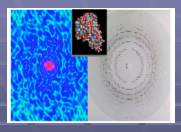
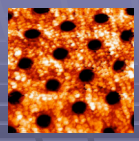
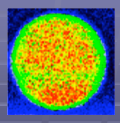
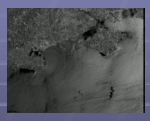





Automation and Semantics: The CombeChem Experience

Jeremy Frey
Informatics & Data Visualisation
Intech Centre Oct 2004

Talk: Workflow

- Introduction to e-Science & the Combechem Project
- Smart Labs
- Semantics & Databases
- Publication@Source
- Conclusions

e-Science

- 'e-Science is about global collaboration in key areas of science, and the next generation of infrastructure that will enable it.'
- 'e-Science will change the dynamic of the way science is undertaken.' 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?
- Publication@Source
  - trace all the way back from publication to the original data – provenance CombeChem
- Who needs provenance? Bush, Blair & Hutton 2004

The CombeChem Project

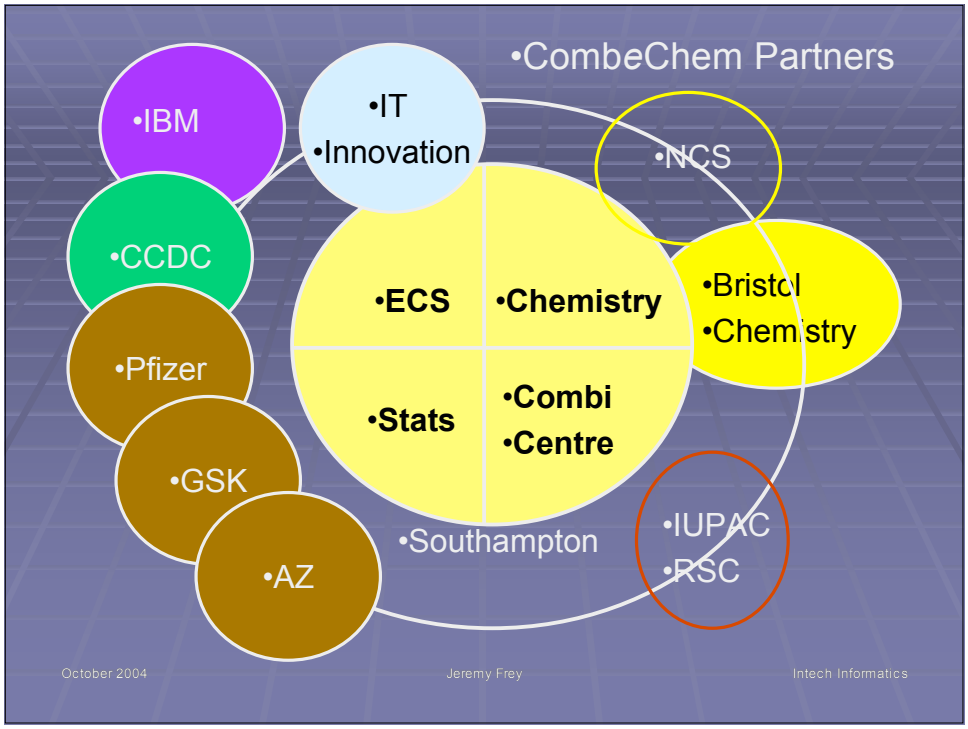
- The exponential world of combinatorial synthesis and high throughput analysis meets the exponentially growing power of computing
  - Automation, Semantics & the Grid"
- End to End linking of data and information
  - In chemistry this can be a very long chain -from a lab to inside a mouse

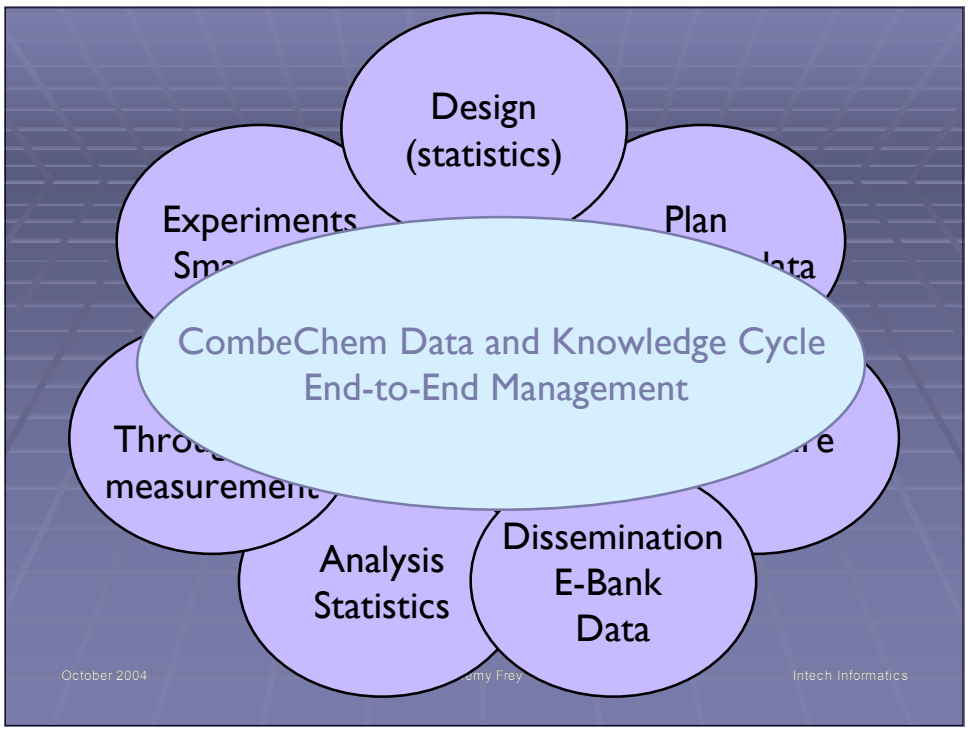
The CombeChem Project

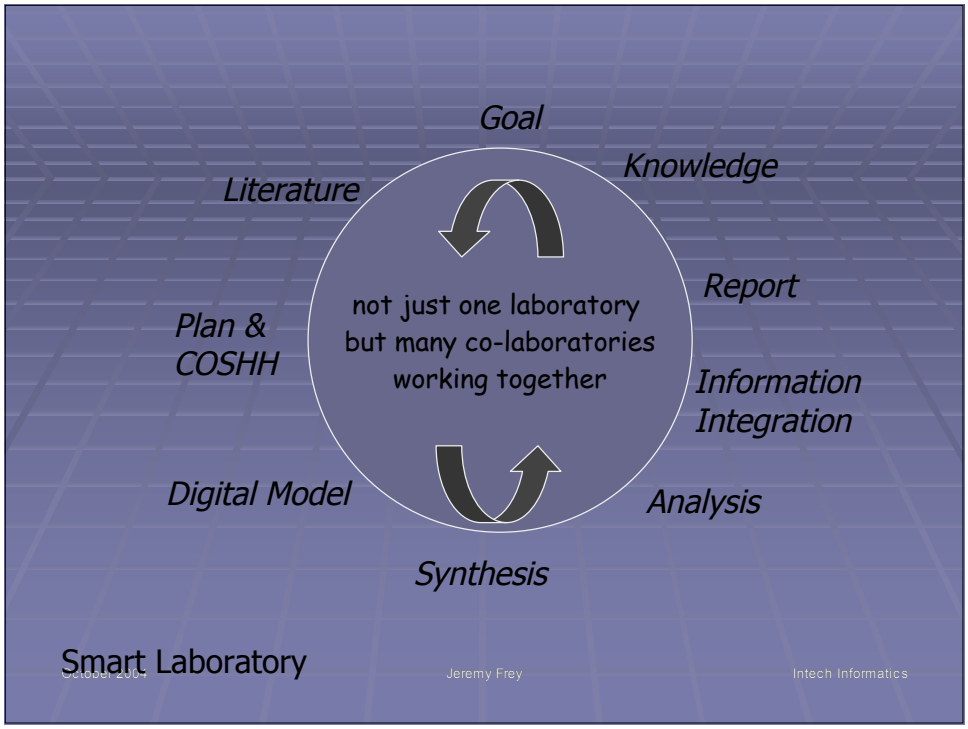
- Collect data with regard to how it could eventually be used
  - Make sure the metadata is of high quality
  - Record properly at source
- The Chemistry Lab
  - People & Machines working together

People

- Chemistry (Southampton & Bristol)
  - Mike Hursthouse, Chris Frampton, Jon Essex, Jeremy Frey, Guy Orpen, Stephan Christensen, Thomas Gelbrich, Sam Peppe, Hongchen Fu, Graham Tizard, Suzanna Ward, Lefteris Danos, Jamie Robinson, Kieron Taylor
- National Crystallography Service (NCS)
  - Simon Coles, Mark Light, Ann Bingham
- Electronics and Computer Science (Southampton)
  - Dave De Roure, Luck Moreau, Mike Luck, Hugo Mills, Graham Smith, Simon Miles, Nicky Harding, Gareth Hughes, monica Schraefel, Terry Payne
- It-Innovation (Southampton)
  - Mike Surridge, Ken Meacham, Steve Taylor, Daren Marvin
- Statistics (Southampton)
  - Alan Welsh, Sue Lewis, Ralph Manson, Dave Woods
- Rutherford Appleton Laboratory –Atlas Datastore

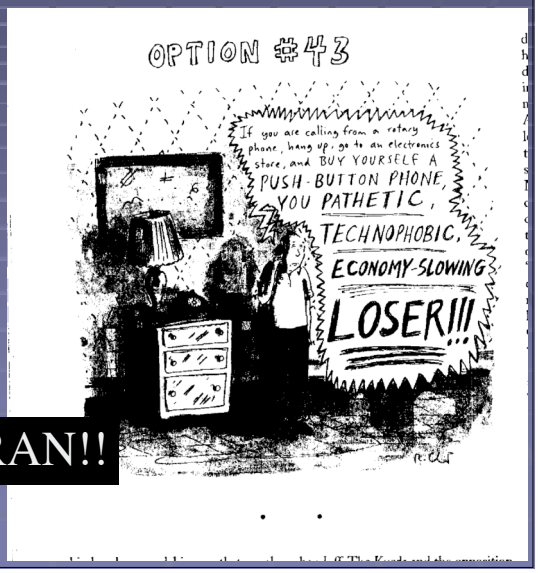

•CombeChem Partners
•IBM
•IT
•Innovation
•NCS
•CCDC
•ECS
•Chemistry
•Bristol
•Chemistry
•Pfizer
•Stats
•Combi
•Centre
•GSK
•AZ
•Southampton
•IUPAC
•RSC
October 2004
Jeremy Frey
Intech Informatics


Design
(statistics)
Experiments
Plan
CombeChem Data and Knowledge Cycle
End-to-End Management
measurement
Analysis
Statistics
Dissemination
E-Bank
Data
October 2004
Intech Informatics



Chemists and programming

- Many Chemists think that they can program



You still use FORTRAN!!

October 2004

e-Workflow

Some Chemists can

What about that! His brain still uses perl scripts

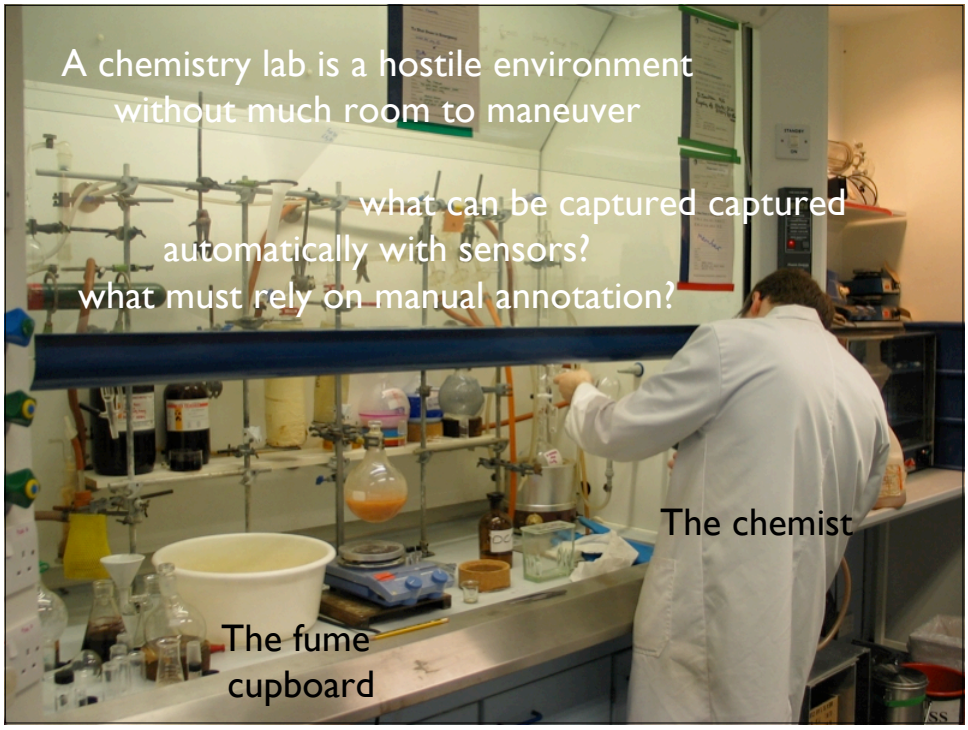
Plans

Small set of fixed plans — NCS

Variable plans, written by chemist (difficult!) — Tea

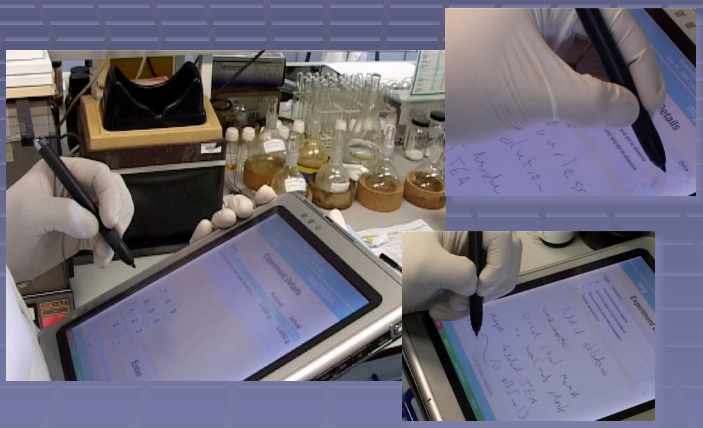
Ad-hoc, implied by process execution — SHG

Continuum of plan types


A chemistry lab is a hostile environment
without much room to maneuver
what can be captured captured
automatically with sensors?
what must rely on manual annotation?
The chemist
The fume
cupboard


Industrial supp
Big block to publication@source: if it's
not digital, it's difficult to share
critical data
entry

Getting real

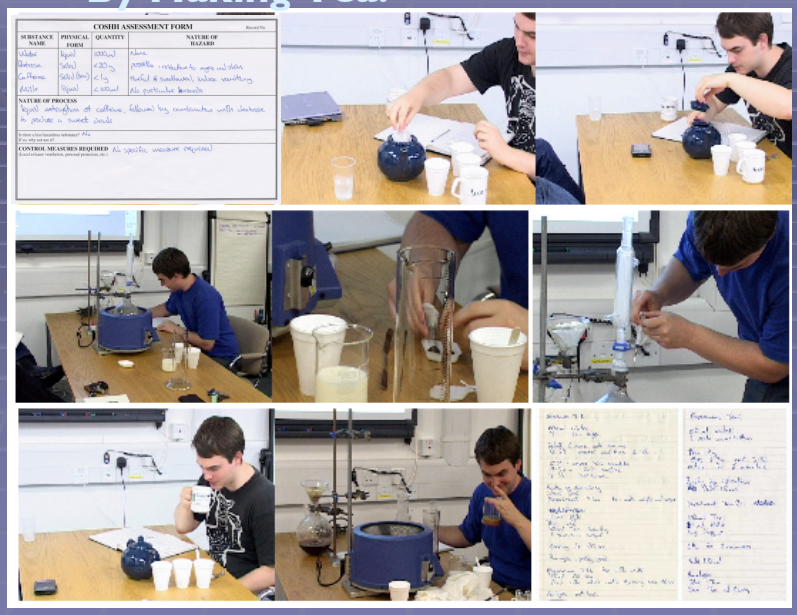


- Functional prototype for in-lab, real use testing

Much more automation in modern chemistry

"That is so cool Dave, you only need a palm pilot"

By Making Tea!



COSHH ASSESSMENT FORM

SUBSTANCE NAME	PHYSICAL FORM	QUANTITY	NATURE OF HAZARD
Water	liquid	1000ml	None
Caffeine	Solid (tablets)	420mg	Harmful if swallowed, irritant
Kettle	liquid	1 litre	No particular hazards

NATURE OF PROCESS: Liquid extraction of caffeine followed by recrystallisation with sucrose to produce a sweet drink

CONTROL MEASURES REQUIRED: No specific measures required

Getting not just the what and how, but the *why*

Data model

Increasing detail

Plan

Intended actions:
guide to chemist,
or [later] workflow

Process record

Measurements
Processes
Annotations

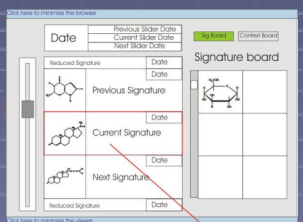
Provenance record

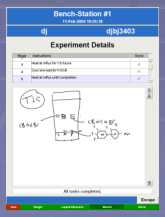
Service invocations
Secure time-stamps
etc…

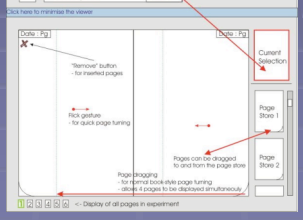
Review over Tea

- We ran through our lo-fi prototypes with chemists by running the tea experiment
  - They knew what was going on and could comment on veracity, features, process

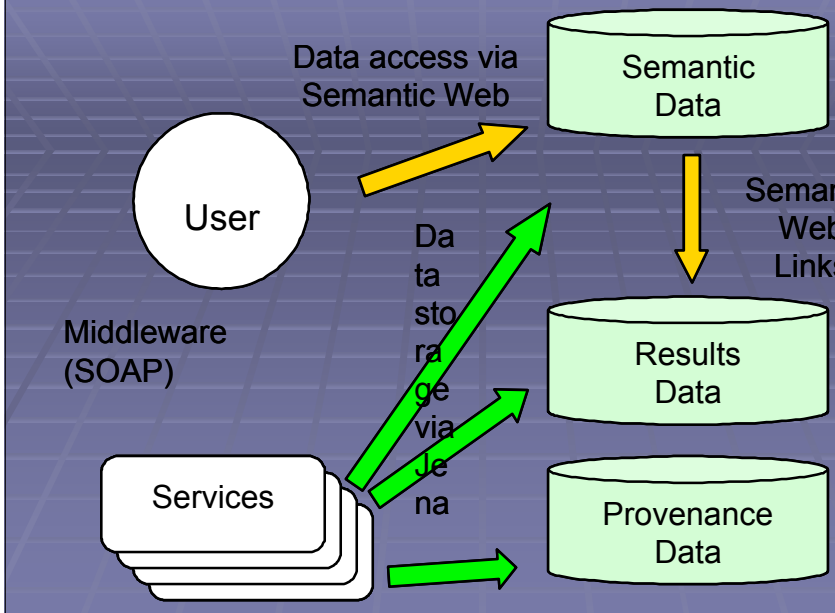
Extensions:





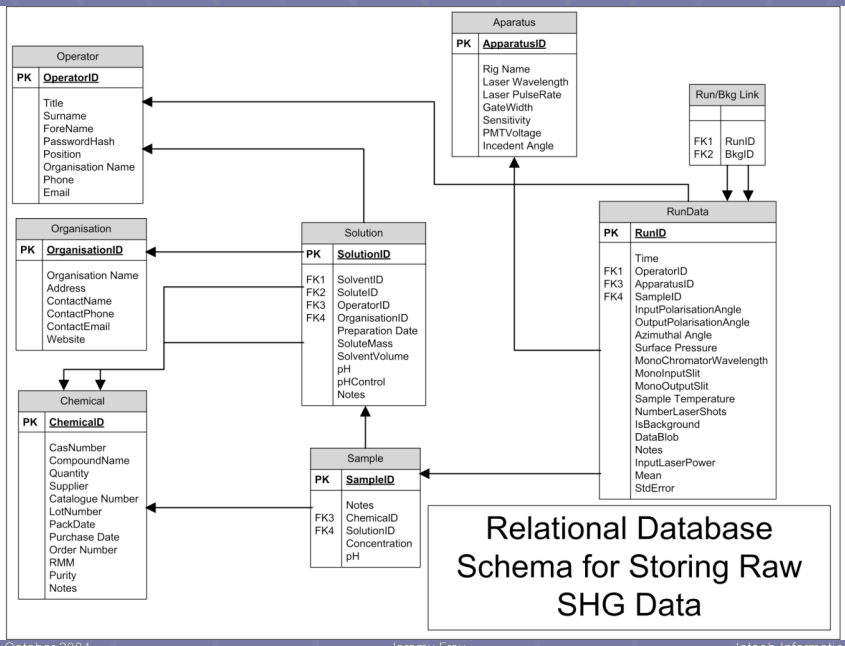


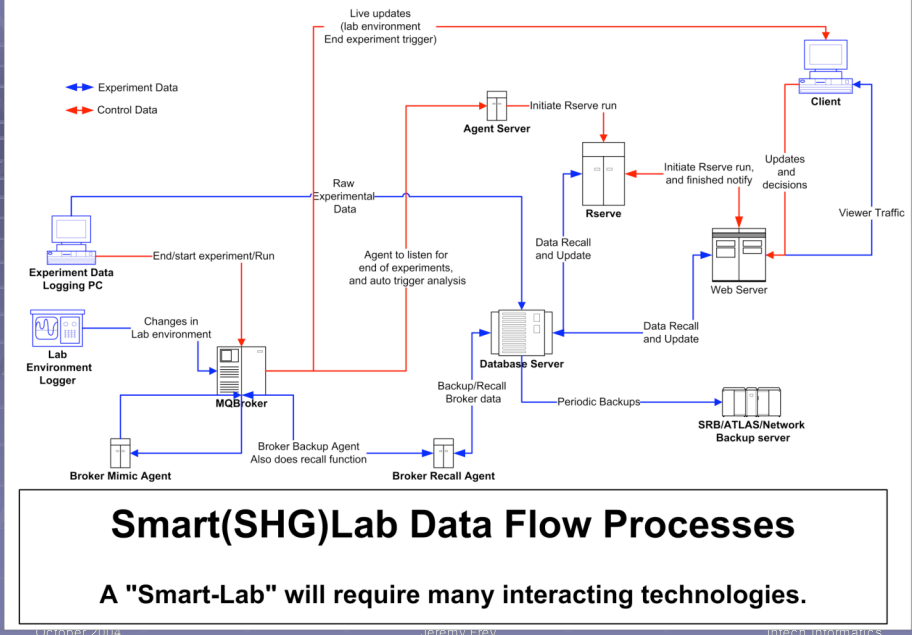
- Ray Cooke
  - Scrolling through lab books
- Will Davies
  - Automating TLC plate capture for record and annotation



Databases

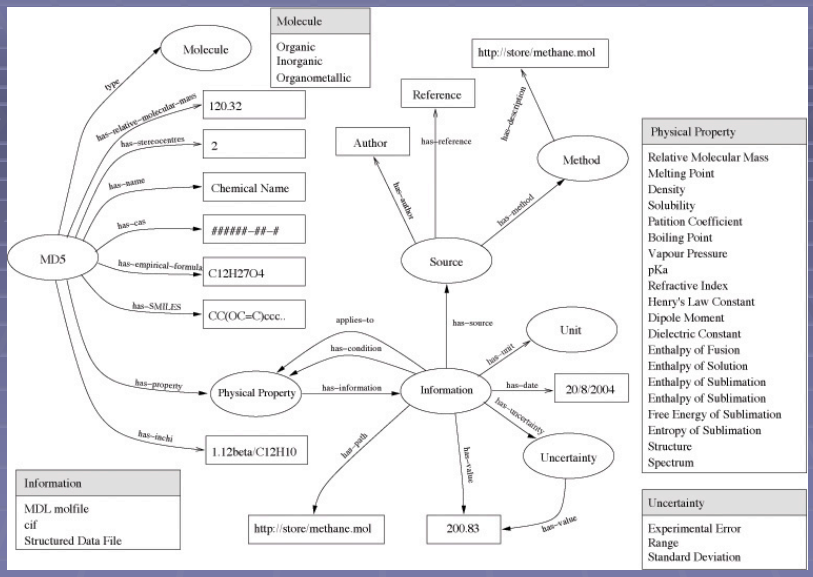
- Database will become the key method of handling all data
- Metadata must be generated at inception and added as data traverses the workflow
- Version control, audit and backup handled at the database level.





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



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>

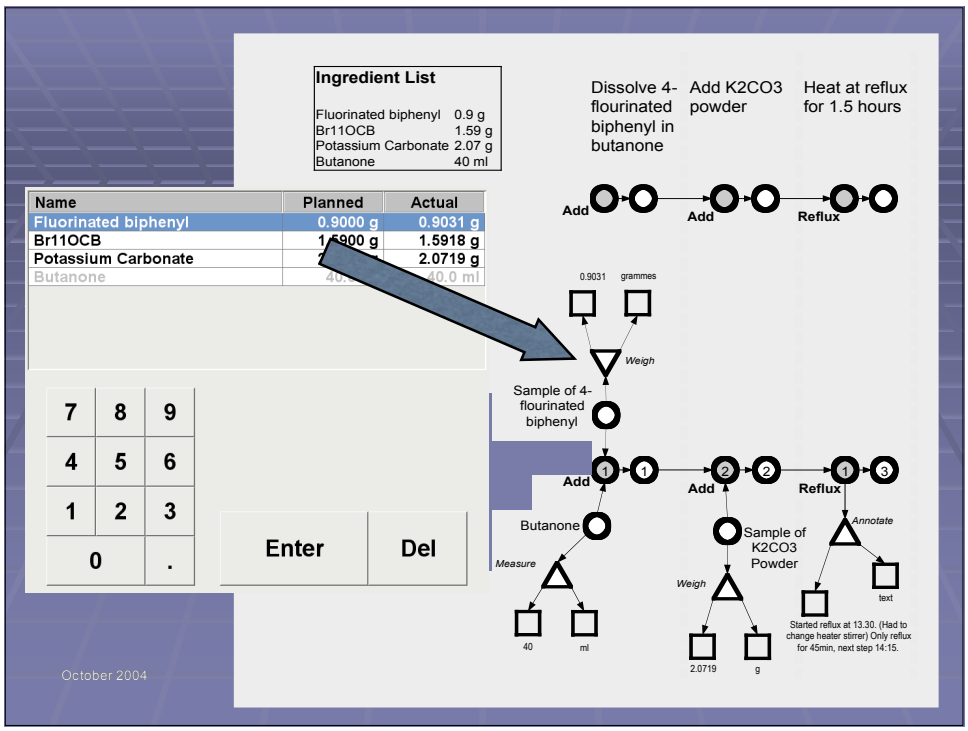
Schema

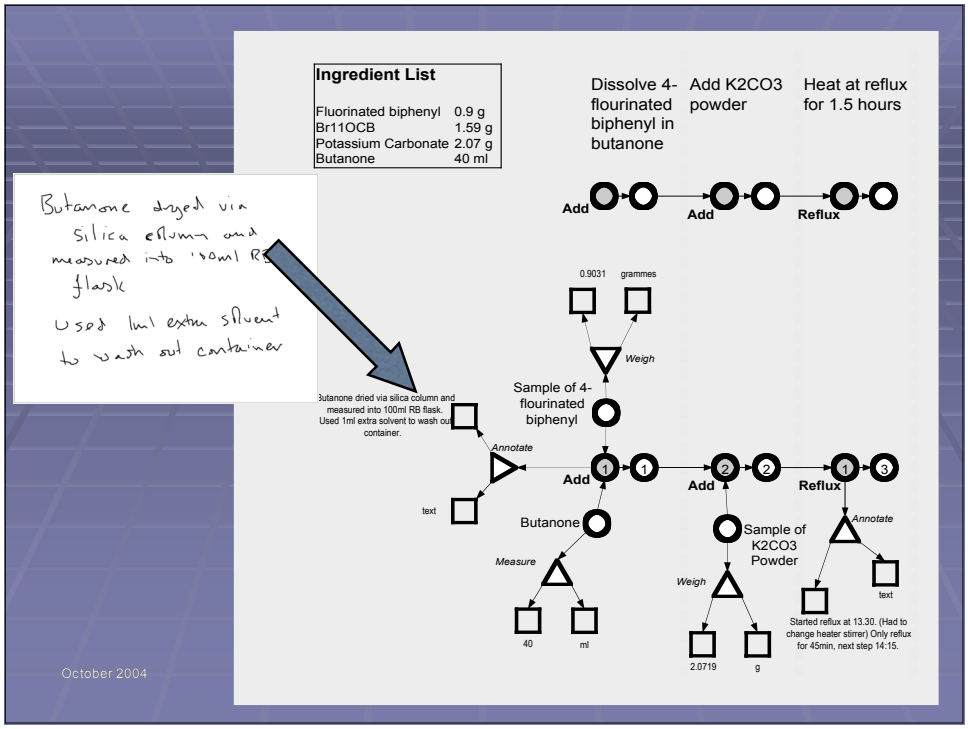
- <rdfs:Class rdf:about="&c;OrganicMolecule">
- <rdfs:label>Organic Molecule</rdfs:label>
- <rdfs:subClassOf rdf:resource="&c;Molecule" />
- </rdfs:Class>

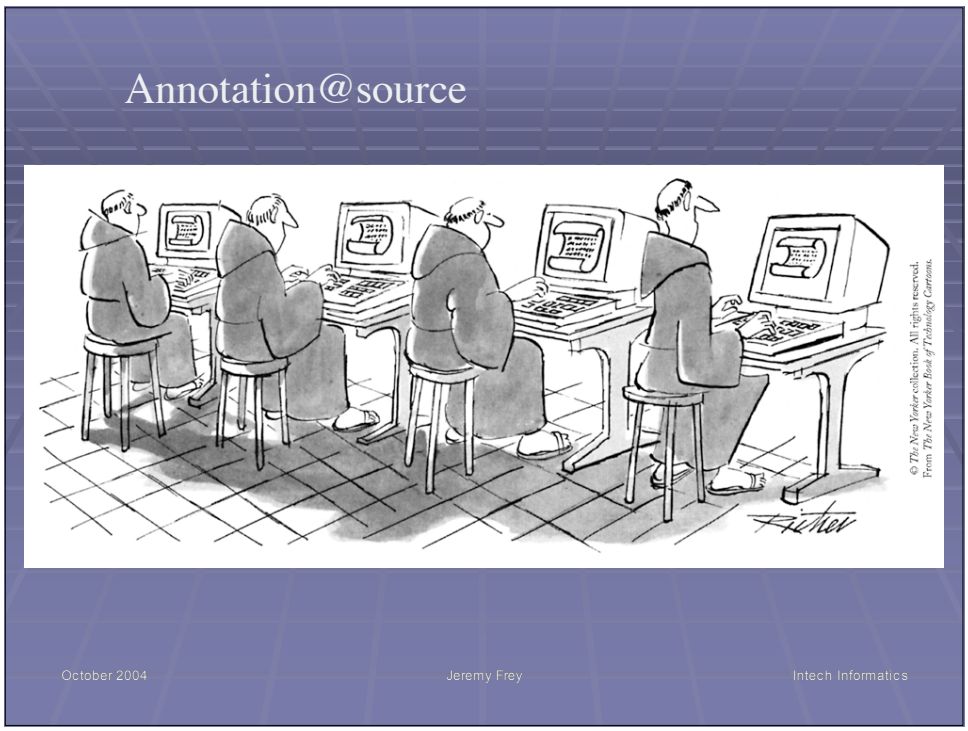
- <rdfs:Class rdf:about="&c;PhysicalProperty">
- <rdfs:label>Property</rdfs:label>
- </rdfs:Class>

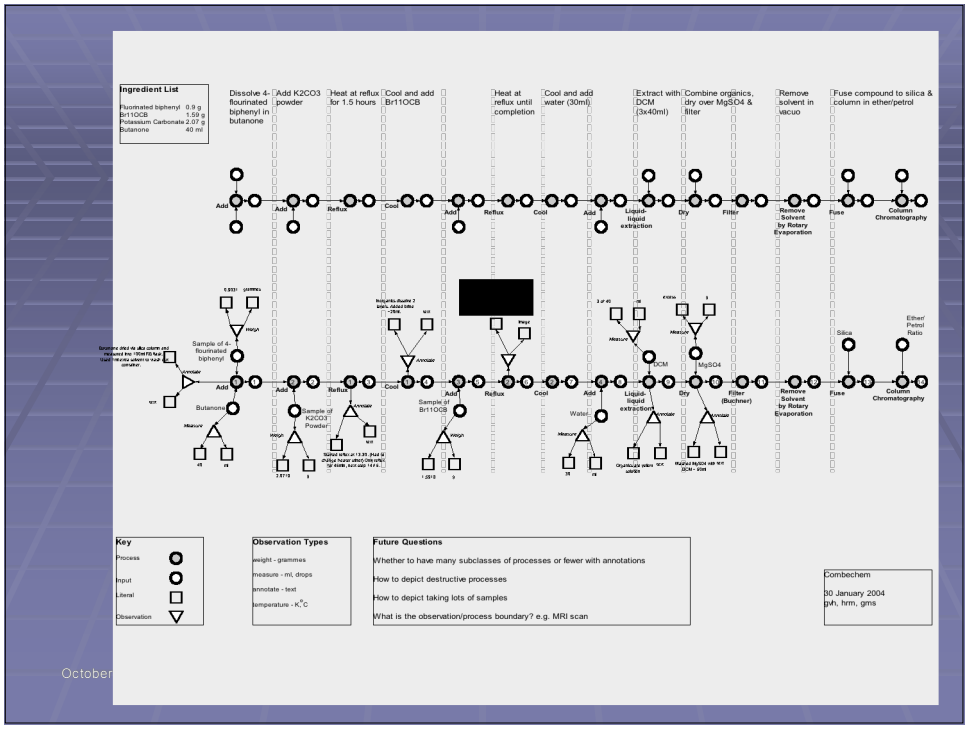
- <rdfs:Class rdf:about="&c;PartitionCoefficient">
- <rdfs:label>Paritition Coefficient</rdfs:label>
- <rdfs:subClassOf rdf:resource="&c;PhysicalProperty" />
- <rdfs:description>Ratio of substance dissolved in octan-1-ol and water
  </rdfs:description>
- </rdfs:Class>

This turns out to be a very flexible approach




Ingredient List
Fluorinated biphenyl 0.9 g
Br11OCB 1.59 g
Potassium Carbonate 2.07 g
Butanone 40 ml
Butanone dryed via silica column and measured into 100ml RB flask. Used 1ml extra solvent to wash out container.
Dissolve 4-flourinated biphenyl in butanone
Add K2CO3 powder
Heat at reflux for 1.5 hours
Add
Add
Reflux
0.9031
grammes
Weigh
Sample of 4-flourinated biphenyl
Annotate
text
Butanone
Measure
40
ml
Sample of K2CO3 Powder
Weigh
2.0719
g
Started reflux at 13.30. (Had to change heater stirrer) Only reflux for 45min, next step 14:15.
October 2004


Annotation@source
© The New Yorker collection. All rights reserved.
From The New Yorker Book of Technology Cartoons.
October 2004
Jeremy Frey
Intech Informatics



Lessons

- That we need two related ontologies
  - Plan – that are going to be done
  - Record – what was done
- Not necessarily the same thing
  - Steps are added/repeated during the experiment
  - Different annotations required for each

Experiments on the Grid

National Crystallography Grid Service

No, the computers are up, We're down

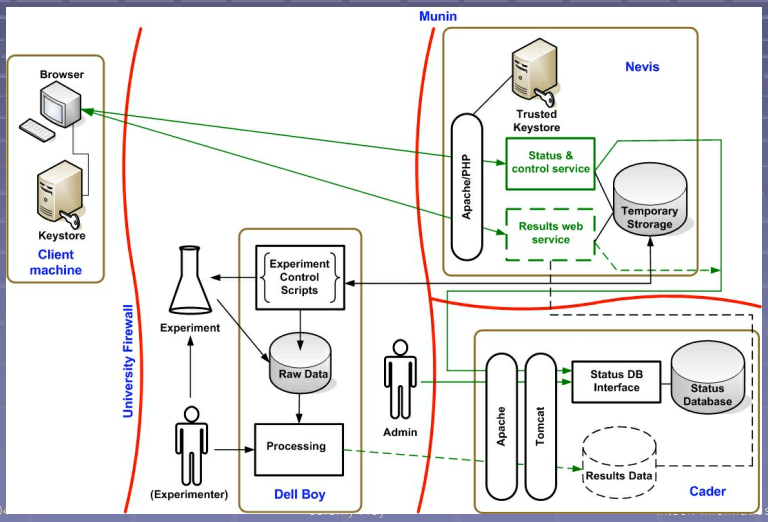
Security and trust for experiments and data

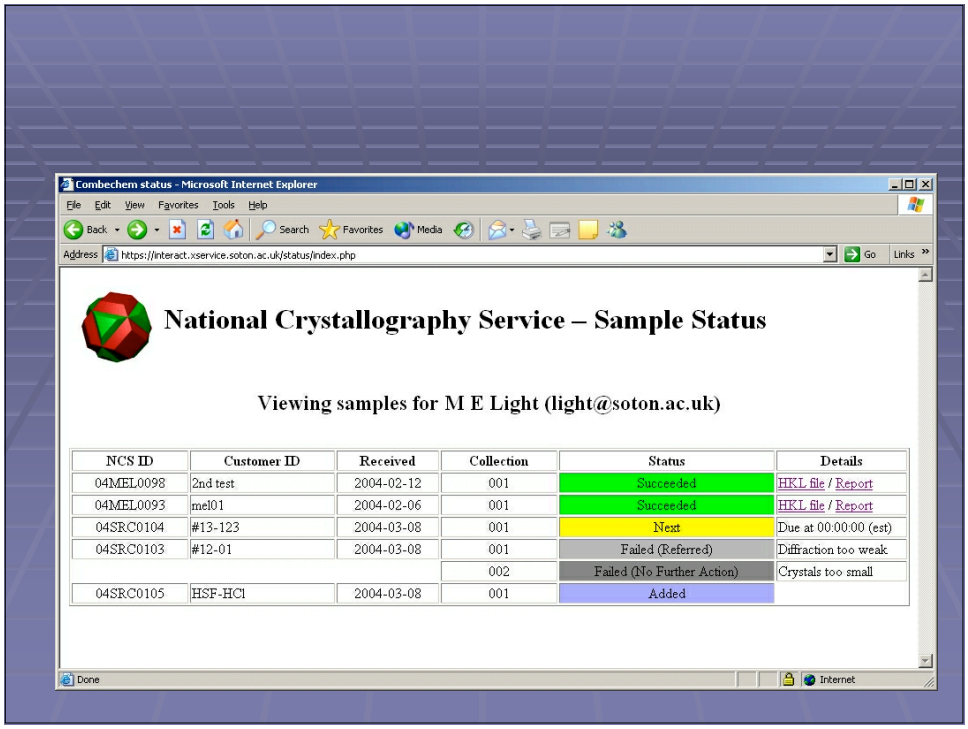
"On the Internet, nobody knows you're a dog."

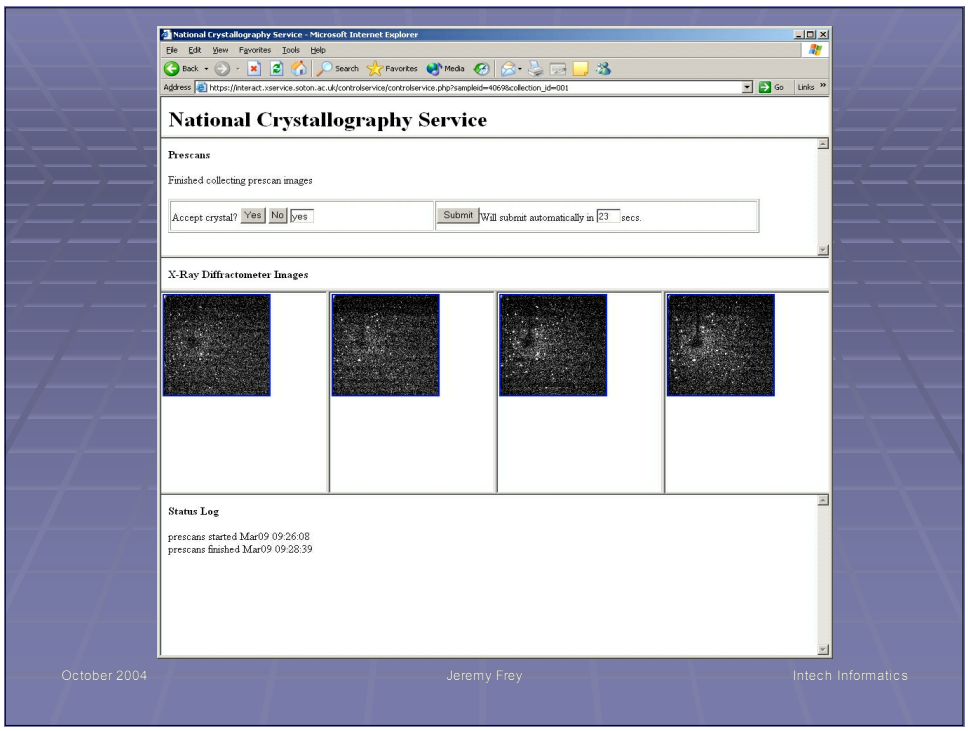
The "Grid Zone"

- Security is fundamental
- Who is using our experiments
- Insulate them from each other and from the rest of our institution
- Process & Role based security
- Use DMZ
- This combination creates a "Grid Zone"

NCS Grid Service Architecture



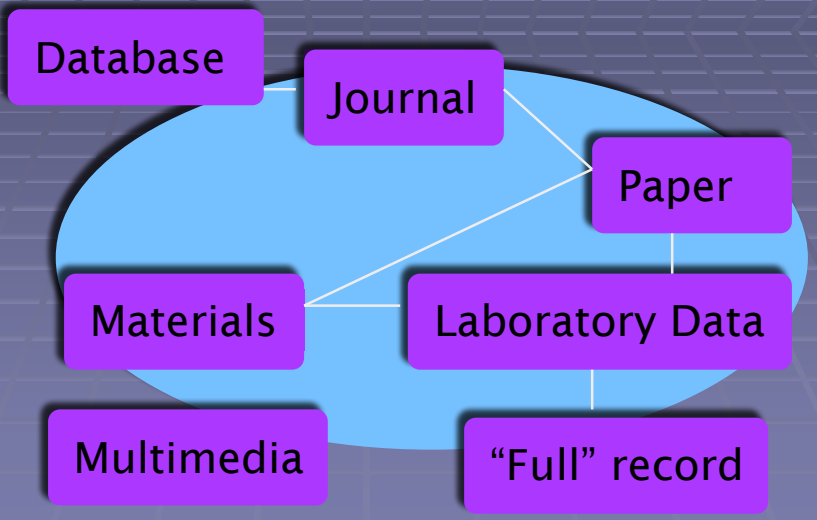

Combechem status - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://interact.xservice.soton.ac.uk/status/index.php
Go Links
National Crystallography Service – Sample Status
Viewing samples for M E Light (light@soton.ac.uk)
NCS ID | Customer ID | 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
Done
Internet

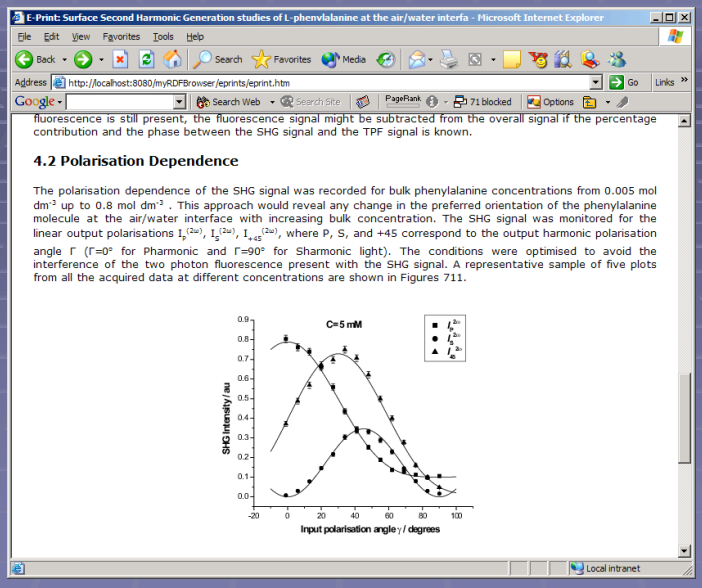

National Crystallography Service - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address https://interact.xservice.soton.ac.uk/controlservice/controlservice.php?sampleid=40699collection_id=001
Go Links
National Crystallography Service
Prescans
Finished collecting prescan images
Accept crystal? Yes No yes
Submit Will submit automatically in 23 secs.
X-Ray Diffractometer Images
Status Log
prescans started Mar09 09:26:08
prescans finished Mar09 09:28:39

Dissemination & Publication

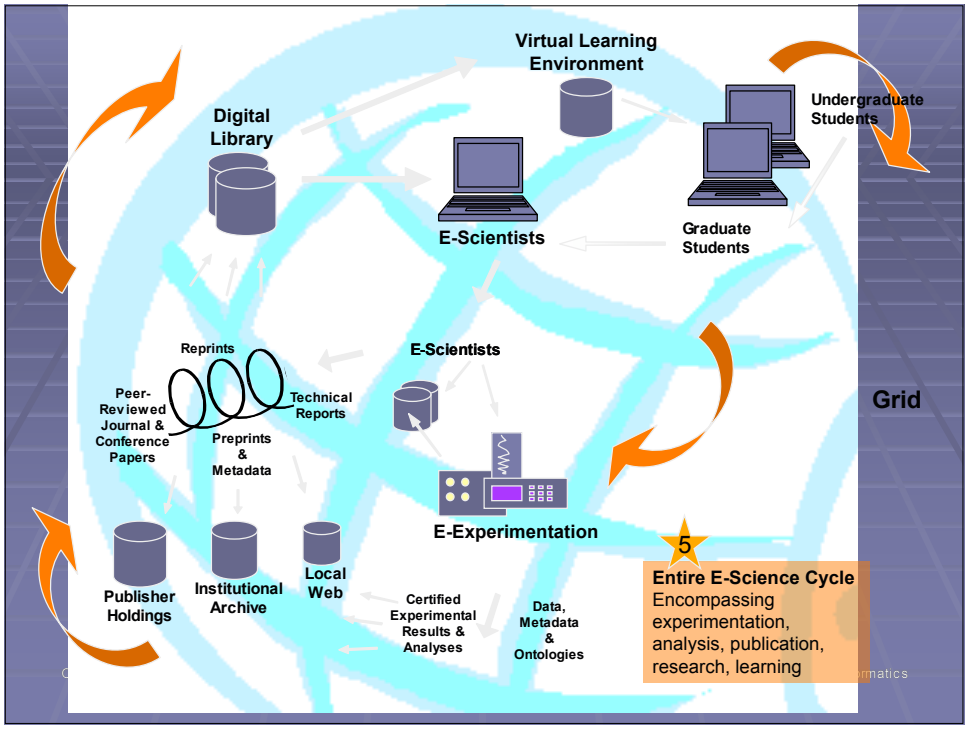
- A different approach is required to provide data to the community
- The grid provides the necessary medium
- What & how do we want to make available

Journals: Publication @ source

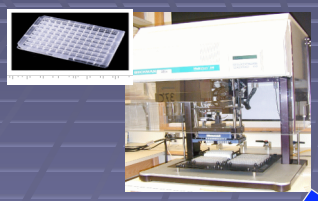


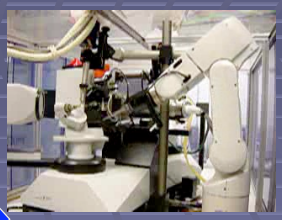


SVG active graphics



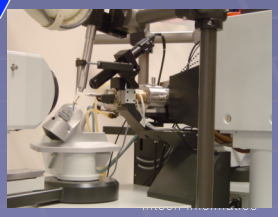
The need for *xtl*-Prints





100's of structures

How do we disseminate?

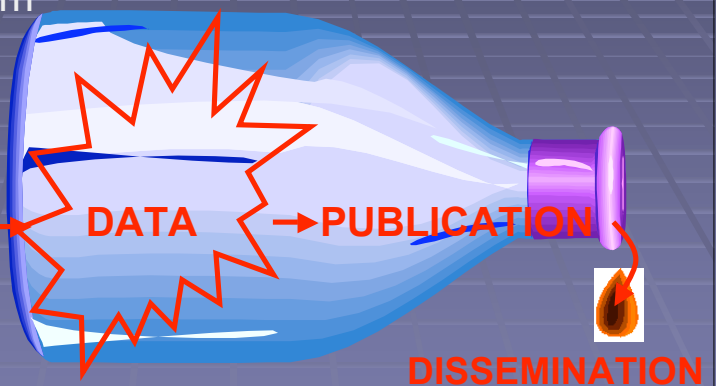


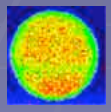
EPSRC Engineering and Physical Sciences Research Council

National Crystallography Service

The need for *xtl*-Prints

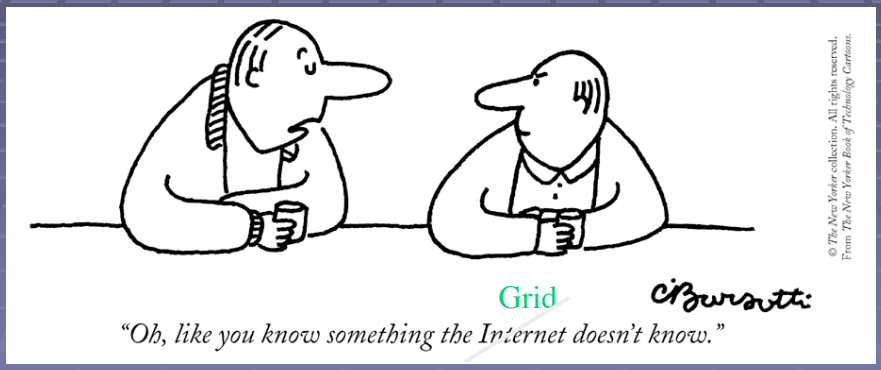
Combechem





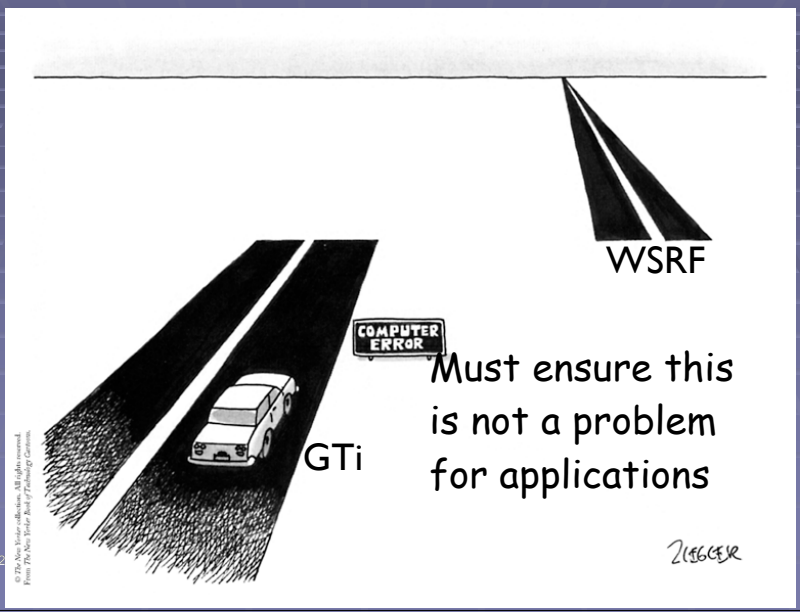
Combichem

Semantic (Pervasive) Grid



e-worries

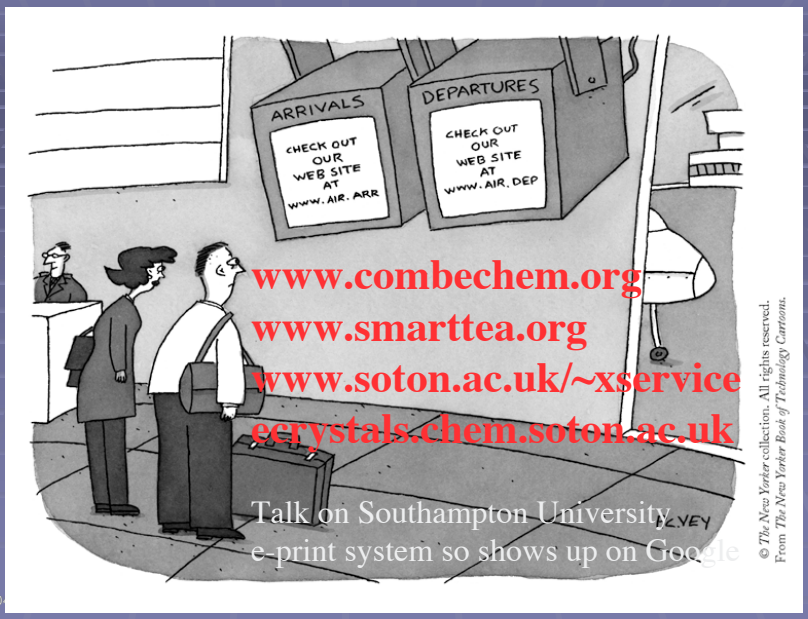
Standards – now not just at the data level but metadata level as well



Making sure other people can re-use your data easily and with confidence

Even when there is a huge amount of it!

Web sites?



Changing the way we work