School of Electronics and Computer Science

Web Intelligence

Professor Nigel Shadbolt

Why distributed and digital is good…

The work of many people…

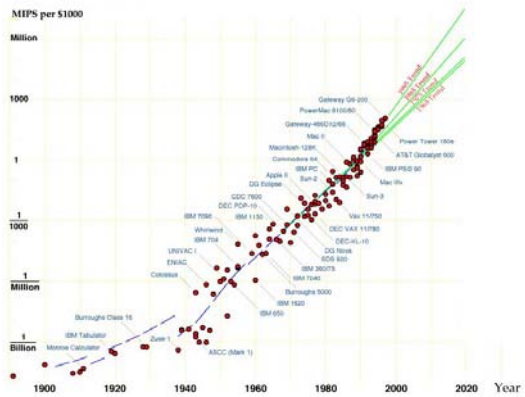
- Harith Alani
- Steve Harris
- Nick Gibbins
- Yannis Kalfoglou
- Kieron O'Hara
- David Dupplaw
- Bo Hu
- Paul Lewis
- Srinandan Dashamapatra
- Duncan Macrae-Spencer
- Hugh Glaser
- Les Carr
- David de Roure
- Wendy Hall
- Mike Brady
- David Hawkes
- Yorick Wilks
- Enrico Motta
- Carole Goble
- Simon Cox
- Andy Keane
- :

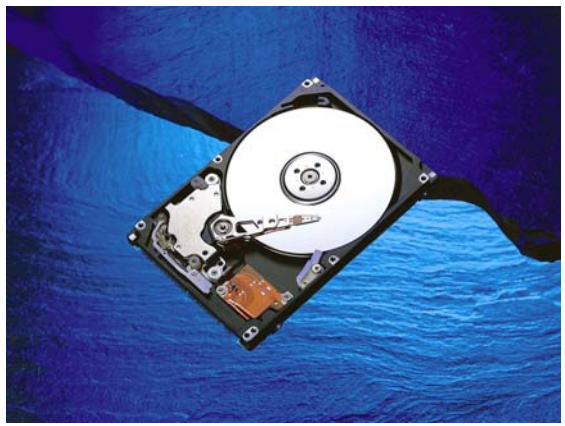
Drivers

- 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





Alters our Perspectives

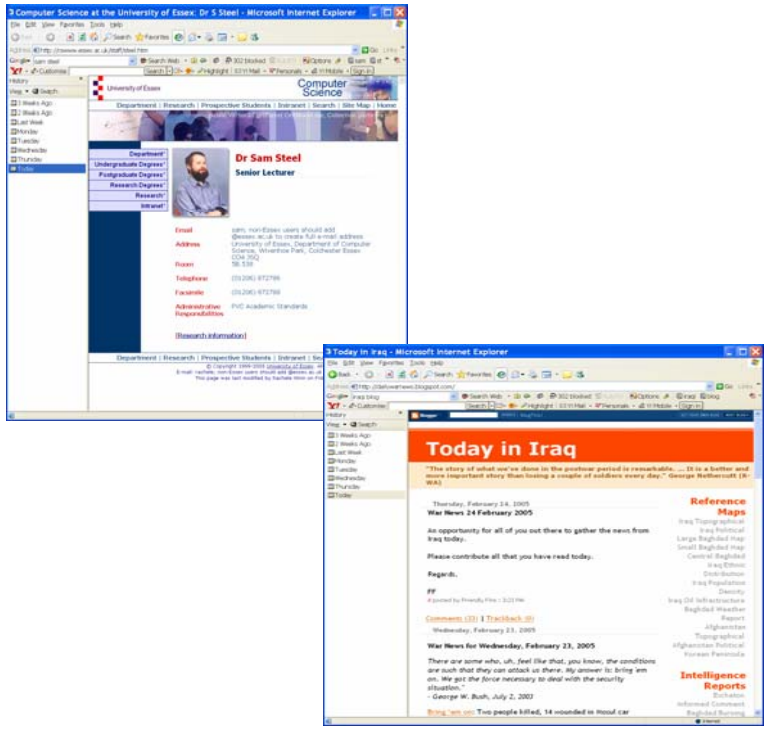
- In May 1997 Deep Blue beat world chess champion Kasparov
- It won the six-game rematch 3.5-2.5
- The system derives its playing strength mainly out of brute force computing power
- It is a massively parallel, 32-node, RS/6000, SP-based computer system enhanced with 256 special purpose VLSI chess processors
- 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…

http://
www2002.org

WWW2002

THE ELEVENTH INTERNATIONAL
WORLD WIDE WEB CONFERENCE

Sheraton Waikiki Hotel
Honolulu, Hawaii, USA
7-11 May 2002

CONFERENCE ORGANIZERS
International World Wide Web Conference Committee

2002 HAWAII

1 LOCATION. 5 DAYS. LEARN. INTERACT.

- Conference Proceedings
- Call for Participation
- Program
- Registration Information
- Hotel Accommodation
- Conference Committee
- Sponsorship/ Exhibition Opportunities
- Volunteer Information
- Information about Hawaii
- Previous & Future WWW Conferences

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

REGISTER NOW

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 (IW^3C^2) 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 (IW^3C^2), 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 3Com Founders chair at the Laboratory for Computer Science (LCS) at the Massachusetts Institute of Technology (MIT).

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

Ian Foster, guru of "Grid Computing", associate

McArthur Prize Winner,

That is machine readable….

This is a type of object event and this is its title

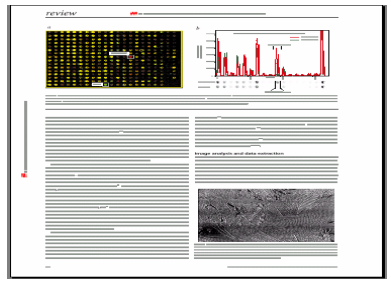
This is the URL of the web page for the event

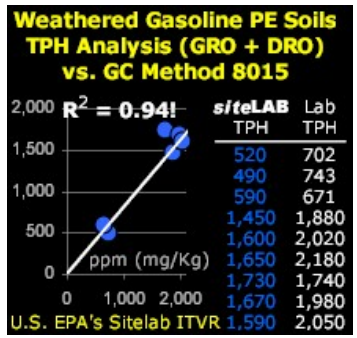
This is a type of object photograph and the photograph is of Tim Berners-Lee

Tim Berners-Lee is an invited speaker at the event

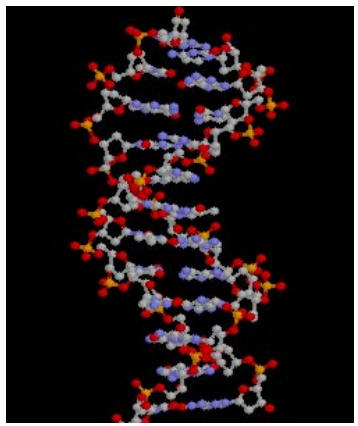
Can Annotate Anything

- Publications…
- Databases…
- Metadata on scientific structures



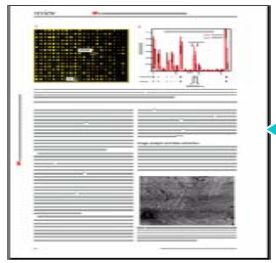


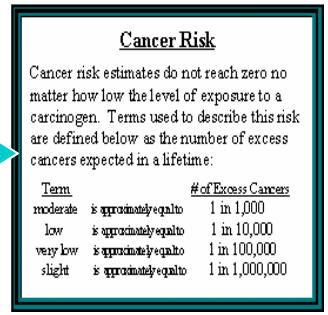
Web data set (XHTML)



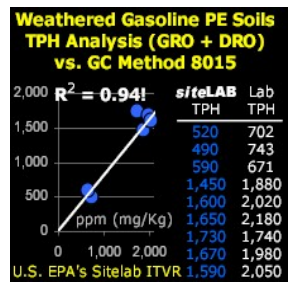
The SW Community: Structured Spaces

- 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).

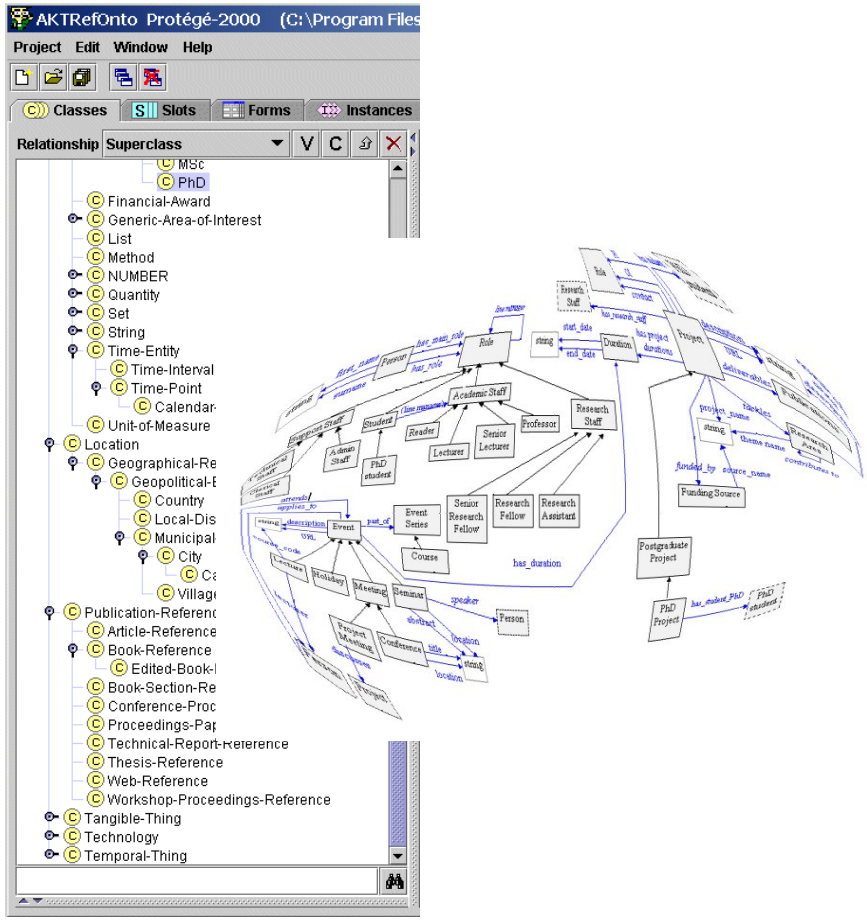


Web data set (XHTML)

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

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

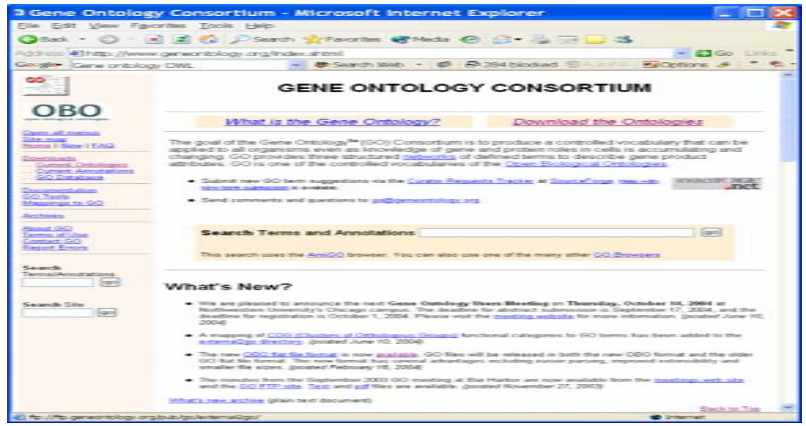


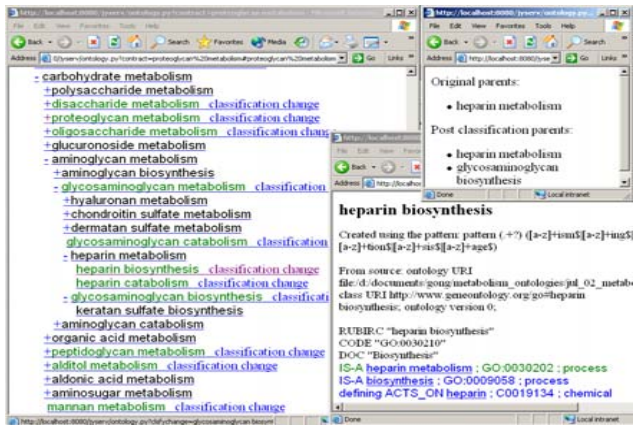
Ontologies offer….

- 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

Genetics: Gene Ontology

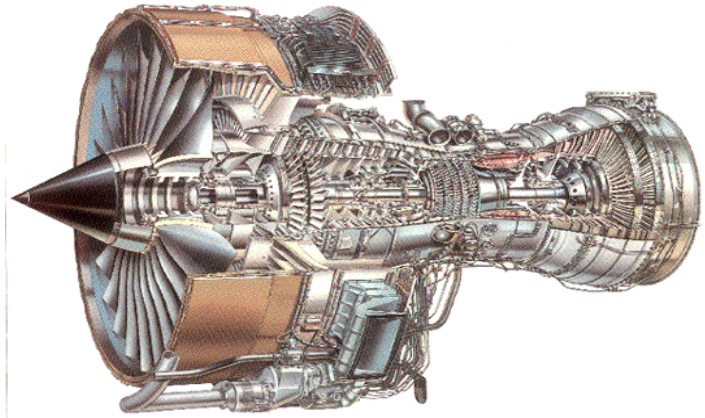
- One of the earliest examples of the benefits of ontologies
- Integration and interoperability were big wins
- Specific tool support
- Considerable resources invested and continuing in maintenance
- Translation into DLs
- Spawned more generic biological ontology efforts

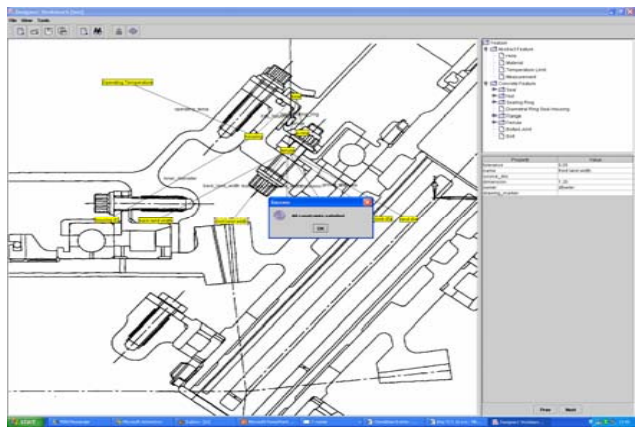




Manufacturing: Aerospace

- 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

OWL

RDF(S)

XOL | Topic Maps | SMIL | RDF

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

Advanced Knowledge Technologies IRC

University of Southampton

The Open University

1495 UNIVERSITY OF ABERDEEN

THE UNIVERSITY OF EDINBURGH

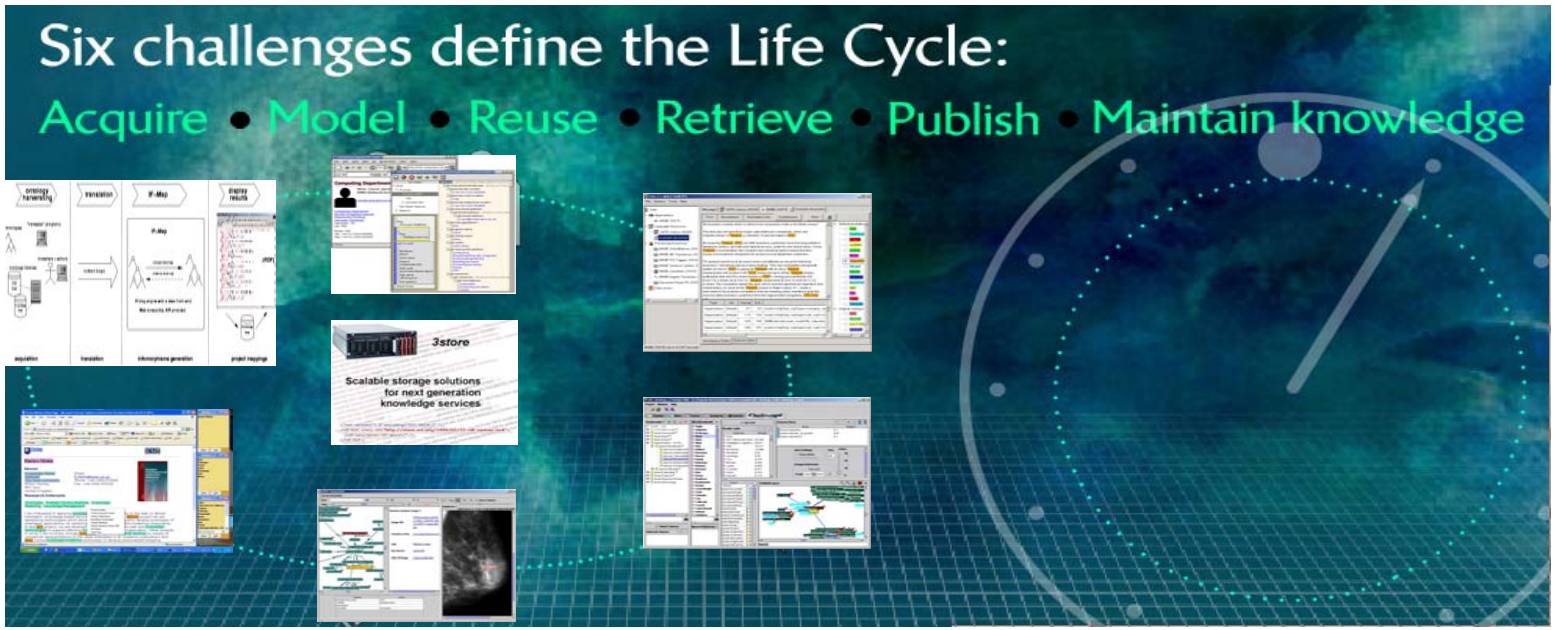
AKT started Sept 00, 6 years, £8.8 Meg, EPSRC

www.aktors.org

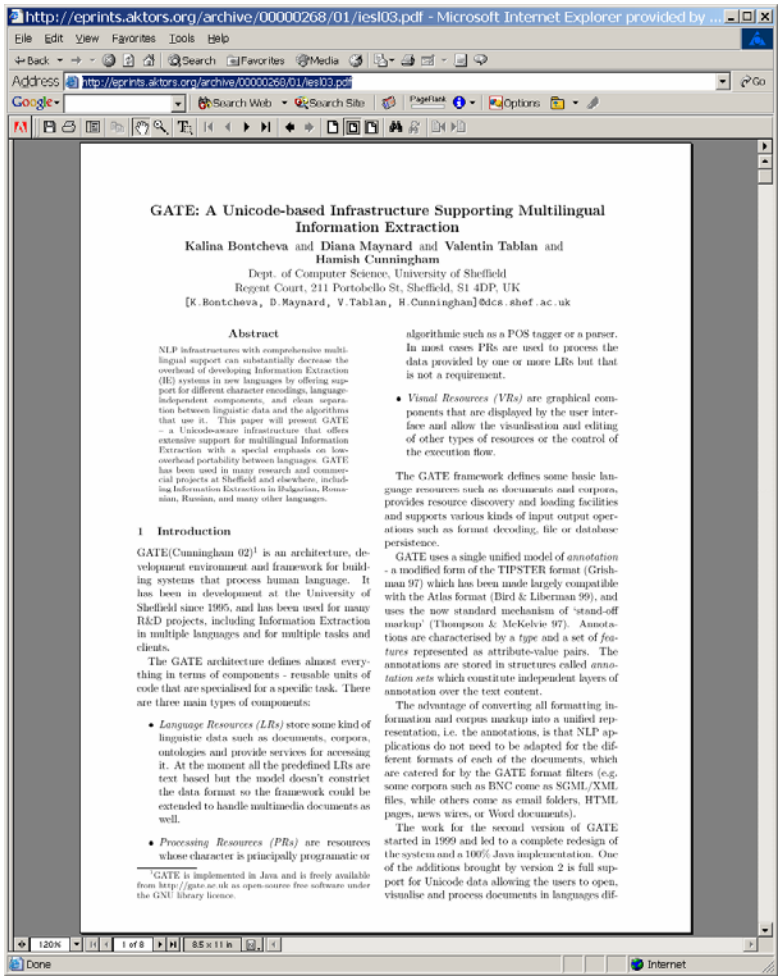
Around 65 investigators and research staff

Infrastructures and Components

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



Exemplar Technology: ClassAKT



http://eprints.aktors.org/archive/00000268/01/iesl03.pdf - Microsoft Internet Explorer provided by ...

File Edit View Favorites Tools Help

Address http://eprints.aktors.org/archive/00000268/01/iesl03.pdf

GATE: A Unicode-based Infrastructure Supporting Multilingual Information Extraction

Kalina Bontcheva and Diana Maynard and Valentin Tablan and Hamish Cunningham

Dept. of Computer Science, University of Sheffield
Regent Court, 211 Portobello St, Sheffield, S1 4DP, UK
{K.Bontcheva, D.Maynard, V.Tablan, H.Cunningham}@dcs.shef.ac.uk

Abstract

NLP infrastructures with comprehensive multilingual support can substantially decrease the overhead of developing Information Extraction (IE) systems in new languages by offering support for different character encodings, language-independent components, and clean separation between linguistic data and the algorithms that use it. This paper will present GATE – a Unicode-aware infrastructure that offers extensive support for multilingual Information Extraction with a special emphasis on low-overhead portability between languages. GATE has been used in many research and commercial projects at Sheffield and elsewhere, including Information Extraction in Bulgarian, Romanian, Russian, and many other languages.

1 Introduction

GATE(Cunningham 02)[1] is an architecture, development environment and framework for building systems that process human language. It has been in development at the University of Sheffield since 1995, and has been used for many R&D projects, including Information Extraction in multiple languages and for multiple tasks and clients.

The GATE architecture defines almost everything in terms of components - reusable units of code that are specialised for a specific task. There are three main types of components:

- *Language Resources (LRs)* store some kind of linguistic data such as documents, corpora, ontologies and provide services for accessing it. At the moment all the predefined LRs are text based but the model doesn't constrict the data format so the framework could be extended to handle multimedia documents as well.
- *Processing Resources (PRs)* are resources whose character is principally programatic or algorithmic such as a POS tagger or a parser. In most cases PRs are used to process the data provided by one or more LRs but that is not a requirement.
- *Visual Resources (VRs)* are graphical components that are displayed by the user interface and allow the visualisation and editing of other types of resources or the control of the execution flow.

The GATE framework defines some basic language resources such as documents and corpora, provides resource discovery and loading facilities and supports various kinds of input output operations such as format decoding, file or database persistence.

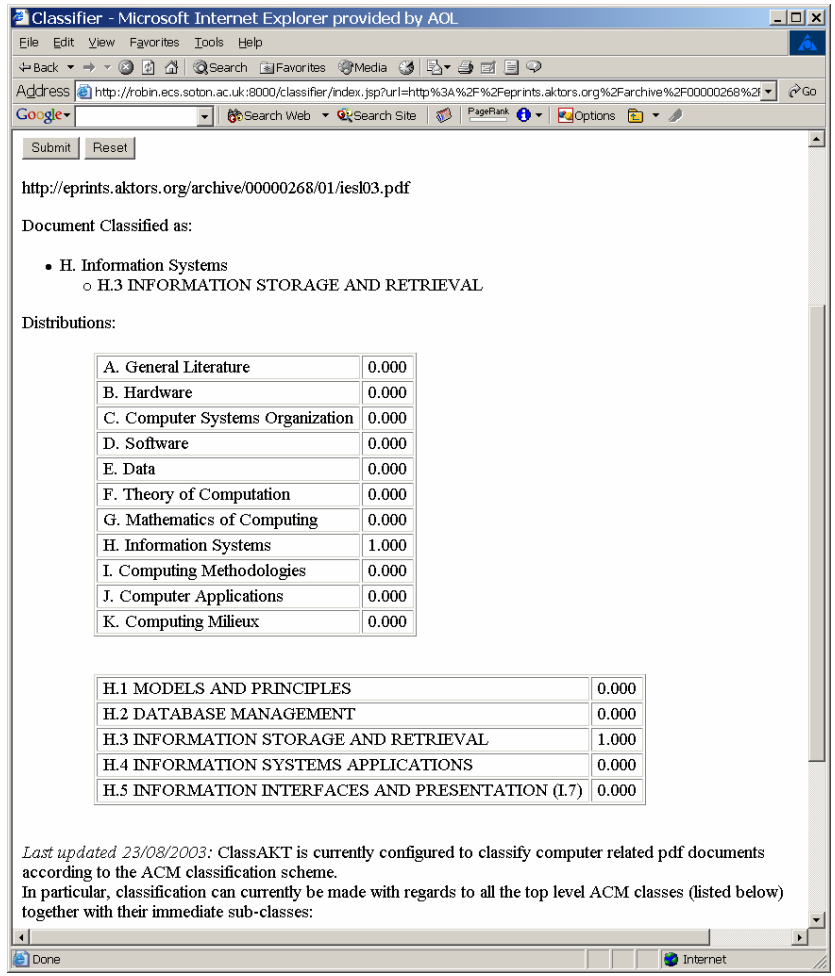
GATE uses a single unified model of *annotation* - a modified form of the TIPSTER format (Grishman 97) which has been made largely compatible with the Atlas format (Bird & Liberman 99), and uses the now standard mechanism of 'stand-off markup' (Thompson & McKelvie 97). Annotations are characterised by a *type* and a set of *features* represented as attribute-value pairs. The annotations are stored in structures called *annotation sets* which constitute independent layers of annotation over the text content.

The advantage of converting all formatting information and corpus markup into a unified representation, i.e. the annotations, is that NLP applications do not need to be adapted for the different formats of each of the documents, which are catered for by the GATE format filters (e.g. some corpora such as BNC come as SGML/XML files, while others come as email folders, HTML pages, news wires, or Word documents).

The work for the second version of GATE started in 1999 and led to a complete redesign of the system and a 100% Java implementation. One of the additions brought by version 2 is full support for Unicode data allowing the users to open, visualise and process documents in languages dif-

[1] GATE is implemented in Java and is freely available from http://gate.ac.uk as open-source free software under the GNU library licence.

Done Internet



Classifier - Microsoft Internet Explorer provided by AOL

File Edit View Favorites Tools Help

Address http://robin.ecs.soton.ac.uk:8000/classifier/index.jsp?url=http%3A%2F%2Feprints.aktors.org%2Farchive%2F00000268%2f

Submit Reset

http://eprints.aktors.org/archive/00000268/01/iesl03.pdf

Document Classified as:

- H. Information Systems
  - H.3 INFORMATION STORAGE AND RETRIEVAL

Distributions:

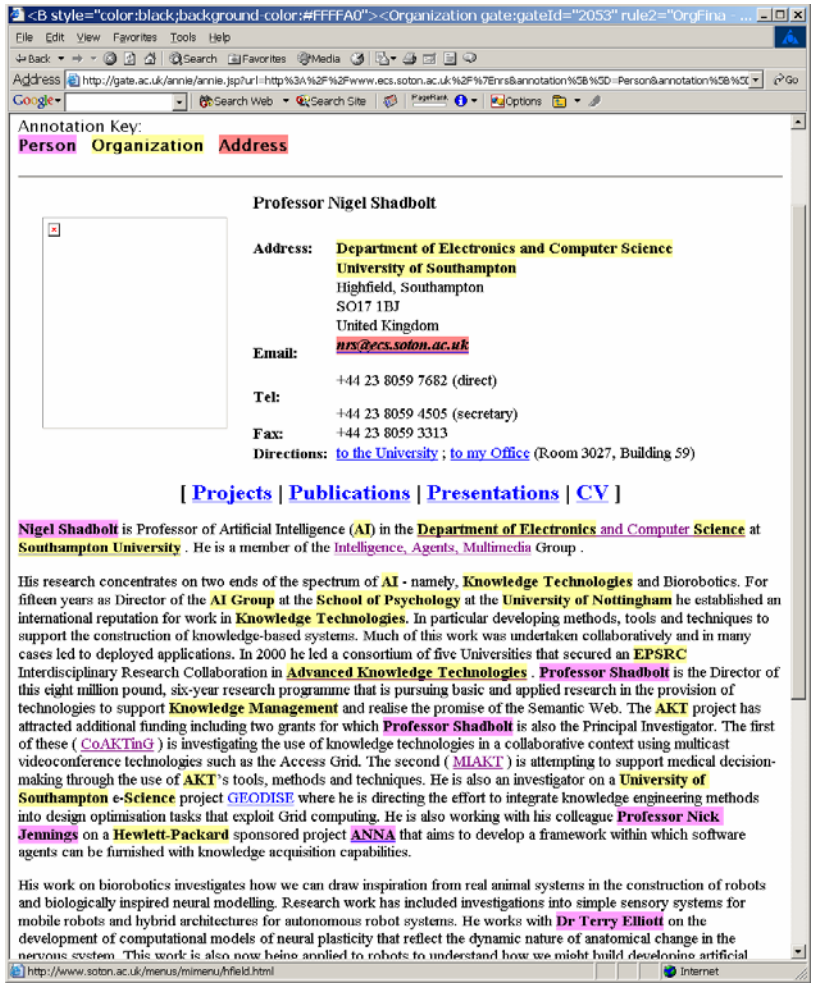
A. General Literature	0.000
B. Hardware	0.000
C. Computer Systems Organization	0.000
D. Software	0.000
E. Data	0.000
F. Theory of Computation	0.000
G. Mathematics of Computing	0.000
H. Information Systems	1.000
I. Computing Methodologies	0.000
J. Computer Applications	0.000
K. Computing Milieux	0.000

H.1 MODELS AND PRINCIPLES	0.000
H.2 DATABASE MANAGEMENT	0.000
H.3 INFORMATION STORAGE AND RETRIEVAL	1.000
H.4 INFORMATION SYSTEMS APPLICATIONS	0.000
H.5 INFORMATION INTERFACES AND PRESENTATION (I.7)	0.000

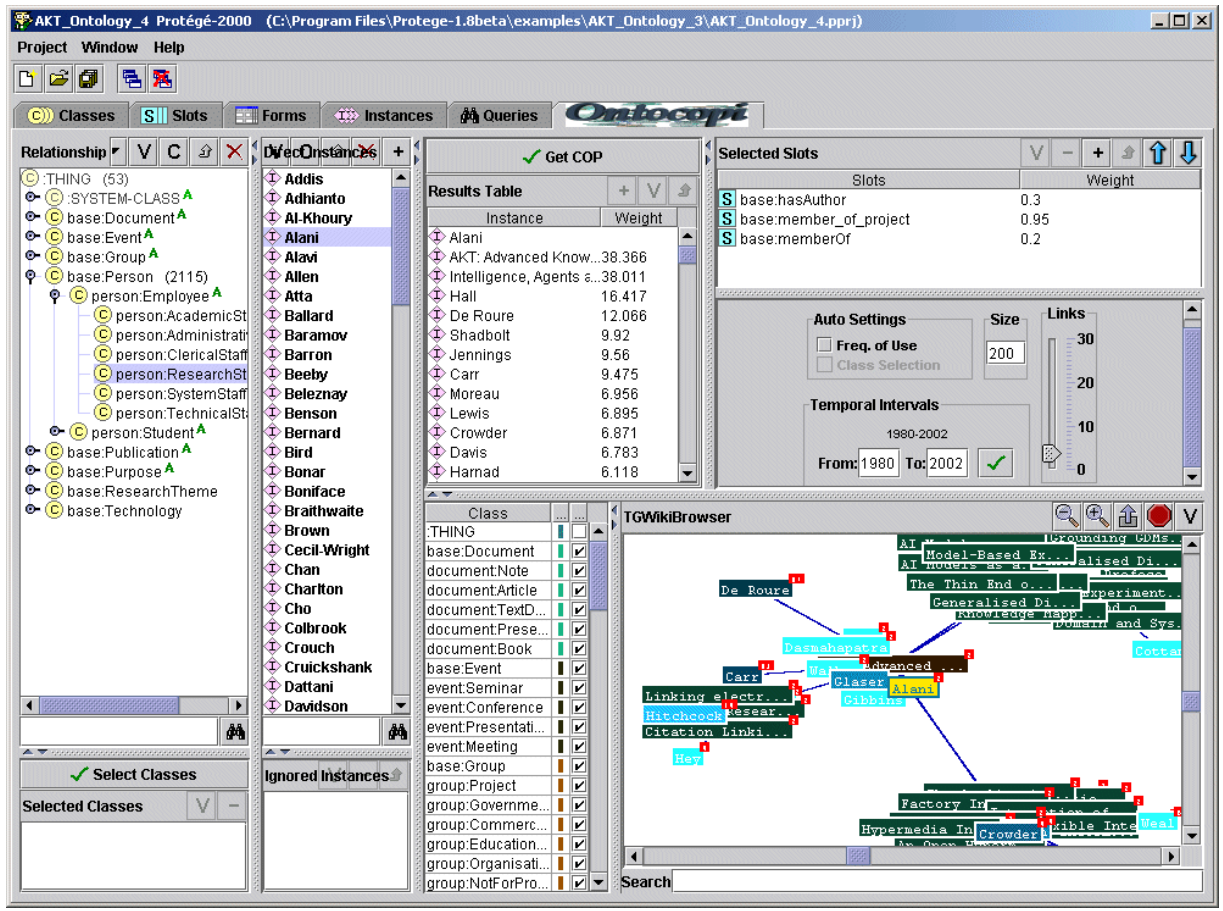
Last updated 23/08/2003: ClassAKT is currently configured to classify computer related pdf documents according to the ACM classification scheme.
In particular, classification can currently be made with regards to all the top level ACM classes (listed below) together with their immediate sub-classes:

Done Internet

Component Technologies: Modelling

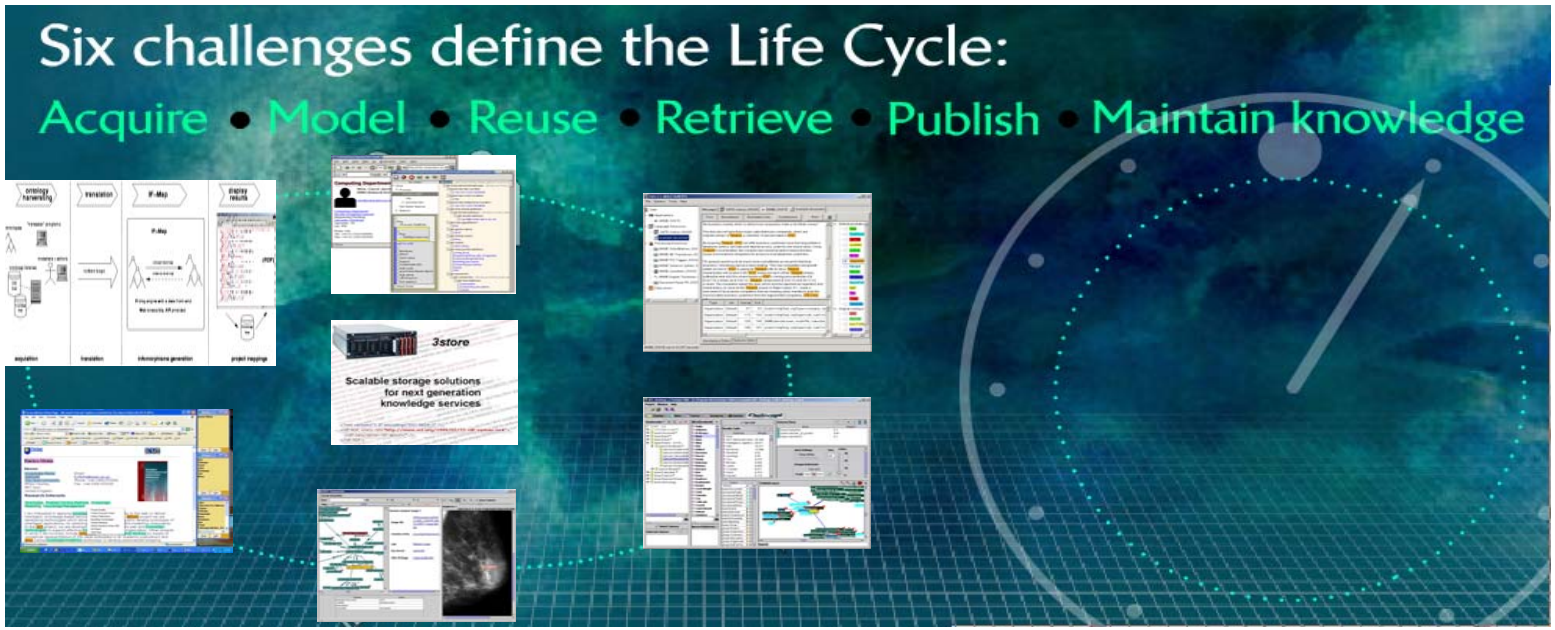


Component Technologies: Modelling



Integrating Semantic Spaces

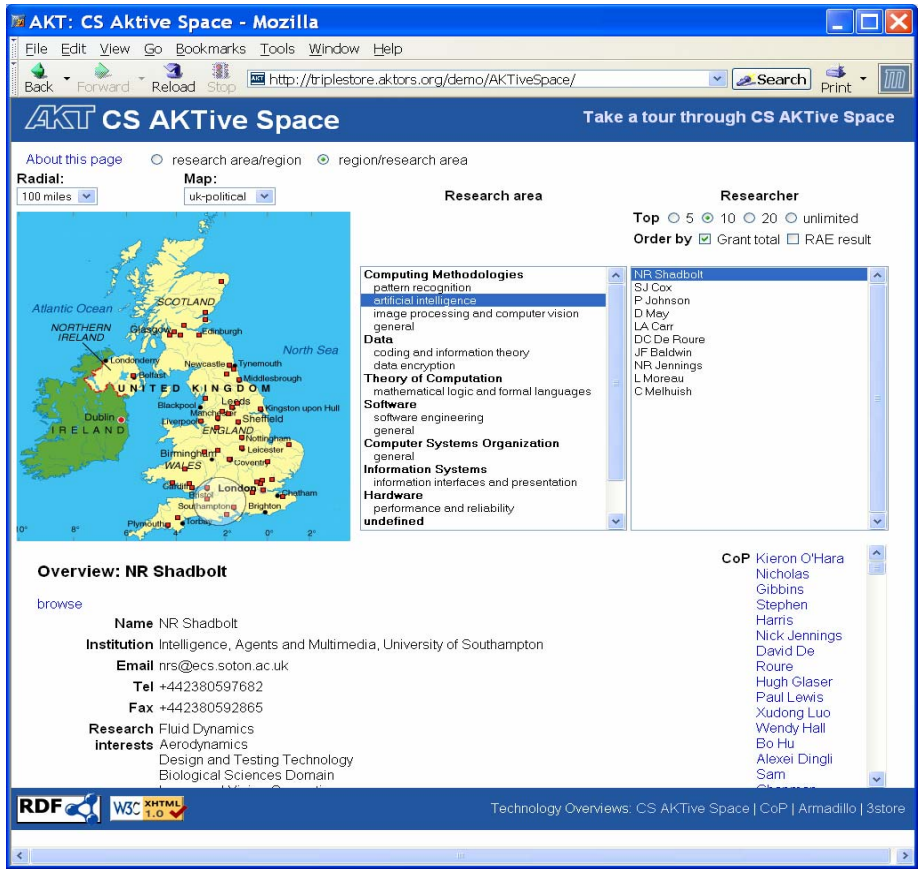
- Exploit core infrastructure
- Integrate component technologies from the knowledge life cycle for an application



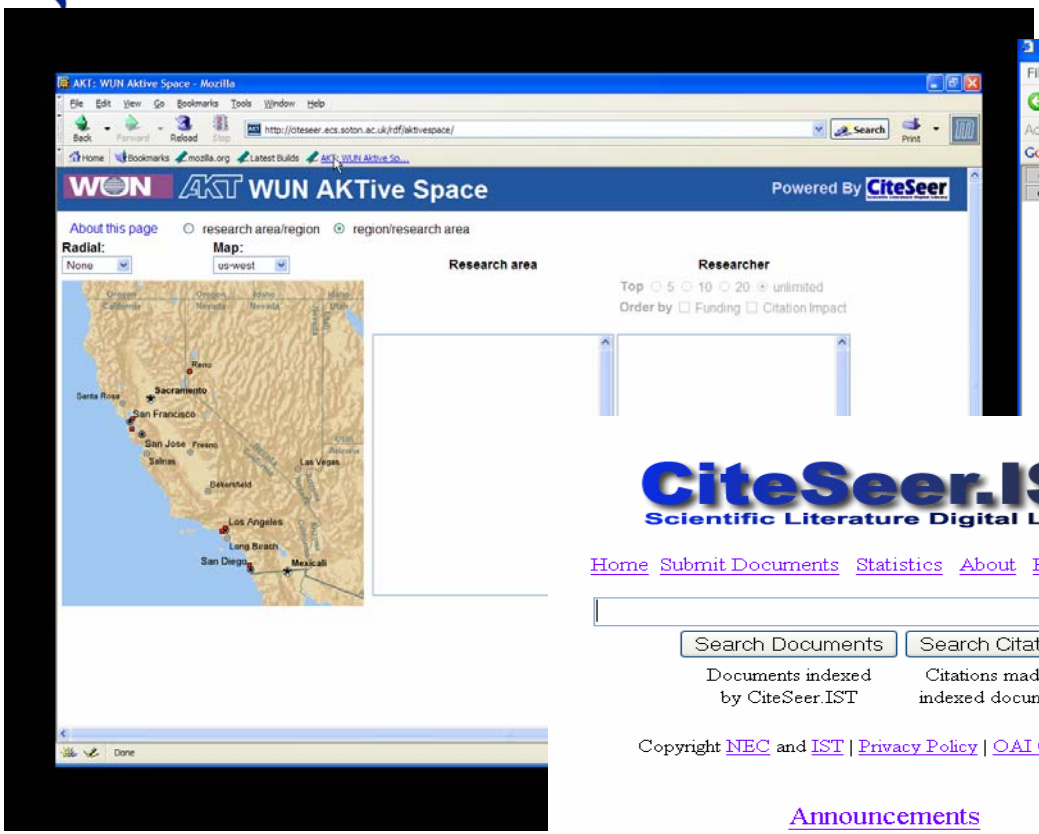
The CS AKTive Space: International Semantic Web Challenge Winner

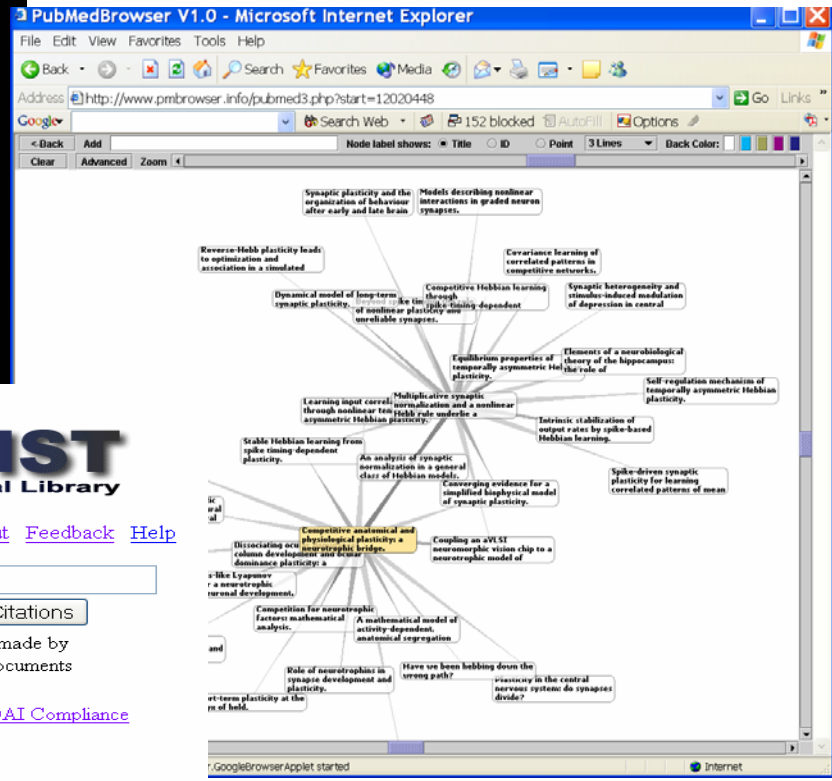
- 24/7 update of content
- Content continually harvested and acquired against community agreed ontology
- Easy access to information gestalts - who, what, where
- Hot spots
  - Institutions
  - Individuals
  - Topics
- Impact of research
  - citation services etc
  - funding levels
  - Changes and deltas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

CS AKTiveSpace



Extending the model





CiteSeer.IST
Scientific Literature Digital Library

Home Submit Documents Statistics About Feedback Help

Search Documents | Search Citations

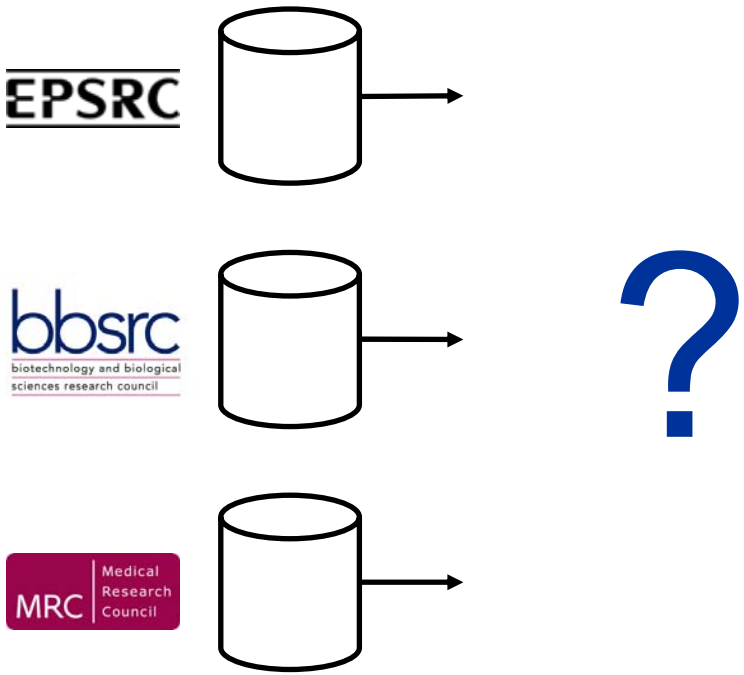
Documents indexed by CiteSeer.IST

Citations made by indexed documents

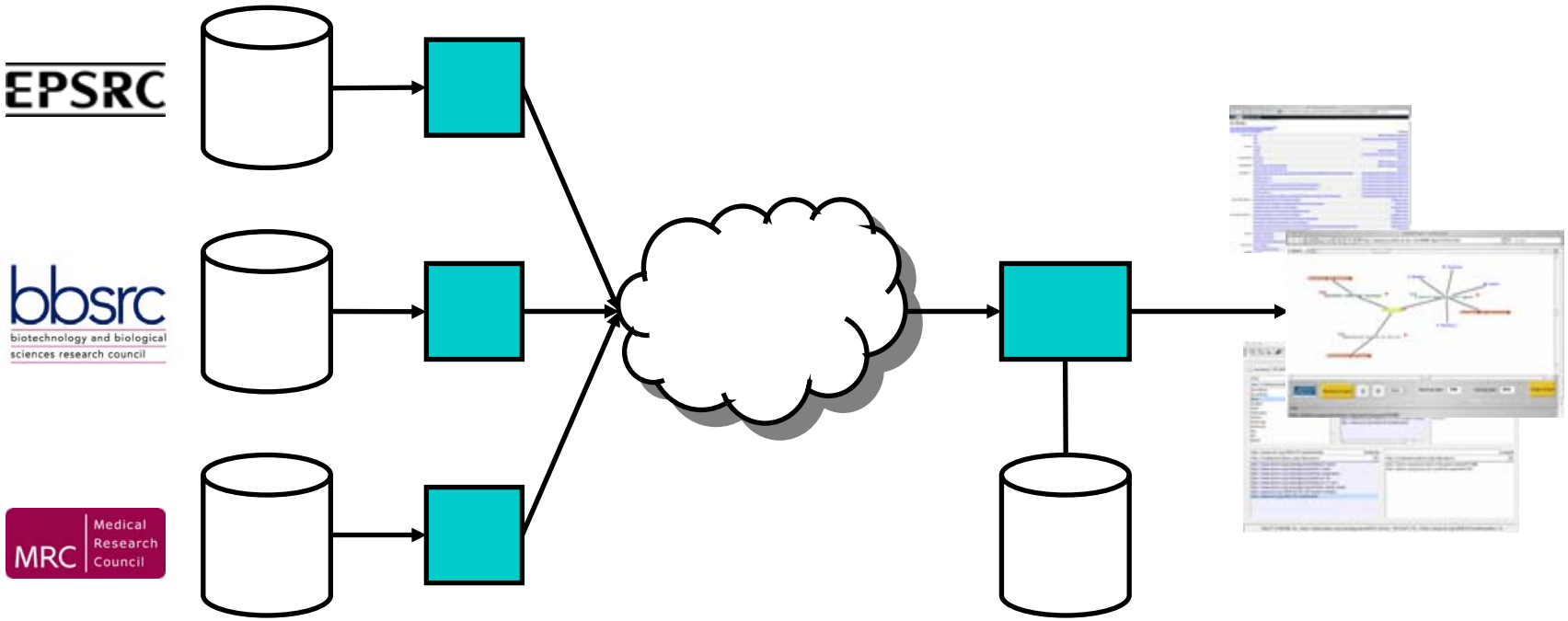
Announcements

National Science Foundation · Microsoft Research · IST · NEC

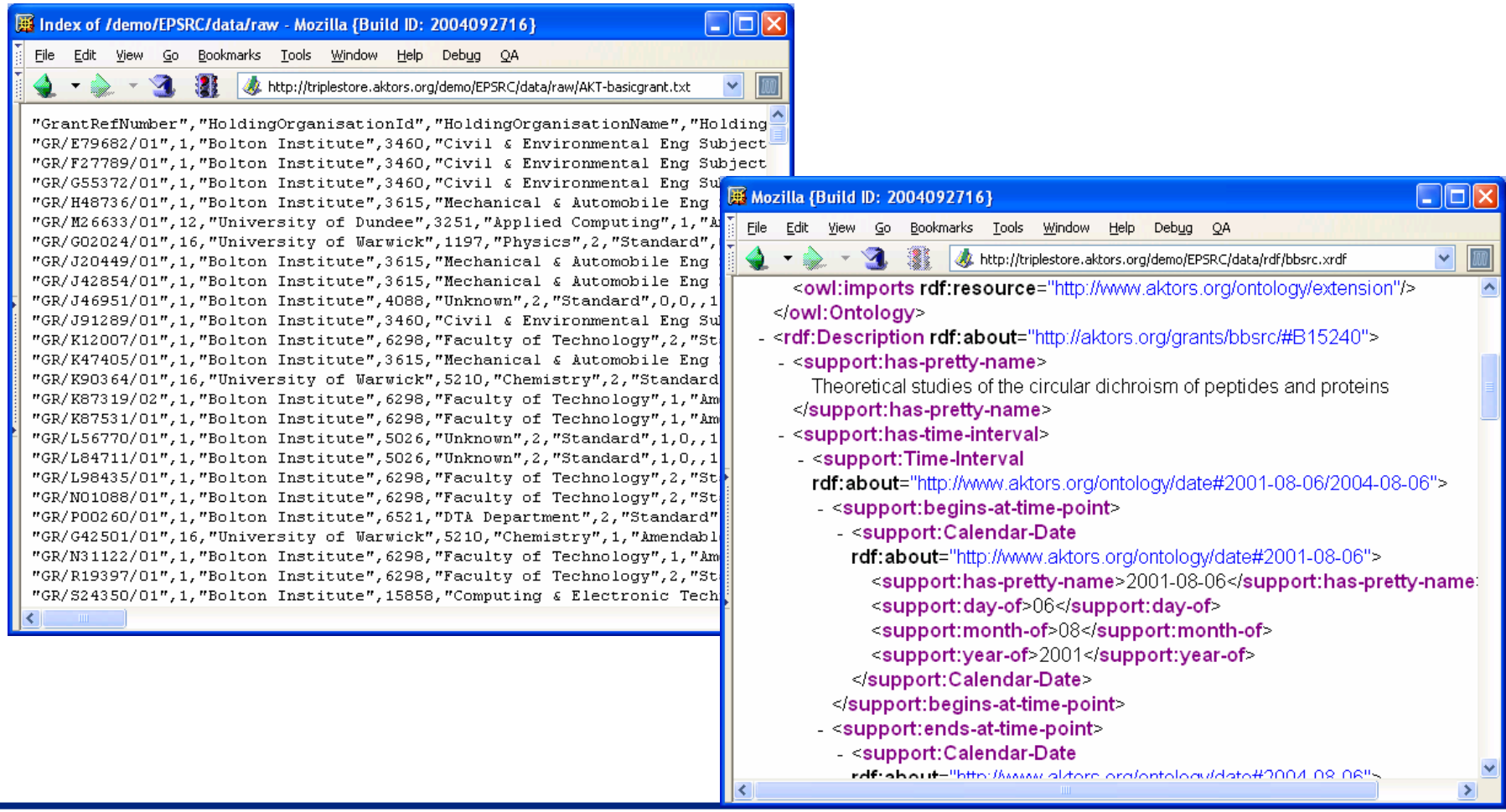
Mediate and Aggregate: UK Research Councils



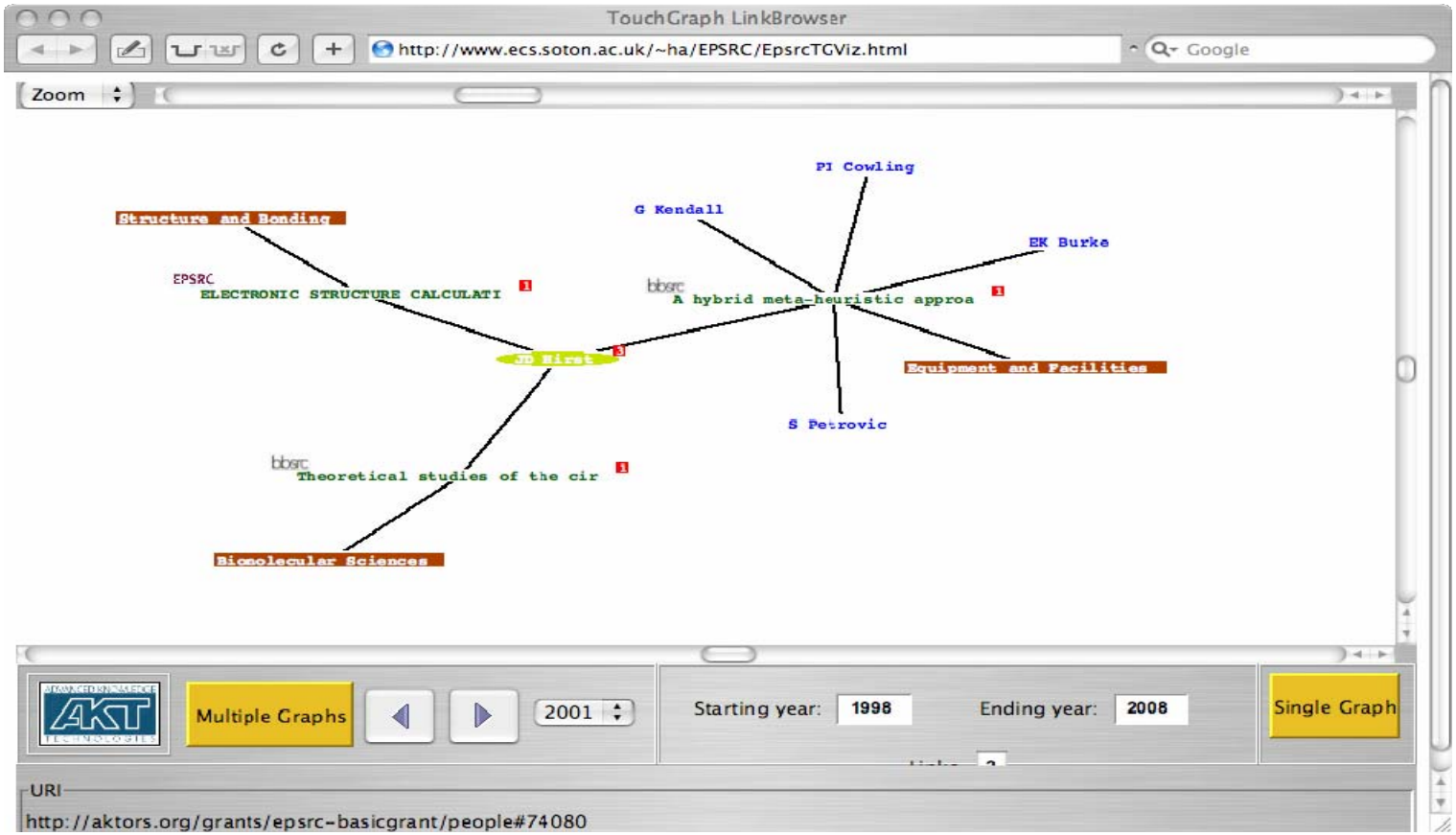
EPSRC: Knowing what they know



Mediate and Aggregate: Ontologies



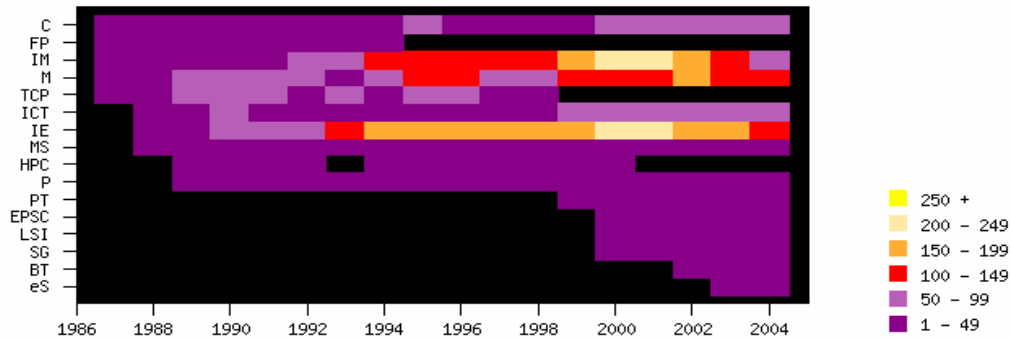
Visualising Interaction



Visualising Interaction: Programmes

Collaboration over time

View: by year by area



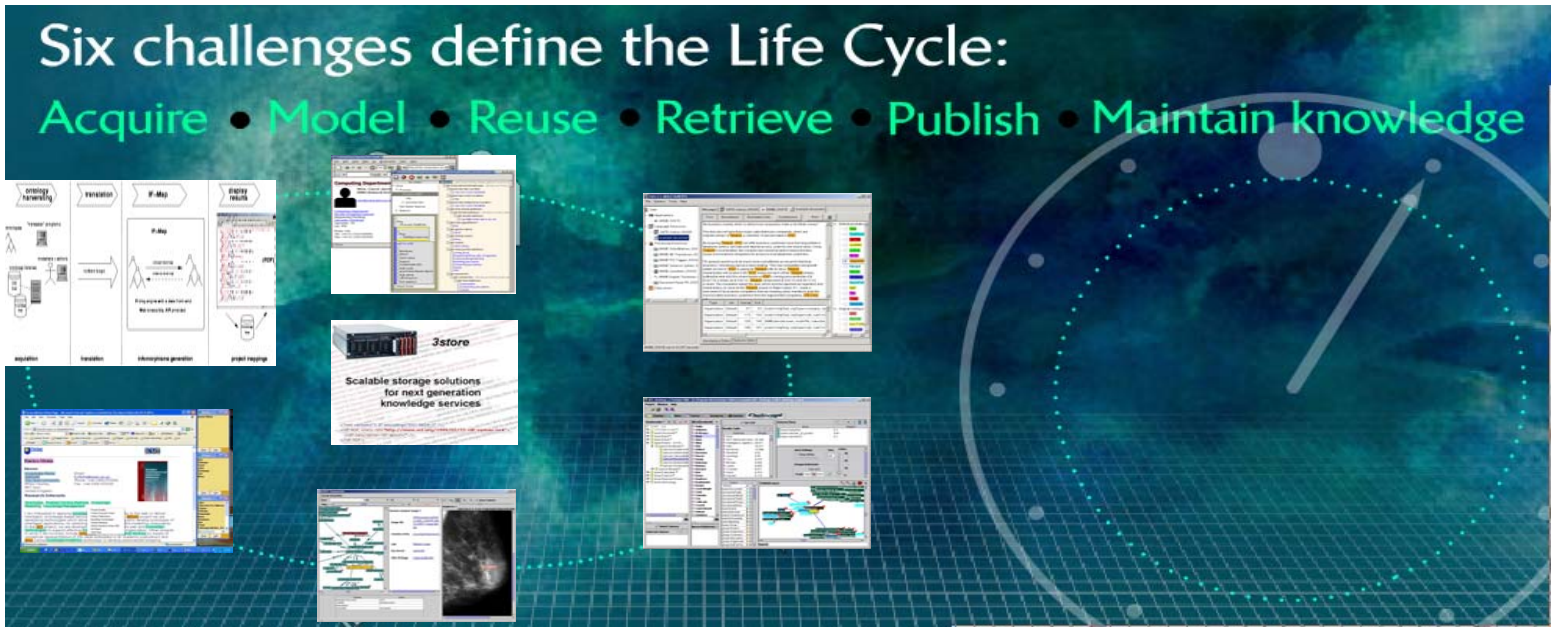
BT C E EPSC FP HPC ICT IE IM LSI M MS P PT SG TCP eS

− +

BT	Basic Technology	LSI	Life Sciences Interface
C	Chemistry	M	Materials
E	Engineering	MS	Mathematical Sciences
EPSC	Engineering & Physical Sciences Council	P	Physics
FP	Facilities Programme	PT	Postgraduate Training
HPC	High Performance Computing	SG	Schemes Group
ICT	Information and Communications Technology	TCP	Teaching Company Project
IE	Infrastructure and Environment	eS	e-Science
IM	Innovative Manufacturing		

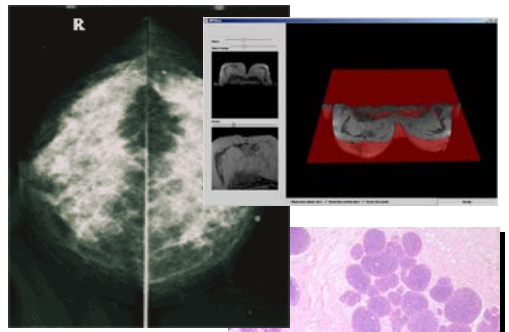
Integrating Semantic Spaces

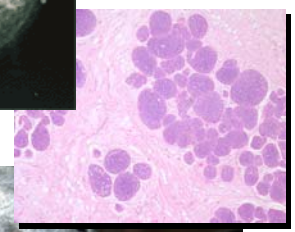
- Exploit core infrastructure
- Integrate component technologies from the knowledge life cycle for an application

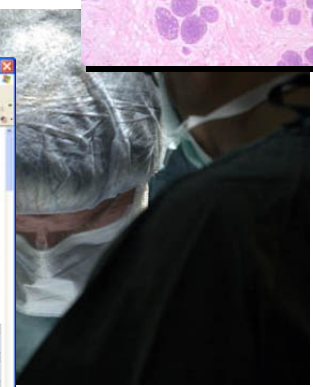


Collaborative Medical Decision Making MIAKT: Integrating Semantic Spaces

- 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





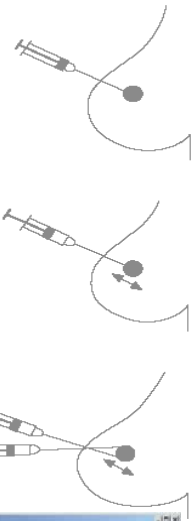


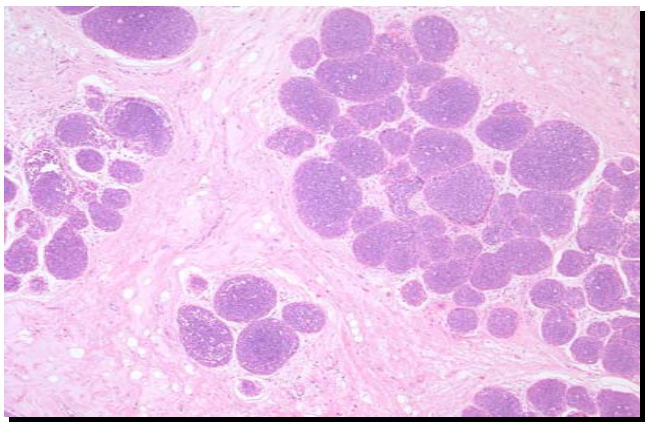
Why This Domain?

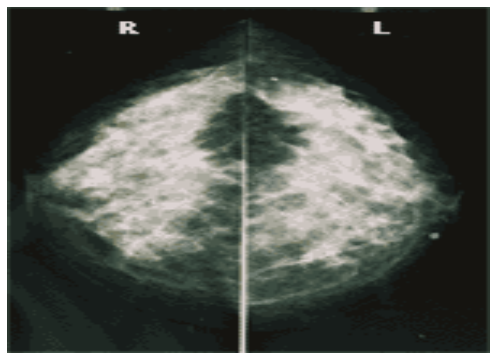
- The user and patient perspective
  - ≈ 10% of women develop breast cancer during their lives
  - 3 million screening cases in UK per year: 8~25% cancers are missed by radiologists; 70~80% biopsies turn out to be benign; inter- and intra-radiologist variability is typically 30%.
- AKT perspective
 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

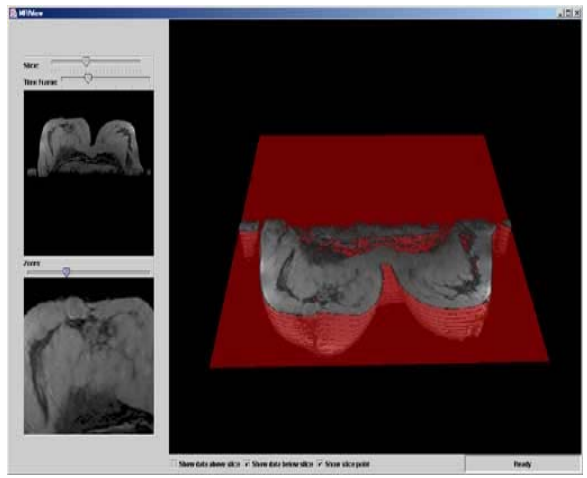
Multi-disciplinary Assessment: The Medics

- 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









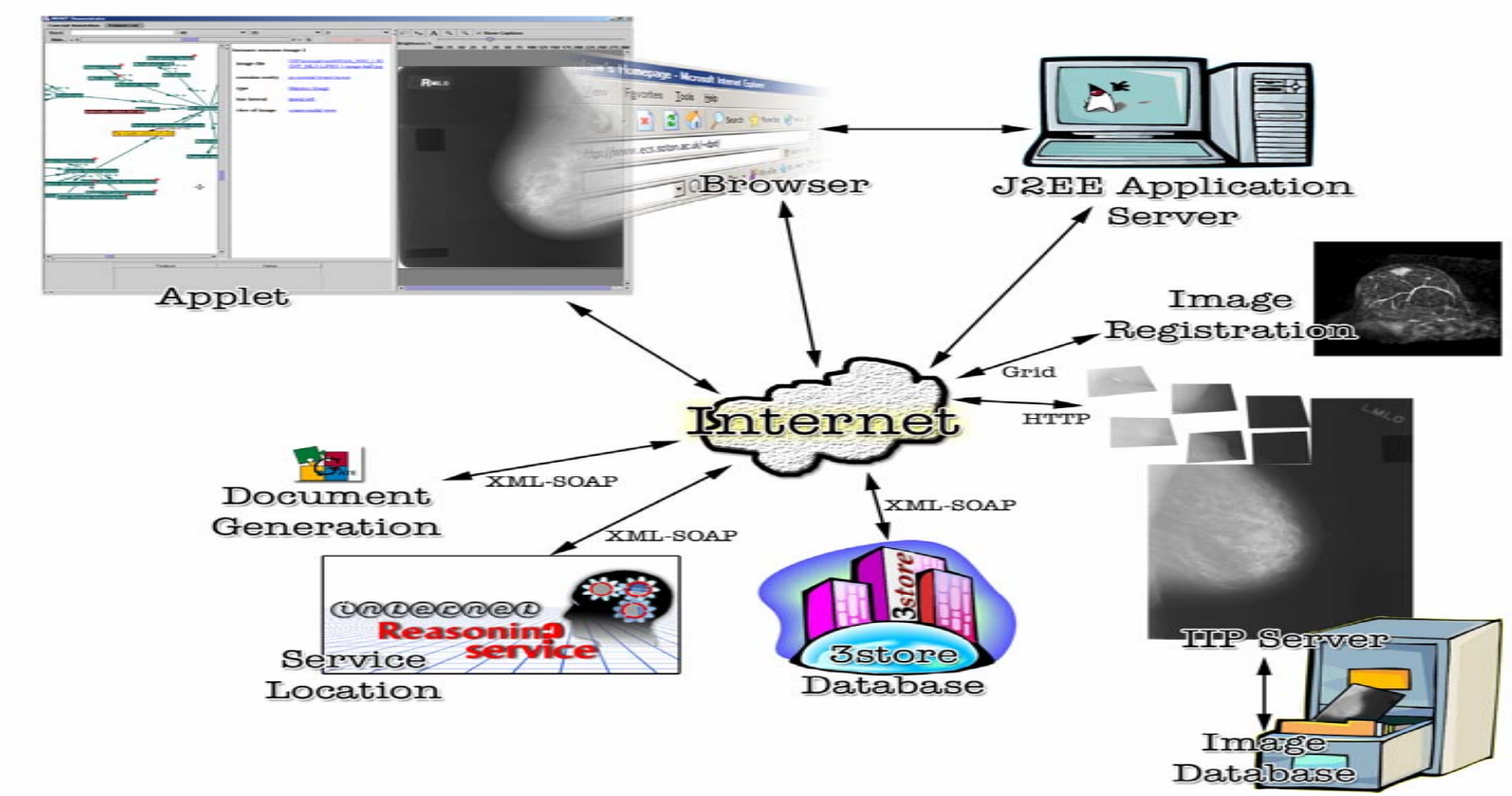
MIAKT Services

- 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

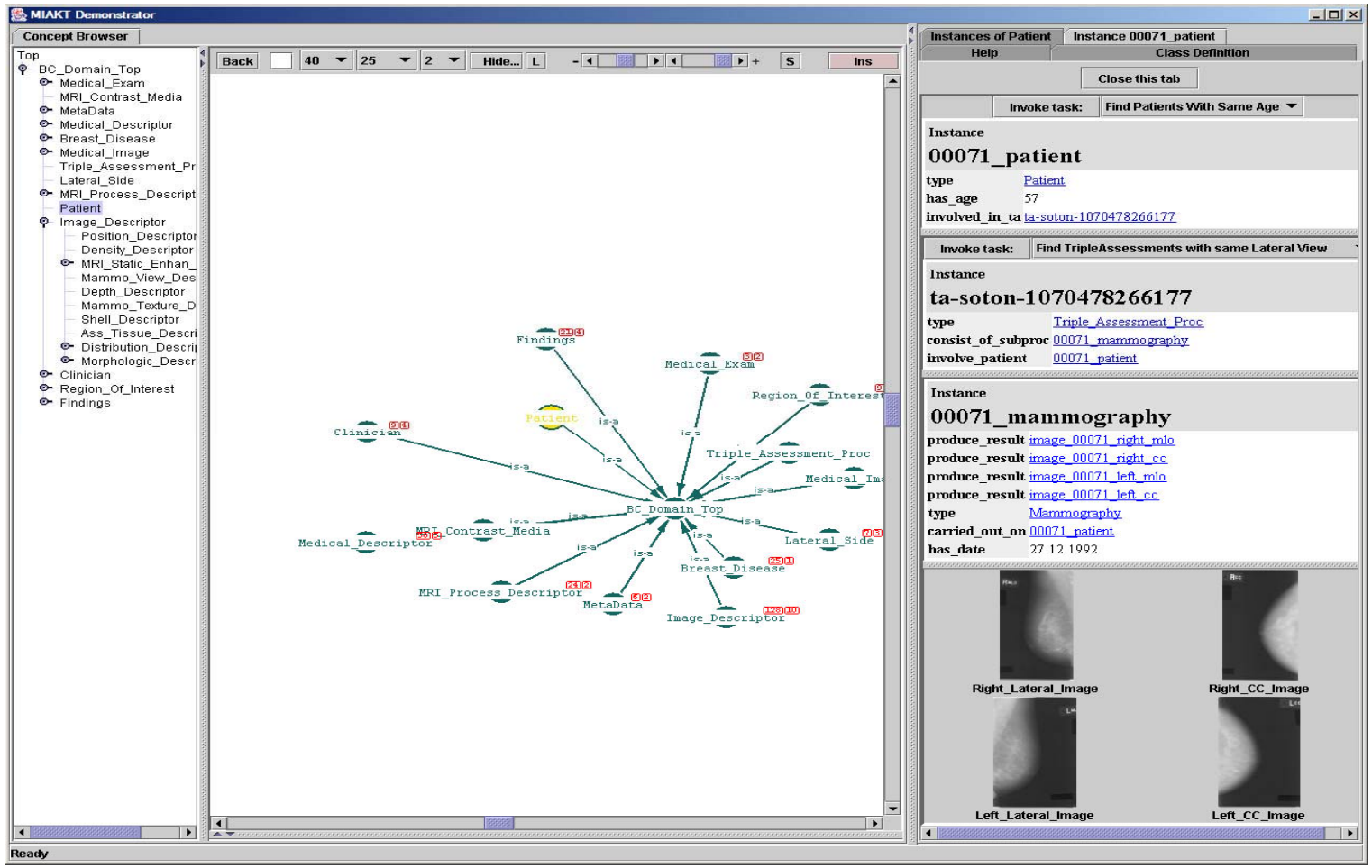
Patient Cases in RDF

```
<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

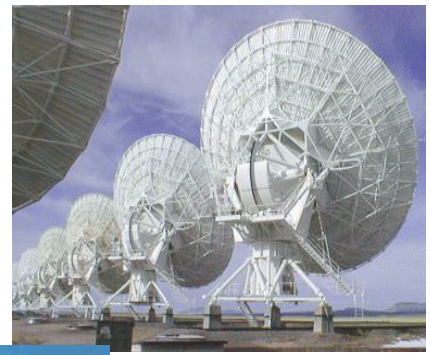


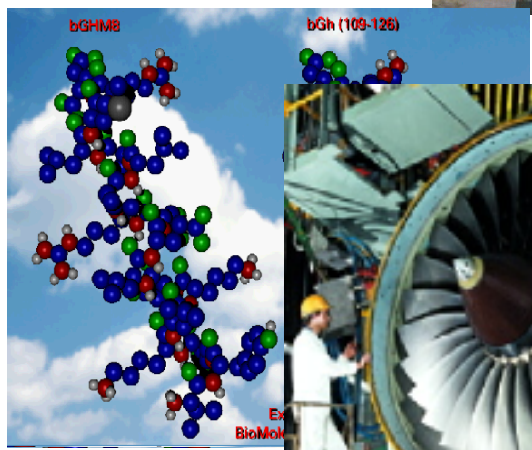
Drivers

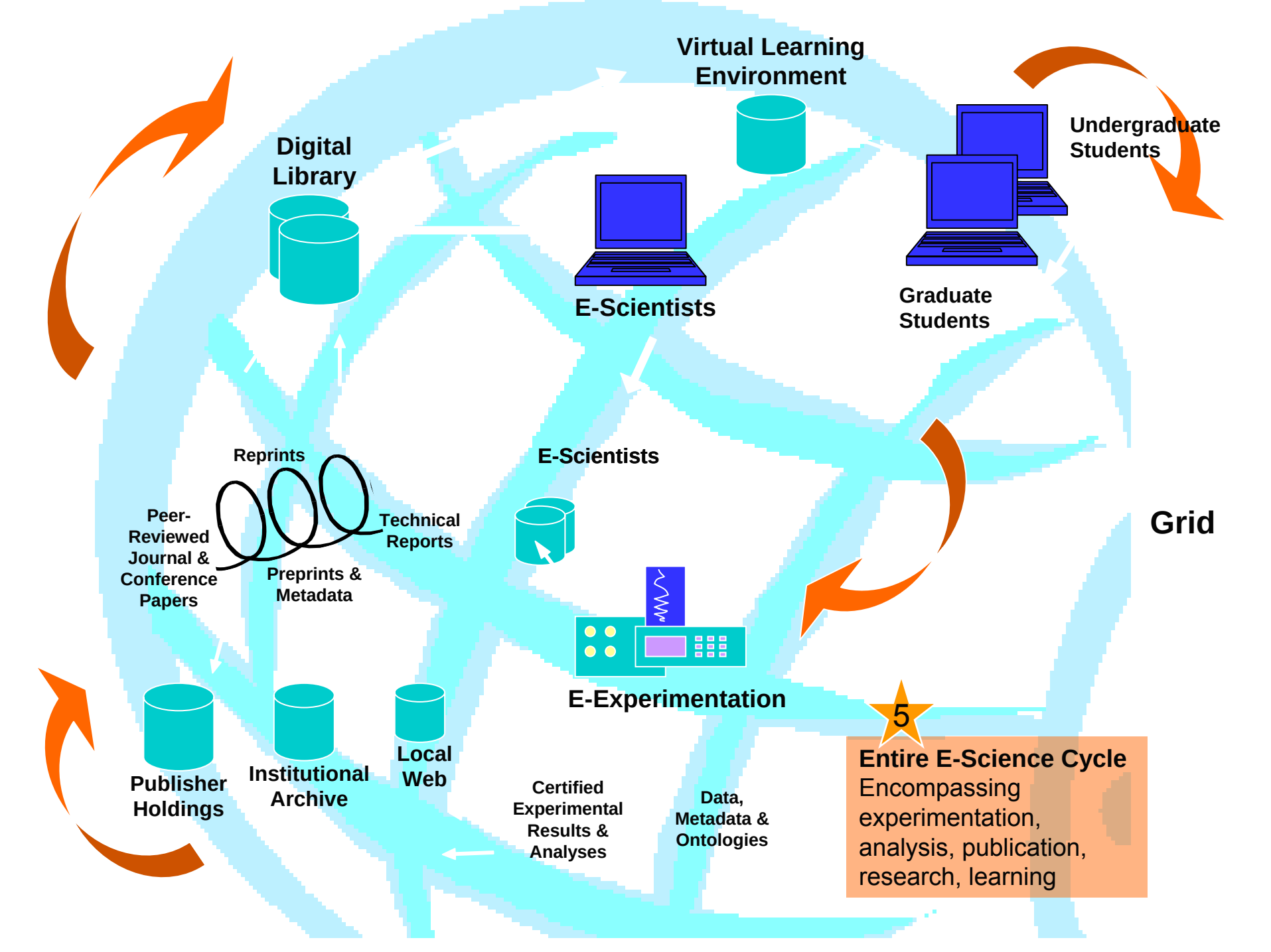
- Moores Law and Powers of 10
- The WWW: Blessing and Curse
- Making the Web Semantic
- Intelligence on the Web
- Research on the Semantic Web
- Futures Challenges

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

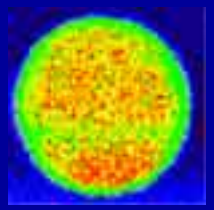





Virtual Learning Environment
Digital Library
Undergraduate Students
E-Scientists
Graduate Students
Reprints
E-Scientists
Peer-Reviewed Journal & Conference Papers
Technical Reports
Preprints & Metadata
Grid
E-Experimentation
Publisher Holdings
Institutional Archive
Local Web
Certified Experimental Results & Analyses
Data, Metadata & Ontologies
5
Entire E-Science Cycle
Encompassing experimentation, analysis, publication, research, learning

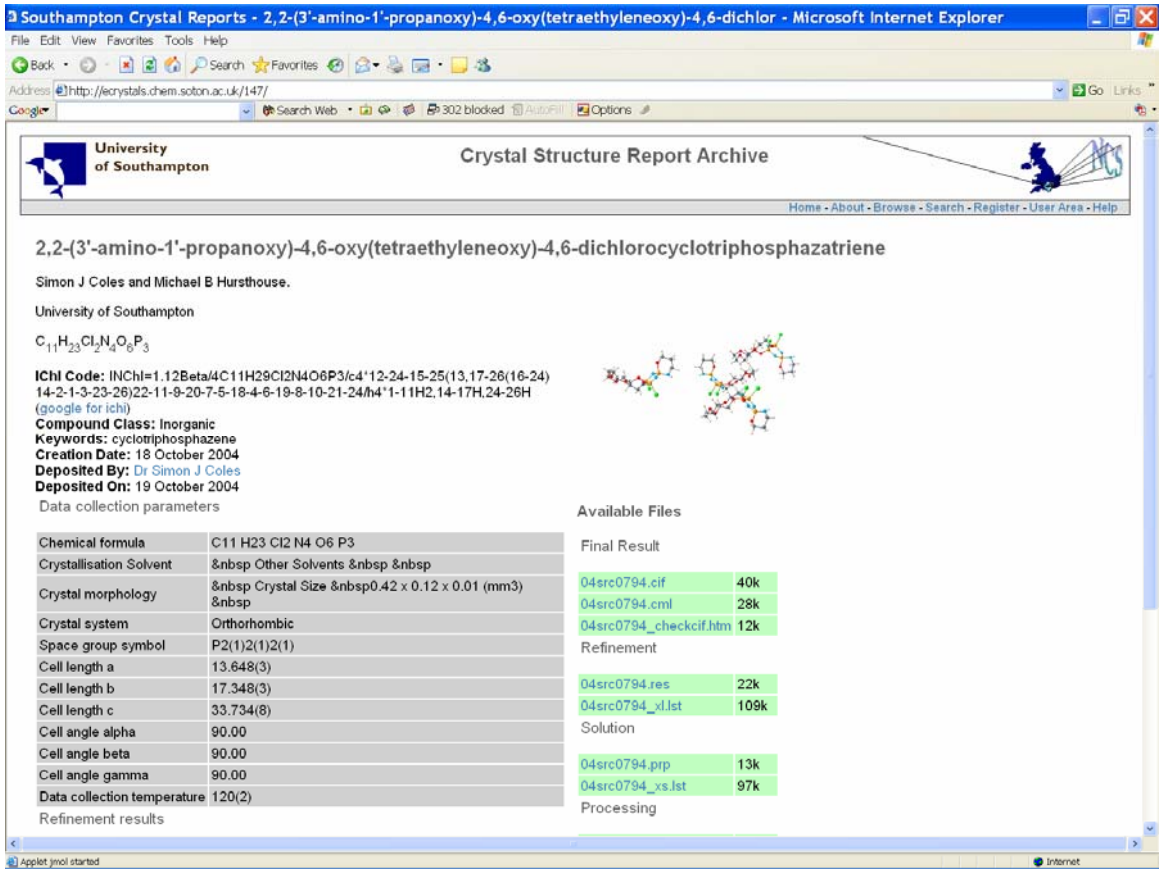
The need for *xtl*-Prints

Combechem

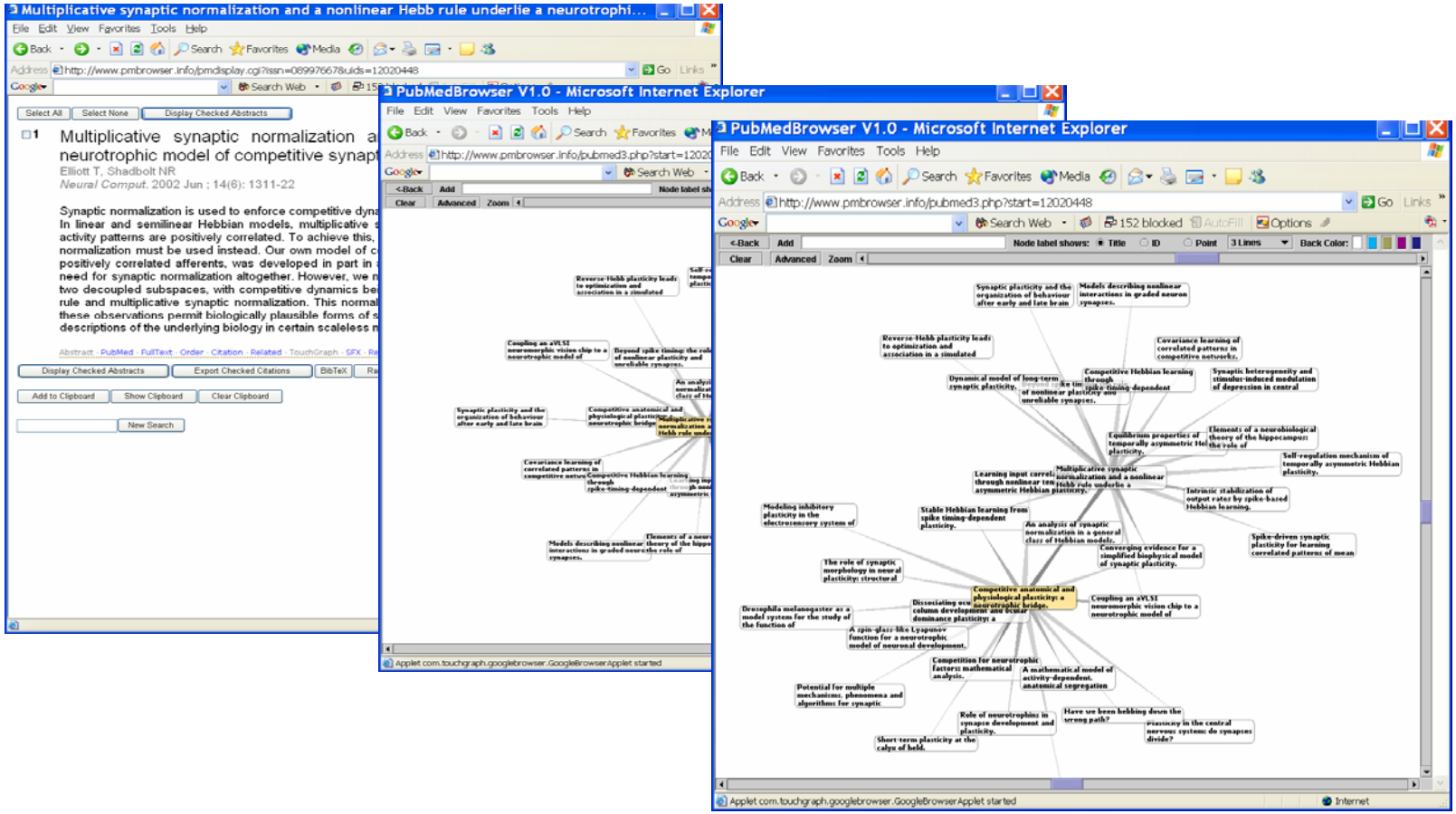


Combichem

Structural Eprints

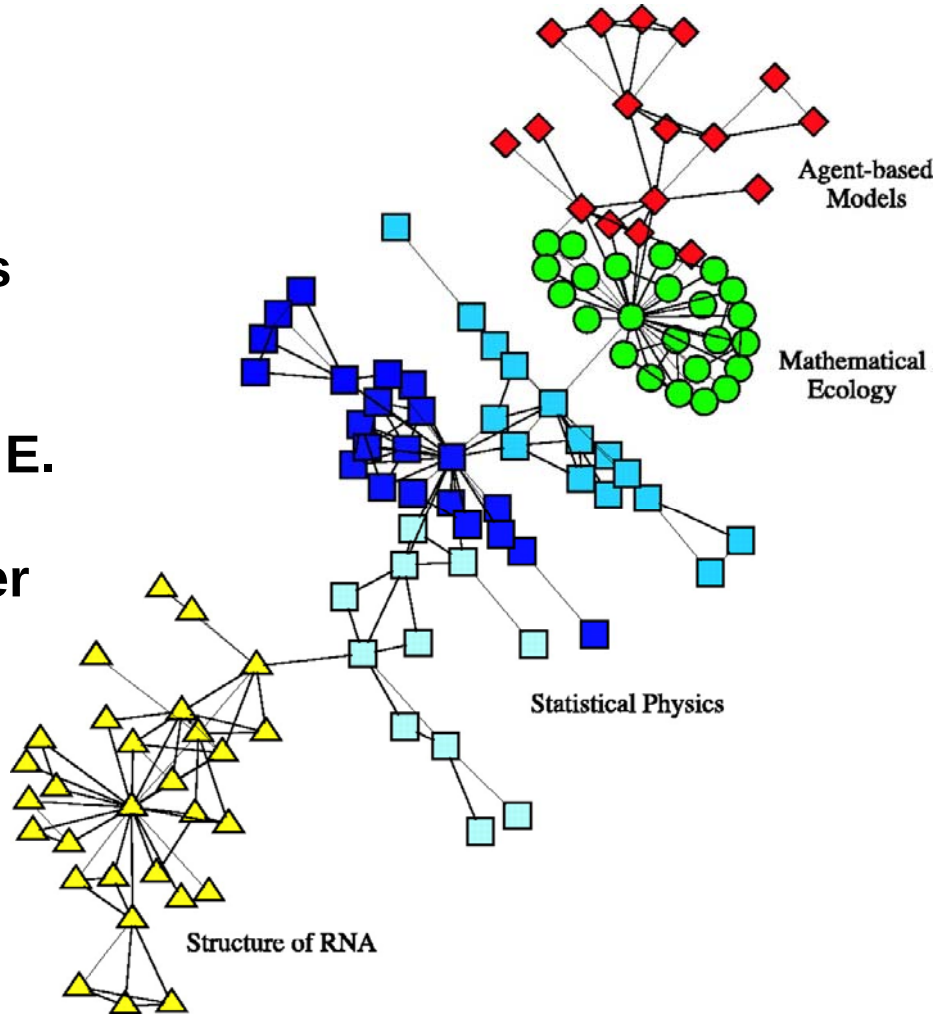


Increasing Use of Value Added Services



Communities of Authors

- **An example of a small coauthorship network depicting collaborations among scientists at a private research institution. Newman, M. E. J. (2004)**
- **Web services to run over archives at varying grainsize**



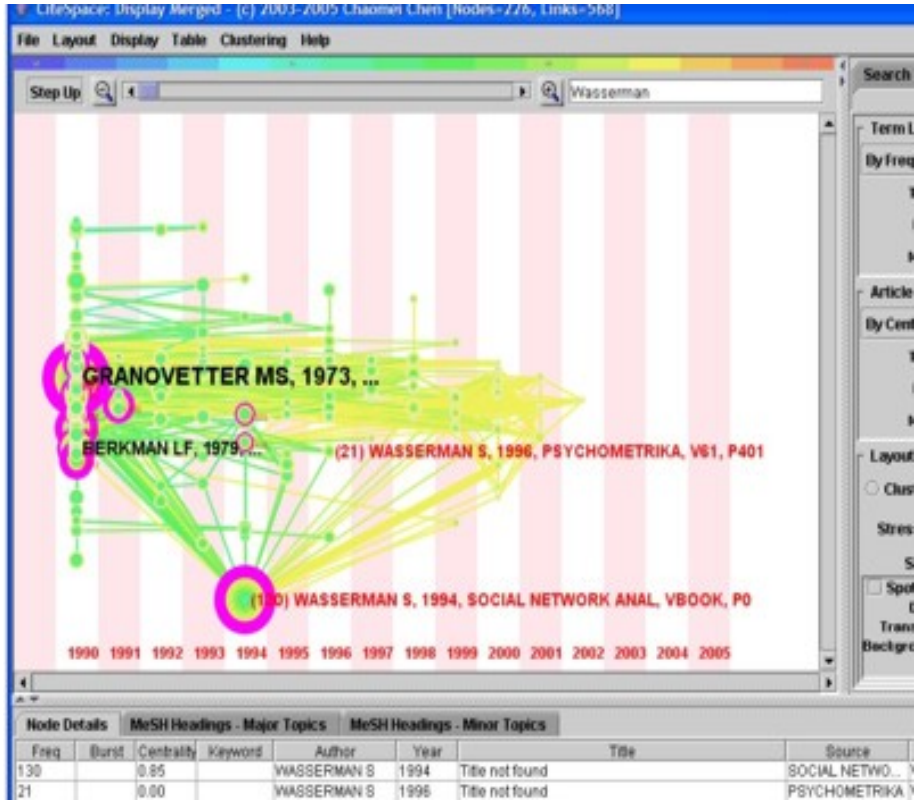
Hubs and Authorities

- Begin with existing measures: document count and citation count.
- Apply Kleinberg (1998) ‘hubs/authorities’ analysis to data.
- Note that higher citation count may not mean higher authority rating: quality citations are what count.

Author Name	Citeseer DocID	Citations	Authority Rating
Anant Agarwal	agarwal91mit	24	6.5161
Anant Agarwal	agarwal95mit	9	6.4675
Anant Agarwal	kranz93integrating	13	5.1927
Anant Agarwal	agarwal88evaluation	21	2.9159

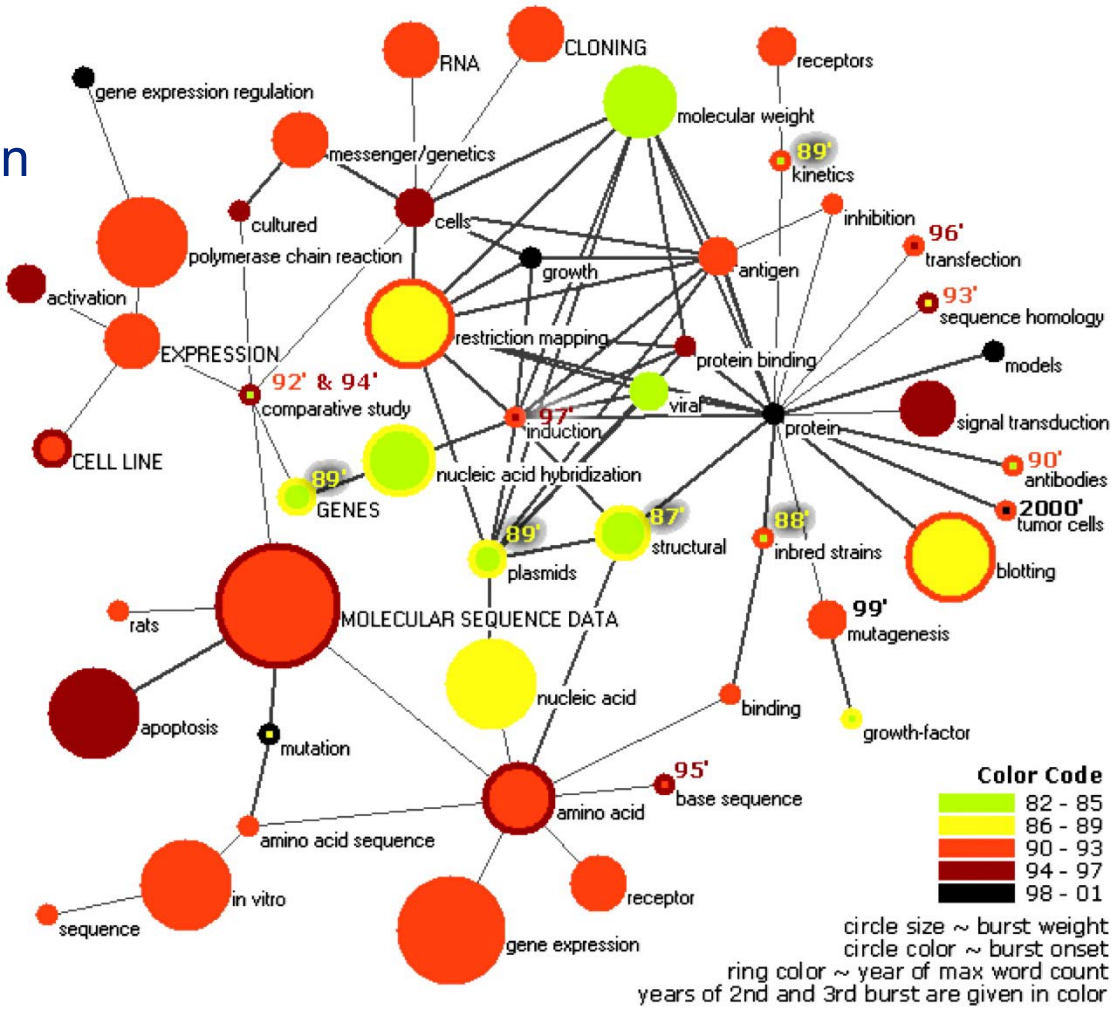
Turning Points and Centrality

- Allows us to find turning points in scientific development: Kuhn's "paradigm shift" moment.
- 'Centrality' measure to be applied to same Citeseer data.



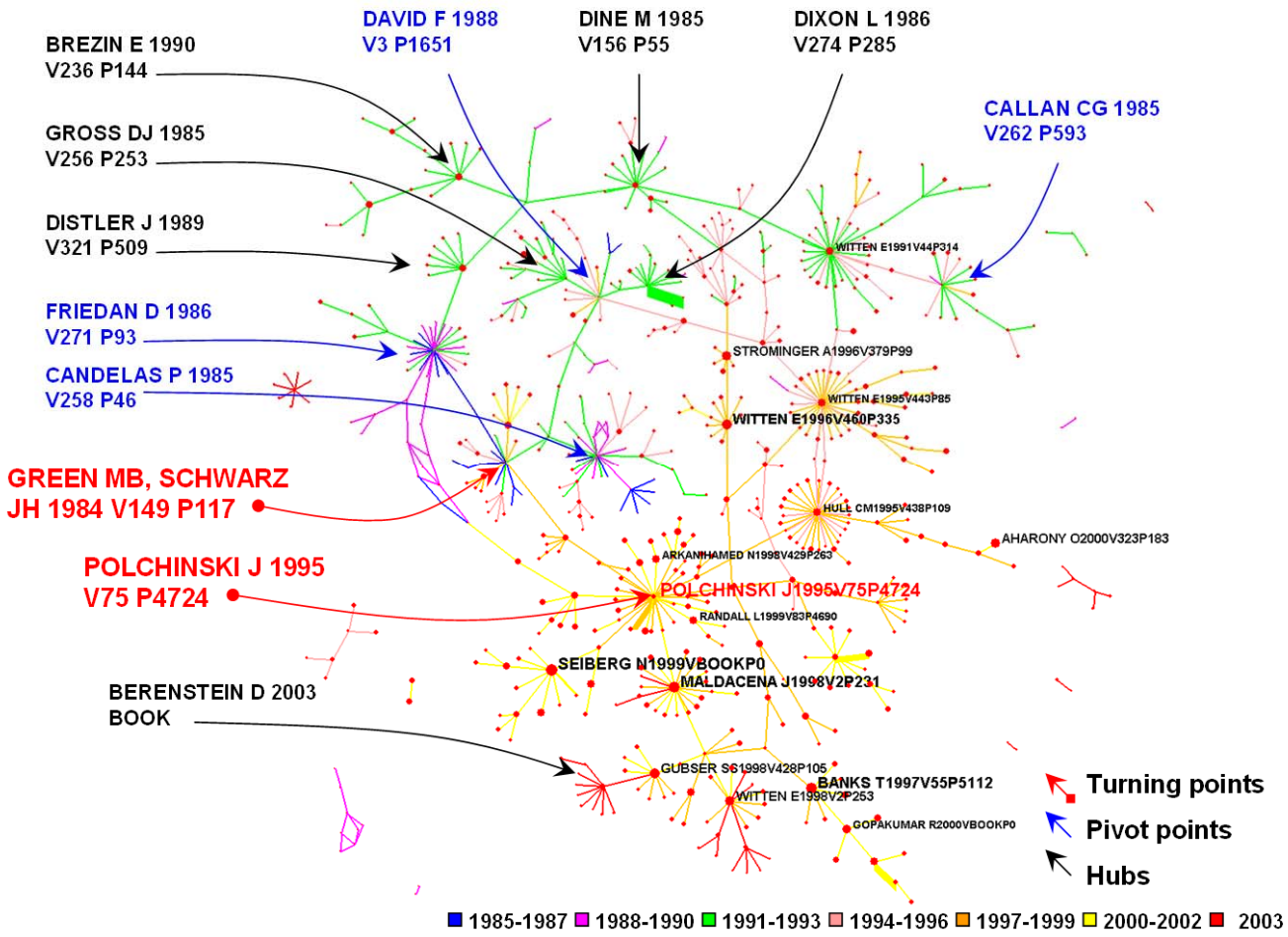
Bursting onto the scene: New Topics

- 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