory
  - Decrease in size
  - Decrease in cost
- Constant migration and obsolescence
  - Our processors will have very limited shelf life
  - Our storage does too
  - Our physics does too
In May 1997 Deep Blue beat world chess champion Kasparov.

It won the six-game rematch 3.5-2.5.

The system derives its playing strength mainly out of brute force computing power.

It is a massively parallel, 32-node, RS/6000, SP-based computer system enhanced with 256 special purpose VLSI chess processors.

Evaluating 100,000,000 positions per second.
Drivers

- Moores Law and Powers of 10
- The WWW
- Making the Web Semantic
- Intelligence on the Web
- Research on the Semantic Web
- Futures Challenges
The Magic of the Network Effect

- On the order of $10^8$ users
- Used in *every* country on Earth
- A tiny percentage is “trained” in any way
- On the order of $10^{10}$ indexed web resources (text) in Google etc
- Massively distributed and open
- A set of protocols and languages driven by a strong standards approach
Exponential Development

- A new era 1990
- nxoc01.cern.ch – the first web server
- http://nxoc01.cern.ch/hypertext/WWW/TheProject.html - the first web page
- In 1992 – 26 web servers
- 1993 – 200 web servers
- 1994 – Mosaic browser – 1000 times more traffic than on the first web server
- 1998 -329 million pages
- The International Telecommunications Union estimated 665 million users 2002
Drivers

- Moores Law and Powers of 10
- The WWW: Blessing and Curse
- Making the Web Semantic
- Intelligence on the Web
- Research on the Semantic Web
- Futures Challenges
Making the Web Semantic…

WWW 2002
THE ELEVENTH INTERNATIONAL
WORLD WIDE WEB CONFERENCE
Sheraton Waikiki Hotel
Honolulu, Hawaii, USA
7-11 May 2002

1 LOCATION. 5 DAYS. LEARN. INTERACT.

Registered participants coming from:
Australia · Canada · Chile · Denmark · France · Germany · Ghana · Hong Kong · India · Italy · Ireland · Japan · Malta · New Zealand · The Netherlands · Norway · Singapore · Switzerland · The United States · Vietnam · Zambia

On 7-11 May 2002, Honolulu, Hawaii will provide the backdrop for The Eleventh International World Wide Web Conference. This prestigious series of the international World Wide Web Conference Committee (W3C) attracts participants from around the world, and it provides a public forum for the World Wide Web Consortium (W3C) through the annual W3C track.

The conference is being organized by the International World Wide Web Conference Committee (W3C), the University of Hawaii and the Pacific Telecommunications Council (PTC).

FEATURED SPEAKERS (CONFIRMED)

Tim Berners-Lee, inventor of the World Wide Web and Director of the W3C who now holds the iCom Founders chair at the Laboratory for Computer Science (LCS) at the Massachusetts Institute of Technology (MIT).

Ian Foster, guru of “Grid Computing”, associate director of the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign.

Richard A. DeMillo, vice president and chief technology officer for Hewlett-Packard Company.

McArthur Prize Winner.
That is machine readable....
Can Annotate Anything

- Publications...
- Databases...
- Metadata on scientific structures

Web data set (XHTML)
The SW Community: Structured Spaces

- Linkage of heterogeneous information
  - web content
  - databases
  - meta-data repository
  - multimedia
- Via ontologies as information mediation structures
- Using Semantic Web languages

Oncogene (MYC):
- Found_In_Organism (Human).
- Gene_Has_Function (Transcriptional_Regulation).
- Gene_Has_Function (Gene_Transcription).
- In_Chromosomal_Location (8q24).
- Gene_Associated_With_Disease (Burkitts_Lymphoma).

<meta>
<classifications>
<classification type="MYC" subtype="old_arx_id">bcr-2-1-059</classification>
</classifications>
</meta>

Cancer Risk
Cancer risk estimates do not reach zero no matter how low the level of exposure is to a carcinogen. Terms used to describe this risk are defined below as the number of excess cancer expected in a lifetime:
- high risk: 1 x 10^10
- medium risk: 1 x 10^6
- low risk: 1 x 10^3
- very low risk: 1 x 10^0

Weathered Gasoline PE Soils
TPH Analysis (GRO + DRO) vs. GC Method 8015

Web data set (XHTML)
Ontologies: Building Blocks for the SW

→ A shared conceptualisation of a domain

→ Provides the semantic backbone for applications

→ Lightweight and is deployed using a W3C recommended standard language
Ontologies offer:

- **Communication**
  - Normative models
  - Networks of relationships
  - Consistent and unambiguous
  - Integrate multiple perspectives
- **Inter-operability and Integration: Sharing & Reuse**
  - Inter-lingua
  - Specifications
  - Reliability
- **Control**
  - Controlled vocabularies
  - Accurate data collection or retrieval
  - Classification
  - Finding, sharing, discovering, navigation, indexing
One of the earliest examples of the benefits of ontologies
Integration and interoperability were big wins
Specific tool support
Considerable resources invested and continuing in maintenance
Translation into DLs
Spawned more generic biological ontology efforts
Manufacturing: Aerospace

- Considerable work on ontologies for products and components
- Used in all stages of the life cycle, from design to in-service maintenance
- Need for multiple perspectives e.g.
  - Whole engine
  - Heat transfer
  - Cost model
  - Manufacturing
  - Assembling/Maintenance
Ontologies: Observations

- In any domain
  - Usually highly implicit
  - Poorly documented
  - Likely to be ambiguous, vague, inconsistent

- When modelling
  - Interaction Problem: tasks influence ontologies
  - Integration Problem: integrating multiple ontologies
  - Modularity Problem: how to modularise and what grain size?

- Maintenance
  - Ongoing maintenance overhead
  - Ontologies evolve and change
  - Design rationale is important

- Upside
  - They do facilitate interoperability
  - They do enhance reuse
  - They are becoming part of the infrastructure
Standards are fundamental

HTML
XML + Name Space + XML Schema
RDF(S)
RDF
OWL
XOL
Topic Maps
SMIL
HTML
XML + Name Space + XML Schema
Unicode
URI
Drivers

- Moores Law and Powers of 10
- The WWW: Blessing and Curse
- Making the Web Semantic
- Intelligence on the Web
- Research on the Semantic Web
- Futures Challenges
AKT started Sept 00, 6 years, £8.8 Meg, EPSRC

www.aktors.org

Around 65 investigators and research staff
Infrastructures and Components

- Built core infrastructures
- Constructed component technologies that cover the knowledge life cycle in a number of applications

Six challenges define the Life Cycle:
Acquire • Model • Reuse • Retrieve • Publish • Maintain knowledge
Exemplar Technology: ClassAKT
Component Technologies: Modelling

Professor Nigel Shadbolt

Address: Department of Electronics and Computer Science, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom

Email: nshadbolt@ecs.soton.ac.uk

Tel: 44 23 8059 7602 (direct) 44 23 8059 5015 (secretary)

Fax: 44 23 8059 3313

Directions: to the University - to my office (Room 227, Building S9)

Projects | Publications | Presentations | CV

Nigel Shadbolt is Professor of Artificial Intelligence (AI) in the Department of Electronics and Computer Science at Southampton University. He is a member of the Intelligent Agents, Multimedia Group.

His research concentrates on two ends of the spectrum of AI - namely, Knowledge Technologies and Bioinformatics. For fifteen years he was Director of the AI Group at the School of Psychology at the University of Nottingham. He established an international reputation for work in Knowledge Technologies, in particular developing methods, tools and techniques to support the construction of knowledge-based systems. Much of this work was undertaken collaboratively and in many cases led to deployed applications. In 2000 he led a consortium of the Universities that secured an EPSRC Interdisciplinary Research Collaboration in Advanced Knowledge Technologies. Professor Shadbolt is the Director of this eight million pound, six-year research programme that is pursuing basic and applied research in the provision of technologies to support Knowledge Management and realise the promise of the Semantic Web. The AKT project has attracted additional funding including two grants for which Professor Shadbolt is also the Principal Investigator. The first of these (CovACT) is investigating the use of knowledge technologies in a collaborative context using multi-modal videoconference technologies such as the Access Grid. The second (MAKT) is attempting to support medical decision-making through the use of AKT’s tools, methods and techniques. He is also an investigator on a University of Southampton e-science project ORIONE where he is directing the effort to integrate knowledge engineering methods into design optimisation tools that exploit Grid computing. He is also working with his colleague Professor Nick Jennings on a Hewlett-Packard sponsored project RNSQ that aims to develop a framework within which software agents can be furnished with knowledge acquisition capabilities.

His work on bio-informatics investigates how we can draw inspiration from real animal systems in the construction of robots and biologically-inspired neural modelling. Research work has included investigations into simple sensory systems for mobile robots and hybrid architectures for autonomous robot systems. He works with Dr Jerry Krams on the development of computational models of neural plasticity that reflect the dynamic nature of statistical change in the nervous system. This work is also now being used in robots to understand how we could build flatten humanoid robots.
Component Technologies: Modelling
Integrating Semantic Spaces

- Exploit core infrastructure
- Integrate component technologies from the knowledge life cycle for an application

Six challenges define the Life Cycle:
Acquire • Model • Reuse • Retrieve • Publish • Maintain knowledge
The CS AKTive Space: International Semantic Web Challenge Winner

- 24/7 update of content
- Content continually harvested and acquired against community agreed ontology
- Easy access to information gestalts - who, what, where
- Hot spots
  - Institutions
  - Individuals
  - Topics
- Impact of research
  - citation services etc
  - funding levels
  - Changes and deltas
- Dynamic Communities of Practice…
Extending the model
Mediate and Aggregate: UK Research Councils

EPSRC

BBSRC

MRC

?
EPSRC: Knowing what they know

data sources

gatherers and mediators

EPSRC

ontology

knowledge repository (triplestore)

bbsrc

applications

MRC

Medical Research Council
Mediate and Aggregate: Ontologies
Visualising Interaction
Integrating Semantic Spaces

- Exploit core infrastructure
- Integrate component technologies from the knowledge life cycle for an application

Six challenges define the Life Cycle:
Acquire • Model • Reuse • Retrieve • Publish • Maintain knowledge
Collaborative Medical Decision Making MIAKT: Integrating Semantic Spaces

- Diverse and heterogeneous content
- Clinical examination
  - Notes
- Imaging
  - X-ray,
  - Ultrasound
  - MRI
- Microscopy
  - Histopathology
- Treatment
  - Protocol Records
  - Re-assessment
- Medical Records
  - Case sets
  - Individual patient records
- Published background
  - Epidemiology
  - Medical Abstracts
Why This Domain?

The user and patient perspective
- ≈ 10% of women develop breast cancer during their lives
- 3 million screening cases in UK per year: 8~25% cancers are missed by radiologists; 70~80% biopsies turn out to be benign; inter- and intra-radiologist variability is typically 30%.

AKT perspective
- Large amounts of data, information and knowledge
- Collaboration among different domain experts but little technical support for the collaborative elements of the problem and little technical support for the semantics of the domain

MIAS perspective
- Deployment of methods as web and grid services
- Need for information management
Multi-disciplinary Assessment: The Medics

- Different domains of expertise overlaid
- Breast imaging – X-ray, ultrasound, MRI
- Clinical examination
- Non-aggressive methods – Fine Needle Aspiration
- Histopathology Slides
- Microscopy – cells and tissues (also, hormone receptors)
- Prognosis, staging and statistical windows into the future
MIAKT Services

- **Image Analysis Services**
  - Oxford’s XRay Mammogram Analyser
  - KCL MRI Mammogram Analyser/Classifier

- **Classification Services**
  - Abnormality Naïve Bayes Classifier (Soton)
  - MRI Lesion Classifier (KCL)

- **Patient Data Retrieval Services (OU)**
  - For example, “Find Patients With Same Age”

- **Image Registration (KCL)**
  - GRID service invoked via web-service

- **Natural Language Report Generation (Sheffield)**
  - Generate a patient report from RDF description

- **UMLS Lookup (Sheffield)**
  - Lookup term definitions in the UMLS

- **Patient Records also accessed through web-service (Soton)**
  - Web-service enabled AKT 3store
<rdf:Description rdf:about="#g1p78_patient">
    <rdf:type rdf:resource="#Patient"/>
    <NS2:has_date_of_birth>01.01.1923</NS2:has_date_of_birth>
    <NS2:involved_in_ta rdf:resource="#ta_soton_000130051992"/>
</rdf:Description>

<rdf:Description rdf:about="#ta_soton_000130051992">
    <rdf:type rdf:resource="#Multi_Disciplinary_Meeting_TA"/>
    <NS2:involve_patient rdf:resource="#g1p78_patient"/>
    <NS2:consist_of_subproc rdf:resource="#oe_00103051992"/>
    <NS2:consist_of_subproc rdf:resource="#hp_00117051992"/>
    <NS2:consist_of_subproc rdf:resource="#ma_00127051992"/>
    <NS2:has_overall_impression rdf:resource="#assessment_b5_malignant"/>
    <NS2:has_overall_diagnosis>invasive carcinoma</NS2:has_overall_diagnosis>
</rdf:Description>

<rdf:Description rdf:about="#oe_00103051992">
    <rdf:type rdf:resource="#Physical_Exam"/>
    <NS2:has_date>03.05.1992</NS2:has_date>
    <NS2:produce_result rdf:resource="#oereport_glp78_1"/>
    <NS2:carried_out_on rdf:resource="#g1p78_patient"/>
</rdf:Description>

<rdf:Description rdf:about="#oereport_glp78_1">
    <NS2:type rdf:resource="#Lateral_OE_Report"/>
    <NS2:contains_roi rdf:resource="#oe_roi_00103051992"/>
    <NS2:has_lateral rdf:resource="#lateral_left"/>
</rdf:Description>
The MIAKT Framework
Demonstration
Drivers

- Moores Law and Powers of 10
- The WWW: Blessing and Curse
- Making the Web Semantic
- Intelligence on the Web
- Research on the Semantic Web
- Futures Challenges
New ways of discovery: e-Science

- A large part of scientific discovery is now a joint human machine endeavour.
- Without considerable compute power no hope of progress.
- Examples from physics, astronomy, biology, chemistry and engineering.
Entire E-Science Cycle
Encompassing experimentation, analysis, publication, research, learning
The need for *xtl*-Prints

Combechem

DATA → PUBLICATION → DISSEMINATION

Combichem
2,2-(3'-amino-1'-propanoxy)-4,6-oxy(tetraethylenoxy)-4,6-dichlorocyclotriphosphazatriene

Simon J Coles and Michael B Headzhouse.
University of Southampton

C81H82Cl4N4O6P3

Code: ICD41-1,12Beta4-4C1H2OC2H4O6P3Mo4-12C4-15-25(13,17-26-10-24)

Compounds Class: Energetic

Keywords: cyclotriphosphazene

Crystal Data: 18 October 2004

Deposited By: Dr Simon J Coles

Data collection parameters

<table>
<thead>
<tr>
<th>Chemical formula</th>
<th>C11 H23 Cl2 N4 O6 P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crystallographic solvent</td>
<td>Acetone Other Solvents Acetone</td>
</tr>
<tr>
<td>Crystal morphology</td>
<td>Single Crystal Size 0.42 x 0.12 x 0.01 mm65</td>
</tr>
<tr>
<td>Crystal system</td>
<td>Triclinic</td>
</tr>
<tr>
<td>Space group symbol</td>
<td>P21(2121)</td>
</tr>
<tr>
<td>Cell length a</td>
<td>13.64(3)</td>
</tr>
<tr>
<td>Cell length b</td>
<td>17.34(3)</td>
</tr>
<tr>
<td>Cell length c</td>
<td>33.734(8)</td>
</tr>
<tr>
<td>Cell angle alpha</td>
<td>90.00</td>
</tr>
<tr>
<td>Cell angle beta</td>
<td>90.00</td>
</tr>
<tr>
<td>Cell angle gamma</td>
<td>90.00</td>
</tr>
<tr>
<td>Date collection temperature</td>
<td>150(2)</td>
</tr>
<tr>
<td>Refinement results</td>
<td></td>
</tr>
</tbody>
</table>

Available files

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>04cc0794.cf</td>
<td>60k</td>
</tr>
<tr>
<td>04cc0794.cel</td>
<td>28k</td>
</tr>
<tr>
<td>04cc0794_checksum.dat</td>
<td>12k</td>
</tr>
<tr>
<td>04cc0794.hkl</td>
<td>62k</td>
</tr>
<tr>
<td>04cc0794_hKL</td>
<td>12k</td>
</tr>
<tr>
<td>04cc0794.pcp</td>
<td>13k</td>
</tr>
<tr>
<td>04cc0794.spc</td>
<td>97k</td>
</tr>
<tr>
<td>04cc0794.xrl</td>
<td>97k</td>
</tr>
<tr>
<td>Solution</td>
<td></td>
</tr>
<tr>
<td>Processing</td>
<td></td>
</tr>
</tbody>
</table>
Increasing Use of Value Added Services
Communities of Authors

- Web services to run over archives at varying grain size
Hubs and Authorities

- Begin with existing measures: document count and citation count.
- Note that higher citation count may not mean higher authority rating: quality citations are what count.

<table>
<thead>
<tr>
<th>Author Name</th>
<th>Citeseer DocID</th>
<th>Citations</th>
<th>Authority Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anant Agarwal</td>
<td>agarwal91mit</td>
<td>24</td>
<td>6.5161</td>
</tr>
<tr>
<td>Anant Agarwal</td>
<td>agarwal95mit</td>
<td>9</td>
<td>6.4675</td>
</tr>
<tr>
<td>Anant Agarwal</td>
<td>kranz93integrating</td>
<td>13</td>
<td>5.1927</td>
</tr>
<tr>
<td>Anant Agarwal</td>
<td>agarwal88evaluation</td>
<td>21</td>
<td>2.9159</td>
</tr>
</tbody>
</table>
Turning Points and Centrality

- Allows us to find turning points in scientific development: Kuhn’s “paradigm shift” moment.
- ‘Centrality’ measure to be applied to same Citeseer data.
Bursting onto the scene: New Topics

New topics in research literatures
Detecting Key Moments

→ Cause and effect chains in content sets
Future Challenges

- Developing, Deploying, Managing and Reusing Ontologies
- Co reference resolution and referential integrity on the SW
- “Inference” on the web
- The annotation bottleneck
- Annotating Multimedia Content
- Composing Knowledge Services
- Trust
  - Representing provenance
  - Mechanisms for trust
More Information

www.aktors.org