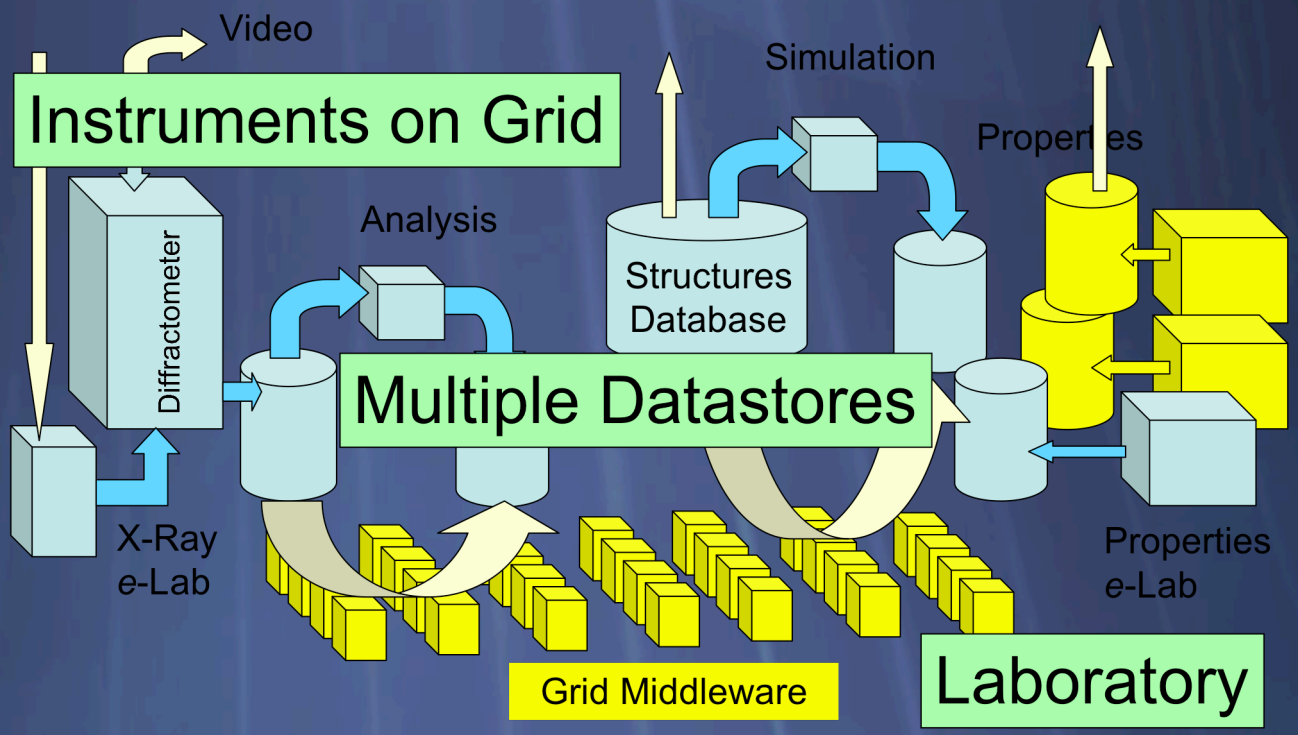
The CombeChem Project

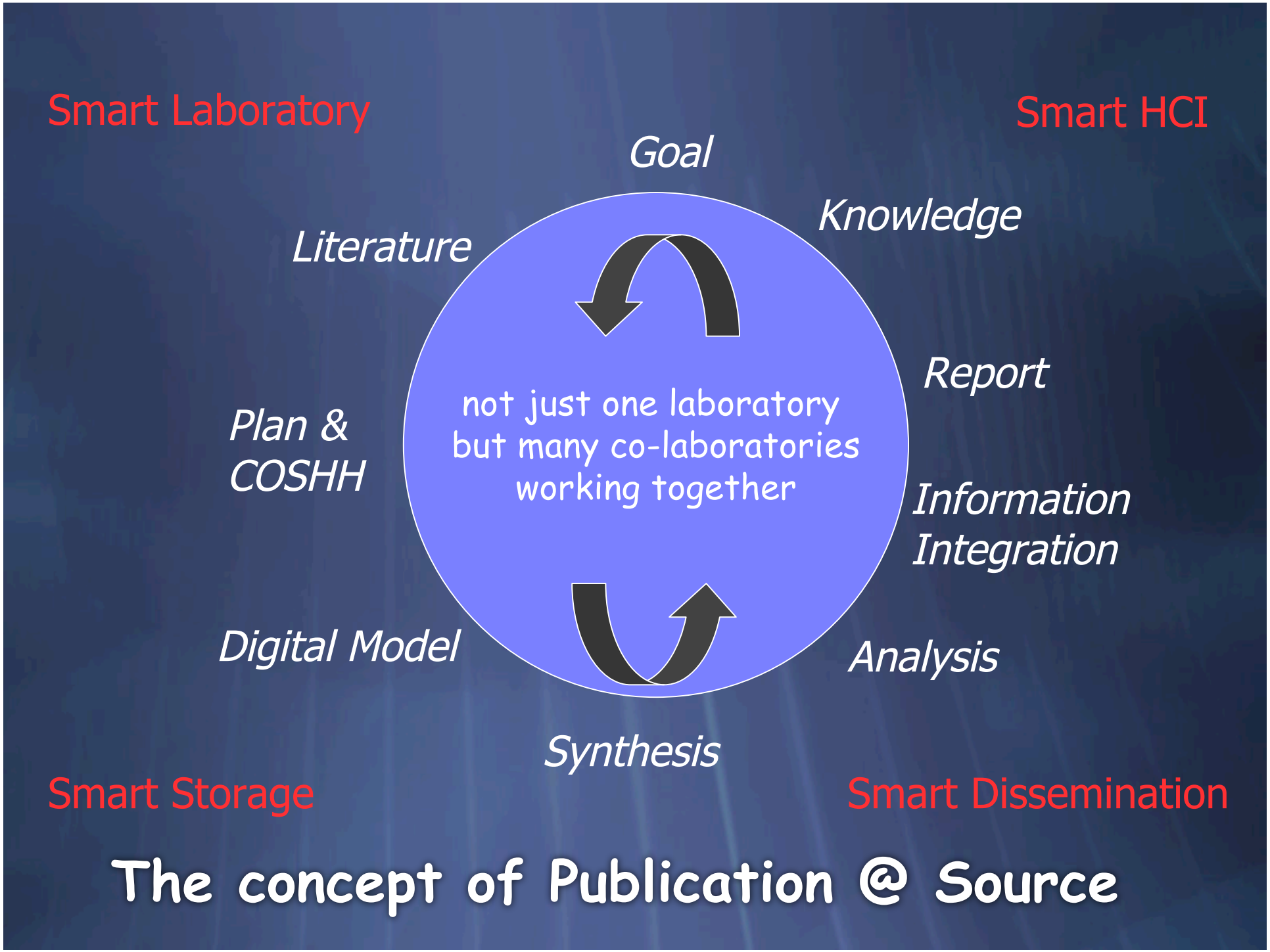
Semantic Support for the Chemistry Life cycle

Jeremy Frey Dave De Roure
Schools of Chemistry and
Electronics and Computer Science
University of Southampton

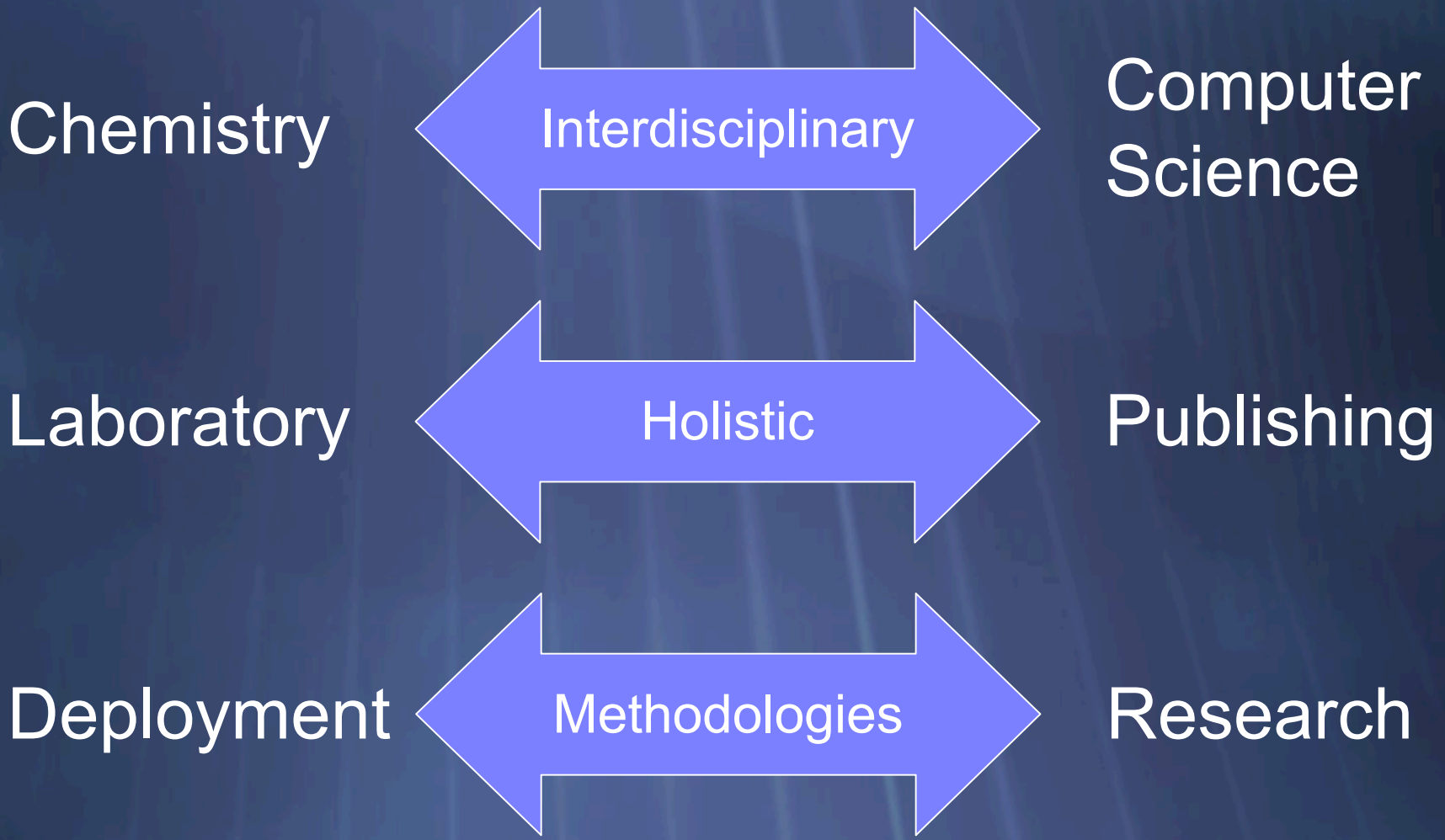
The CombeChem Project

- End to End linking of data and information: Laboratory to publication and back again
- The exponential world of combinatorial synthesis and high throughput analysis meets the exponentially growing power of computing



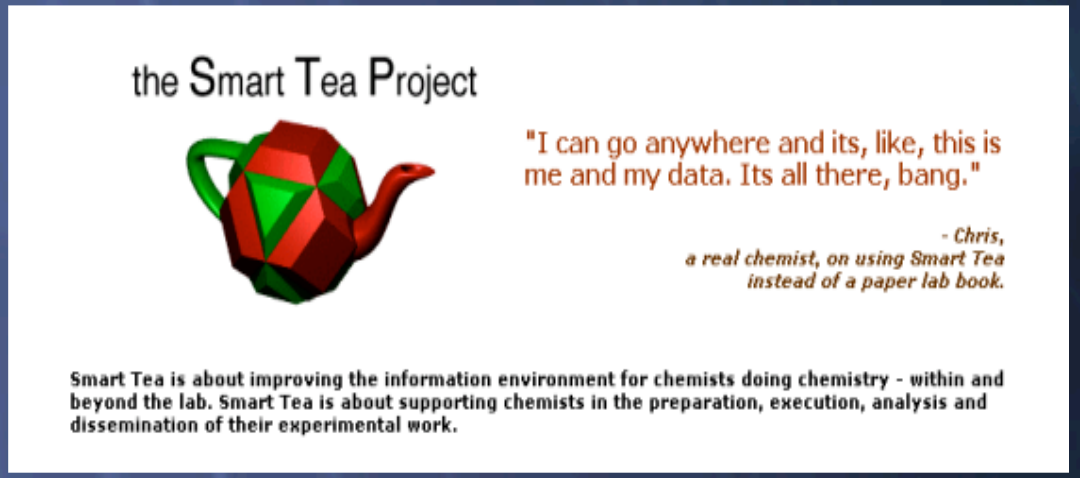

Smart Laboratory
Smart HCI
Goal
Knowledge
Literature
Report
Plan &
COSHH
not just one laboratory
but many co-laboratories
working together
Information
Integration
Digital Model
Analysis
Synthesis
Smart Storage
Smart Dissemination
The concept of Publication @ Source

The Stretches of CombeChem



CombeChem Smart Tea

- Ethnography
- Electronic Lab Notebook
- Capture in RDF (Resource Description Framework)

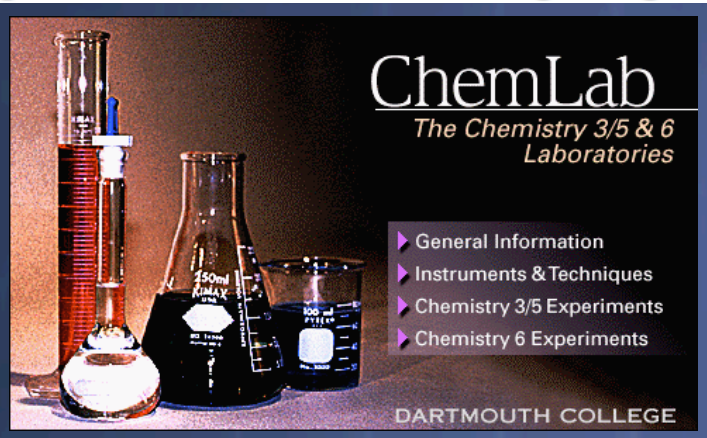


necessary if a calculation or discussion is changed; the section to be deleted is simply removed by drawing a neat "x" through it.

In view of the fact that a notebook is a primary record, data are not copied into it from other sources (such as this manual or a lab partner's notebook, in a joint experiment) without clear acknowledgment of the source. Observations are never collected on note pads, filter paper, or other temporary paper for later transfer into a notebook. If you are caught using the "scrap of paper" technique, your improperly recorded data may be confiscated by your TA or instructor at any time. It is important to develop a standard approach to using a notebook routinely as the primary receptacle of observations.

Each week at the beginning of lab lecture, you will turn in your prelab problems from the manual for grading. P[illegible] problems not turned in at the beginning of lab lecture [illegible] be

Observations are never collected on note pads, filter paper or other temporary paper for later transfer into a notebook

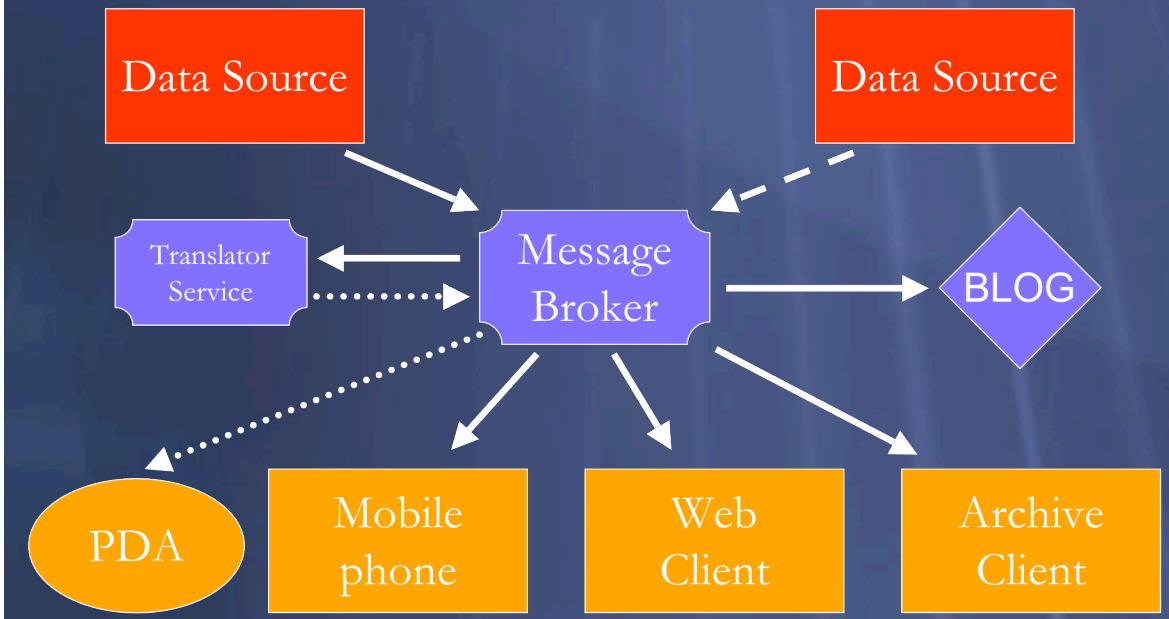


If you are caught using the "scrap of paper" technique, your improperly recorded data may be confiscated by your TA

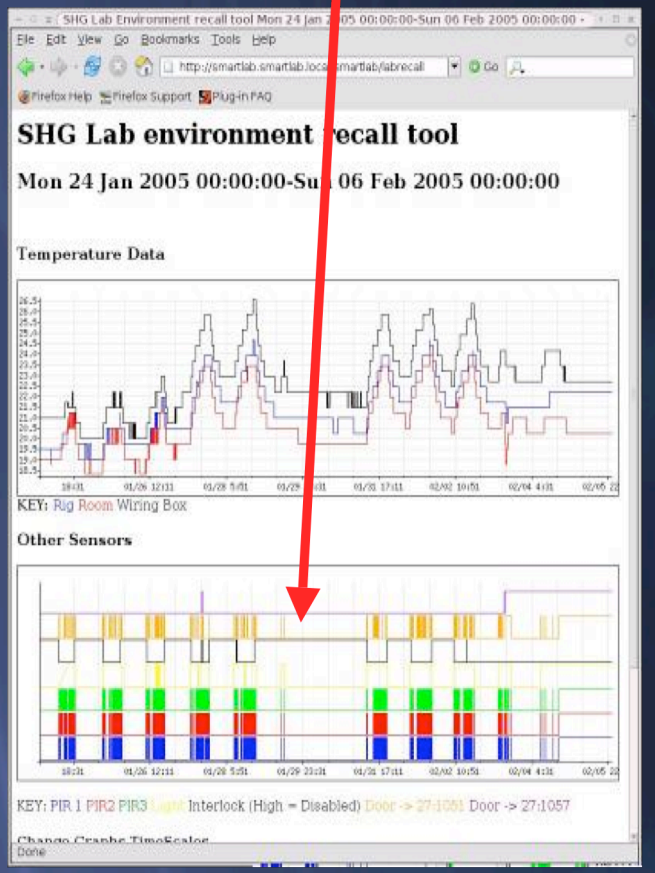
PLANS

- Make use of Plans to inform the digital context - metadata in advance
- Have concern for the "End-to-End life cycle" of chemistry information from the start
- Understanding Usability and Human Computer Interaction is vital for adoption

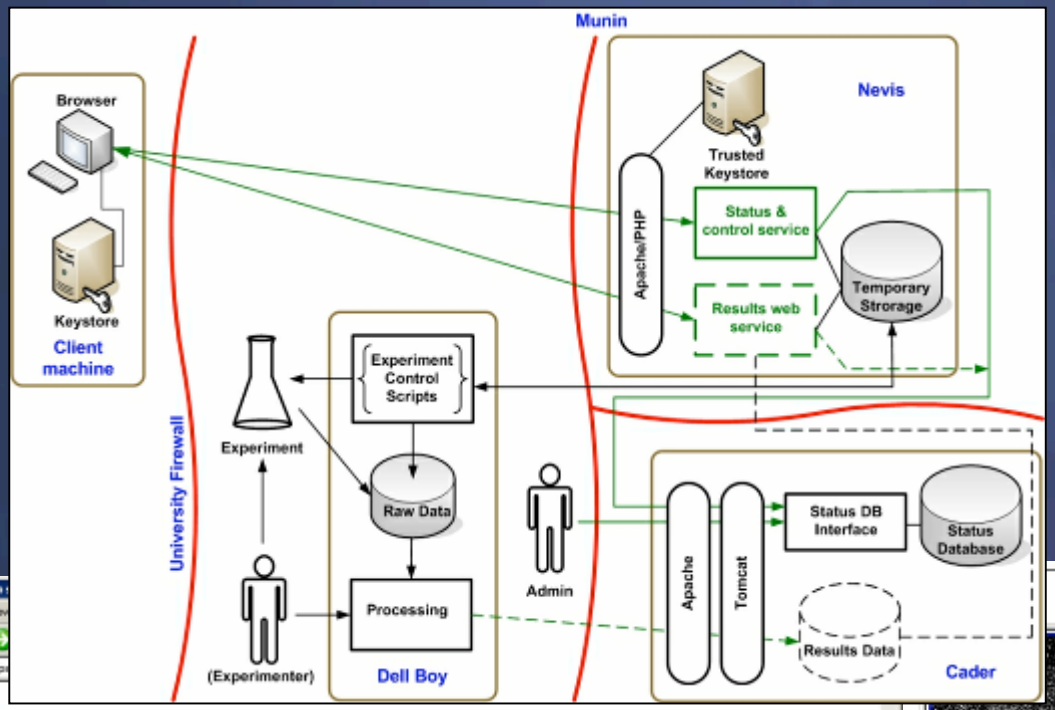
Pub-Sub systems provide the flexible & extensible approach to distribution



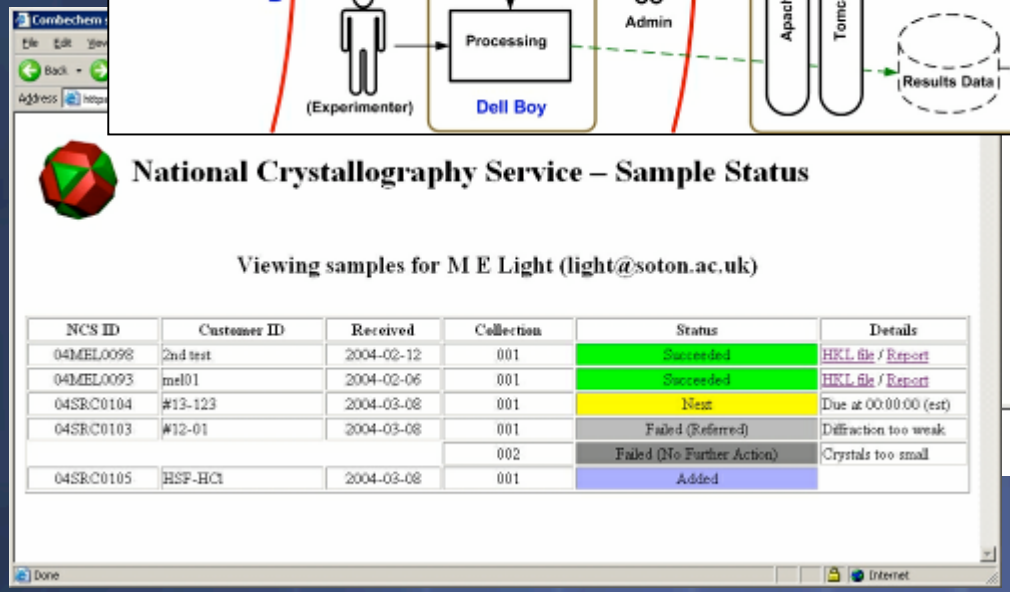
Air Conditioning failed



National Crystallography Service Grid Service Architecture



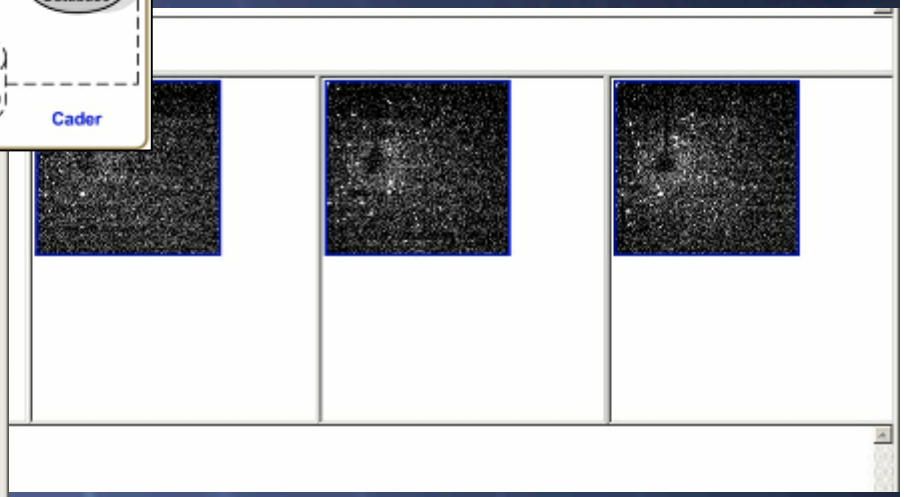
Users can follow and interact with experiment



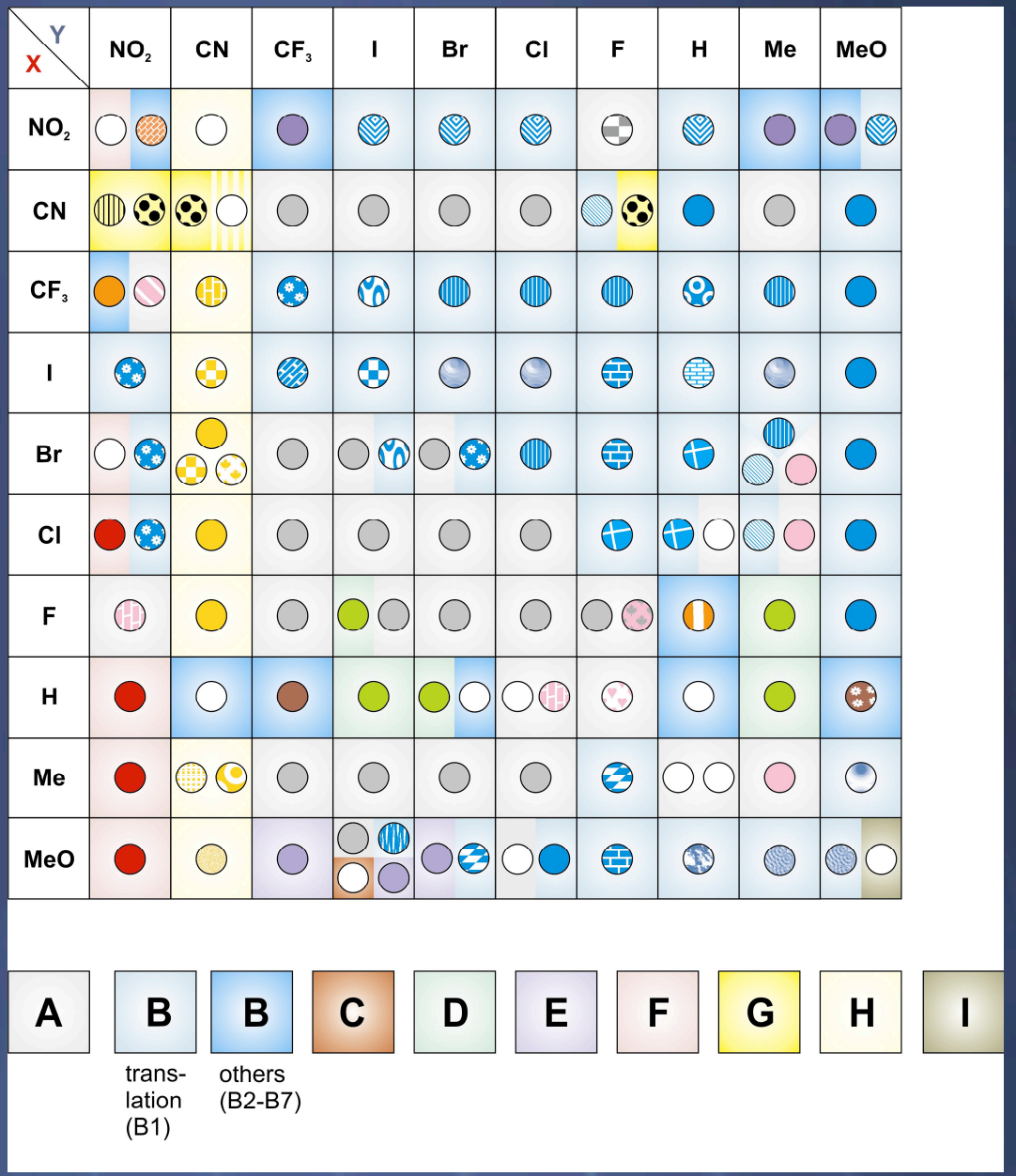
National Crystallography Service – Sample Status

Viewing samples for M E Light (light@soton.ac.uk)

NCS ID	Customer ID	Received	Collection	Status	Details
04MEL0098	2nd test	2004-02-12	001	Succeeded	HKL file / Report
04MEL0093	mel01	2004-02-06	001	Succeeded	HKL file / Report
04SRC0104	#13-123	2004-03-08	001	Next	Due at 00:00:00 (est)
04SRC0103	#12-01	2004-03-08	001	Failed (Referred)	Diffraction too weak
			002	Failed (No Further Action)	Crystals too small
04SRC0105	HSF-HCl	2004-03-08	001	Added	

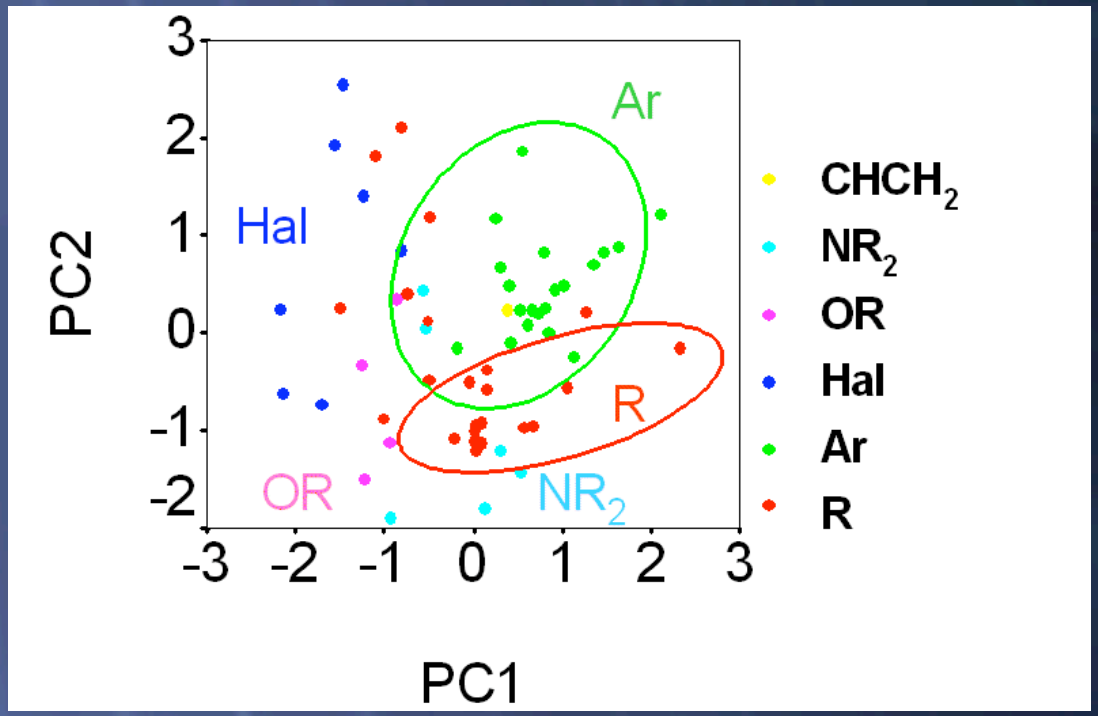


Chemical families polymorph similarities



Statistics

- Collect information about ligands and their (transition metal) complexes
- Calculate descriptors with standard computational approach (DFT)
- Robustness (computational, chemical, statistical)
- Overlap with available experimental data



Map of Ligand Space for Monodentate Phosphorus(III) Ligands

Ligand Knowledge Base

Dave's Chemistry Experiment

1. Take a building full of chemists
2. Add RDF tools
3. Stir occasionally
4. See what's been made

- A very big chunk of Semantic Web
- An ontology for units

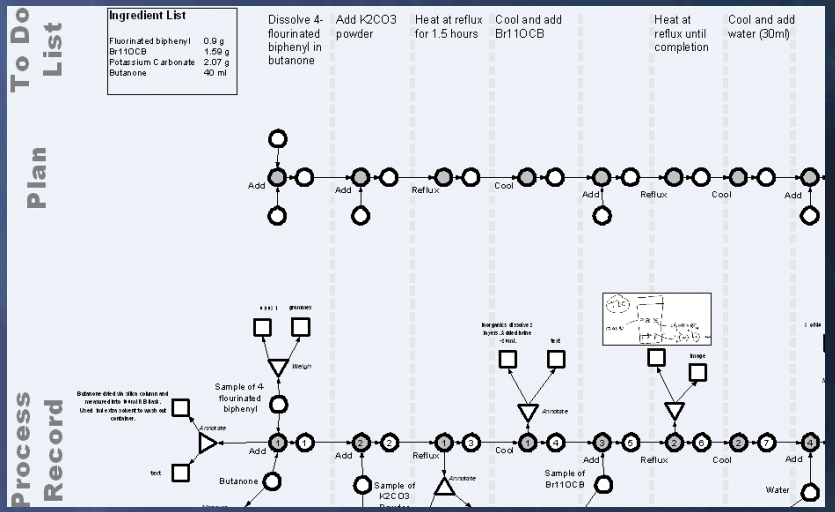
Statistics on Green Triplestore

Thu Apr 21 11:37:17 2005

models	2573454
triples	84188993
inferred (FC)	24269915
ground facts	59919078
resources	9987377
literals	7974229
classes	88
properties	49

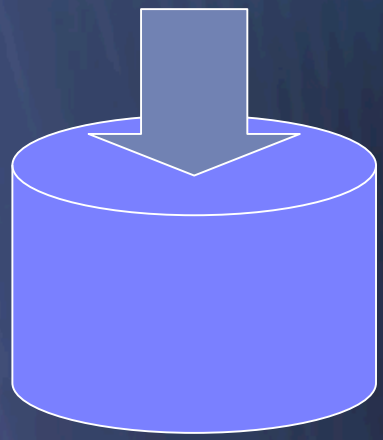
Semantic DataGrid

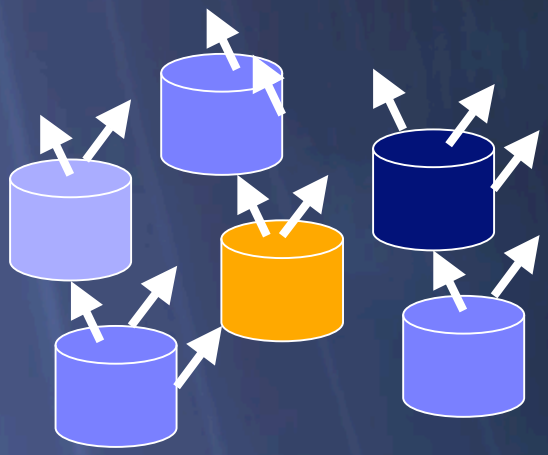
- CombeChem uses Semantic Web for
  - Enhanced (annotated) DataGrid over multiple diverse stores
  - Some Data Storage
  - Storage of Provenance Information
  - Annotated multimedia streams



Triplestores

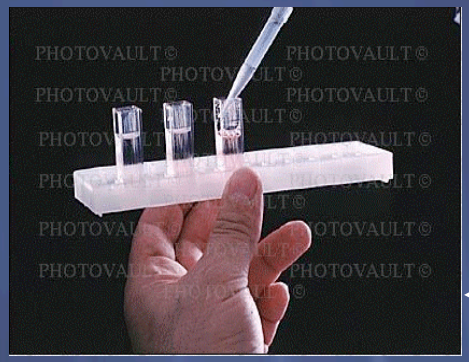
- Started with the data hoarding approach of CSAKTive Space, using 3store from the AKT IRC
- Scalability, lifecycle and the CombeChem sharing and publishing ethos led to the use of multiple triplestores to cache and query rather than store



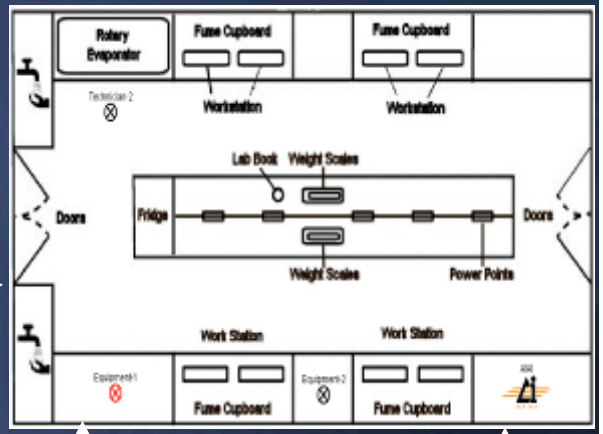


Comb-e-Chem: Facility e-Science in Action

Presence Awareness + Remote Participation



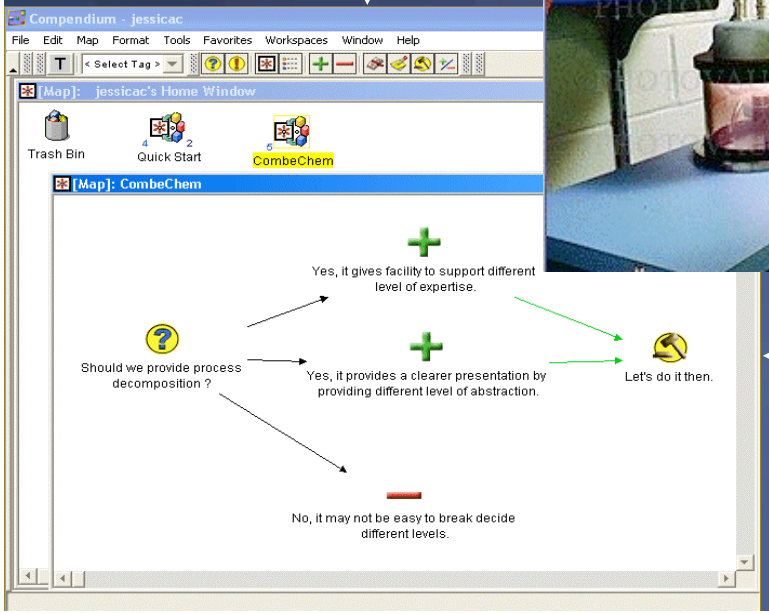
Resource + Floor Management

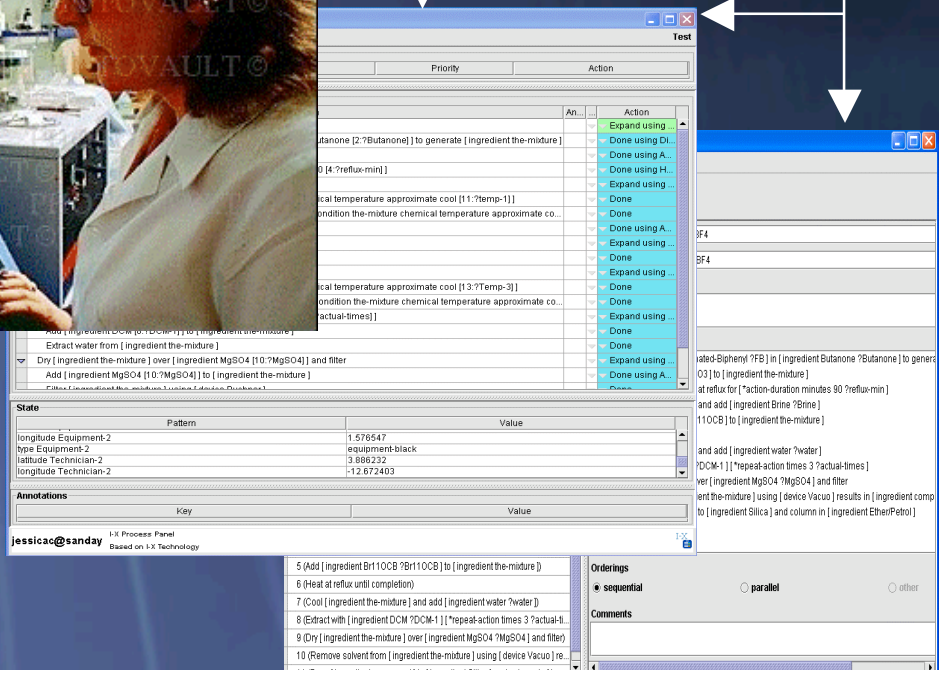


Run-time tracking and control

Workflow Execution And Monitoring

Instrument Tracking And utilisation

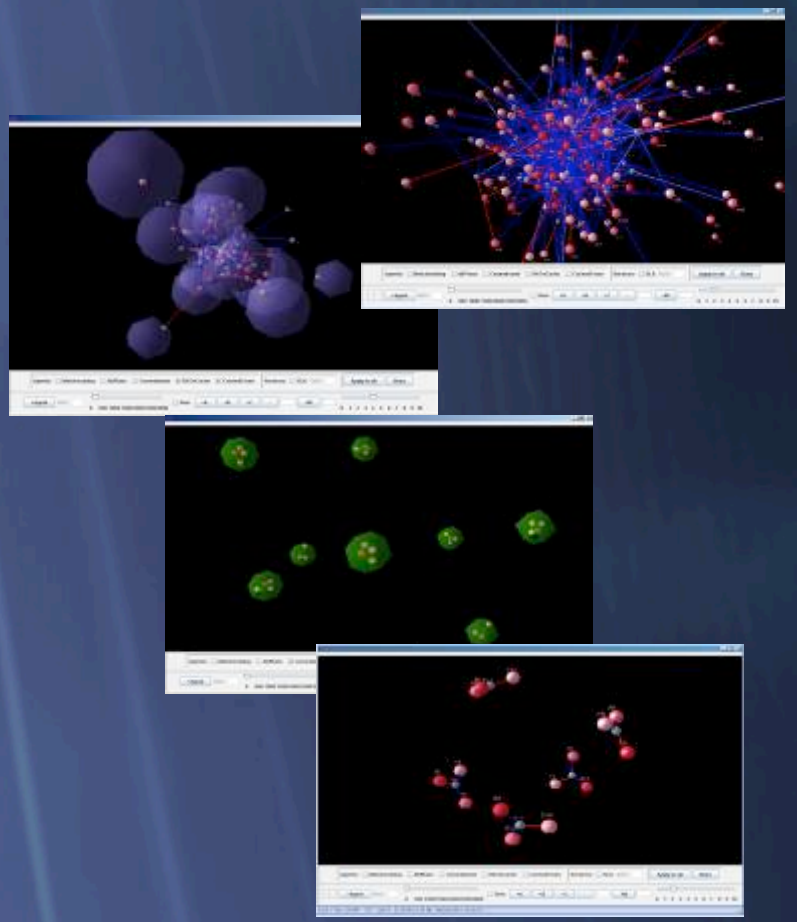


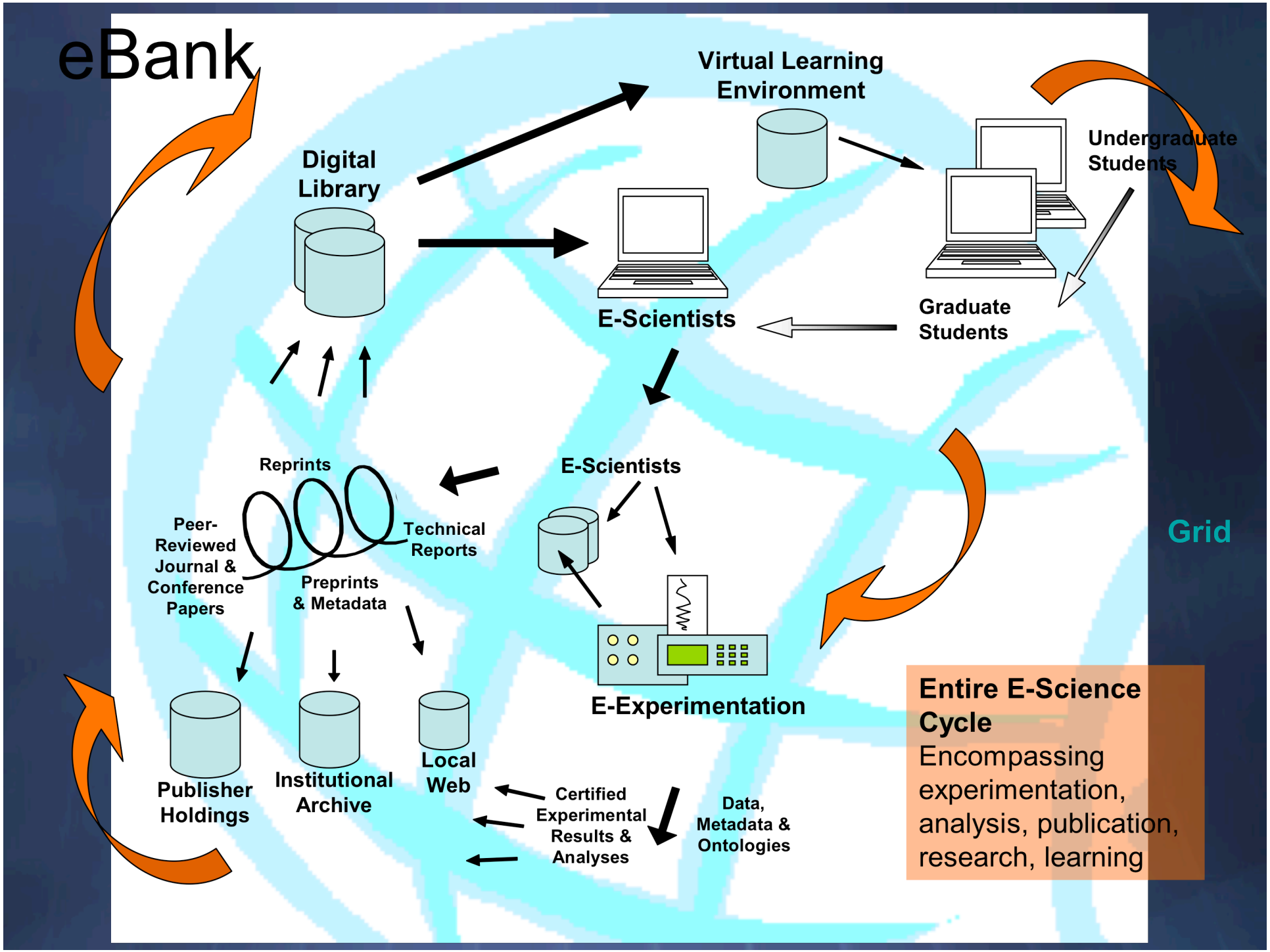


Action Realisation + Rationale Feedback

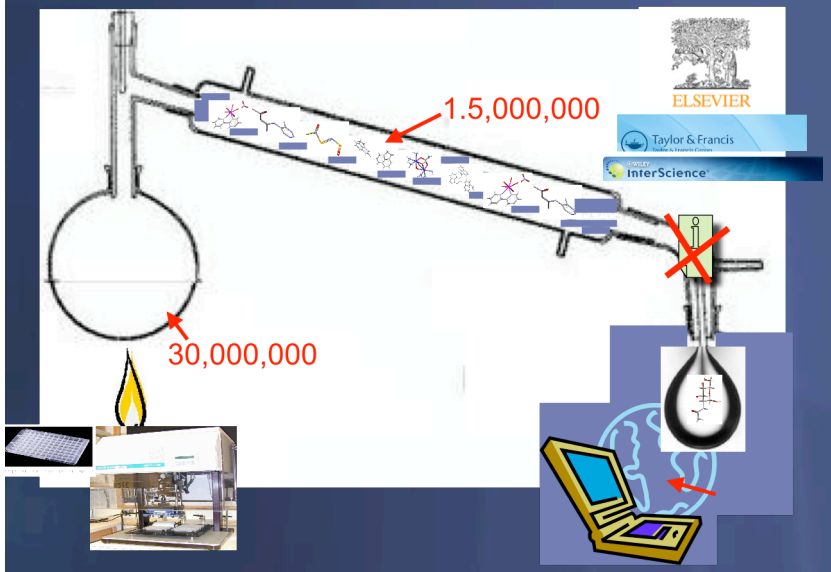
Autonomic e-Science

- Built simulator of a future combechem in which 1000s of services are negotiating and self-organising
- Informed by combechem experience
- Article in IEEE Intelligent Systems on the Self-Organising Semantic Grid
- Raises questions about the future role of the scientist

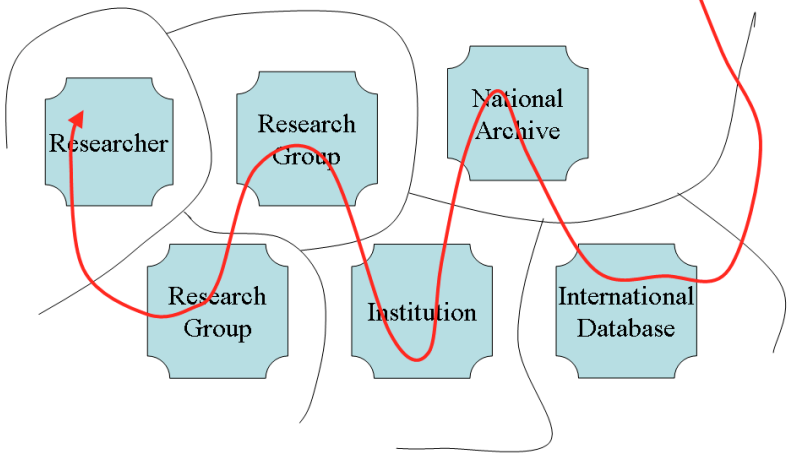


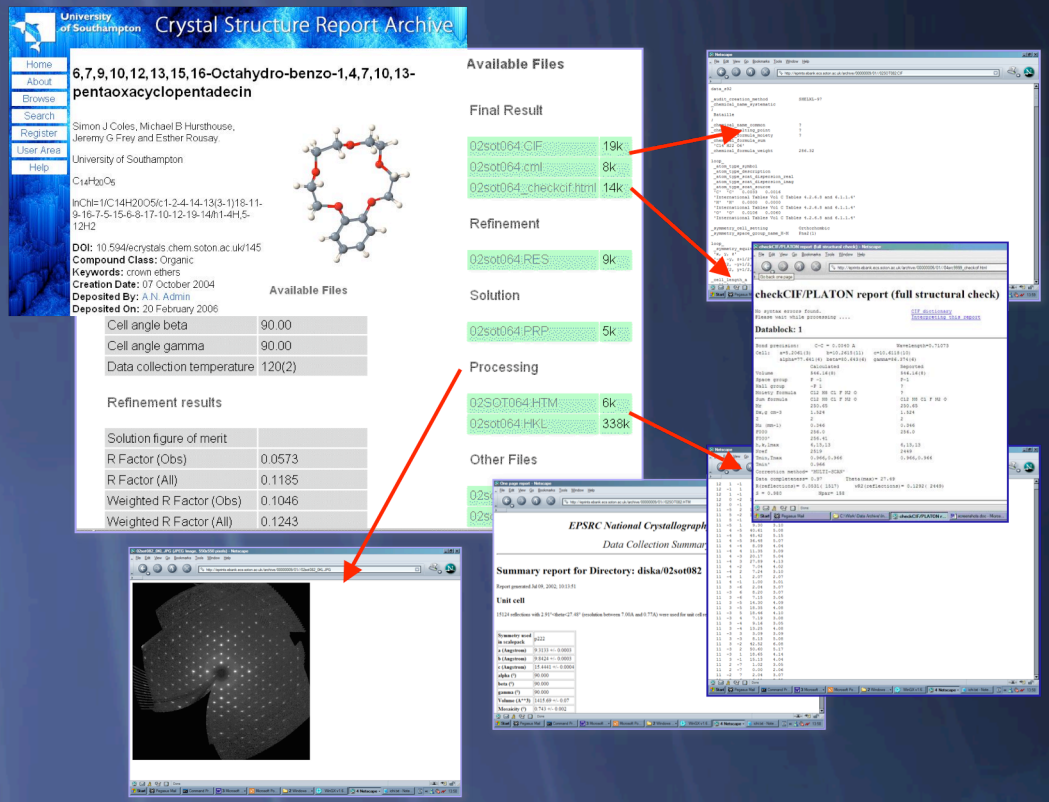

eBank
Virtual Learning Environment
Undergraduate Students
Digital Library
E-Scientists
Graduate Students
Reprints
E-Scientists
Peer-Reviewed Journal & Conference Papers
Technical Reports
Preprints & Metadata
Grid
E-Experimentation
Publisher Holdings
Institutional Archive
Local Web
Certified Experimental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

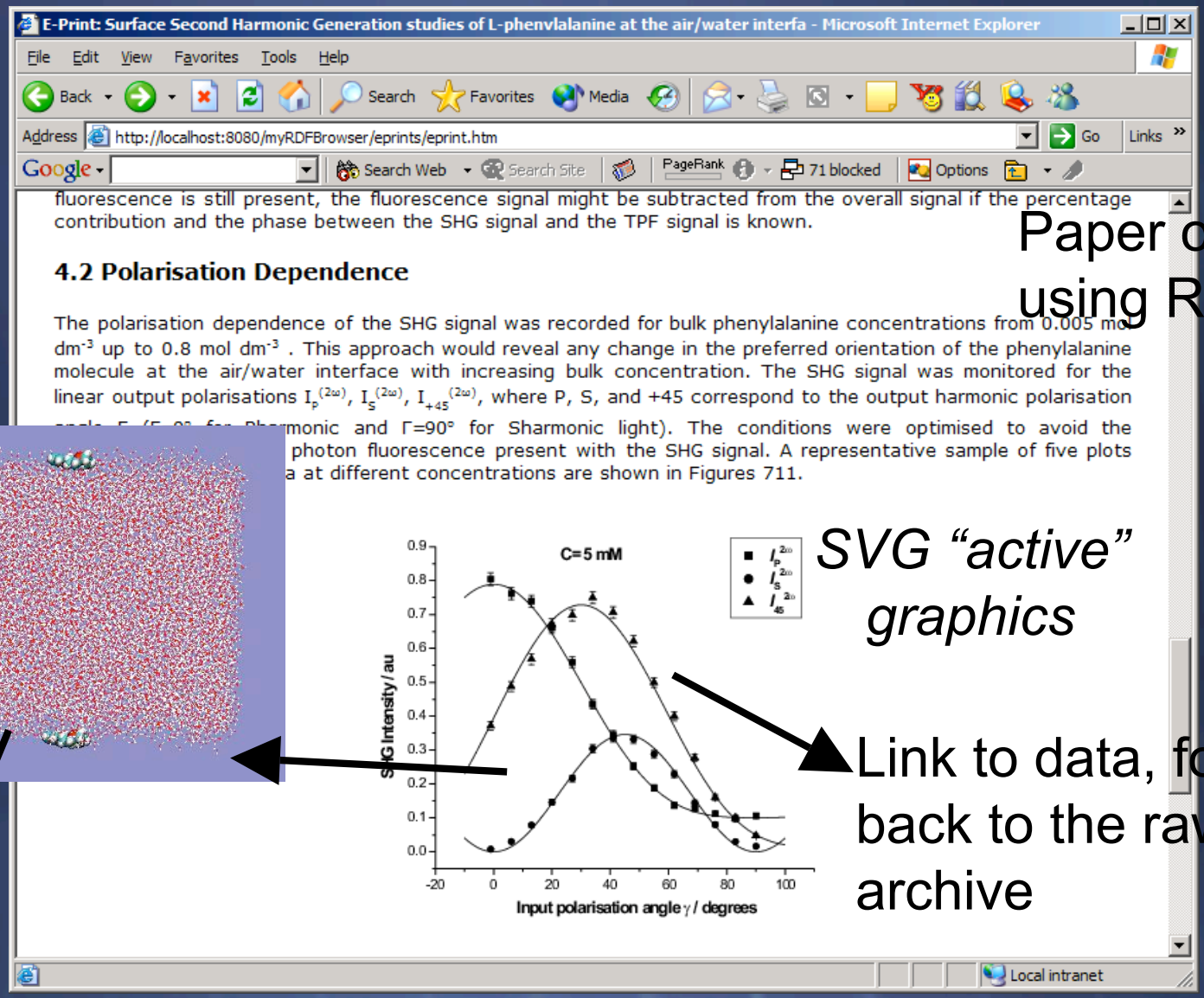
Access to the underlying data



Several groups making and analysing the library Administrative Domains transfer or share the data







Paper organized using RDF

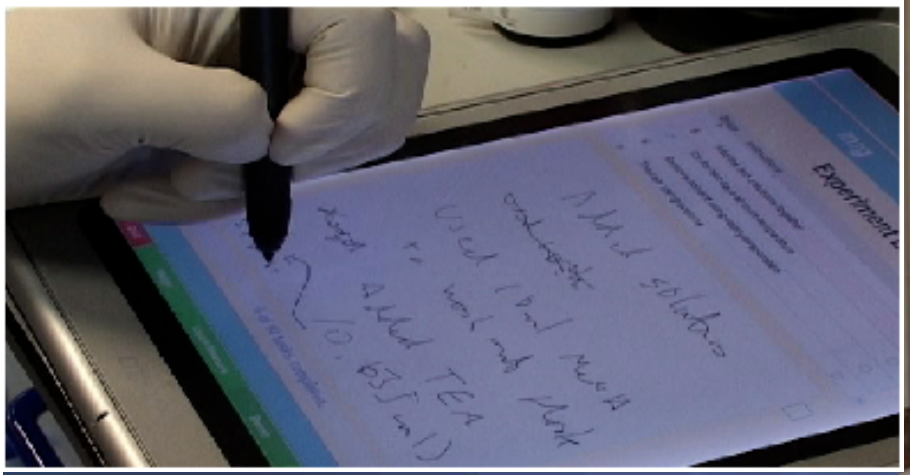
SVG "active" graphics

Link to data, follow links back to the raw data archive

Link to simulation, full simulation data archived in BioSimGrid

R4L

Data capture



R4L Repository for the Laboratory

the Smart Tea Project

"I can go anywhere and its, like, this is me and my data. Its all there, bang."

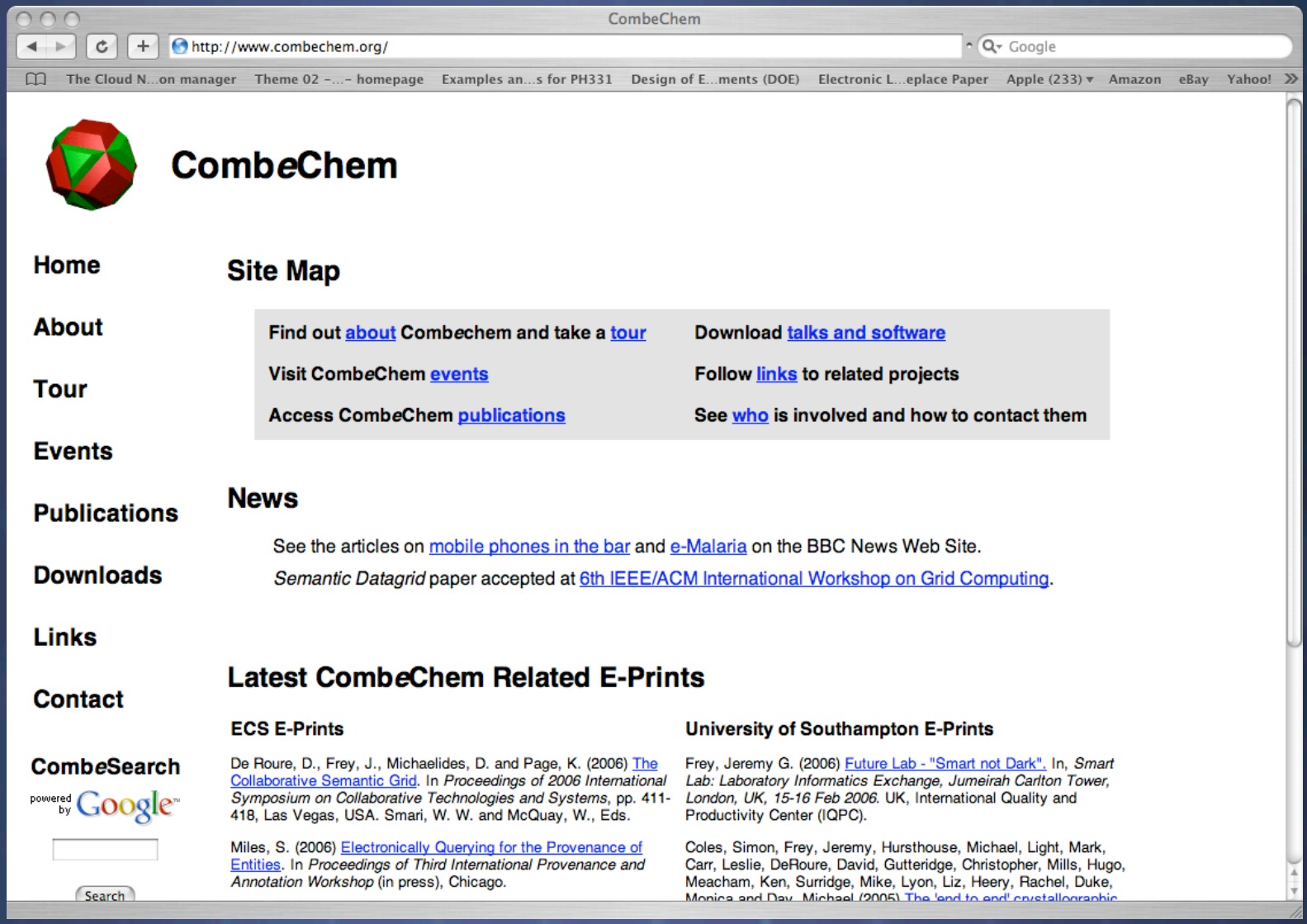
- Chris,
a real chemist, on using Smart Tea instead of a paper lab book.

Smart Tea is about improving the information environment for chemists doing chemistry - within and beyond the lab. Smart Tea is about supporting chemists in the preparation, execution, analysis and dissemination of their experimental work.

Take Homes

- Making sure other people can find, understand and re-use your data easily and with confidence (even when there is a huge amount of it!)
- Whole lifecycle approach from lab to publication
- Significant rollout of next generation Web technologies – Semantic DataGrid
- Distinctive in e-Science for focusing on laboratory, usability and collaboration
- Agent of culture shift in publishing and open access to data
- Outreach including schools
- Platform and agenda for future research in the Pervasive Semantic Grid

www.combechem.org



Questions

Grid Innovation

- CombeChem has focused on accelerating science by accelerating the process and not necessarily the computation
  - Uses existing cluster and grid techniques
  - Early focus on security for National Crystallographic Service
  - Adopted Web Services from the outset
  - Uses asynchronous message passing for integration
- Semantic DataGrid

Middleware Outputs

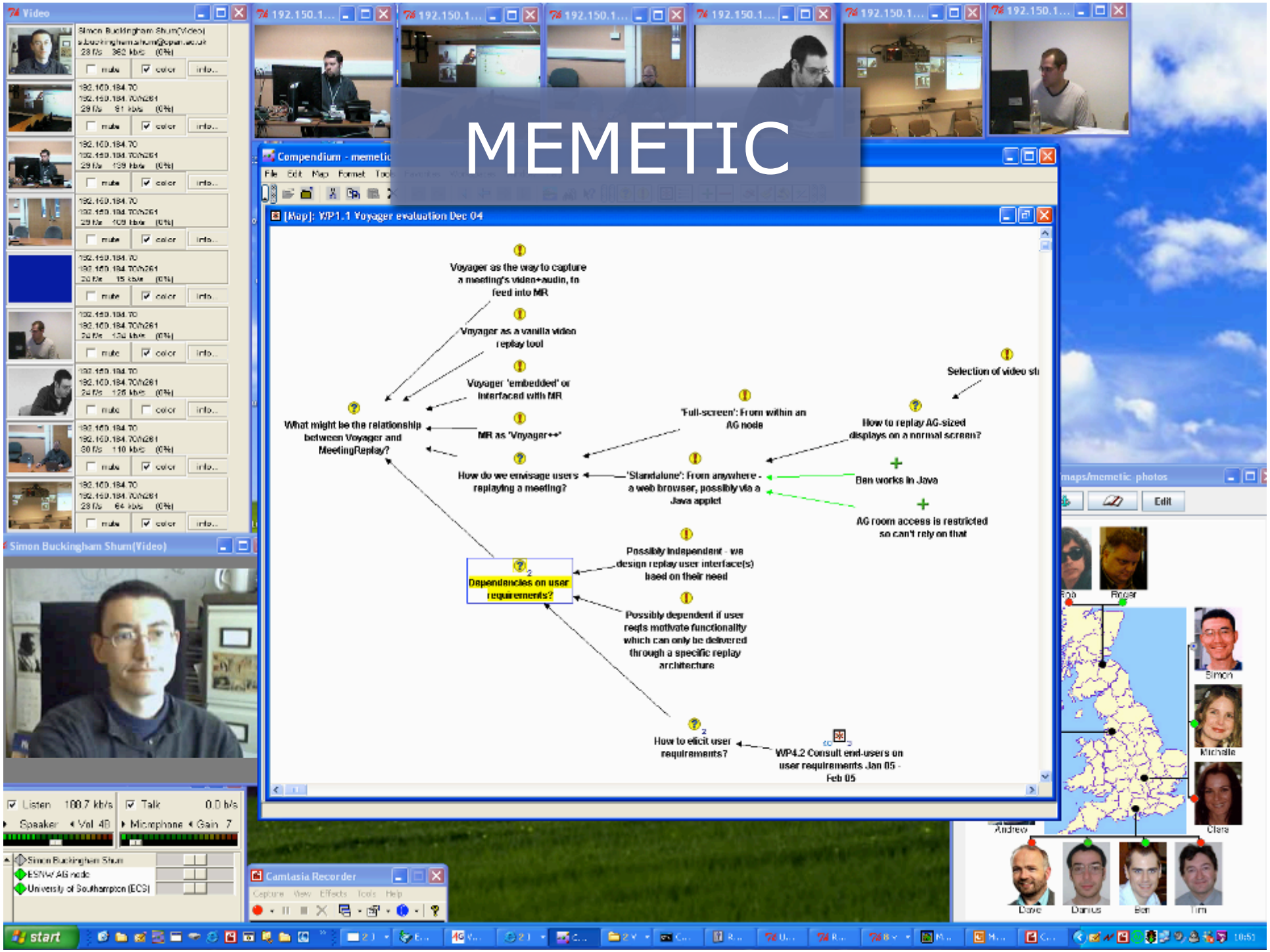
- Security and access control developed for NCS
- Software written by IT Innovation for CombeChem fed into the software distribution for the EU Grid for Industrial Applications (GRIA) Project
- It forked!
  - GRIA now on release 5, good adoption by industrials in EU projects (e.g. SIMDAT)
  - Solutions evolved with current Grid standards

Other Outputs

- Security and Access Control in GRIA 5
- Statistics software:
  - Design search algorithms for Generalized Linear Models
  - Design of experiments eLearning module
  - Elicitation in Chemistry Investigations (EliCIT)
- RDF streaming tools
- Units Ontology

Staffing

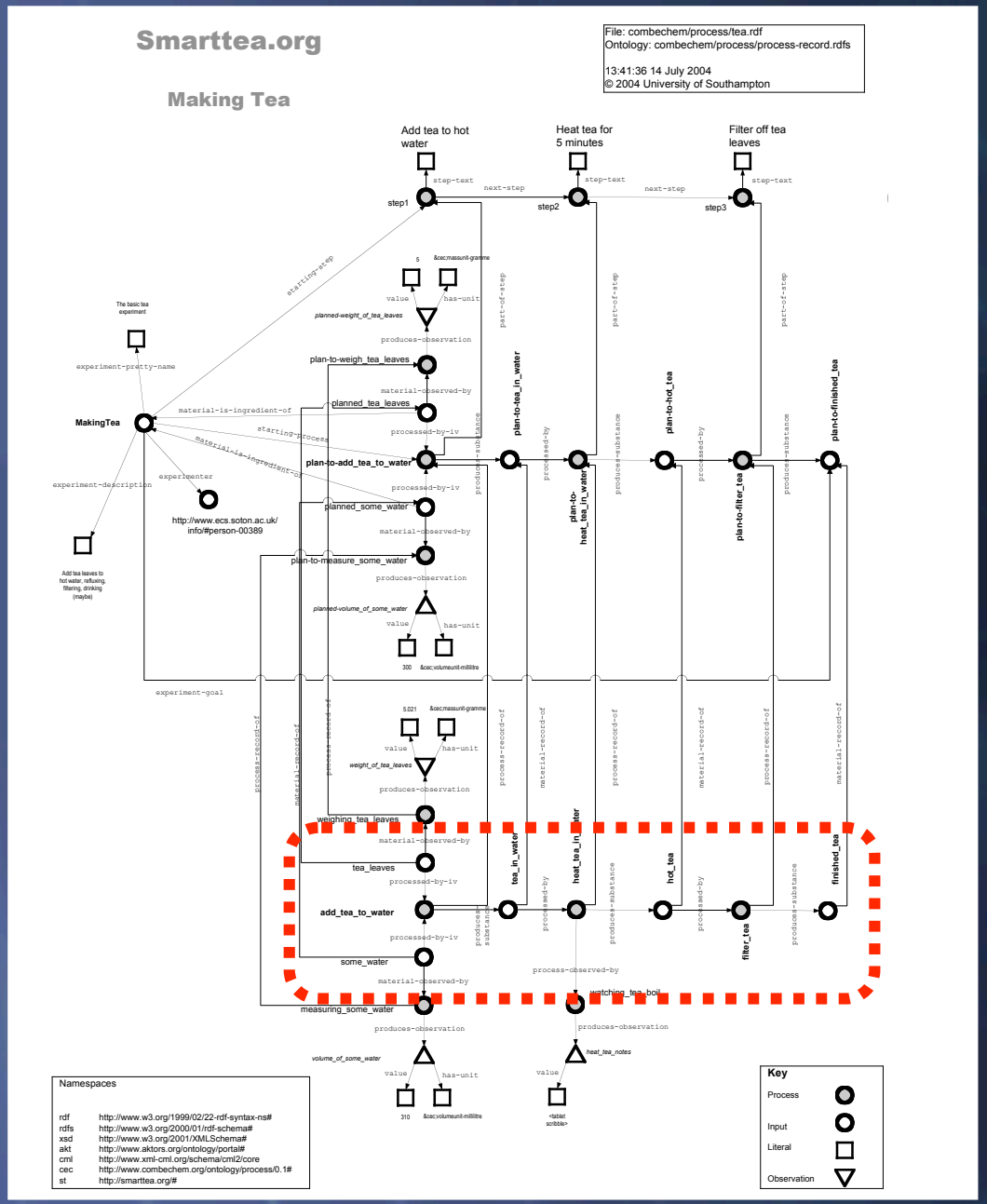
- Deploy-then-research strategy
- Core team persisted through most of project and developed interdisciplinary knowledge
- Brought in additional staff for specific tasks


MEMETIC
[Map]: WP1.1 Voyager evaluation Dec 04
Voyager as the way to capture a meeting's video+audio, to feed into MR
Vmyager as a vanilla video replay tool
Voyager 'embedded' or interfaced with MR
MR as 'Vmyager++'
What might be the relationship between Voyager and MeetingReplay?
How do we envisage users replaying a meeting?
'Full-screen': From within an AG node
'Standalone': From anywhere - a web browser, possibly via a Java applet
How to replay AG-sized displays on a normal screen?
Selection of video st
Ben works in Java
AG room access is restricted so can't rely on that
Possibly independent - we design replay user interface(s) based on their need
Dependencies on user requirements?
Possibly dependent if user reqts motivate functionality which can only be delivered through a specific replay architecture
How to elicit user requirements?
WP4.2 Consult end-users on user requirements Jan 05 - Feb 05
Simon Buckingham Shum(Video)
Camtasia Recorder
maps/memetic photos
Edit
Simon
Michelle
Clara
Andrew
Dave
Darius
Ben
Tim
start

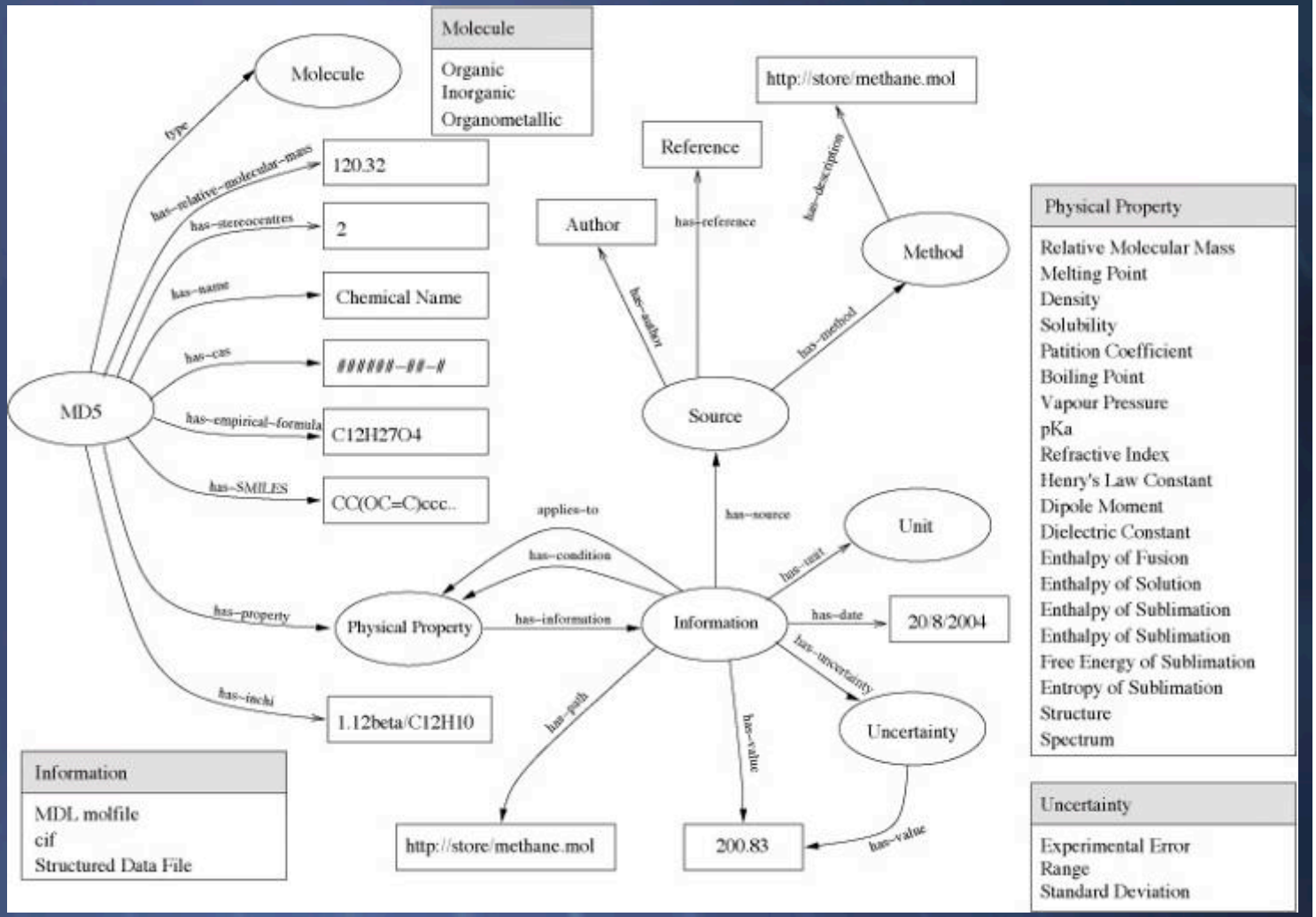
`getRecord()`

There is a potential containment problem in pulling back partial RDF graphs from the triple store.

Solved by using multiple triple stores but boundaries are a major issue for the future.



RDF/RDFS High level Schema for chemical properties

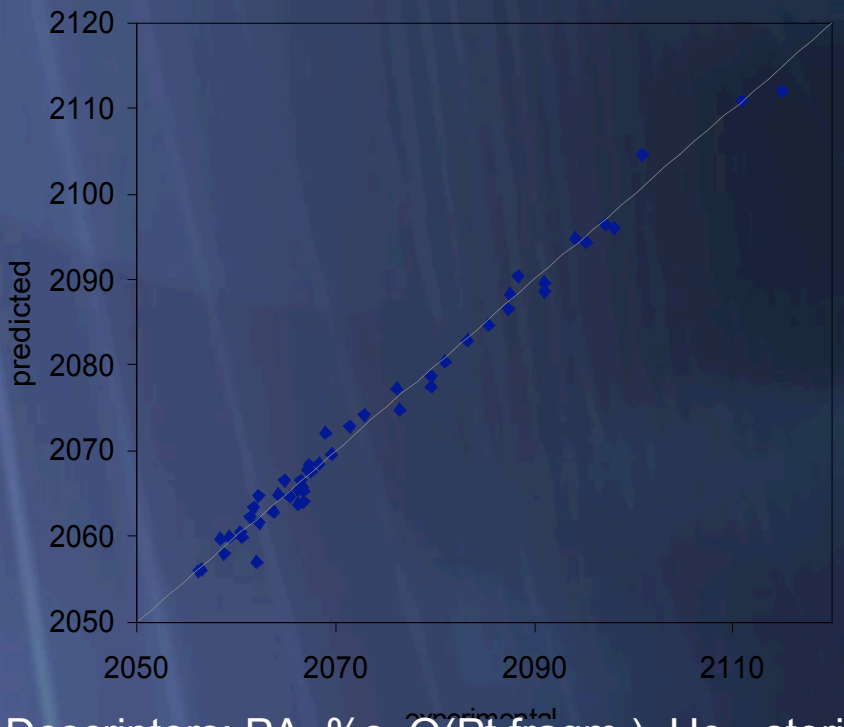


Statistics

1. Identify and screen new catalysts *in silico* –prediction of desirable properties.
2. Direct experimental screening (high-throughput).
3. Detect and quantify ligand similarities/differences.
4. Add to chemical knowledge – interpret ligand contributions to experimental observations.

Tolman Electronic Parameter (cm^{-1})

(ν_{CO} in $Ni(CO)_3L$)

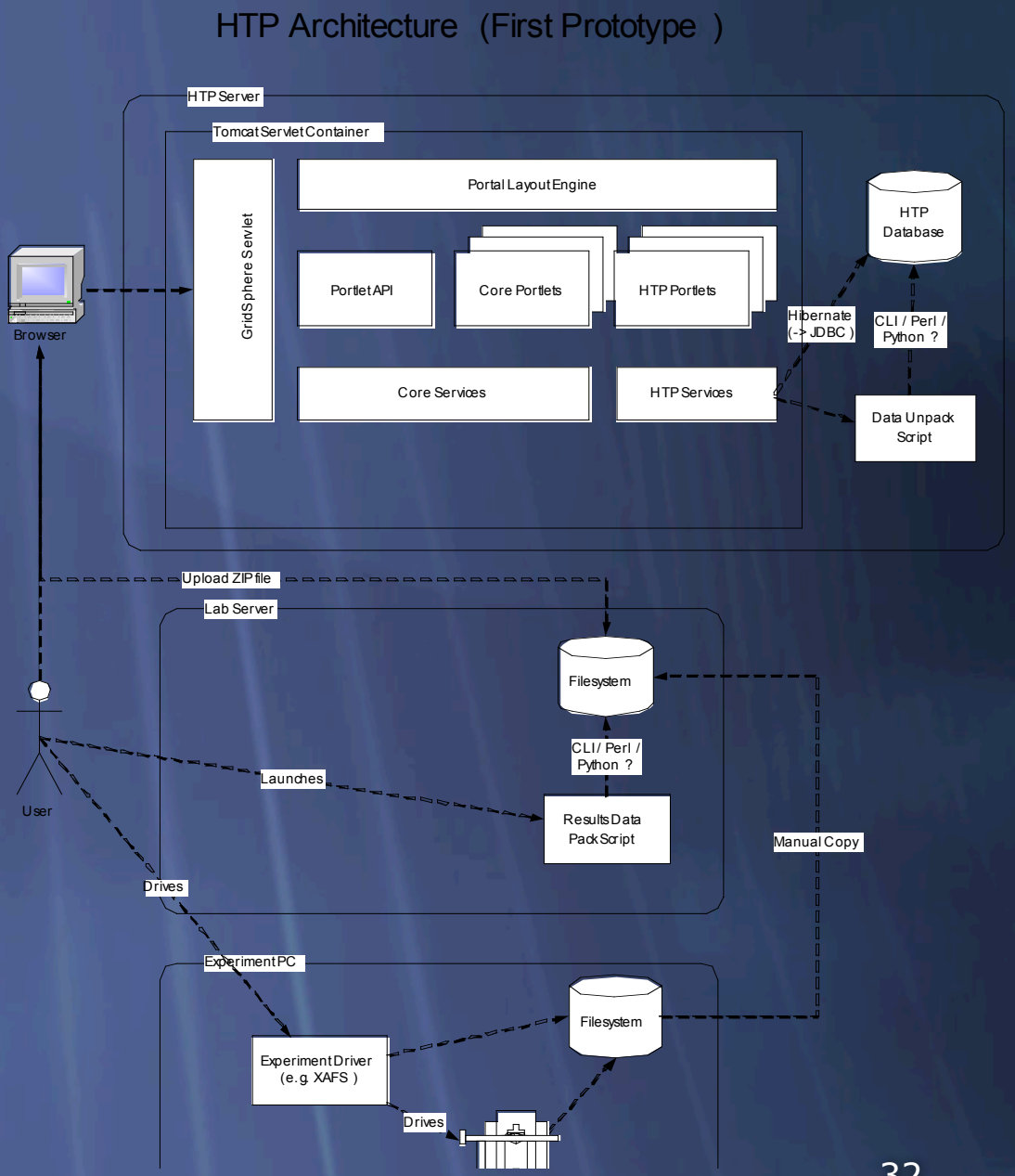


Descriptors: PA, %s, Q(Pt fragm.), He_8_steric, P-B, P-Pt, ΔR-P-R(Pd)

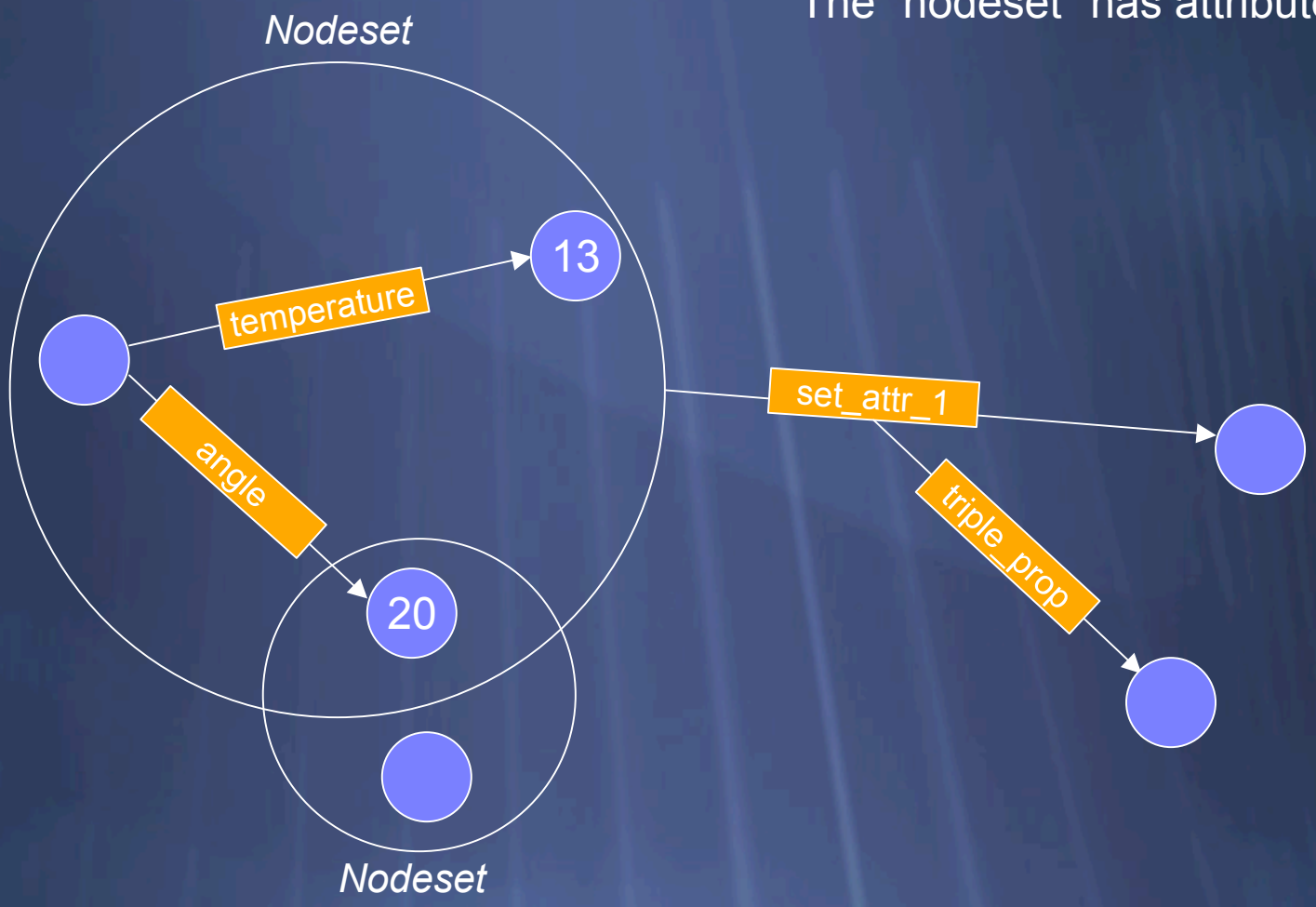
Potential applications of ligand maps

HTP Sample Tracking

Using ideas from the NCS Grid Service we have produced a prototype for a high throughput catalyst experiment involving array samples investigated by Raman, MS, EXAFS with the samples manufactured at one site and tested at several others



The “nodeset” has attributes



- The edge with the attribute name set_attr_1 is an attribute of a nodeset.
- The edge with the attribute name triple_prop is an attribute of the above edge.

Grid and Pervasive Computing

- Electronic Lab Notebook
- Lab Environment
- Mobile Devices
- Semantic throughout