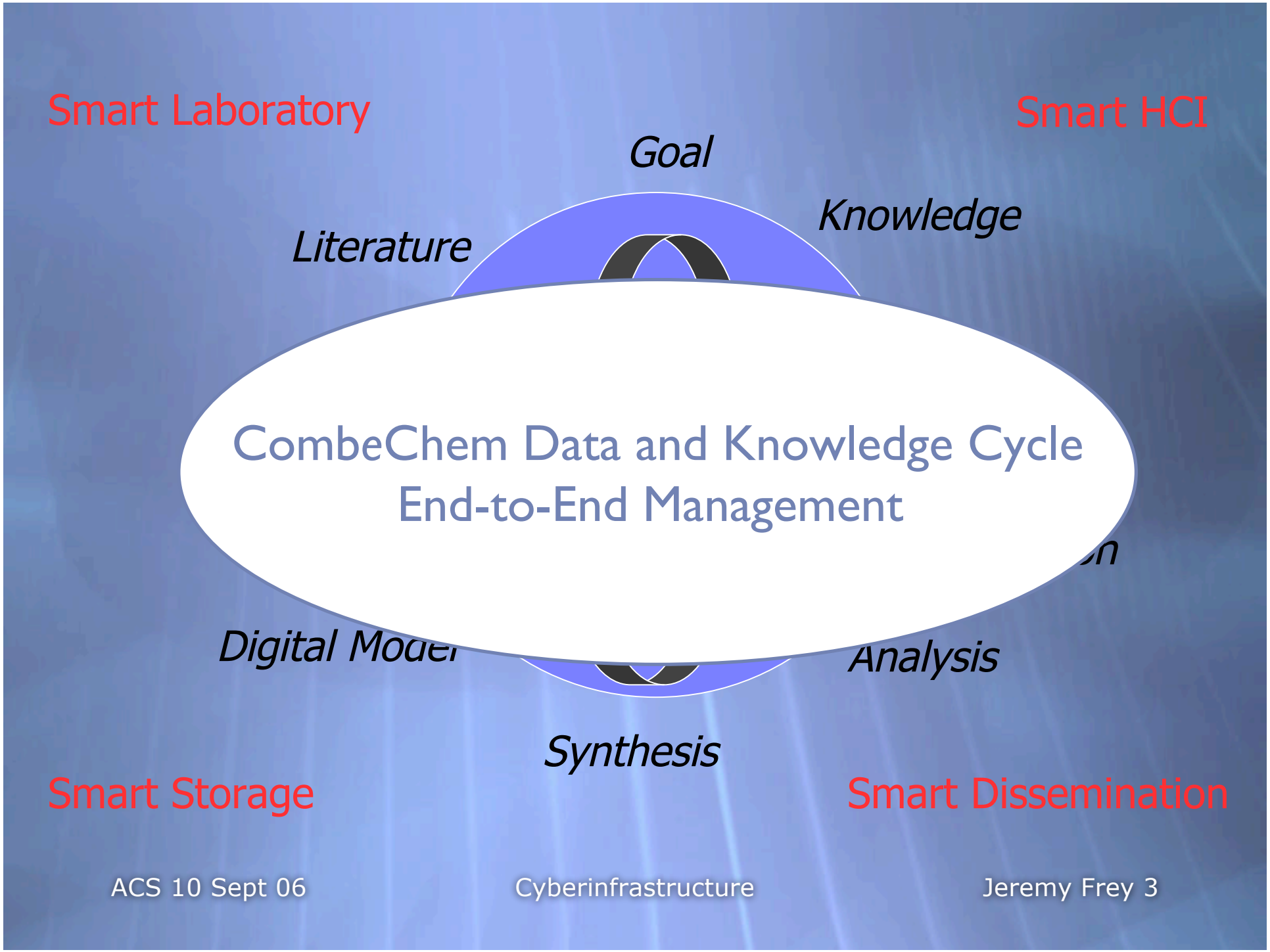
The "CombeChem" Project

Semantic Support for the Chemical Information Life cycle

Jeremy Frey
School of Chemistry
University of Southampton

The CombeChem Project

- End to End linking of data and information
  - Laboratory to publication and back again
  - Very long data chains can be involved e.g. from a chemistry lab to mouse genetic expression
- The exponential world of combinatorial synthesis and high throughput analysis meets the exponentially growing power of computing
  - "Automation, Semantics & the Grid"


Smart Laboratory
Smart HCI
Goal
Knowledge
Literature
CombeChem Data and Knowledge Cycle
End-to-End Management
Digital Model
Analysis
Synthesis
Smart Storage
Smart Dissemination

e-Science

- 'e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it.'

  John Taylor, DG of UK OST

- '[The Grid] intends to make access to computing power, scientific data repositories and experimental facilities as easy as the Web makes access to information.'

  Tony Blair, 2002

- What is the web?

The concept of Publication@Source

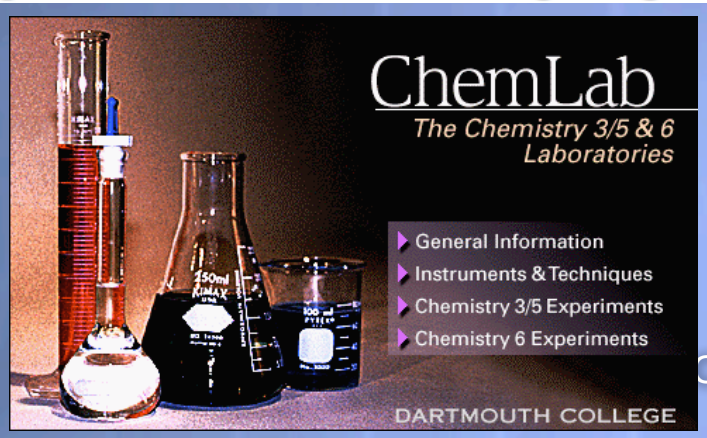
- Trace all the way back from publication to the original data – provenance
- **The data is the key** - DataGrid
- Start as you mean to go on – ELNs are a necessity

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

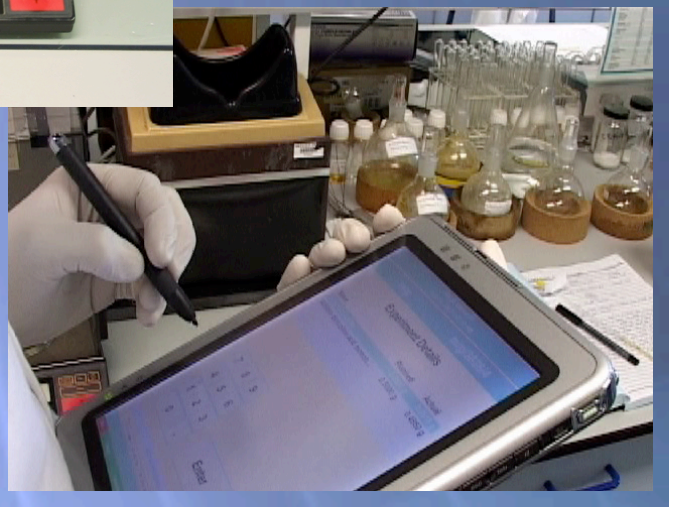


If you are caught using the "scrap of paper" technique, your improperly recorded data may be confiscated by your TA

Lab books are a big block to publication@source: if it's not digital, it is more difficult to share

Only some equipment is networked

This is where it all starts: The Lab & The Lab Book



Need a usable digital lab book. Design by analogy to help Chemists and Computer Scientists work together.

COSHH

leverage off things we already have to do

COSHH ASSESSMENT FORM Record No.

SUBSTANCE NAME	PHYSICAL FORM	QUANTITY	NATURE OF HAZARD
Water	liquid	1000ml	None
Dextrose	Solid	<20 g	possible irritation to eyes and skin
Caffeine	Solid (tea)	<1g	Harmful if swallowed, induce vomiting.
Milk	liquid	<100ml	No particular hazards

NATURE OF PROCESS
liquid extraction of caffeine, followed by combination with dextrose to produce a sweet drink

Is there a less hazardous substance? No
If so, why not use it?

CONTROL MEASURES REQUIRED No specific measure required
(Local exhaust ventilation, personal protection, etc.)

Perhaps computational chemists should be made to do the same!
Statistical modelling and design requires planning to be most effective

PLAN

Process Record

Ingredient List

Fluorinated biphenyl	0.9 g
Br11OCB	1.59 g
Potassium Carbonate	2.07 g
Butanone	40 ml

Dissolve 4-flourinated biphenyl in butanone

Add K2CO3 powder

Heat at reflux for 1.5 hours

Cool and add Br11OCB

Heat at reflux until completion

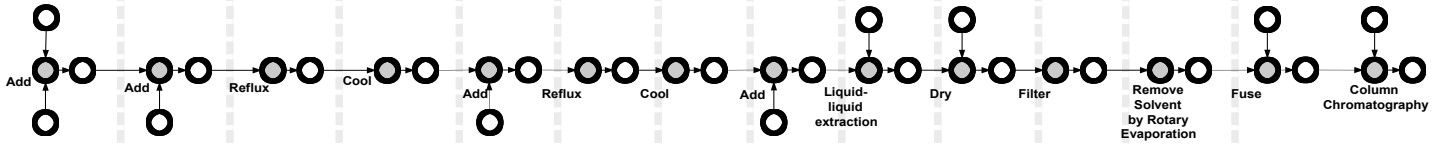
Cool and add water (30ml)

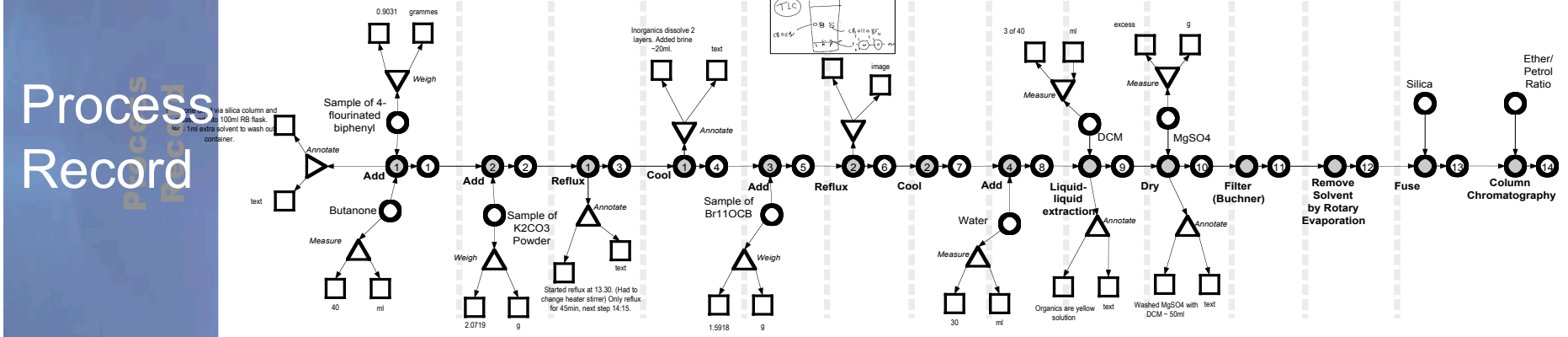
Extract with DCM (3x40ml)

Combine organics, dry over MgSO4 & filter

Remove solvent in vacuo

Fuse compound to silica & column in ether/petrol





Key

Process	○
Input	○
Literal	□
Observation	▽

Observation Types

- weight - grammes
- measure - ml, drops
- annotate - text
- temperature - K, °C

Future Questions

- Whether to have many subclasses of processes or fewer with annotations
- How to depict destructive processes
- How to depict taking lots of samples
- What is the observation/process boundary? e.g. MRI scan

Combechem

30 January 2004
gvh, hrm, gms

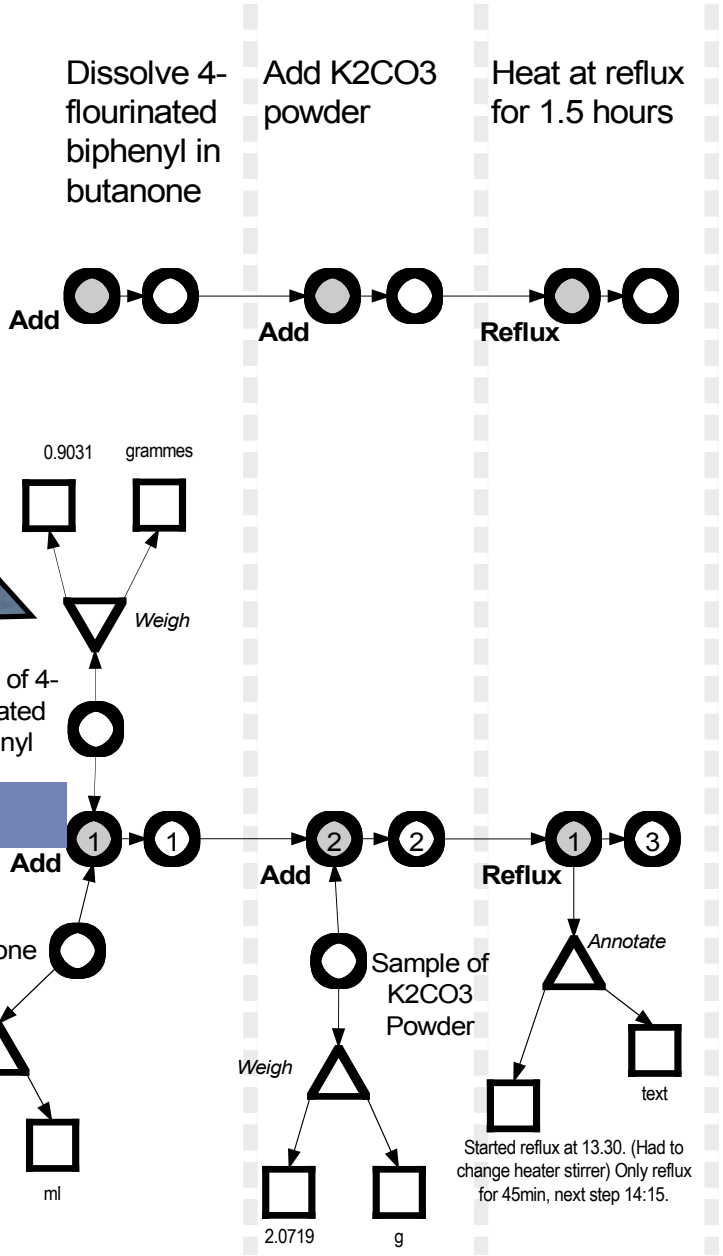
ACS

Ingredient List

Fluorinated biphenyl	0.9 g
Br11OCB	1.59 g
Potassium Carbonate	2.07 g
Butanone	40 ml

Name	Planned	Actual
Fluorinated biphenyl	0.9000 g	0.9031 g
Br11OCB	1.5900 g	1.5918 g
Potassium Carbonate	2.0700 g	2.0719 g
Butanone	40.0 ml	40.0 ml

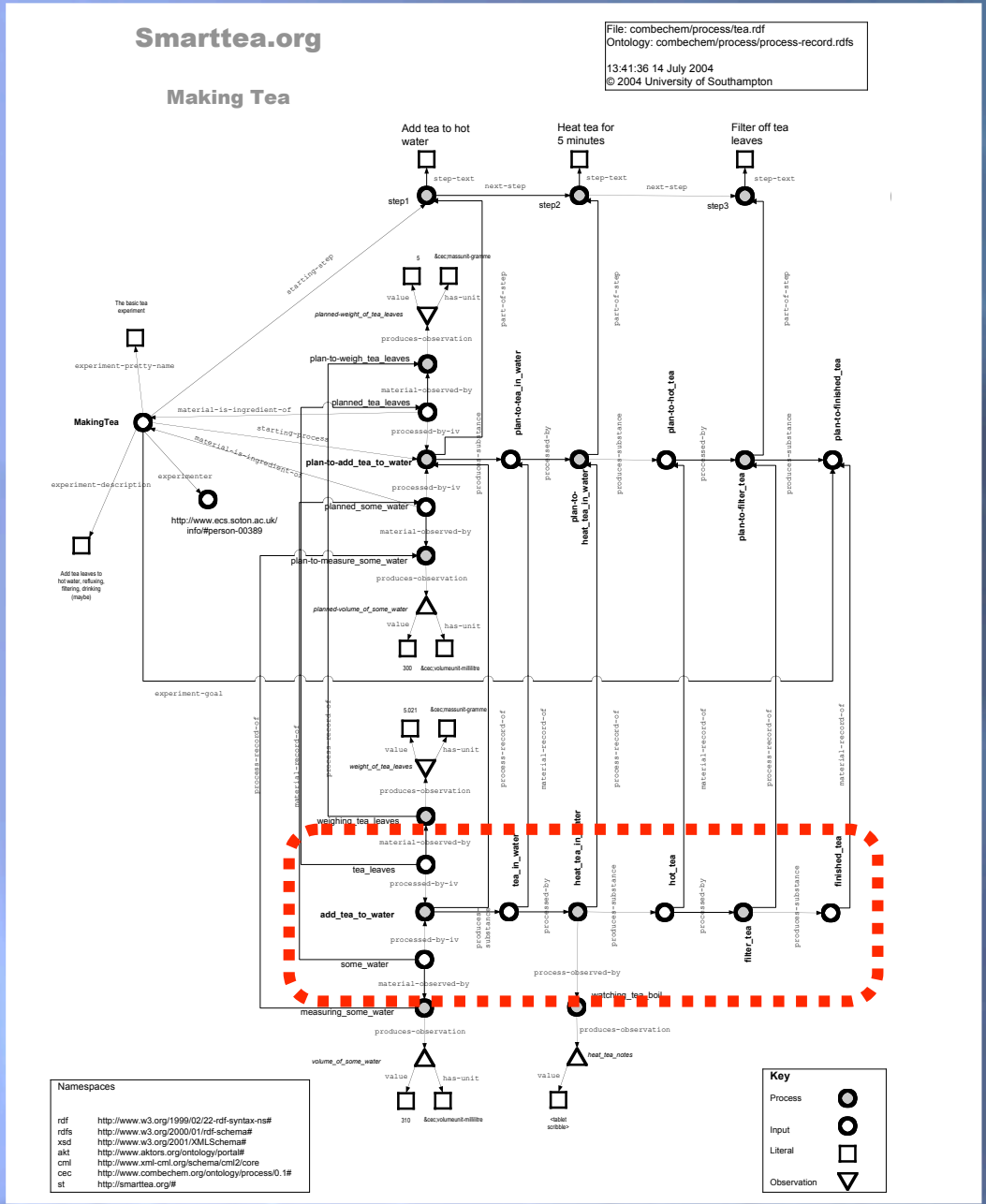
7 8 9
4 5 6
1 2 3
0 .
Enter Del



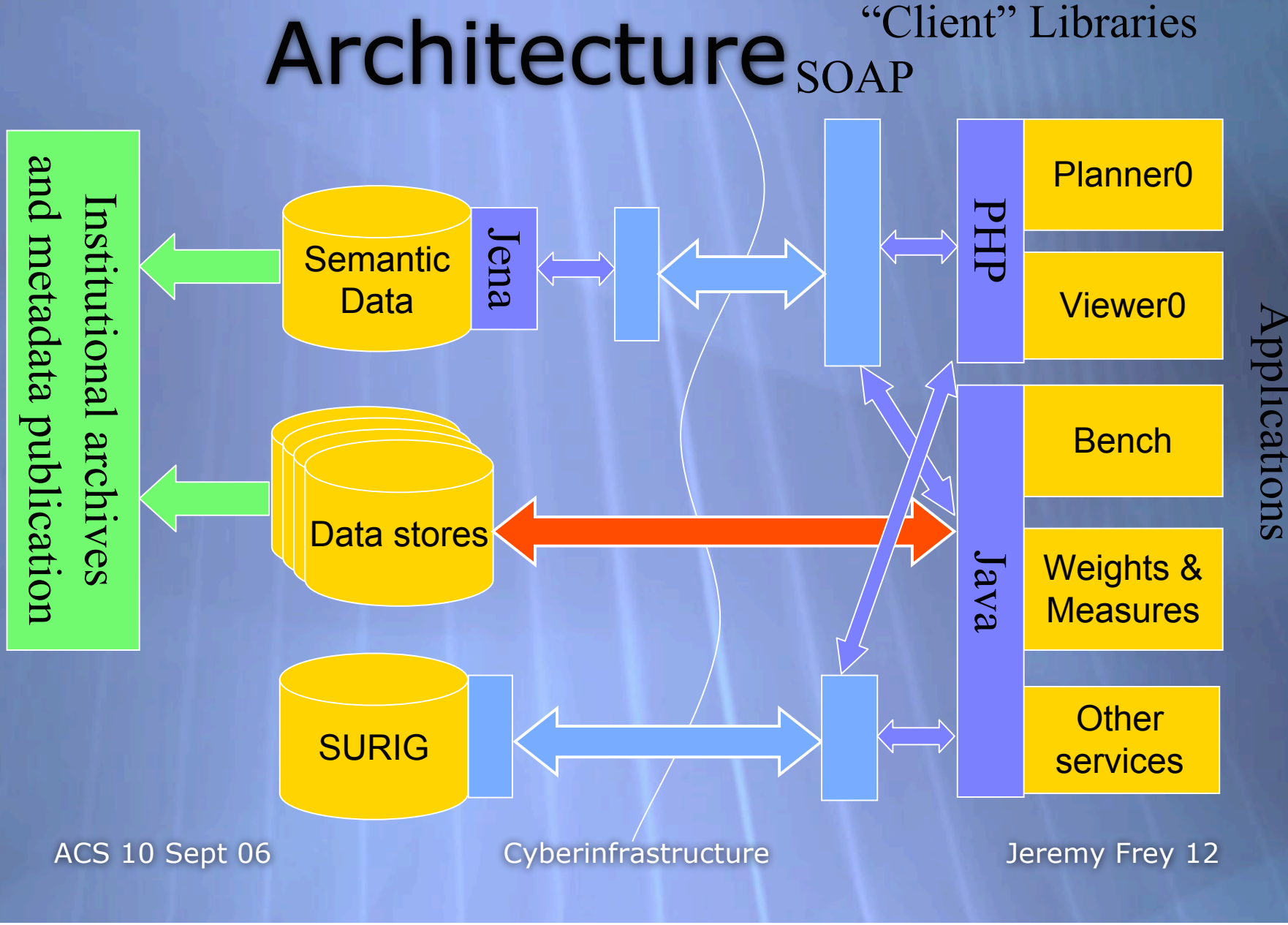
getRecord()

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.

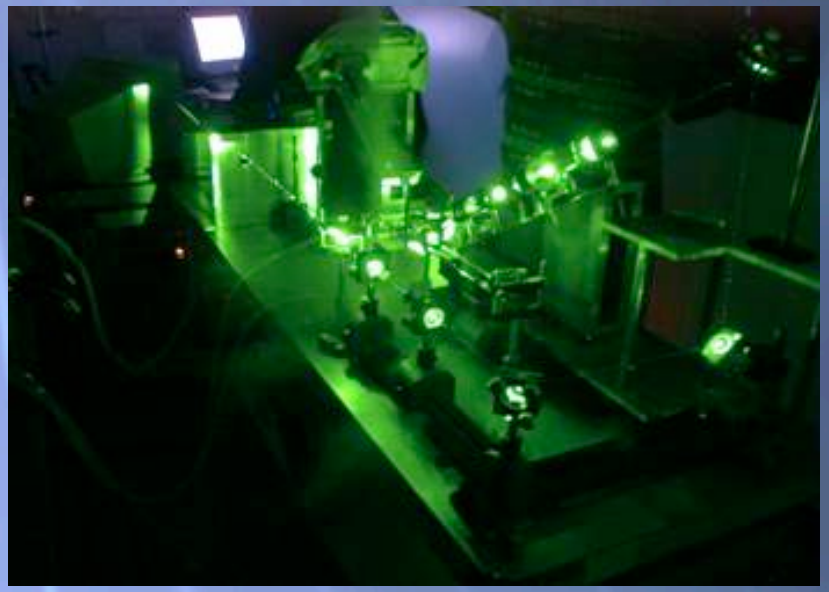


Architecture

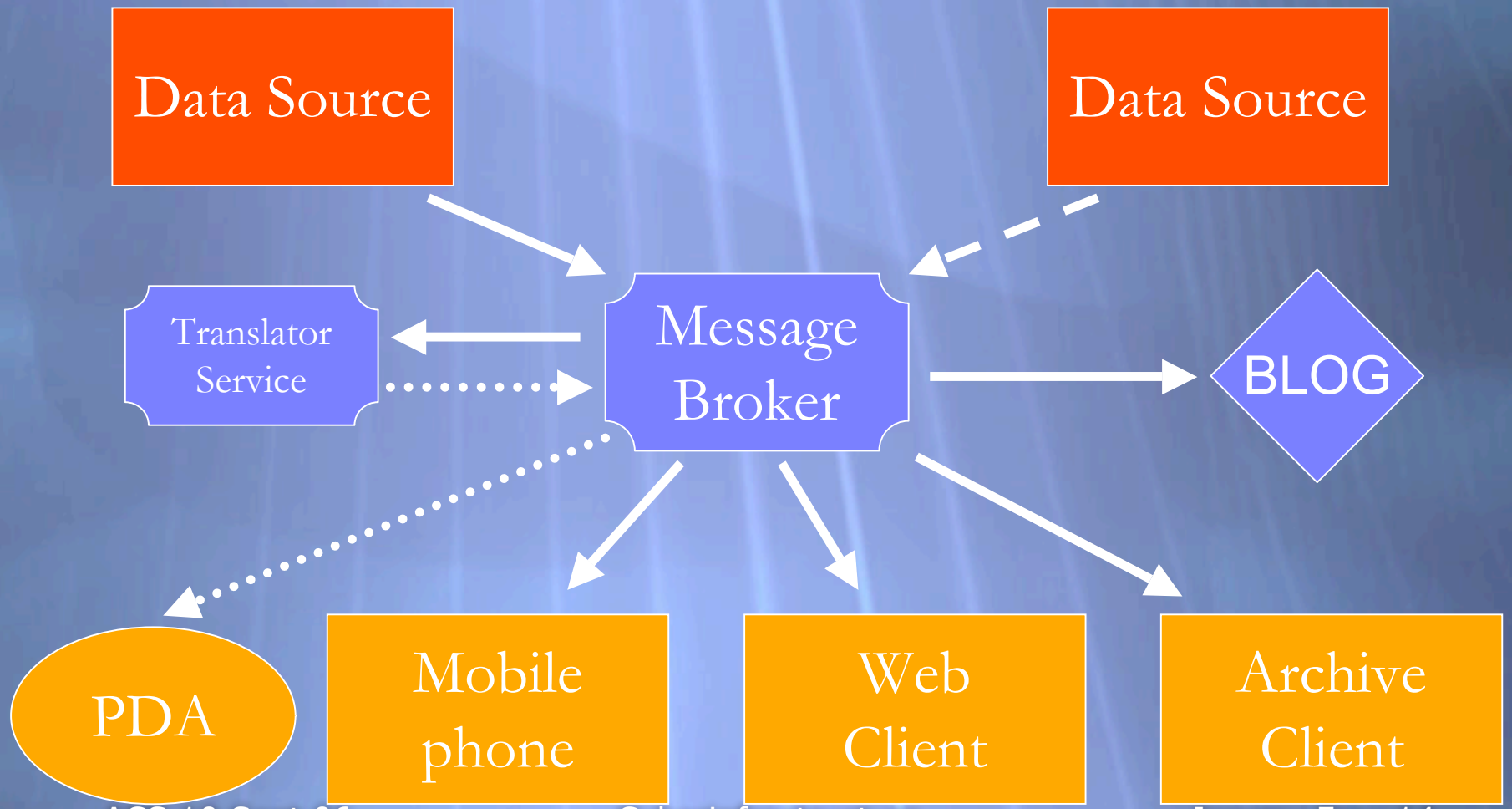


The Laboratory

- Capture information from places you would not want to put your eyes
- Capture environmental data automatically
- Capture people and movements
- Provide this information in real time as well as for the laboratory record

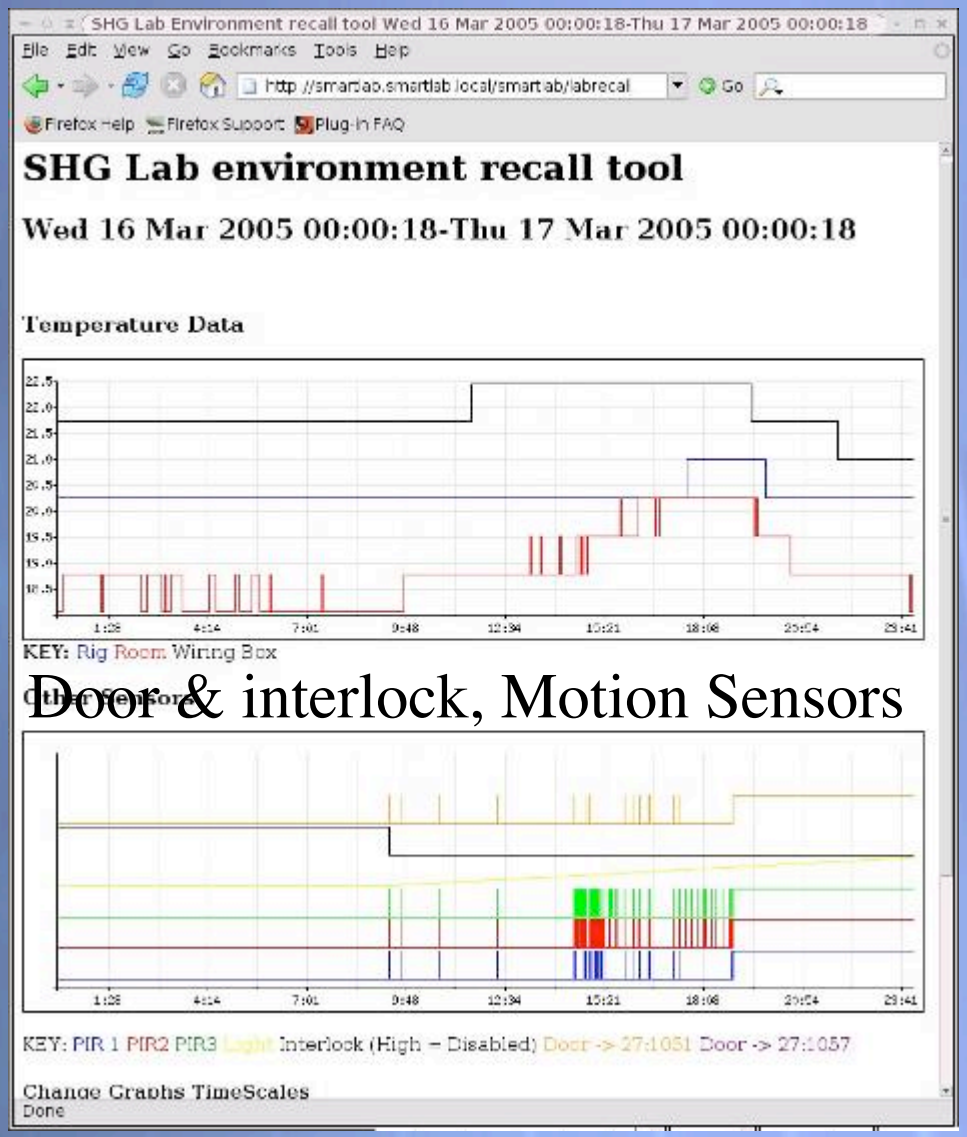


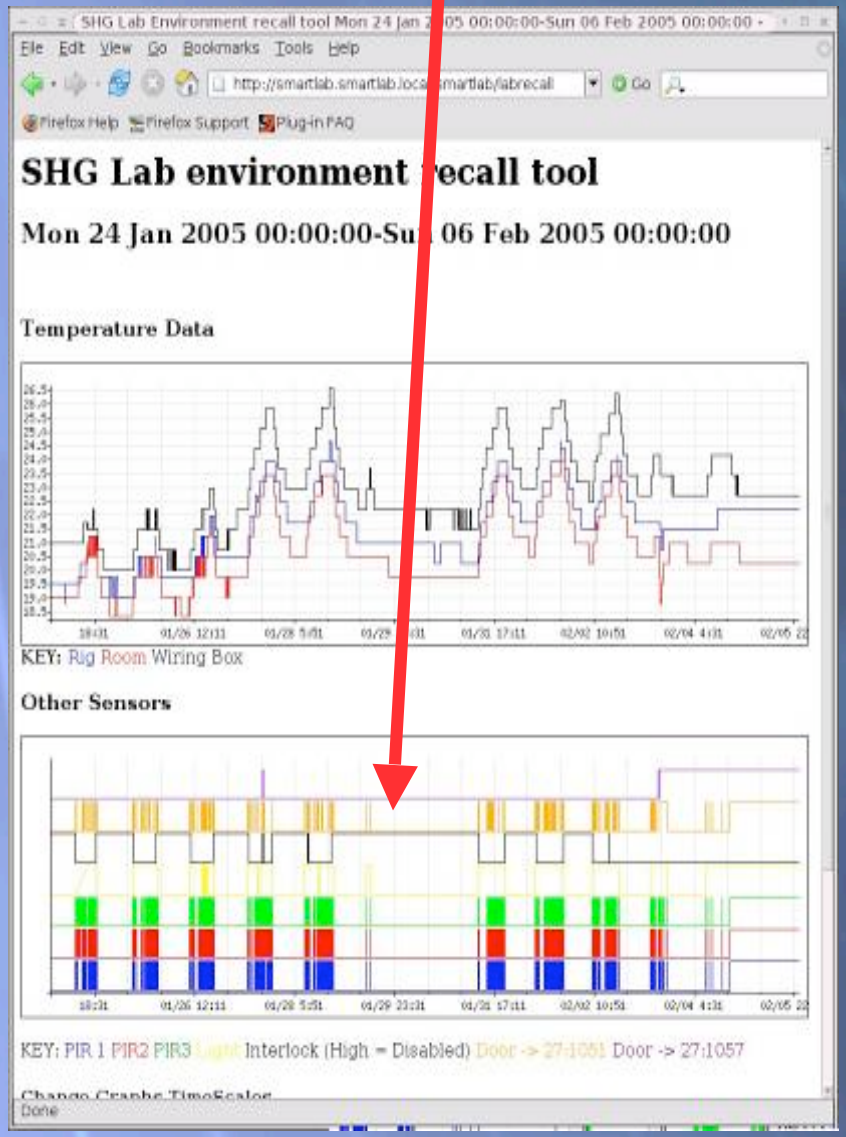
Pub-Sub systems provide the flexible & extensible approach to distribution



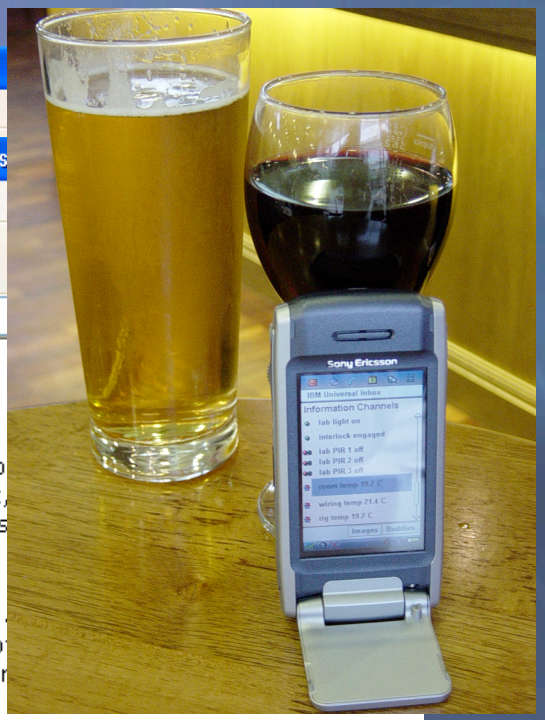
Temperature – room, laser

Air Conditioning failed





Door & interlock, Motion Sensors



BBC NEWS | Science/Nature | Chemists escape labs via mobiles - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back

Address http://news.bbc.co.uk/1/hi/sci/t

bbc.co.uk

Low Graphics version | Change edition

BBC

News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Education
Science/Nature
Technology
Entertainment

Have Your Say
Magazine
In Pictures
Week at a Glance
Country Profiles
In Depth

Last Updated: Fr

E-mail this t

Chemists

By Jo Twist
BBC News scien

A blend of m
and award-w
is letting scie
escape the la

The software,
"middleware",
computer syst
other securely
instantaneous

As part of a n
project in the
used to let Sc

BBC NEWS | Science/Nature | Chemists escape labs via mobiles - Micros

File Edit View Favorites Tools Help

Back Search Favorites

Address http://news.bbc.co.uk/1/hi/sci/tech/4233757.stm

experiment in more detail.

PDAs everywhere

Members of Dr Frey's non-linear laser spectroscop
group, working as part of the Combechem project,
successfully trying out the system at conferences
the university bar.

The next step is to evolve the system so that at
button on a mobile, scientists will be able to remo
the conditions in the lab, like turning down the ter

It could also realise the promise of controlling home environments from phones easily and seamlessly.

But there are many other potential applications for the monitoring jobs done by the power industries, healthcare professions and other labs.

IBM won the Royal Academy of Engineering's MacRobert prize which rewards technological and engineering innovation for the program in June last year.

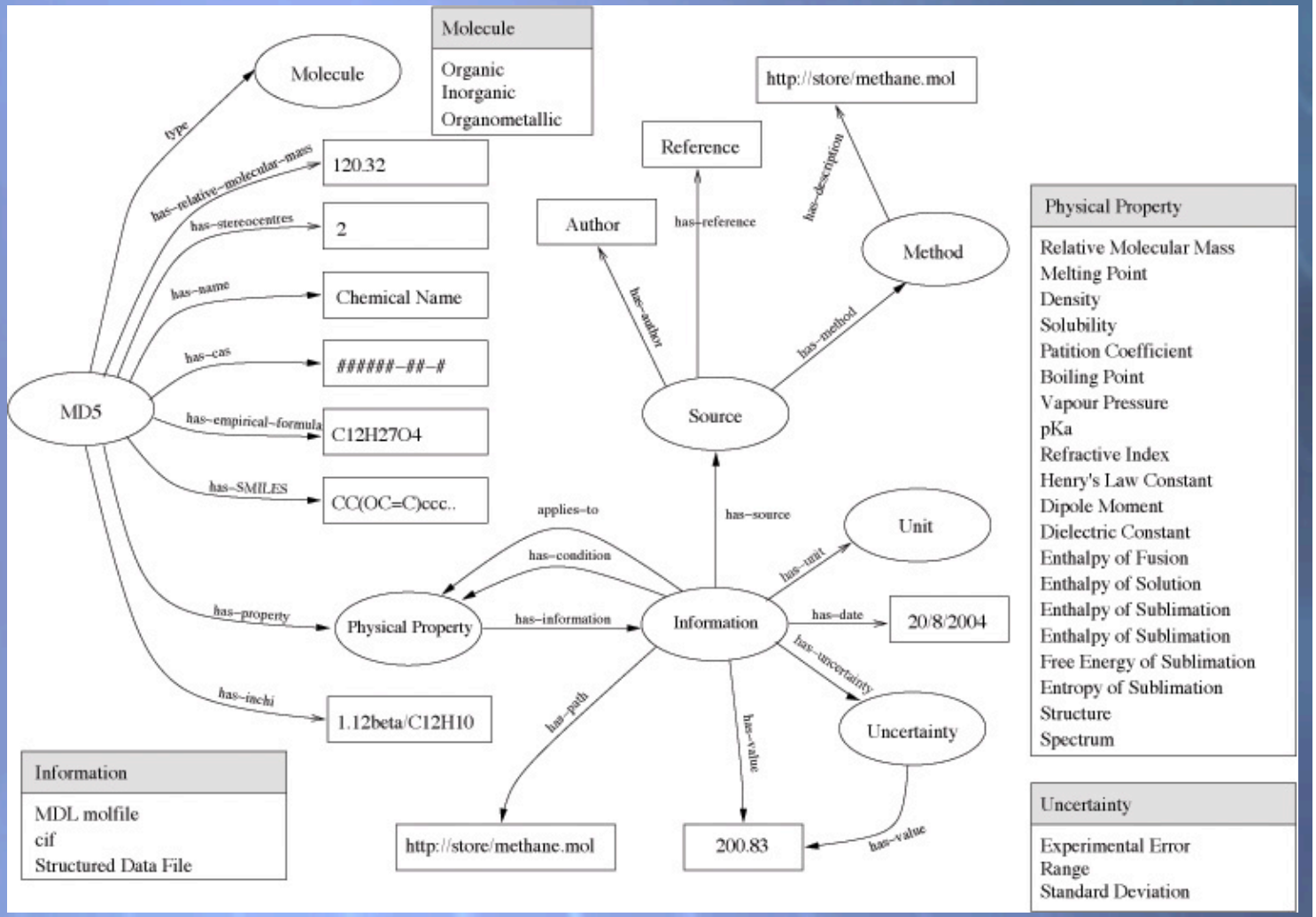
Chemists enjoy a drink at the bar while keeping and eye on the lab

Used by top global banks, the WebSphere MQ family is a decade old.

Databases - Our experience

- What do you do when the actual users keep changing their mind?
- Is a traditional relational database suitable?
- Danger of re-enforcing scientific bias against relational database for laboratory data.
- RDF & Triple stores were again the solution

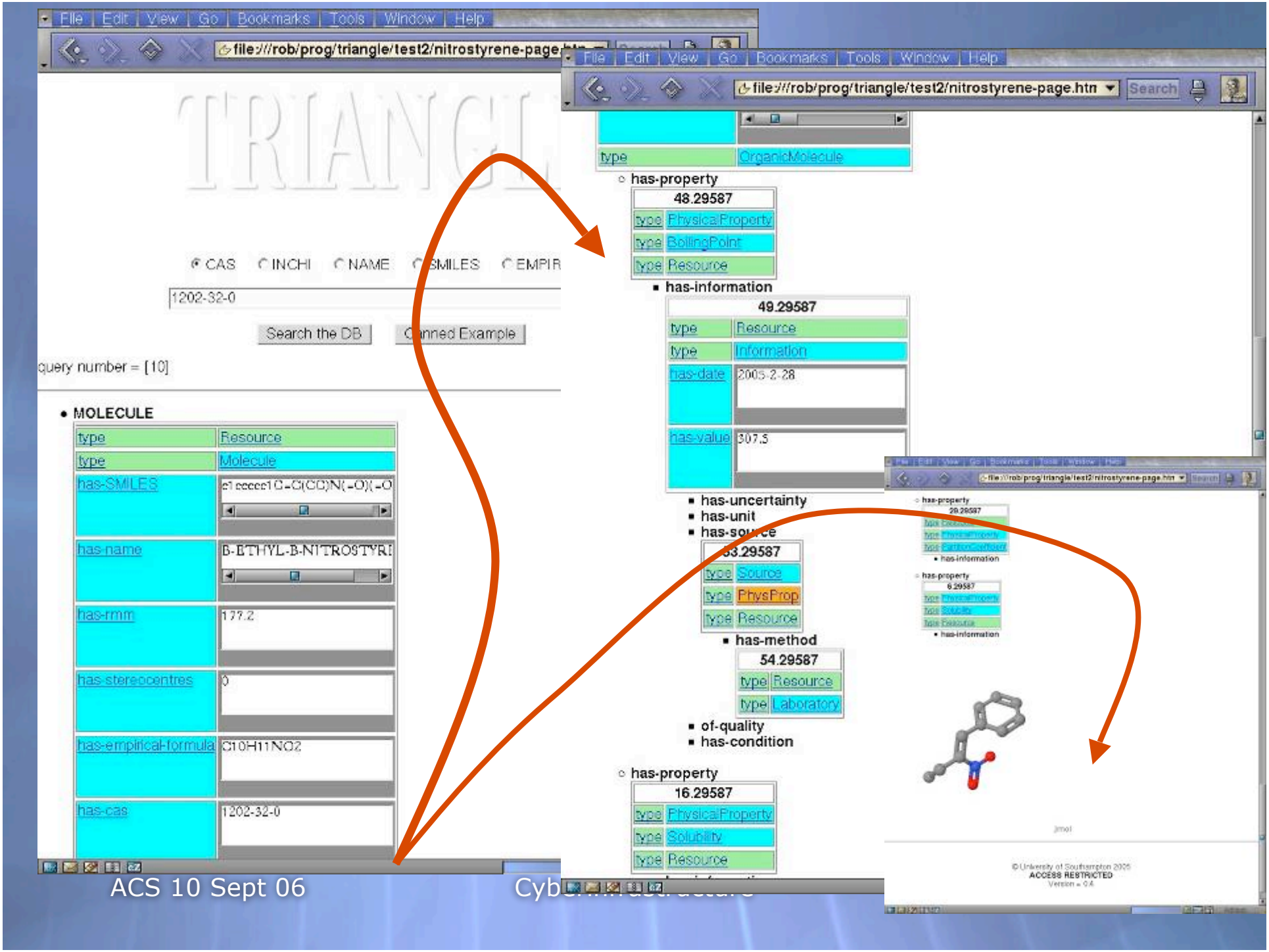
RDF/RDFS High level Schema for chemical properties



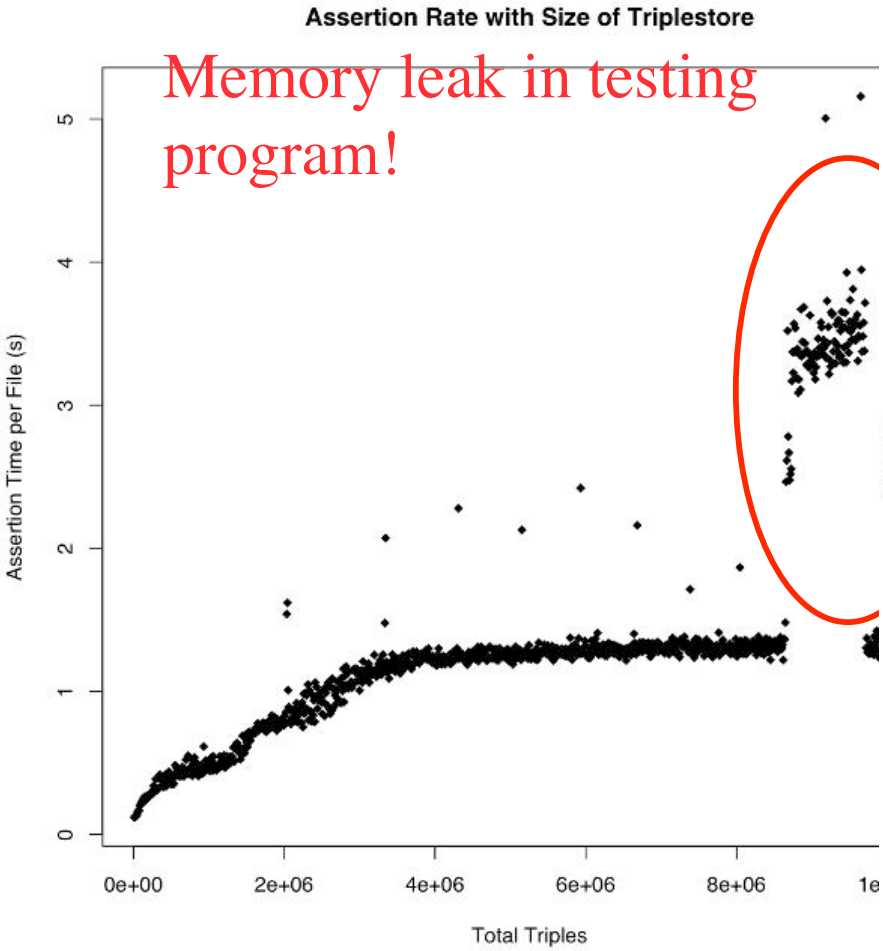
Property in RDF

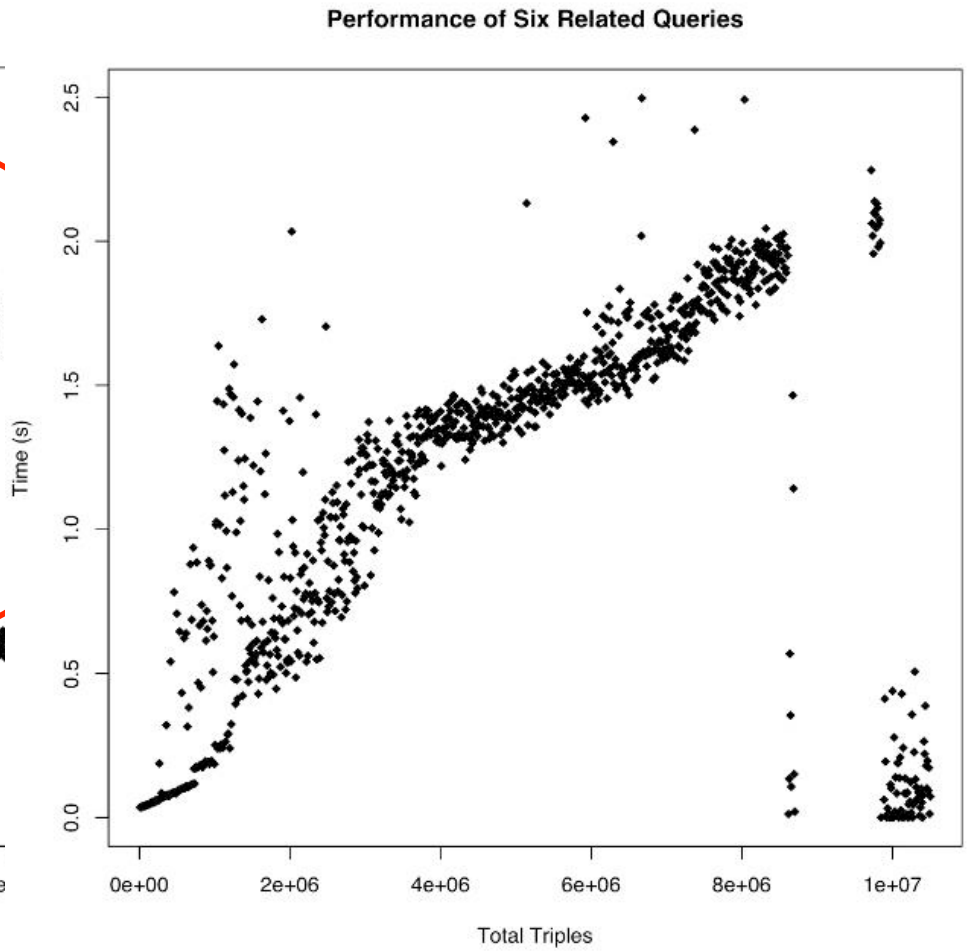
```
<c:OrganicMolecule
rdf:about="file:///storage/ba8efc2ce0edada69d63b02d1b8630c6.rdf">
<c:has-inchi>1.12Beta/C12H13NO2/c1-2-15-8-9-5-6-11(14)12-10(9)4-3-7-
13-12/h1H3,2H2,3-7H,8H2,14H</c:has-inchi>
 <c:has-cas>22049-19-0</c:has-cas>
 <c:has-empirical-formula>C12H13NO2</c:has-empirical-formula>
 <c:has-stereocentres>0</c:has-stereocentres>
 <c:has-property>
  <c:MeltingPoint>
   <c:has-information>
    <c:Information>
     <c:has-value>150
       </c:has-value>
     <c:has-uncertainty>
      <c:Range>
       <c:has-value>16
         </c:has-value>
      </c:Range>
     </c:has-uncertainty>
    </c:Information>
   </c:has-information>
  </c:MeltingPoint>
 </c:has-property>
</c:OrganicMolecule>
```

```
<rdfs:Class rdf:about="&c;OrganicMolecule">
        <rdfs:label>Organic Molecule</rdfs:label>
        <rdfs:subClassOf rdf:resource="&c;Molecule" />
</rdfs:Class>

<rdfs:Class rdf:about="&c;PhysicalProperty">
        <rdfs:label>Property</rdfs:label>
</rdfs:Class>

<rdfs:Class rdf:about="&c;PartitionCoefficient">
        <rdfs:label>Paritition Coefficient</rdfs:label>
        <rdfs:subClassOf rdf:resource="&c;PhysicalProperty" />
        <rdfs:description>Ratio of substance dissolved in octan-
1-ol and water
        </rdfs:description>
</rdfs:Class>
```


File Edit View Go Bookmarks Tools Window Help
file:///rob/prog/triangle/test2/nitrostyrene-page.htm
Search
TRIANGLE
CAS INCHI NAME SMILES EMPIR
1202-32-0
Search the DB
Canned Example
query number = [10]
MOLECULE
type Resource
type Molecule
has-SMILES
has-name
B-ETHYL-B-NITROSTYRE
has-rmm
177.2
has-stereocentres
0
has-empirical-formula
C10H11NO2
has-cas
1202-32-0
type OrganicMolecule
has-property
48.29587
type PhysicalProperty
type BoilingPoint
type Resource
has-information
49.29587
type Resource
type Information
has-date
2005-2-28
has-value
307.5
has-uncertainty
has-unit
has-source
type Source
type PhysProp
type Resource
has-method
54.29587
type Resource
type Laboratory
of-quality
has-condition
has-property
16.29587
type PhysicalProperty
type Solubility
type Resource
ACCESS RESTRICTED

Triple Stores - The Heart of the Semantic Web
Scaling - 3Store response





Scaling the triplestores

Moved from…

- A model of harvesting data from multiple sources into one scalable store

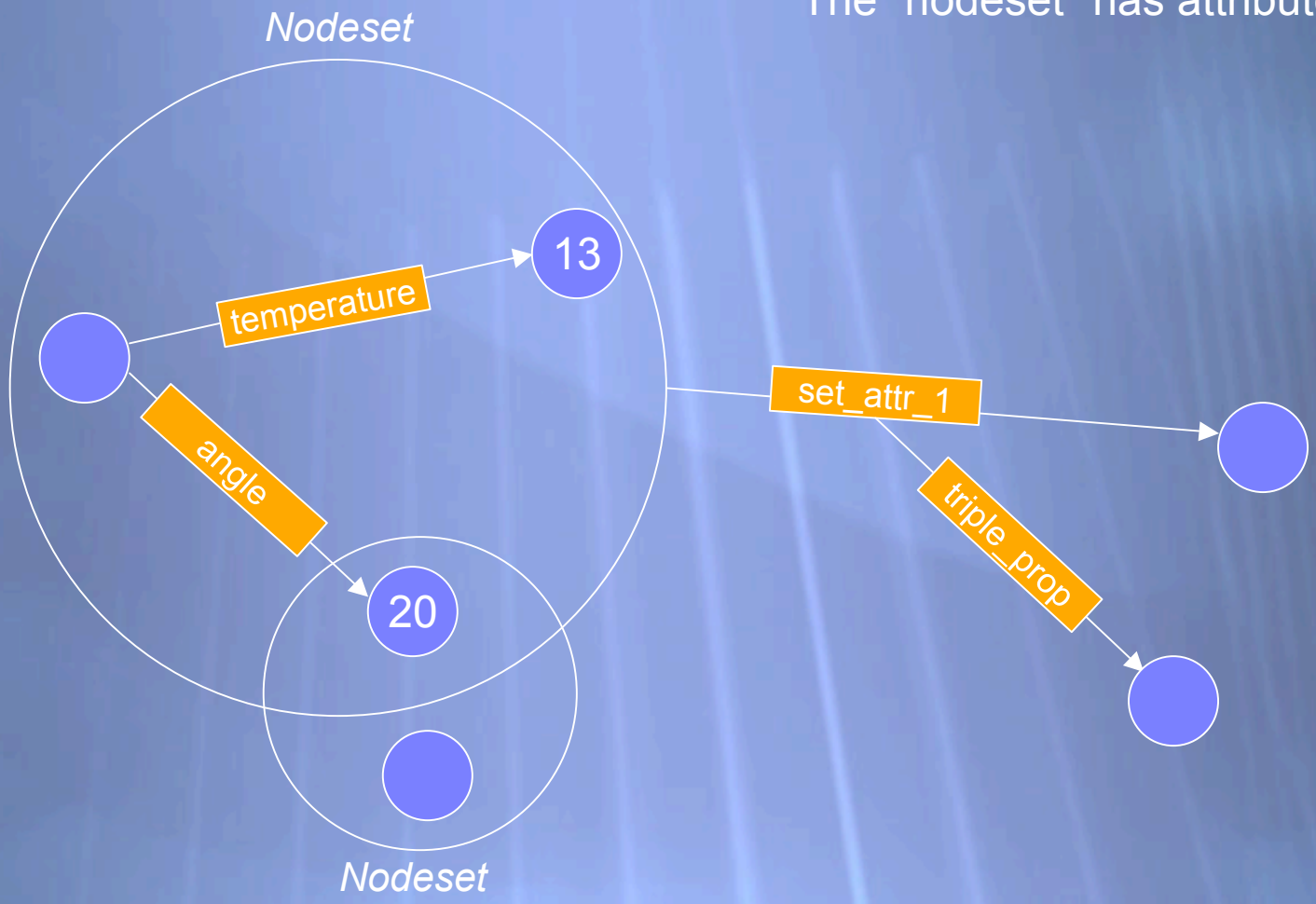
to

- A model of distributed RDF sources and caching what is needed for the task at hand into multiple stores fit-for-purpose

The Semantic Web!

- Very Complex RDF
- Use a graph-based model.
  - Base on RDF ideas
  - Actual data is stored as a graph
    - Contrast with models like E-R, where the graph "models" the data, rather than actually being the data.
    - A node in E-R might be "customer", and represent the class of entities that are customers, rather than any specific customer.
- The pseudograph model:
  - Each node is a datum.
  - Each edge denotes an association/attribute/property.
  - Nodes can be grouped into nodesets, which are also nodes.
    - A node may be in more than one nodeset.
  - A node-edge-node triple can also be a node.
  - Main difference from RDF is an attempt to build reification into the model.

The “nodeset” has attributes



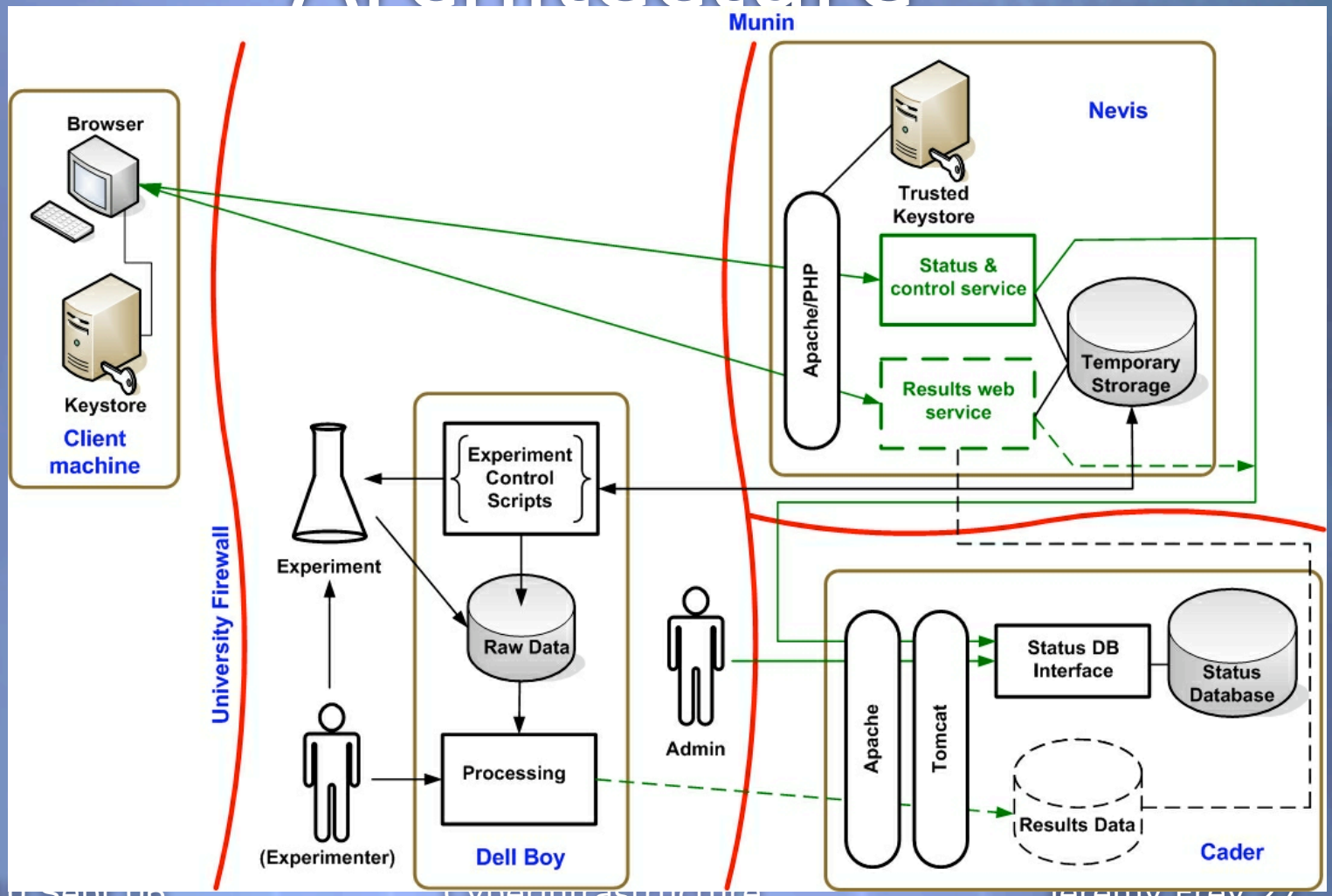
- The edge with the attribute name set_attr_1 is an attribute of a nodeset.
- The edge with the attribute name triple_prop is an attribute of the above edge.

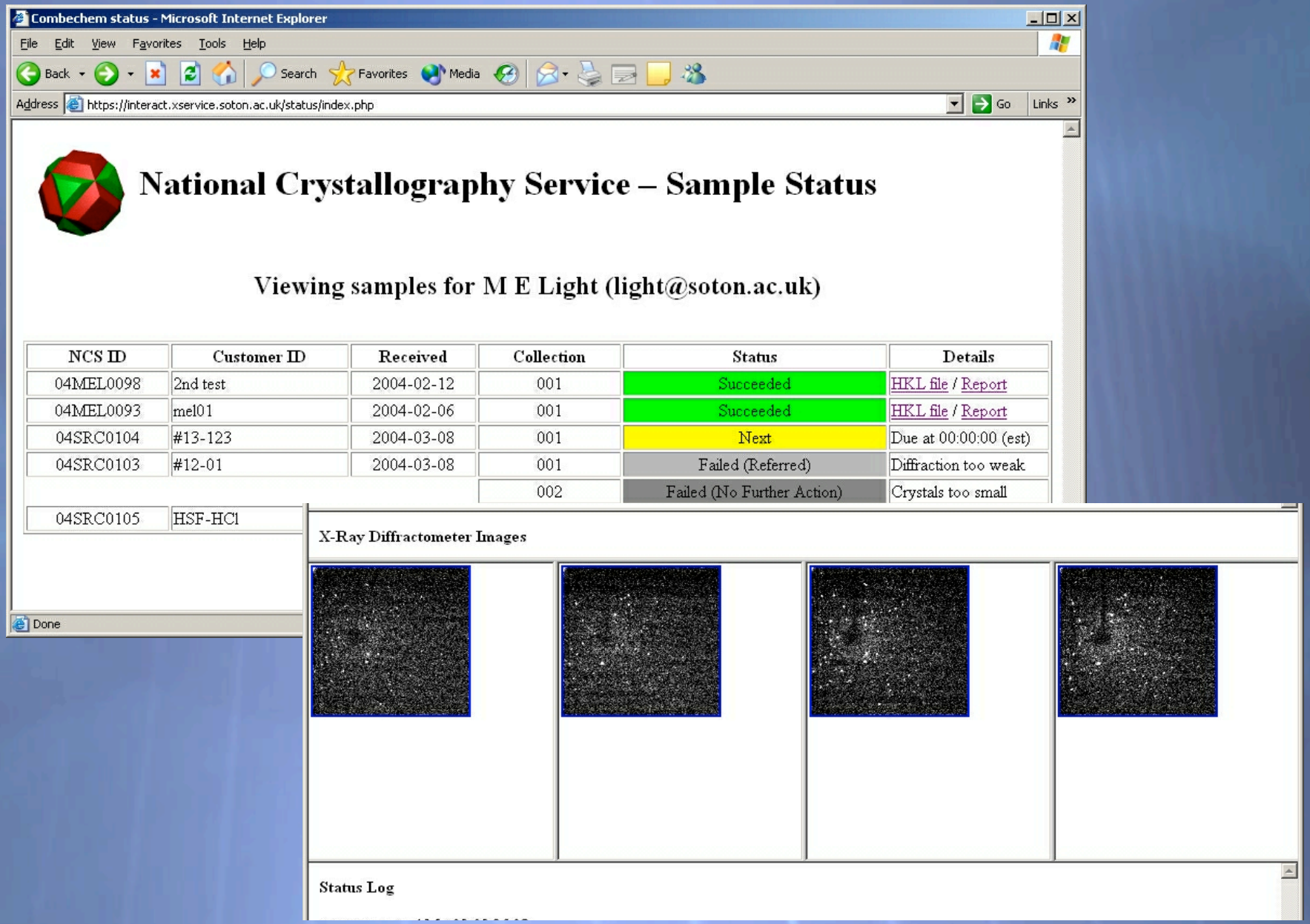
Experiments on the Grid

Security and trust for experiments and data

Role based authorisation needed for experiments not entirely the same as for computers

NCS Grid Service Architecture

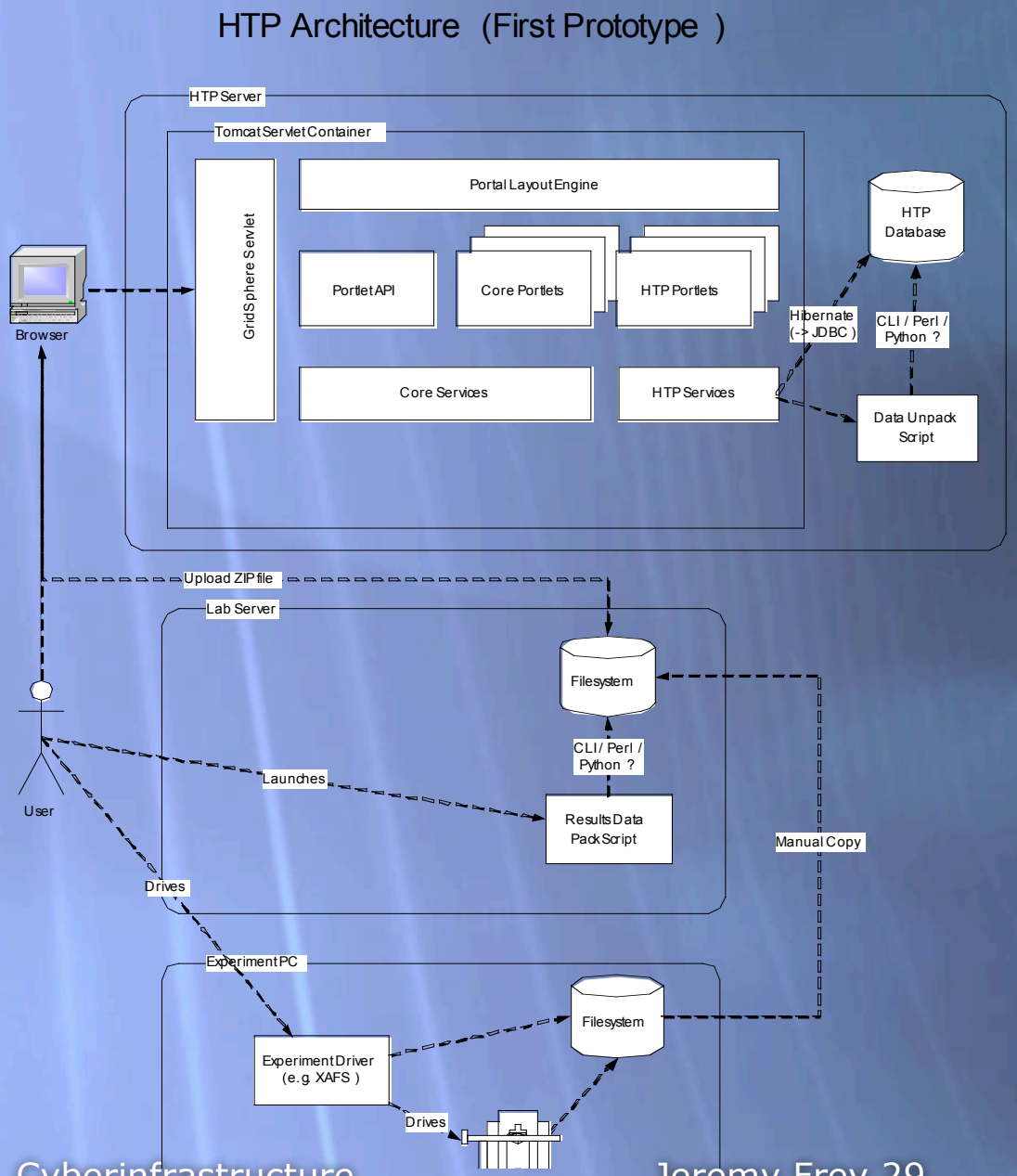




Users can follow and interact with experiment

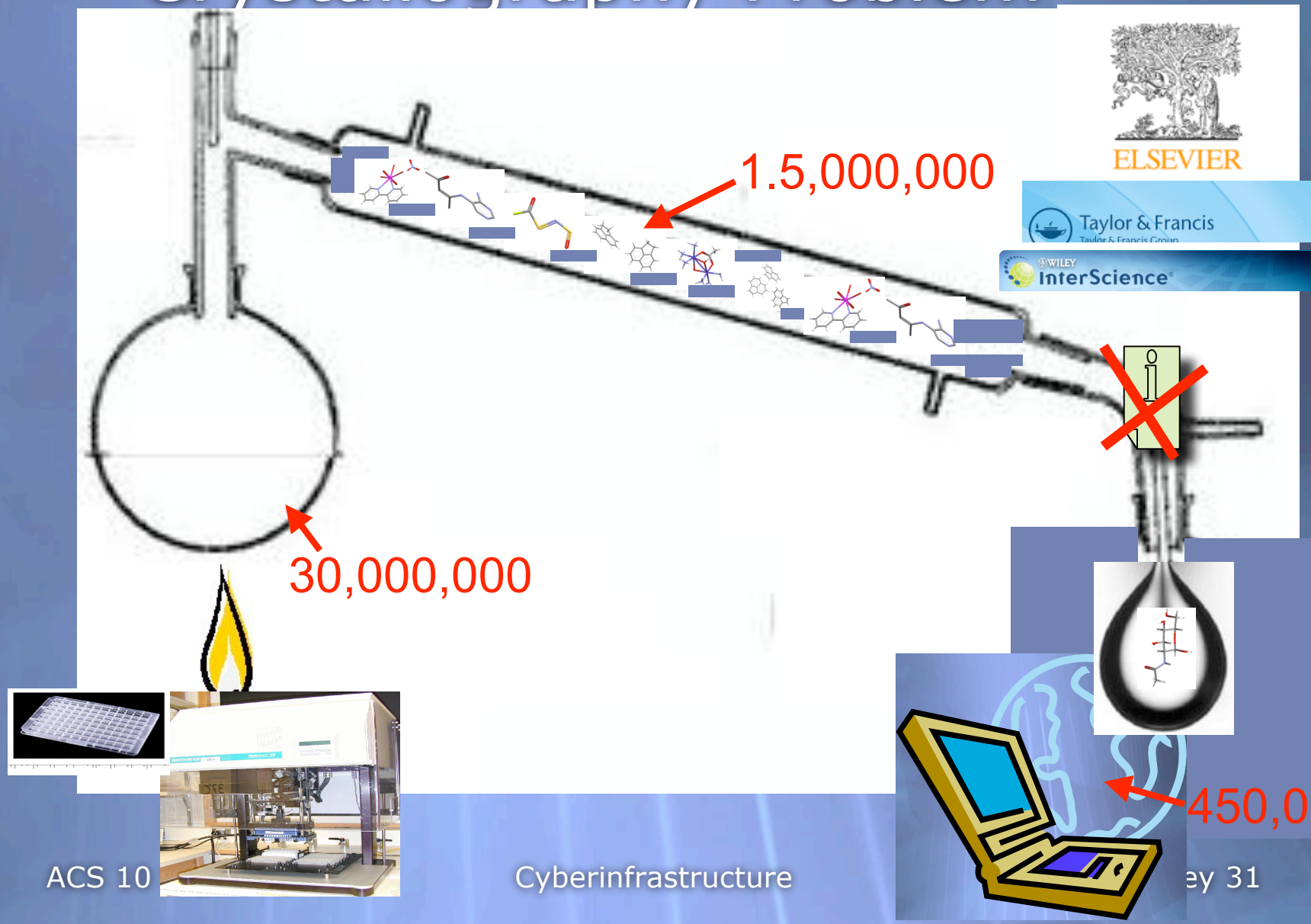
HTP Sample Tracking

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



Dissemination

A Data-Rich Subject – the Crystallography Problem



The eCrystals Data Archive

University of Southampton Crystal Structure Report Archive

Home
About
Browse
Search
Register
User Area
Help

6,7,9,10,12,13,15,16-Octahydro-benzo-1,4,7,10,13-pentaoxacyclopentadecin

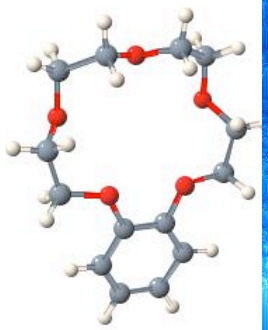
Simon J Coles, Michael B Hursthouse, Jeremy G Frey and Esther Rousay.

University of Southampton

$C_{14}H_{20}O_5$

InChI=1/C14H20O5/c1-2-4-14-13(3-1)18-11-9-16-7-5-15-6-8-17-10-12-19-14/h1-4H,5-12H2

DOI: 10.594/ecrystals.chem.soton.ac.uk/145
Compound Class: Organic
Keywords: crown ethers
Creation Date: 07 October 2004
Deposited By: A.N. Admin
Deposited On: 20 February 2006



Available Files

Depositor Comments

Structure already known, but accurately redetermined for a local research project.

Data collection parameters

Chemical formula	C14 H20 O5
Crystallisation Solvent	
Crystal morphology	Plate
Crystal system	Orthorhombic
Space group symbol	Pbca
Cell length a	16.4963(18)
Cell length b	8.325(3)
Cell length c	20.061(6)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	120(2)

Refinement results

Solution figure of merit	0.0409
R Factor (Obs)	0.0487
R Factor (All)	0.0977
Weighted R Factor (Obs)	0.1008
Weighted R Factor (All)	0.1192

Citation: Coles, S.J., Hursthouse, M.B., Frey, J.G. and Rousay, E. (2004), Southampton, UK, University of Southampton, Crystal Structure Report Archive. (doi:10.1594/ecrystals.chem.soton.ac.uk/145)

Final Result

04sjc0831.cif	13k
04sjc0831.cml	6k

Validation

04sjc0831_checkcif.htm	7k

Refinement

04sjc0831.res	6k
04sjc0831_xl.lst	34k

Solution

04sjc0831.prp	6k
04sjc0831_xs.lst	39k

Processing

04sjc0831.hkl	702k
04sjc0831.htm	10k
04sjc0831_0kl.jpg	57k
04sjc0831_h0l.jpg	85k
04sjc0831_hk0.jpg	88k

Data Collection

04sjc0831_crystal.jpg	17k

Other Files

04sjc0831.doc	78k
04sjc0831.fcf.txt	155k

http://ecrystals.chem.soton.ac.uk

Access to the underlying data

Available Files

Data collection parameters

Chemical formula	C30 H26 Fe N2 O3
Crystallisation Solvent	
Crystal morphology	
Crystal system	Orthorhombic
Space group symbol	Pbca
Cell length a	6.0816(4)
Cell length b	24.8503(16)
Cell length c	31.120(3)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	120(2)

Refinement results

Solution figure of merit	
R Factor (Obs)	0.0573
R Factor (All)	0.1185
Weighted R Factor (Obs)	0.1046
Weighted R Factor (All)	0.1243

Final Result

02sot064.CIF	19k
02sot064.cml	8k
02sot064_checkcif.html	14k

Refinement

02sot064.RES	9k

Solution

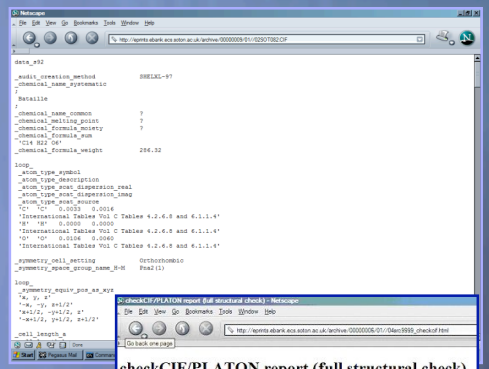
02sot064.PRP	5k

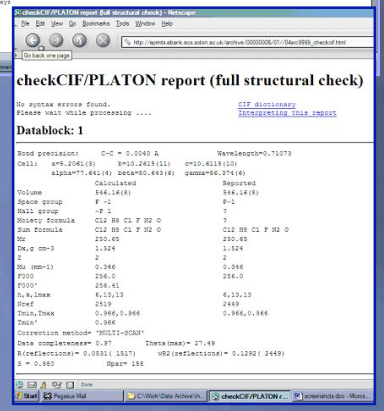
Processing

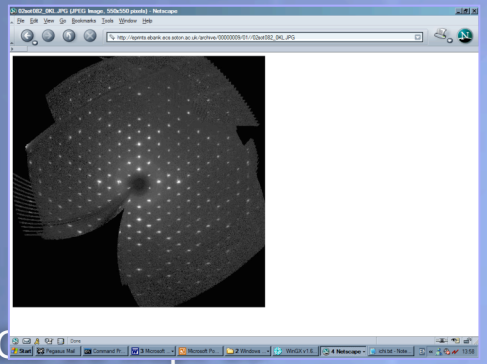
02SOT064.HTM	6k
02sot064.HKL	338k

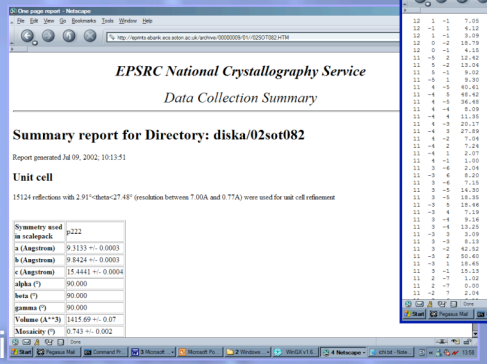
Other Files

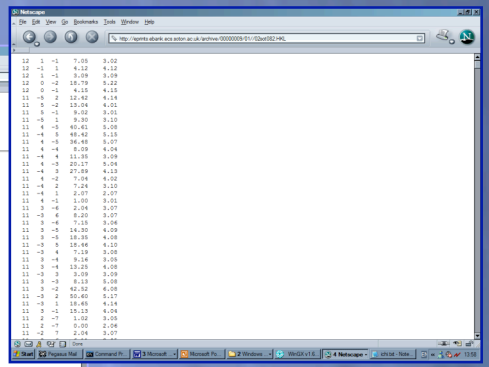
02sot064.DOC	113k
02sot064.LST	49k

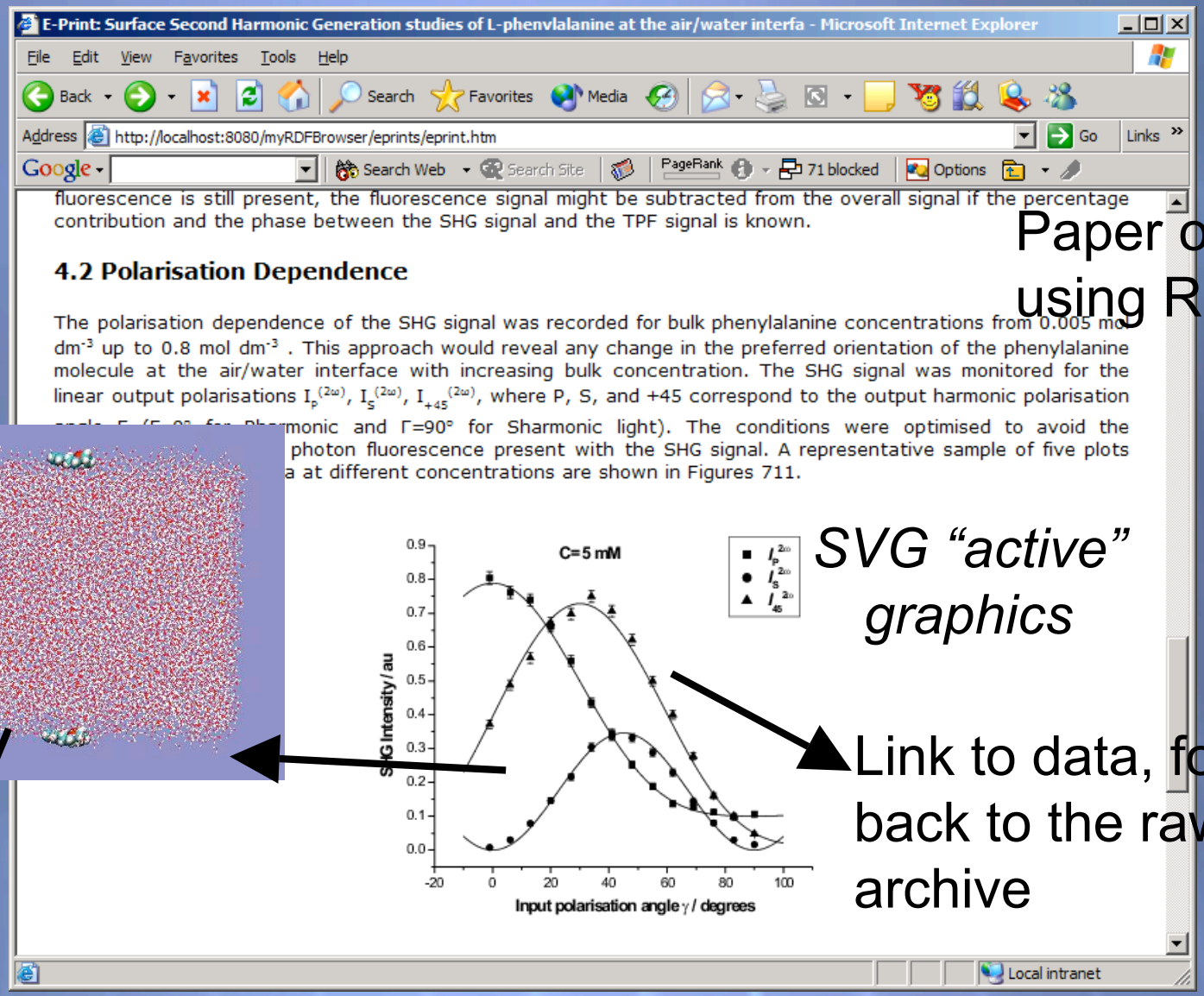












Paper organized using RDF

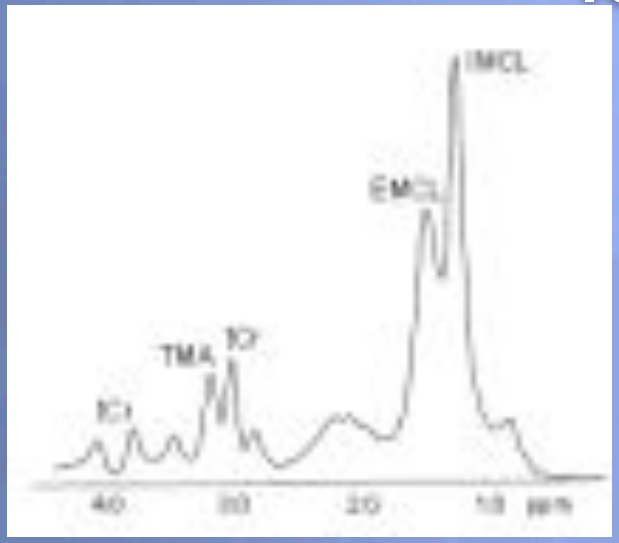
SVG "active" graphics

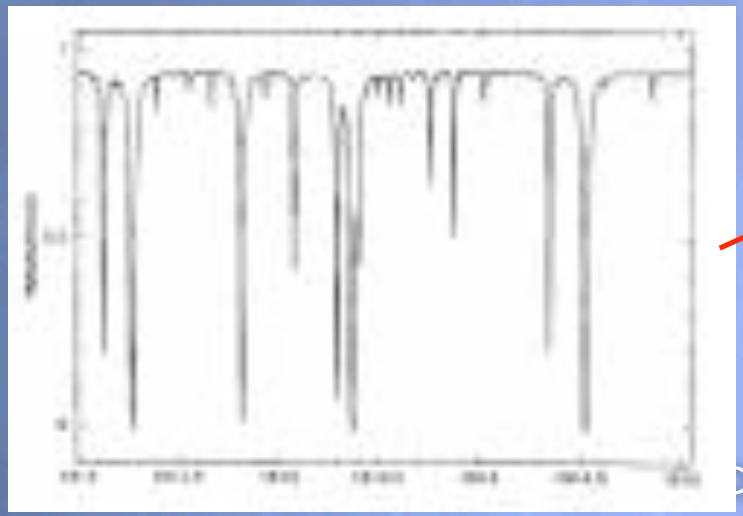
Link to data, follow links back to the raw data archive

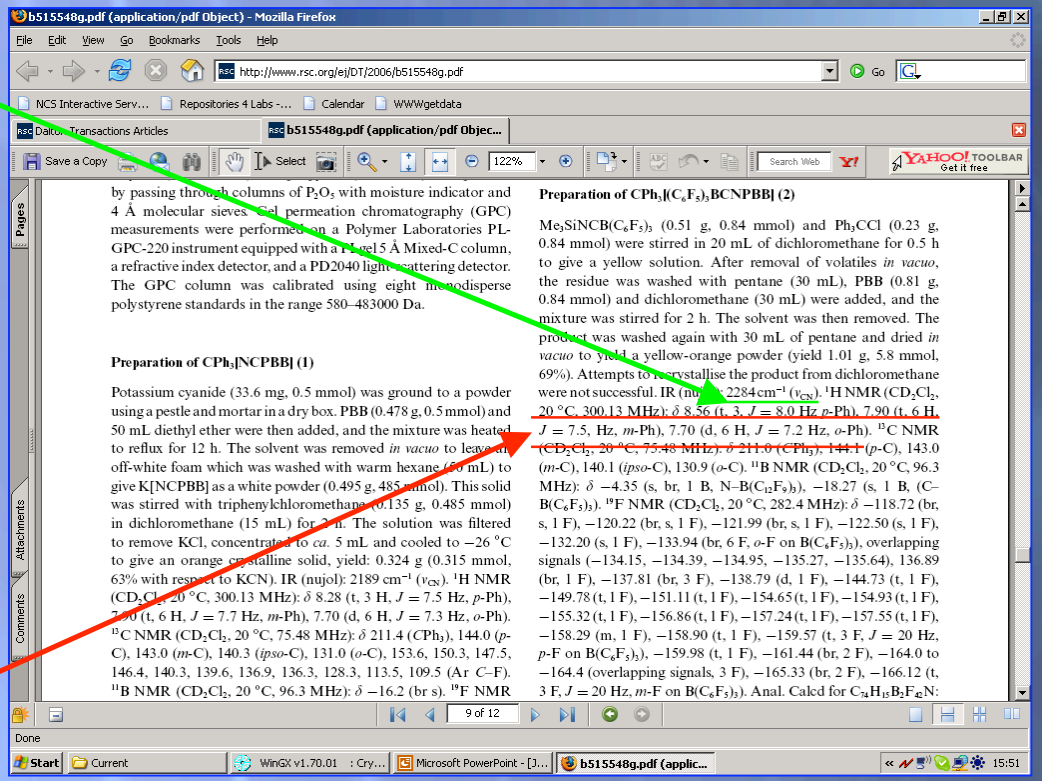
R4L

Link to simulation, full simulation data archived in BioSimGrid

Need for a data archive in the laboratory



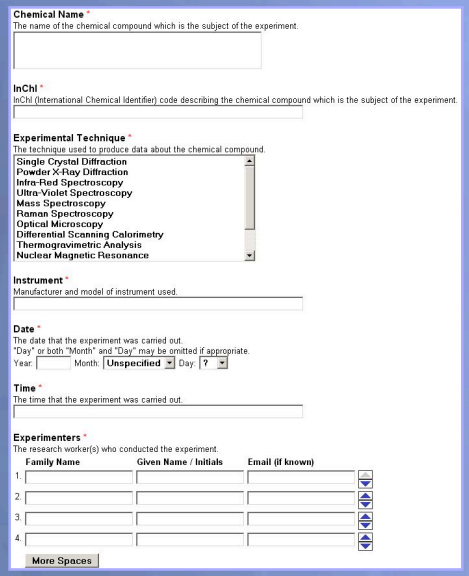


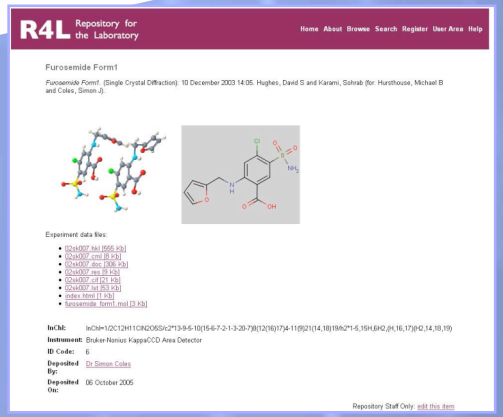


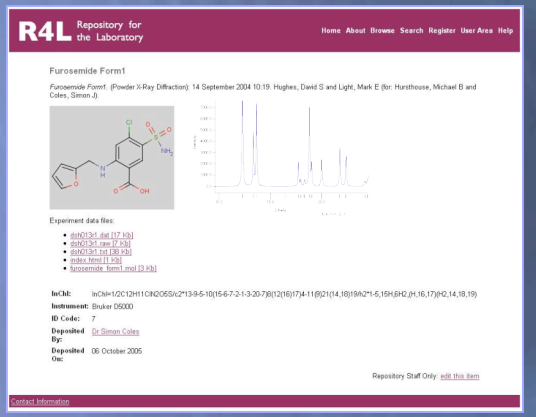
Not just the published spectra!

The R4L Repository

Deposit

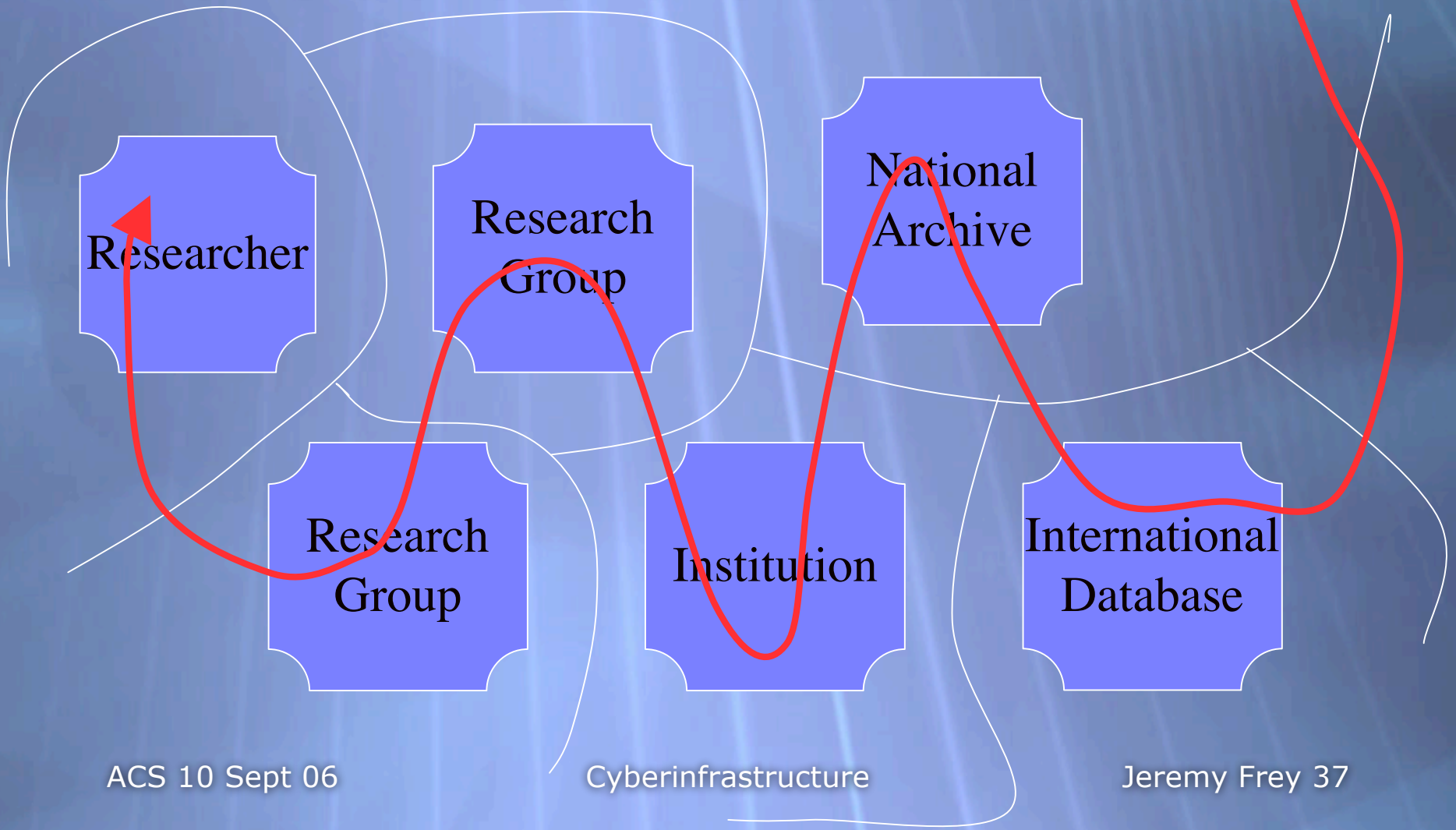






Search

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



Summary:

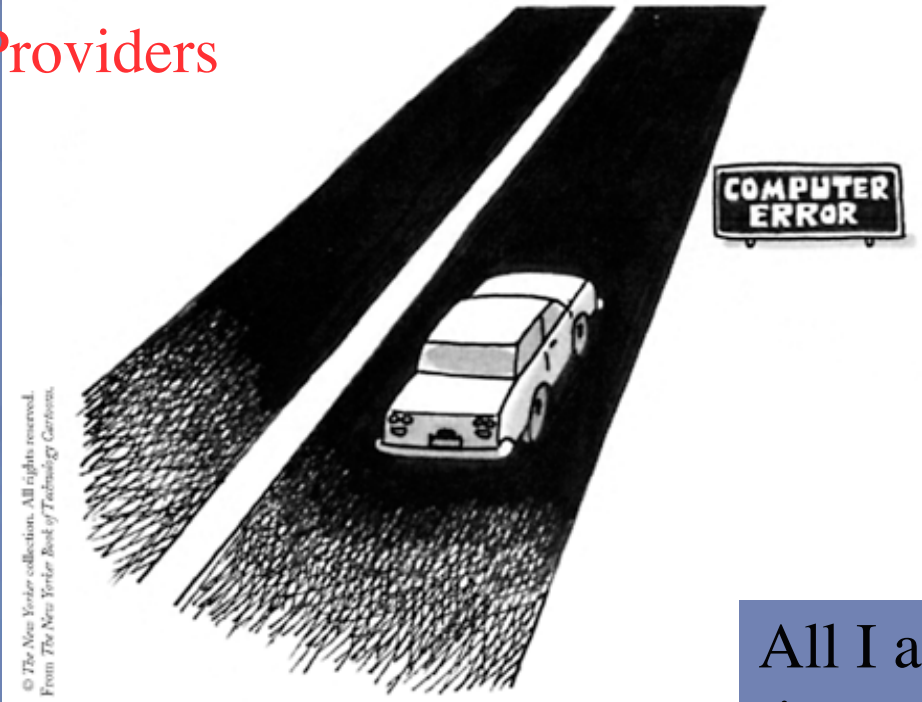
- Making sure other people can find, understand and re-use your data easily and with confidence (even when there is a huge amount of it!)
- Make use of Plans to inform the digital context - metadata in advance
- Have concern for the "End-to-End life cycle" of chemistry information from the start.
- Understanding Usability and Human Computer Interaction is vital for adoption

Summary:

- Is the semantic web/grid enough
- It is sufficient to evolve and improve the practice of chemistry
- It is not enough - 'Pictures are worth a thousand words' more like 10,000 words of metadata
- The Semiotic Web will be needed to improve the communication of models

Information Consumers

Information Providers



All I am saying is that now is the time to develop the technology to deflect an asteroid

www.combechem.org