

The curation of laboratory experimental data as part of the overall data lifecycle

Jeremy G.Frey
**School of Chemistry, University of
Southampton, UK**

21 Nov 2006
DCC Conference, Glasgow



If you do things right at the start then all the following processes are much easier!

Exponentially growing amount of data - the future overwhelms the past



The CombeChem Project

- End to End linking of data and information
 - Publication@Source
- So collect data with regard to how it could eventually be used
 - Make sure the metadata is of high quality
 - Record properly at source in Digital Form
- The Chemistry Lab
 - People & Machines working together



Combechem

Smart Lab

Instruments on the Grid

Statistics

R4L

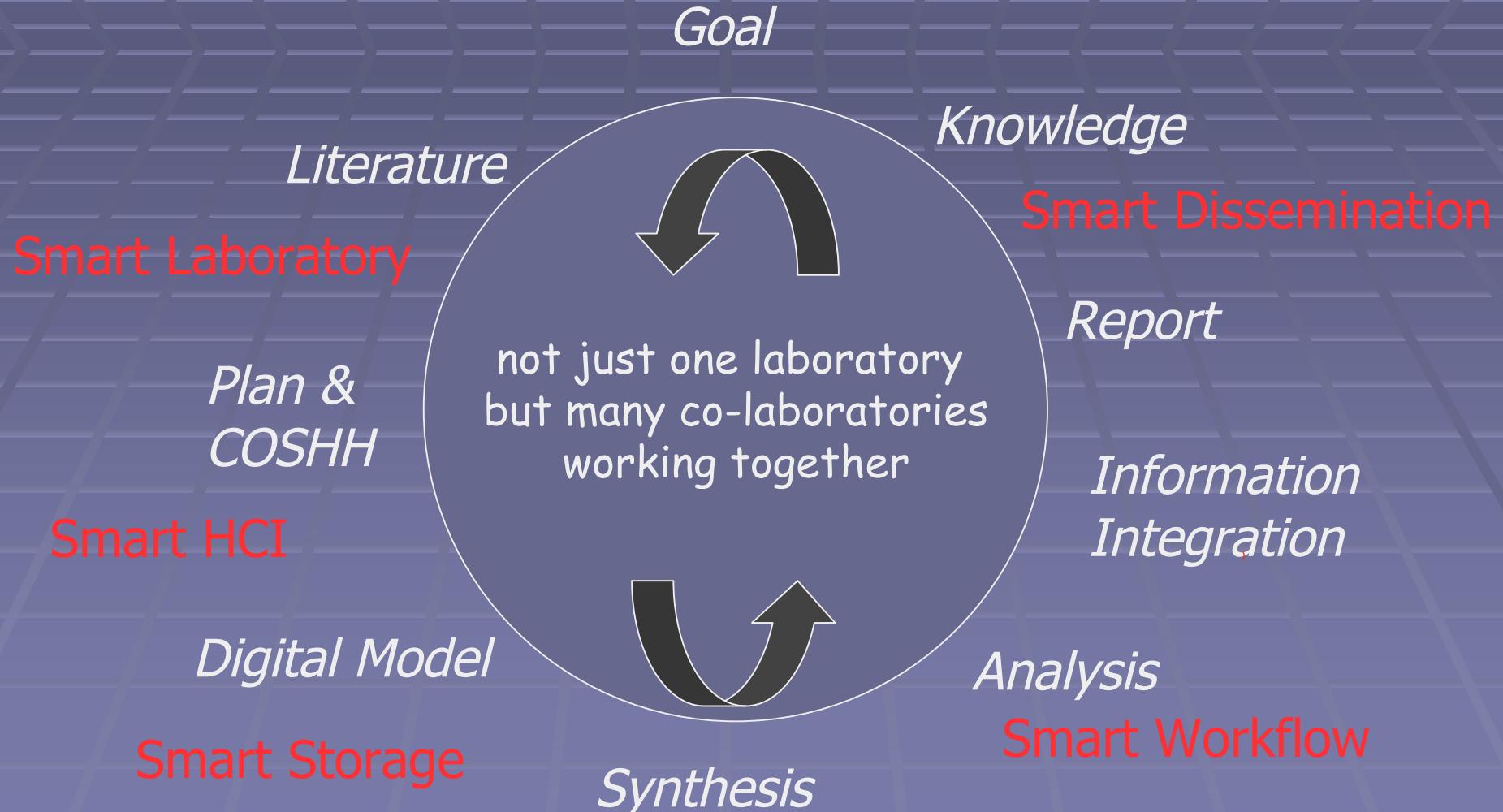
BioSimGrid

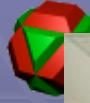
E-Malaria

e-Bank



The concept of Publication @ Source





If only I knew exactly
how she did this
experiments

I wish I had
recorded things at
the start the way I
do now.....

I wish I could get
the numbers from
this graph - the pdf
is not much use.

I know all this supplementary
information could be useful but
will people really remember the
format? Is it worth all the
hassle?

Typical Laboratory



Need to make
the data
available

Need to be
able to find it

But how to
expose it?



© The New Yorker collection. All rights reserved.
From The New Yorker Book of Technology Cartoons.

First, they do an online search



I am sure we collected that information a few years ago...

The details should be in her thesis....

Can you read what he says here....?

Can you find the file of data that were used to make the plot?

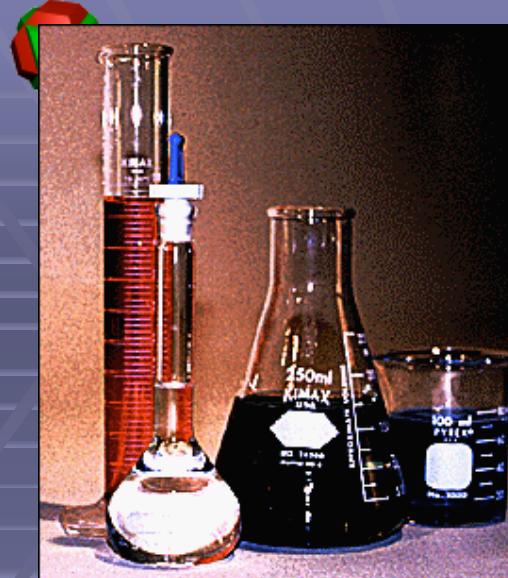
Some of these problems are due to the lack of information recorded at the time. Others are due to loss of information over time.



What are the people up to?

- Capture Data and Context
 - People
 - Process
 - Environment





ChemLab

The Chemistry 3/5 & 6 Laboratories

- ▶ General Information
- ▶ Instruments & Techniques
- ▶ Chemistry 3/5 Experiments
- ▶ Chemistry 6 Experiments

DARTMOUTH COLLEGE

Safety

- [General Rules](#)
- [Safety Equipment](#)
- [Safety Hazards](#)
- [Emergency Procedures](#)

Resources

- [Applets](#)
- [General FAQ](#)
- [Uncertainty](#)

[ChemLab Home](#)

Permanent,
documented
and primary
record of
laboratory
observations

21 Nov 2006

| [Info](#) | [Techniques](#) | [Chem 3/5](#) | [Chem 6](#)

How to Keep a Notebook

One of the most useful skills you will acquire in the laboratory is the proper use of a laboratory notebook. Notebooks, or other formally kept records, are an essential tool in many careers, ranging from that of the research scientist to that of the practicing physician. The effort invested in developing good habits of notebook use will be amply repaid for students who pursue a future in the basic or applied sciences. Experience has indicated that skillful notebook use is developed by most students only through continued special effort--it does not come naturally. Some of the main principles of sound notebook use are outlined below.

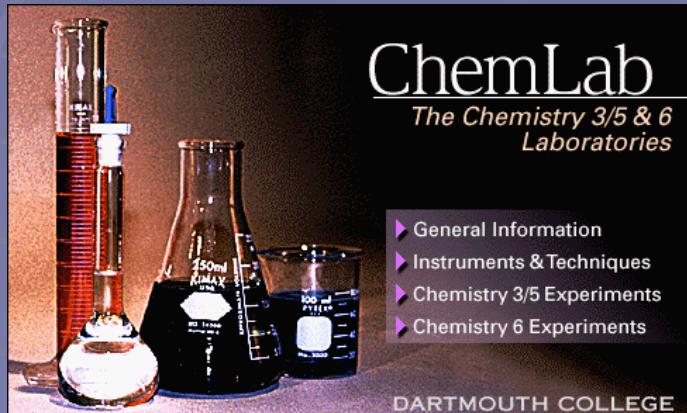
The laboratory notebook is a permanent, documented, and primary record of laboratory observations. Therefore, your notebook will be a bound journal with pages that should be numbered in advance and never torn out. A notebook will be supplied to you before the first laboratory period. Write your name, the name of your TA, and your lab section on the cover of your notebook. All notebook entries must be in ink and clearly dated. No entry is ever erased or obliterated by pen or "white out". Changes are made by drawing a single line through an entry in such a way that it can still be read and placing the new entry nearby. If it is a primary datum that is changed, a brief explanation of the change should be entered (e.g. "balance drifted" or "reading error"). No explanation is necessary if a calculation or discussion is changed; the section to be deleted is simply removed by drawing a neat "x" through it.



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. Prelab problems not turned in at the beginning of lab lecture will 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



COSHH

Leverage off things we already have to do – “We have a cunning plan”



COSHH ASSESSMENT FORM				Record No.
Substance Name	Physical Form	Quantity	Nature of Hazard	
Water	liquid	1000ml	None	
Dextrose	Solid	<20g	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 (Local exhaust ventilation, personal protection, etc.) No specific measure required				

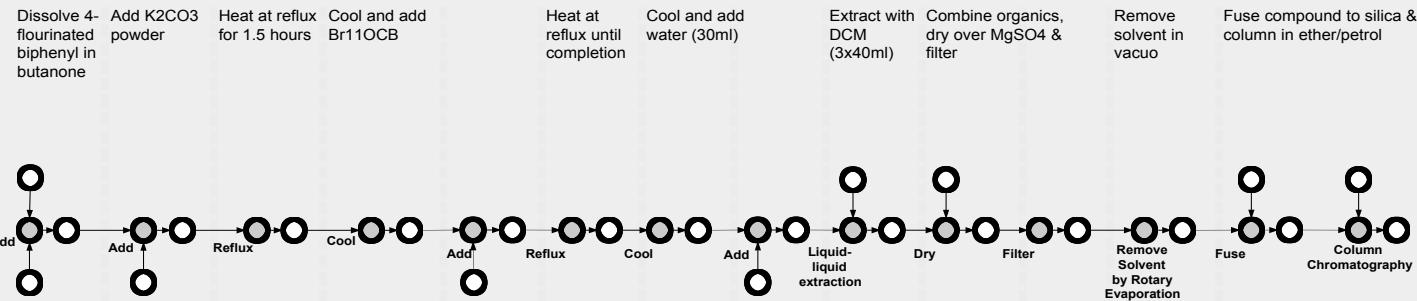


To [
List

TO DO LIST

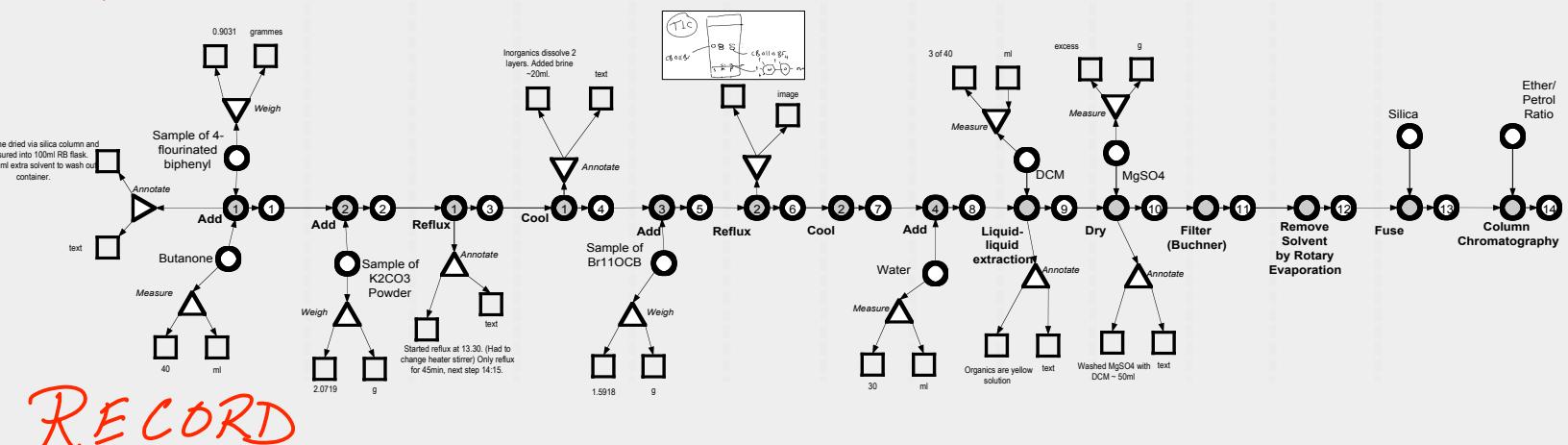
Ingredient List

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

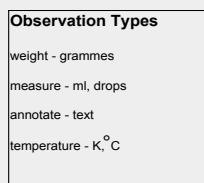
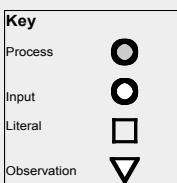


PLAN

Process
Record



RECORD



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



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

Simple Interface

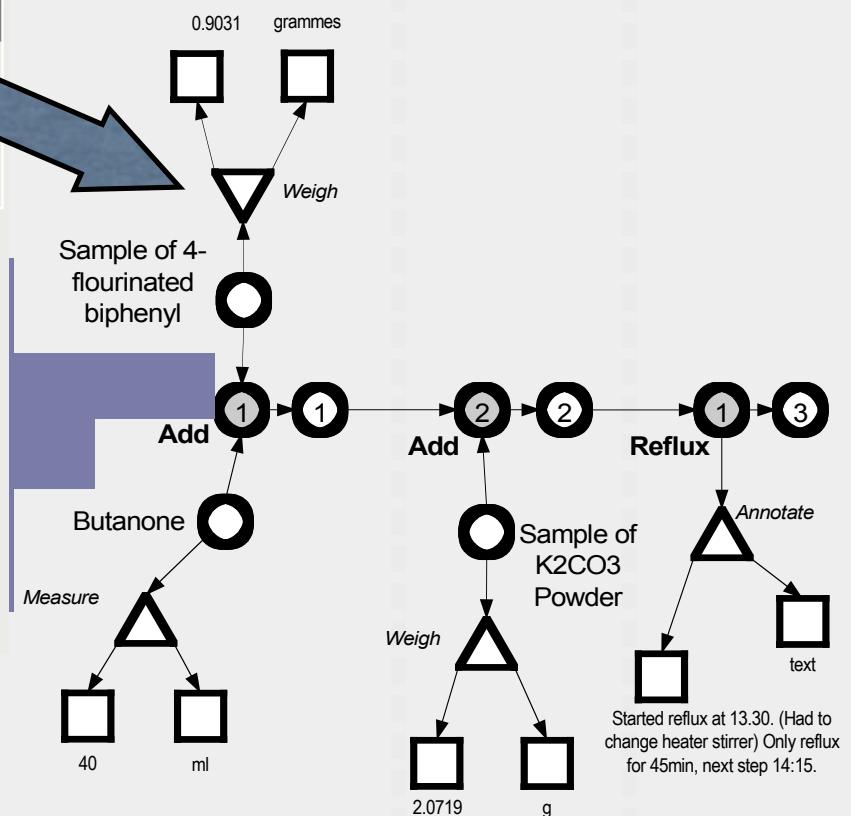
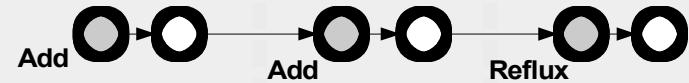
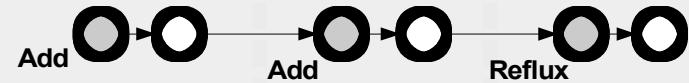
7	8	9
4	5	6
1	2	3
0	.	

Enter

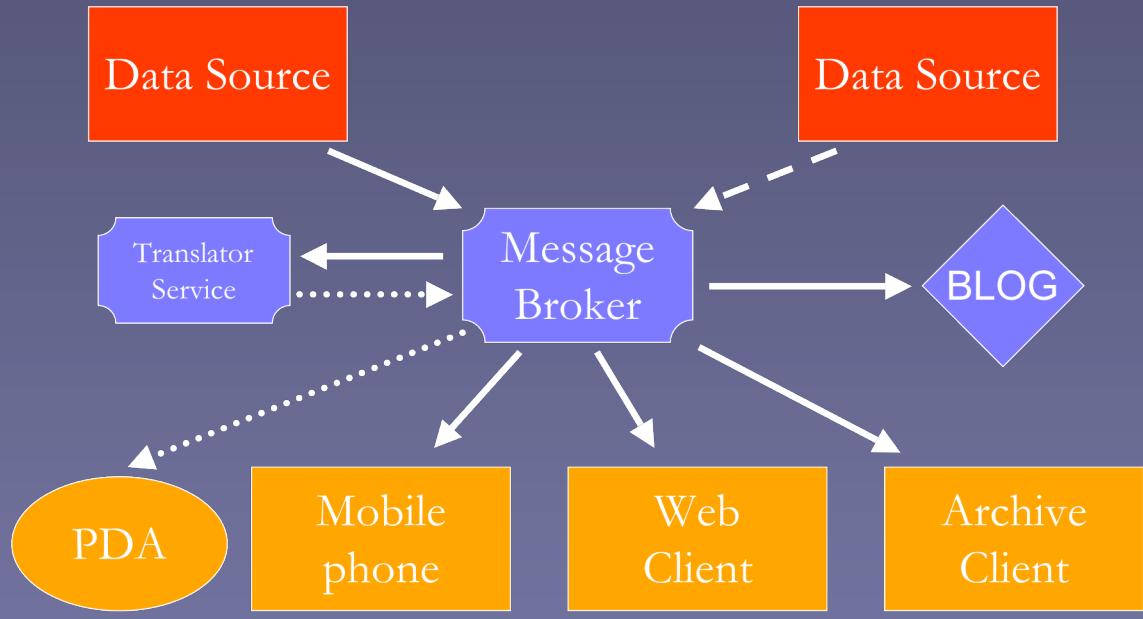
Del

21 Nov 2006

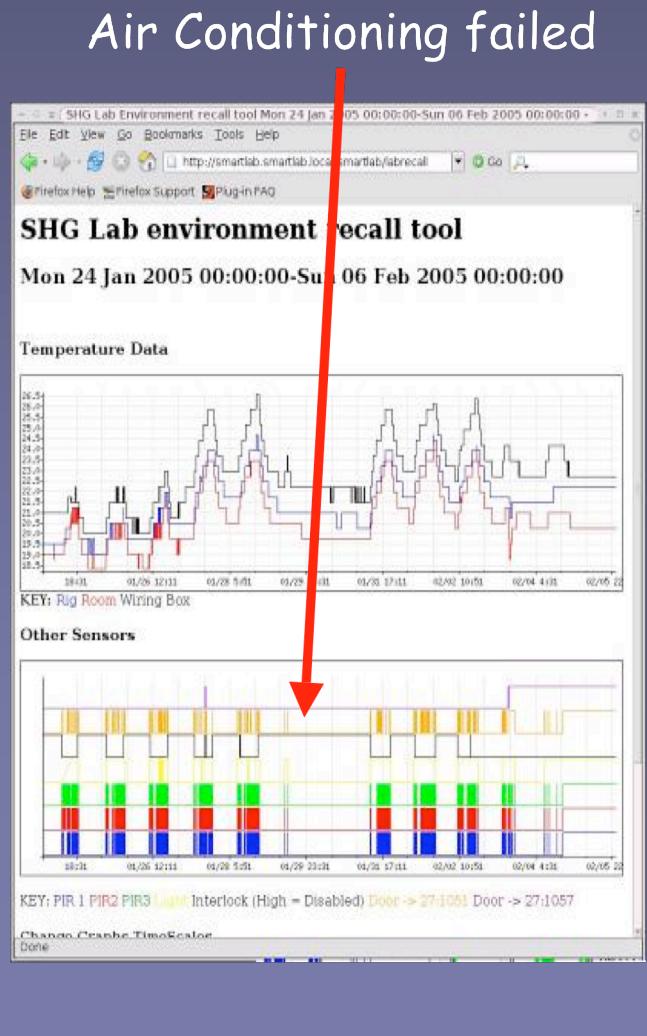
Dissolve 4-fluorinated biphenyl in butanone
Add K₂CO₃ powder
Heat at reflux for 1.5 hours



Pub-Sub systems provide the flexible & extensible approach to distribution of real time laboratory monitoring & archiving



Smart Laboratory Spaces





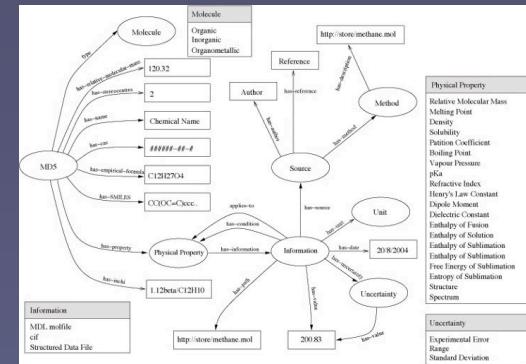
But what
about the
laboratory
environment?



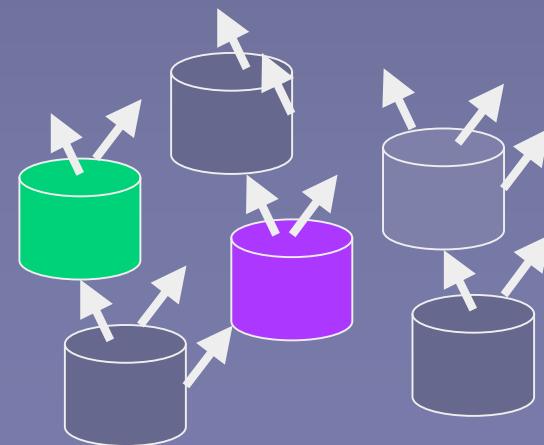
© 2002 The New Yorker collection. All rights reserved.
From *The New Yorker Book of Technology Cartoons*.

Semantic DataGrid

- CombeChem used, tested & strained the Semantic Web for
 - Enhanced (annotated) DataGrid over multiple diverse stores
 - Storage of Provenance Information
 - Some Data Storage
 - Annotated multimedia streams
 - Units & Properties Ontology
 - Multiple Triple Stores



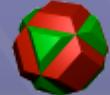
Statistics on Green Triplestore	
Thu Apr 21 11:37:17 2005	
models	2573454
triples	84188993
inferred (FC)	24269915
ground facts	59919078
resources	9987377
literals	7974229
classes	88
properties	49





Laboratory “Blogs”

- Laboratory notebook is a Blog
- Encourage and facilitate collaboration
- Need a data repository behind the Blog
 - R4L
 - E-Bank
- Flexible
 - Service oriented approach being developed
- A VRE



Instrument Blog

[Login](#) [more blogs](#)

MQTT Lego Microscope

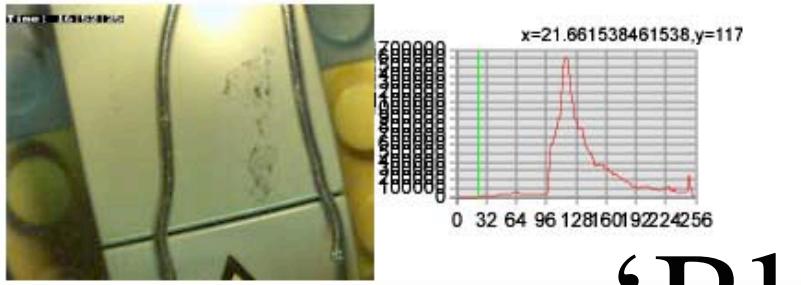
A highly advanced remote control microscope.

Data Collection

15th September 2006 @ 16:52

A data collection was made by Andrew Milsted (ajm3) with sample description: Paper Clip

Data



The image shows a close-up of a paper clip under a microscope. To the right is a graph with a red peak. The x-axis is labeled with values 0, 32, 64, 96, 128, 160, 192, 224, 256. The y-axis has labels 100000, 200000, 300000, 400000, 500000, 600000, 700000. A green vertical line is at x=21.661538461538, and a red line shows a peak at y=117.

ajm3 | [Data Collection](#) | [Comments \(0\)](#)

Archives
[September 2006 \(2\)](#)

Sections
[Data Collection \(2\)](#)

Search

Links

‘Blog-jects’

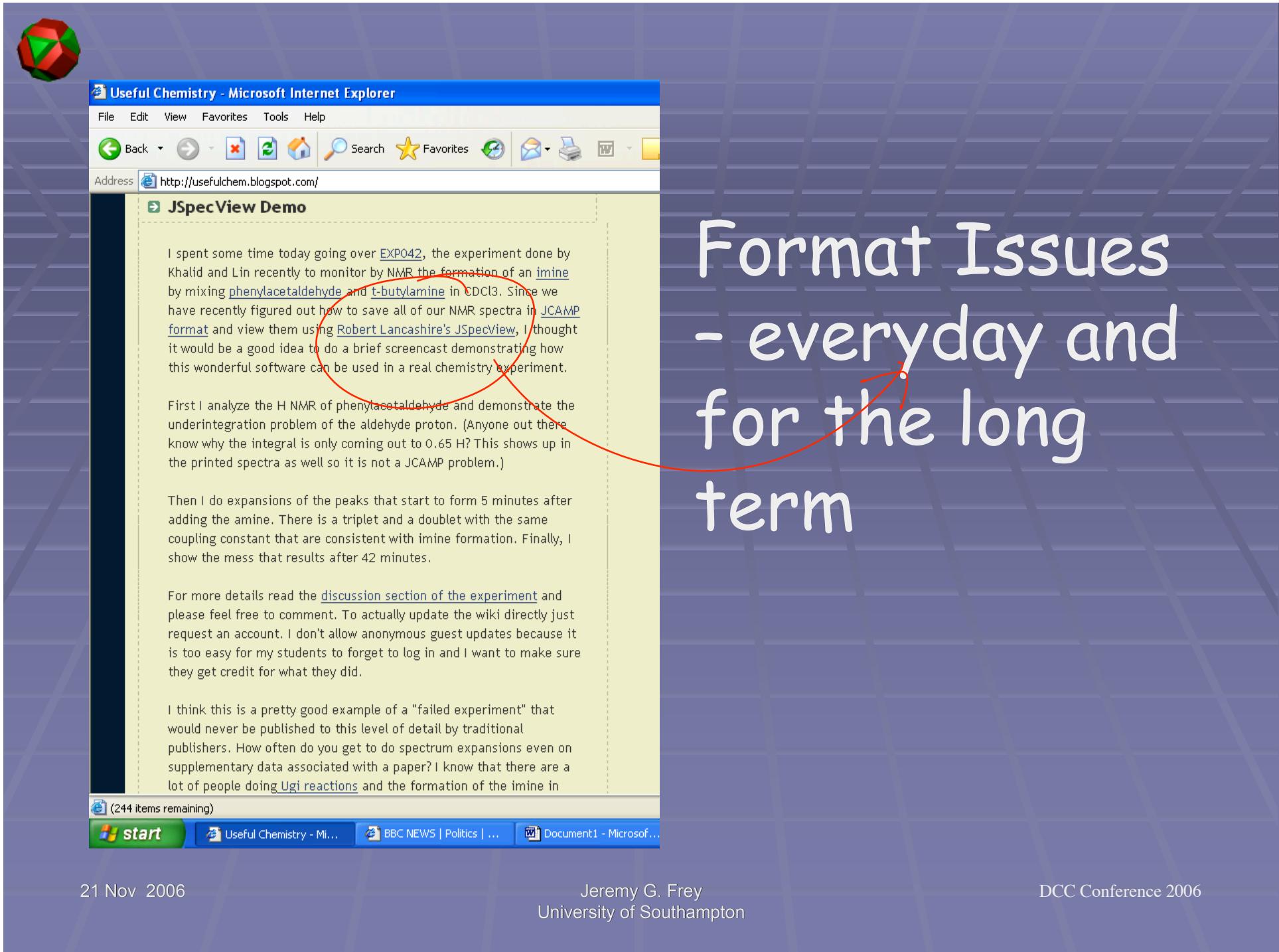
21 Nov 2006

Jeremy G. Frey
University of Southampton

DCC Conference 2006



The 'Scientific Blog' is being tried in an attempt to combine laboratory notebooks and publication



JSpecView Demo

I spent some time today going over EXP042, the experiment done by Khalid and Lin recently to monitor by NMR the formation of an imine by mixing [phenylacetaldehyde](#) and [t-butylamine](#) in [CDCl3](#). Since we have recently figured out how to save all of our NMR spectra in [JCAMP format](#) and view them using [Robert Lancashire's JSpecView](#), I thought it would be a good idea to do a brief screencast demonstrating how this wonderful software can be used in a real chemistry experiment.

First I analyze the H NMR of phenylacetaldehyde and demonstrate the underintegration problem of the aldehyde proton. (Anyone out there know why the integral is only coming out to 0.65 H? This shows up in the printed spectra as well so it is not a JCAMP problem.)

Then I do expansions of the peaks that start to form 5 minutes after adding the amine. There is a triplet and a doublet with the same coupling constant that are consistent with imine formation. Finally, I show the mess that results after 42 minutes.

For more details read the [discussion section of the experiment](#) and please feel free to comment. To actually update the wiki directly just request an account. I don't allow anonymous guest updates because it is too easy for my students to forget to log in and I want to make sure they get credit for what they did.

I think this is a pretty good example of a "failed experiment" that would never be published to this level of detail by traditional publishers. How often do you get to do spectrum expansions even on supplementary data associated with a paper? I know that there are a lot of people doing [Ugi reactions](#) and the formation of the imine in

(244 items remaining)

Format Issues
- everyday and
for the long
term



Useful Chemistry - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: <http://usefulchem.blogspot.com/>

The video ipod compatible podcast is available here.

Here it is on YouTube:

posted by Jean-Claude Bradley @ 4:30 PM 1 comments links to this post

WEDNESDAY, NOVEMBER 15, 2006

(11 items remaining)

start BBC NEWS | Politics | ... Document1 - Microsoft Internet Explorer Microsoft Internet Explorer

Note the use of
"YouTube"

usefulchem » Exp042 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address: <http://usefulchem.wikispaces.com/Exp042>

guest · Join · Help · Sign In ·

UsefulChem Exp042 Protected page discussion history notify me

Actions

Recent Changes Manage Space Search

Navigation

Home

Objective
To study formation of an imine from phenylacetaldehyde and t-butylamine

Procedure
CDCl₃ solutions of phenylacetaldehyde (240 μ L in 2 mL, 1 M) and t-butylamine (146 μ L in 2 mL, 0.7 M) are prepared in separate 1 dram vials. One mL of each solution are used to obtain initial H and C NMR spectra. The remaining 1 mL of each solution are mixed in a 1 dram vial and shaken vigorously. The resulting solution is transferred to an NMR tube and the reaction is monitored by H and C NMR.

Results

t-butylamine solution (BA)
H-NMR 1.27 (br s, NH₂), 1.15 (s, CH₃)

phenylacetaldehyde solution (PA)
H-NMR 9.71 (t, J=2.3 Hz, CHO, **0.65H**), 7.30 (m, 3H), 7.19 (d, J=7.5 Hz, 2H), 3.65 (d, J=2.3 Hz, CH₂, 2H)

42A 5 min
H-NMR new peaks 9.76 (t), 7.79 (d), 7.73 (d), 7.65 (t, J=5.3 Hz), 3.57 (d, J=5.3 Hz), 1.19 (s) and many other peaks in the 2-6 ppm region
From PA and BA: 9.73 (t), 7.1-7.4 (m, more than just PA), 3.67 (d), 1.18 (s)

42B 11 min
42C 28 min

Ads by Google
NMR NMR Analytical Laboratories NMR Laboratory Network www.intertek-cb.com
Nmr Products The UK Specialists in Stable Isotopes www.okgas.com
Mass:Spec MALDI Chips On-target processing makes the most of your MALDI samples www.qiagen.com/protein
NMR Spectroscopy Info On-Line and Analytical NMR Services Consulting and Analysis Process-NMR.Com
NIR Calibration On Demand Continually built from your

An experiment that failed... Publishable?
Useful?



CoAKTing

Memetic

Record the 'Scientific Conversation' -
this part of the record often exists
only in the 'grey literature'

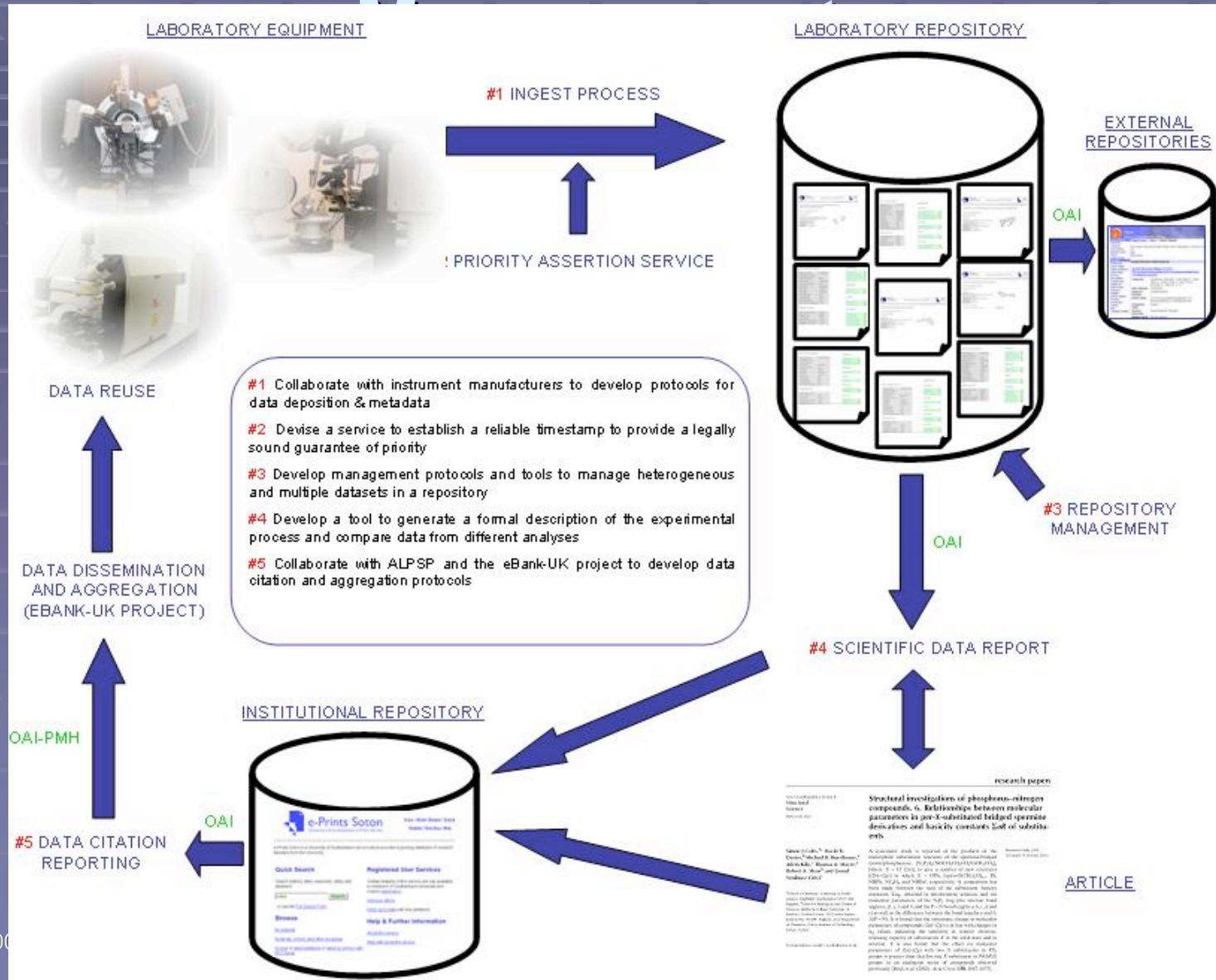
21 Nov 2006

Jeremy G. Frey
University of Southampton

DCC Conference 2006

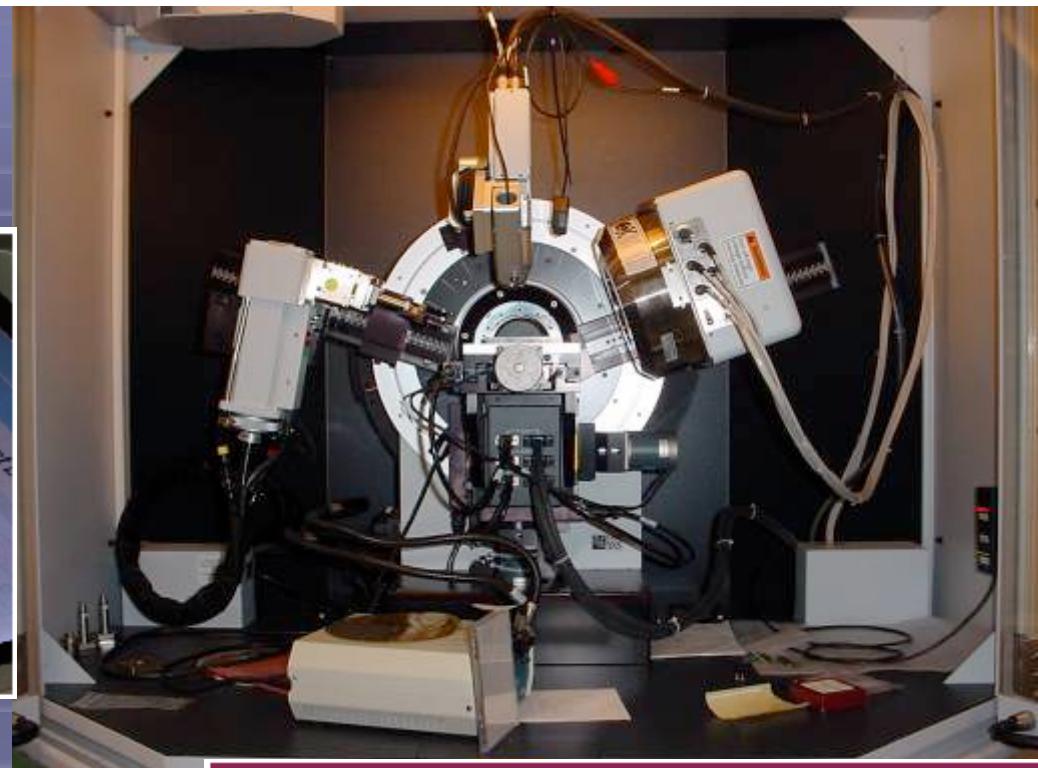
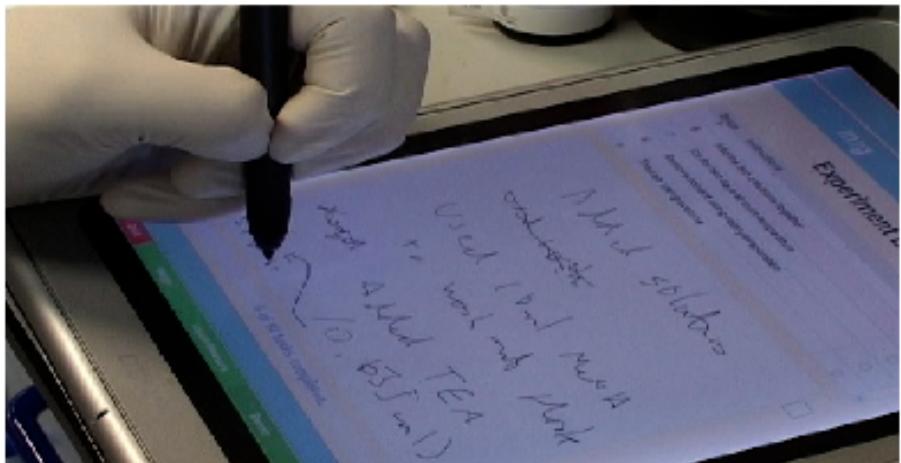


Laboratory IRs and Information





Repositories



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.

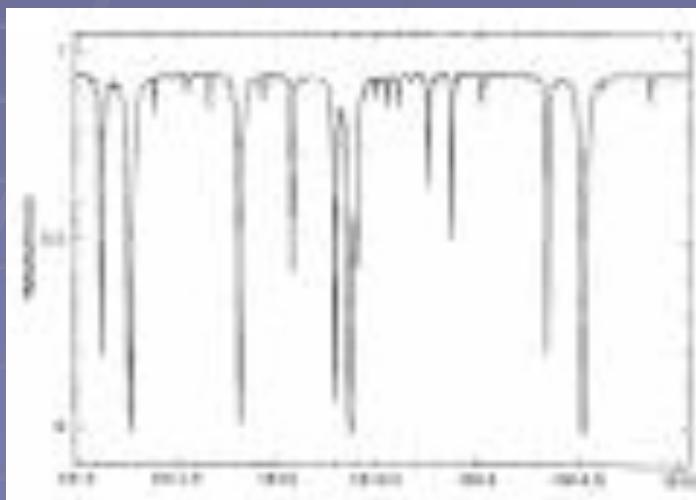
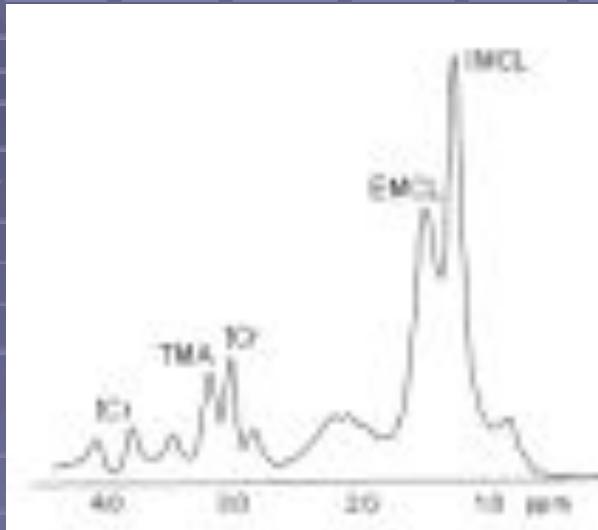


Validation

- Increasing the value of data
- How to bring all the necessary information together to enable appropriate validation
- Increasingly difficult & expensive to achieve
- Need provenance and context
- Essential step otherwise just a collection of items



Why? Publishing Data and Information Loss



Preparation of $\text{CPh}_3[\text{C}_6\text{F}_5]_3\text{BCNPBB}$ (2)

$\text{Me}_3\text{SiNCB}(\text{C}_6\text{F}_5)_3$ (0.51 g, 0.84 mmol) and Ph_3CCl (0.23 g, 0.84 mmol) were stirred in 20 mL of dichloromethane for 0.5 h to give a yellow solution. After removal of volatiles *in vacuo*, the residue was washed with pentane (30 mL), PBB (0.81 g, 0.84 mmol) and dichloromethane (30 mL) were added, and the mixture was stirred for 2 h. The solvent was then removed. The product was washed again with 30 mL of pentane and dried *in vacuo* to yield a yellow-orange powder (yield 1.01 g, 5.8 mmol, 69%). Attempts to crystallise the product from dichloromethane were not successful. IR (nujol): 2284 cm^{-1} ($\nu_{\text{C=C}}$). $^1\text{H NMR}$ (CD_2Cl_2 , 20 °C, 300.13 MHz): δ 8.56 (t, 3, $J = 8.0\text{ Hz}$ *p*-Ph), 7.90 (t, 6 H, $J = 7.5\text{ Hz}$, *m*-Ph), 7.70 (d, 6 H, $J = 7.2\text{ Hz}$, *o*-Ph). $^{13}\text{C NMR}$ (CD_2Cl_2 , 20 °C, 75.48 MHz): δ 211.0 (CPh_3), 144.1 (*p*-C), 143.0 (*m*-C), 140.1 (*ipso*-C), 130.9 (*o*-C). $^{11}\text{B NMR}$ (CD_2Cl_2 , 20 °C, 96.3 MHz): δ -4.35 (s, br, 1 B, $\text{N}(\text{C}_6\text{F}_5)_3$). $^{19}\text{F NMR}$ (CD_2Cl_2 , 20 °C, 282.4 MHz): δ -18.72 (br, s, 1 F), -120.22 (br, s, 1 F), -121.99 (br, s, 1 F), -122.50 (s, 1 F), -132.20 (s, 1 F), -133.94 (br, 6 F, *o*-F on $\text{B}(\text{C}_6\text{F}_5)_3$), overlapping signals (-134.15, -134.39, -134.95, -135.27, -135.64), 136.89 (br, 1 F), -137.81 (br, 3 F), -138.79 (d, 1 F), -144.73 (t, 1 F), -149.78 (t, 1 F), -151.11 (t, 1 F), -154.65 (t, 1 F), -154.93 (t, 1 F), -155.32 (t, 1 F), -156.86 (t, 1 F), -157.24 (t, 1 F), -157.55 (t, 1 F), -158.29 (m, 1 F), -158.90 (t, 1 F), -159.57 (t, 3 F, $J = 20\text{ Hz}$, *p*-F on $\text{B}(\text{C}_6\text{F}_5)_3$), -159.98 (t, 1 F), -161.44 (br, 2 F), -164.0 to -164.4 (overlapping signals, 3 F), -165.33 (br, 2 F), -166.12 (t, 3 F, $J = 20\text{ Hz}$, *m*-F on $\text{B}(\text{C}_6\text{F}_5)_3$). Anal. Calcd for $\text{C}_{24}\text{H}_{15}\text{B}_2\text{F}_{15}\text{N}$:

Paper organized using RDF

Link to simulation, full simulation data archived in BioSimGrid

SVG “active” graphics

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

R4L

E-Print: Surface Second Harmonic Generation studies of L-phenylalanine at the air/water interfa - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media PageRank 71 blocked Options

Address http://localhost:8080/myRDFBrowser/eprints/eprint.htm

Google Search Web Search Site PageRank Options

fluorescence is still present, the fluorescence signal might be subtracted from the overall signal if the percentage contribution and the phase between the SHG signal and the TPF signal is known.

4.2 Polarisation Dependence

The polarisation dependence of the SHG signal was recorded for bulk phenylalanine concentrations from 0.005 mol dm⁻³ up to 0.8 mol dm⁻³. This approach would reveal any change in the preferred orientation of the phenylalanine molecule at the air/water interface with increasing bulk concentration. The SHG signal was monitored for the linear output polarisations $I_p^{(2\omega)}$, $I_s^{(2\omega)}$, $I_{+45}^{(2\omega)}$, where P, S, and +45 correspond to the output harmonic polarisation angles $\Gamma=0^\circ$ for Ramanic and $\Gamma=90^\circ$ for Sharmonic light. The conditions were optimised to avoid the photon fluorescence present with the SHG signal. A representative sample of five plots a at different concentrations are shown in Figures 711.

SHG Intensity / au

Input polarisation angle γ / degrees

C = 5 mM

Legend: $I_p^{(2\omega)}$ (squares), $I_s^{(2\omega)}$ (circles), $I_{+45}^{(2\omega)}$ (triangles)

Input polarisation angle γ / degrees	$I_p^{(2\omega)}$ (au)	$I_s^{(2\omega)}$ (au)	$I_{+45}^{(2\omega)}$ (au)
0	0.25	0.05	0.80
10	0.45	0.05	0.75
20	0.70	0.15	0.65
30	0.55	0.25	0.55
40	0.35	0.35	0.45
50	0.20	0.25	0.35
60	0.10	0.15	0.25
70	0.05	0.05	0.15
80	0.02	0.02	0.08
90	0.01	0.01	0.05

Local intranet

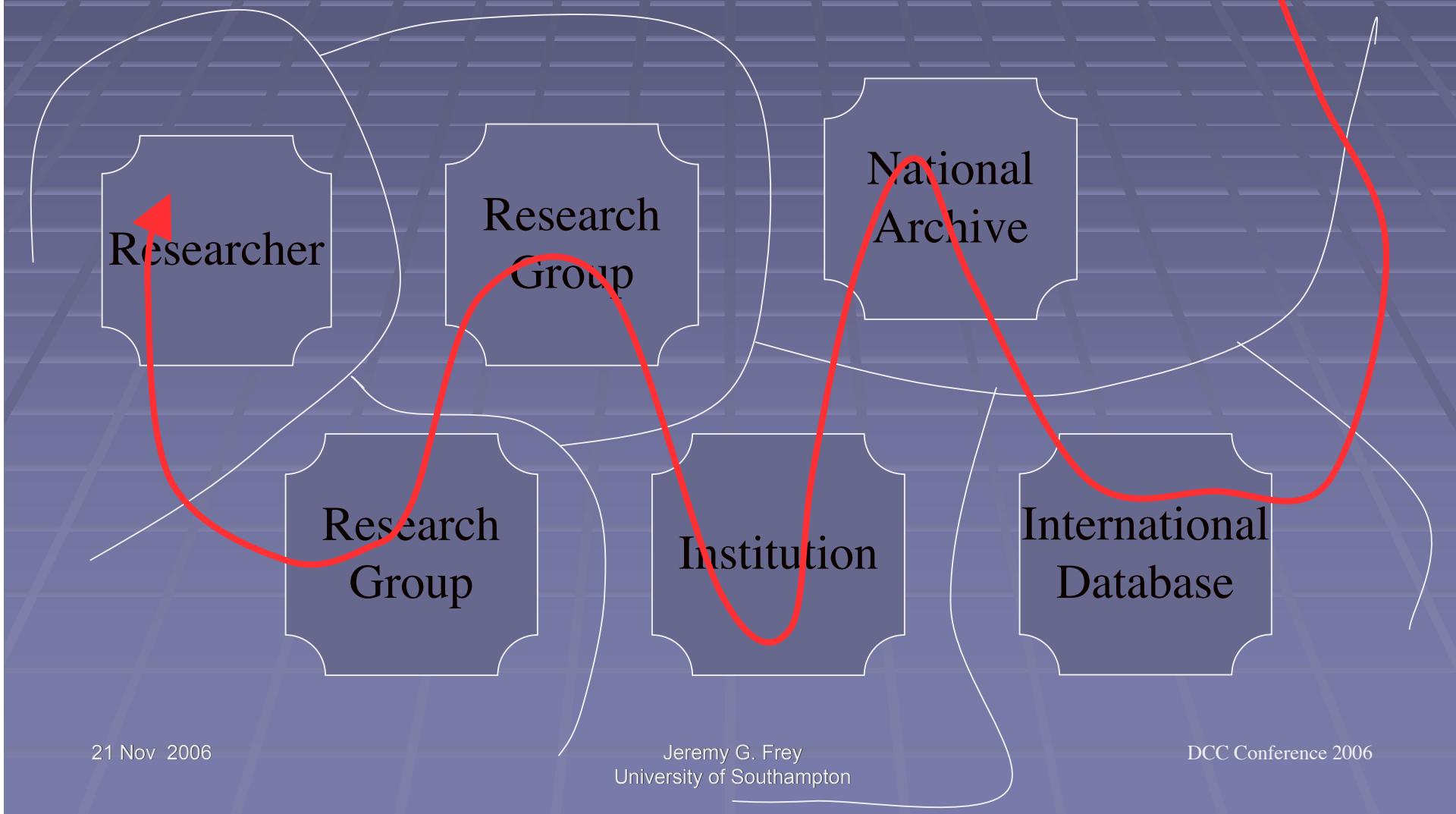
21 Nov 2006

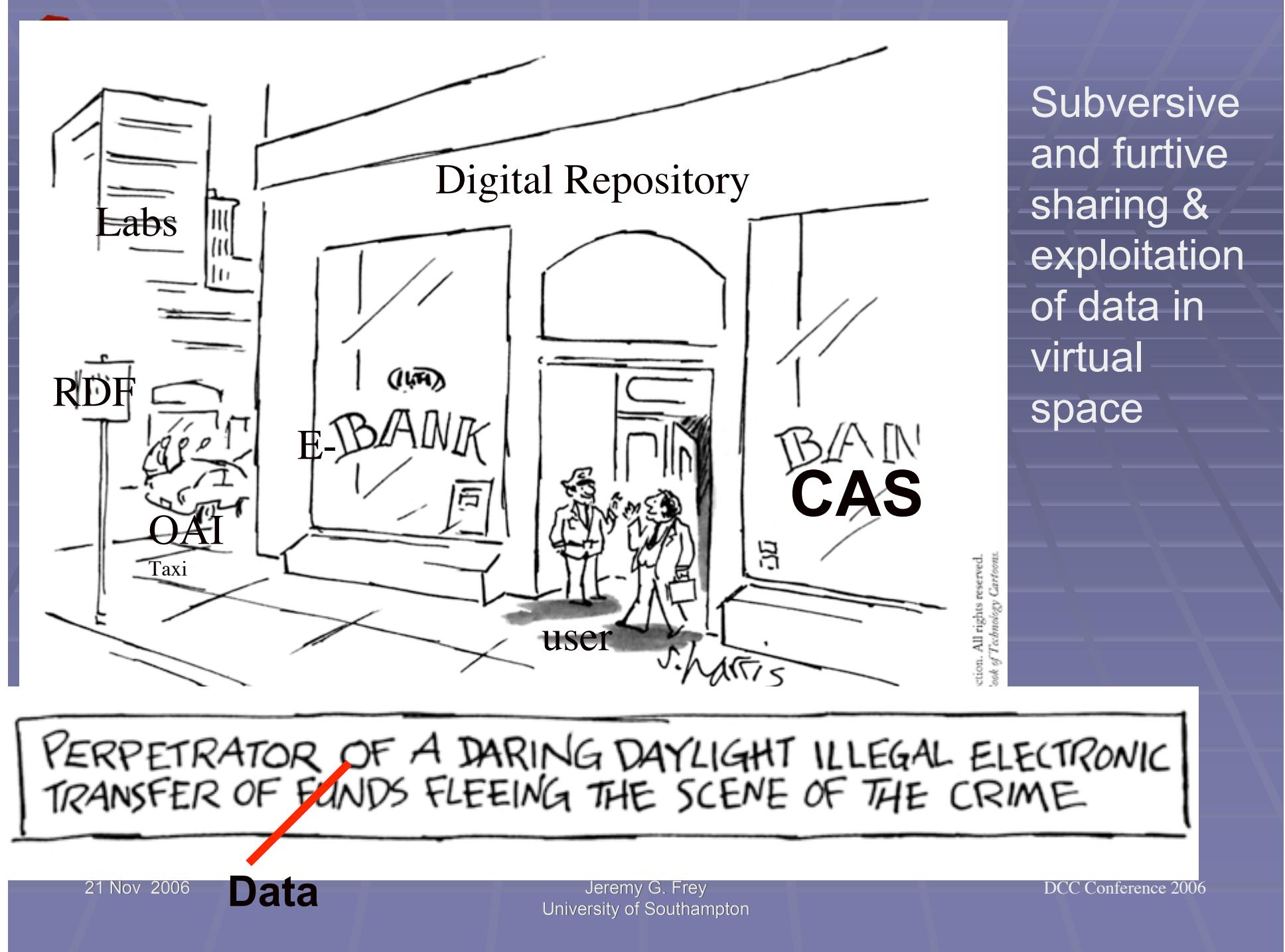
Jeremy G. Frey
University of Southampton

DCC Conference 2006



Access to information requires crossing administrative domains







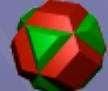
© Yorker collection. All rights reserved.
New Yorker Book of Technology Cartoons.

He is charged with expressing contempt for meta-data



Metadata Lifecycle

- Creation and maintenance of metadata
- Need a metadata infrastructure as well as a data infrastructure
- Capture process as well as results
- Automatic metadata generation when possible
- Human annotation will always be needed



Plans

- Plans are useful
- This is the way things are supposed to be done
- The Plan provides a digital context so increases the value of planning
- Key to our 'Smart Lab' approach....
- Is it the best way?



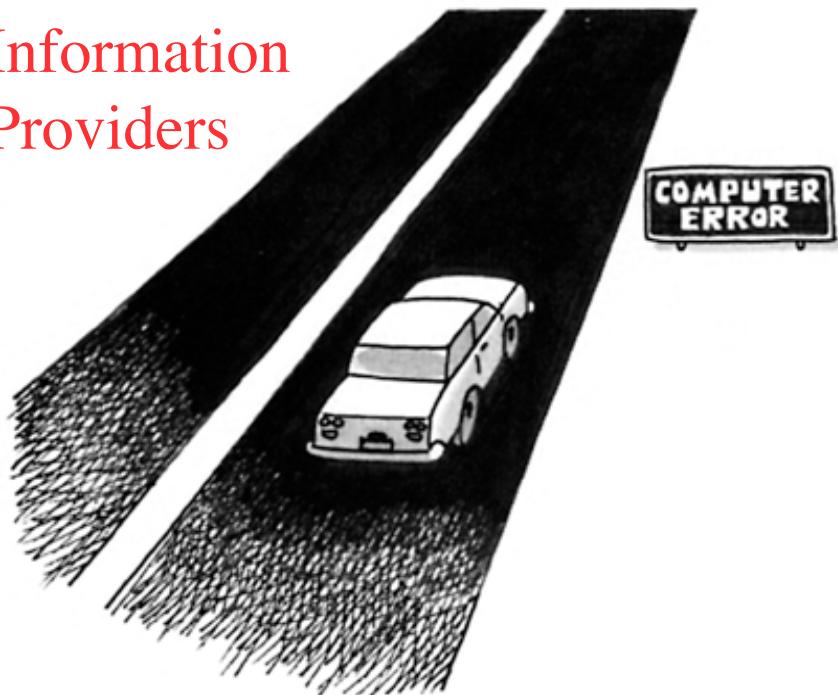
Who is responsible

- Context is crucial for curation
- every person, on each step of the process of converting data to knowledge
- Need to consider the future access to this information by themselves and others.

These are the same people – if we can ‘talk’ to ourselves efficiently over time then that is a good start to be able to ‘talk’ to others



Information
Providers



Information
Consumers

© The New Yorker collection. All rights reserved.
From The New Yorker Book of Technology Cartoons.

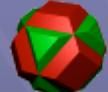


We must speed up the knowledge discovery process



© The New Yorker collection. All rights reserved.
From *The New Yorker Book of Technology Cartoons*.

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



PEOPLE

- Southampton ECS,
MATHS & CHEMISTRY
- IT-INNOVATION
- BRISTOL
- UKOLN
- CCLRC
- INDIANA
- SYDNEY
- MANCHESTER
- EPRSC e-Science &
Chemistry Programmes
- JISC e-Infrastructure
- DTI
- See web site for full
details and links
- www.combechem.org