

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

E-Malaria

R4L

Instruments on the Grid

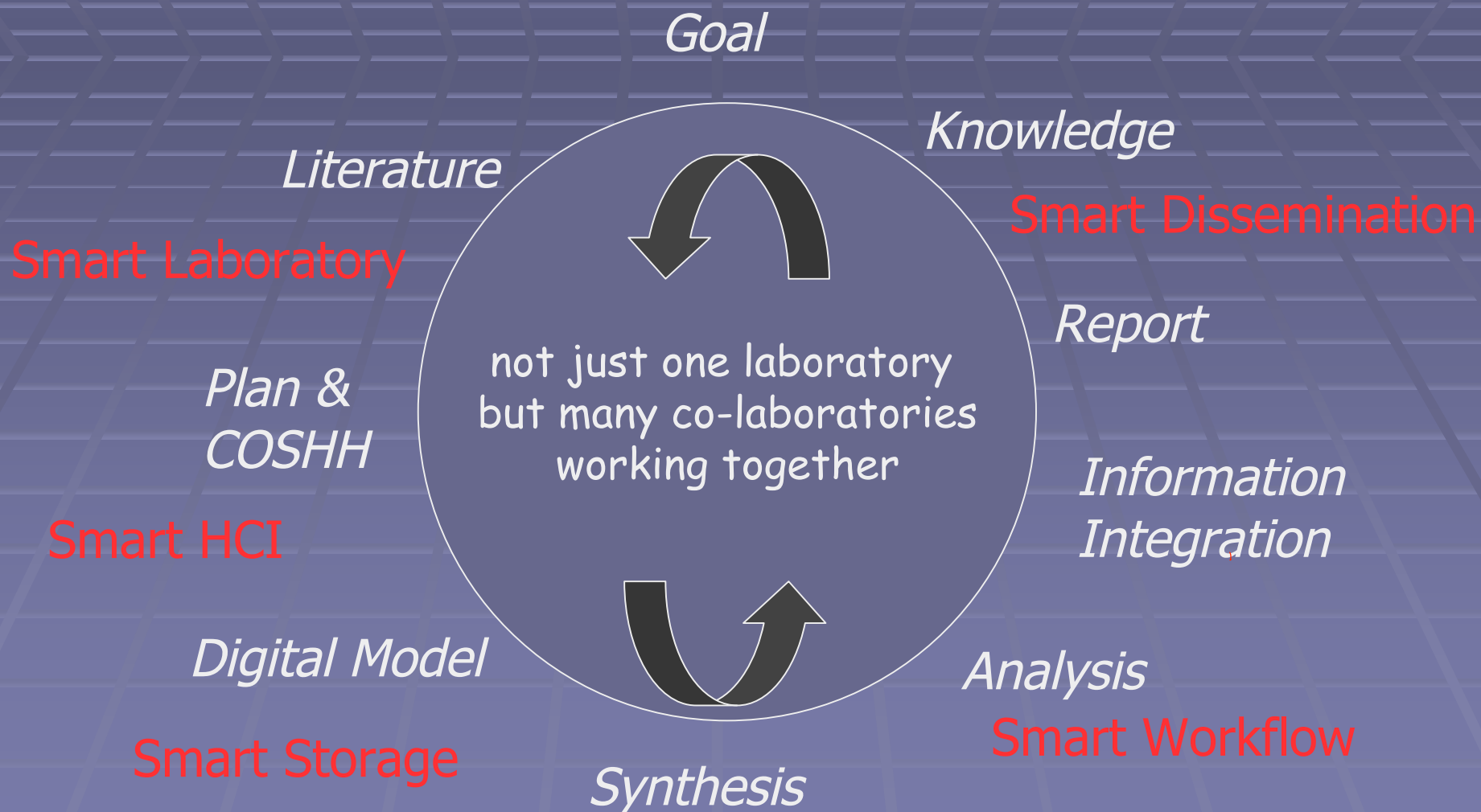
e-Bank

Statistics

BioSimGrid



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

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

Typical Laboratory



Need to make
the data
available

Need to be
able to find it

But how to
expose it?



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

Permanent,
documented
and primary
record of
laboratory
observations

21 Nov 2006

Safety

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

Resources

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

[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. 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



Jeremy G. Frey
University of Southampton

DCC Conference 2006



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	Solids	<20g	possible irritation to eyes and skin	
Caffeine	Solids (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				



21 Nov 2006

Jeremy G. Frey
University of Southampton

DCC Conference 2006



To DO LIST

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

Dissolve 4-fluorinated 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)

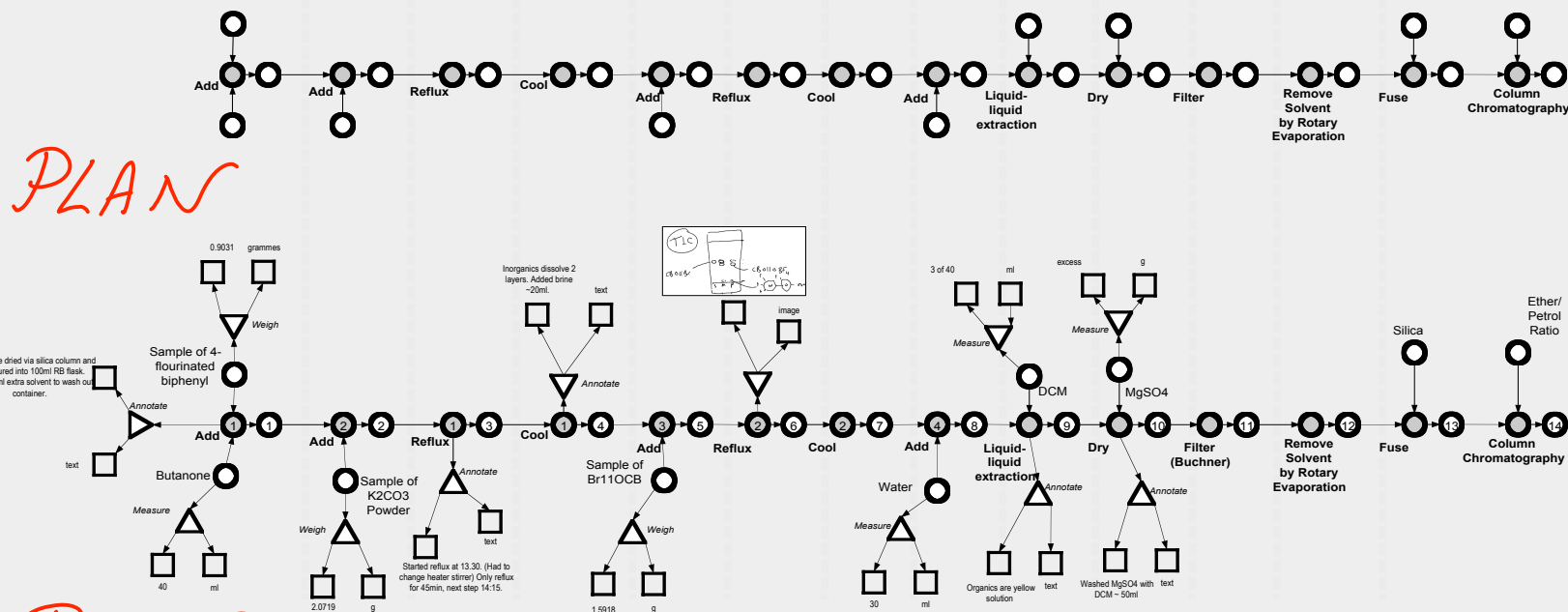
Combine organics, dry over MgSO₄ & filter

Remove solvent in vacuo

Fuse compound to silica & column in ether/petrol

PLAN

RECORD



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



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	2.0719 g	2.0719 g
Butanone	40.0 ml	40.0 ml

Simple Interface

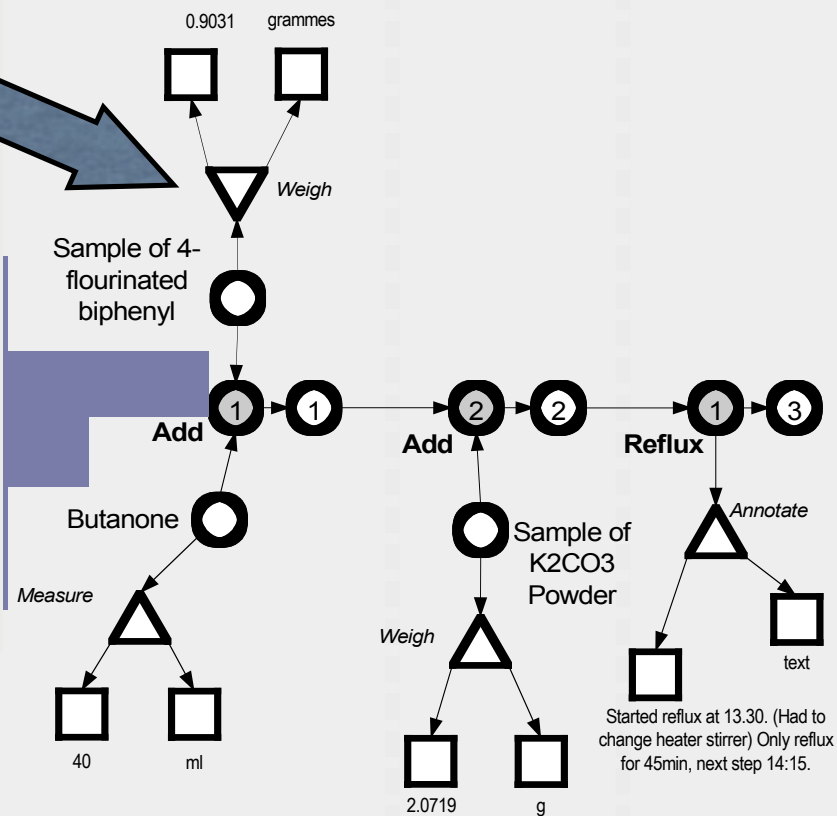
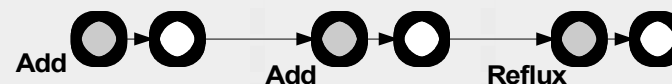
7	8	9
4	5	6
1	2	3
0	.	

Enter	Del
-------	-----

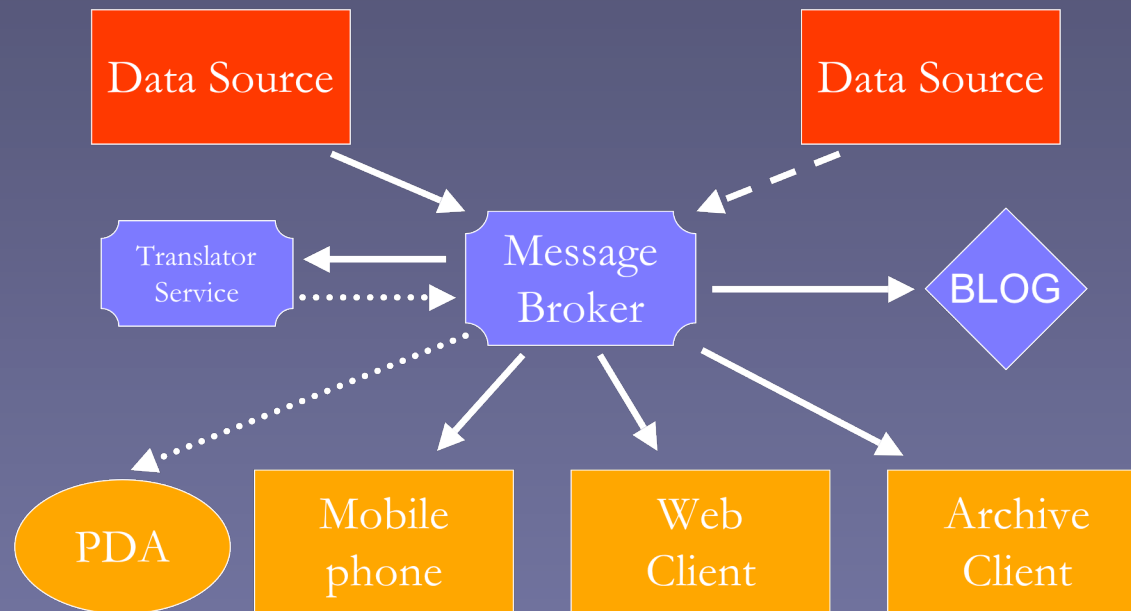
Dissolve 4-flourinated biphenyl in butanone

Add K₂CO₃ powder

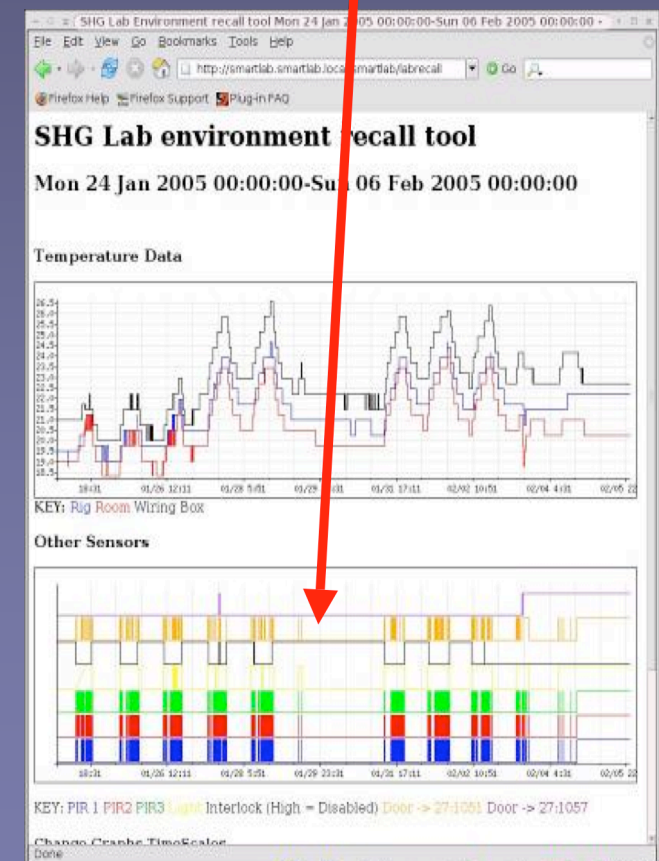
Heat at reflux for 1.5 hours



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



Air Conditioning failed



Smart Laboratory Spaces



But what
about the
laboratory
environment?

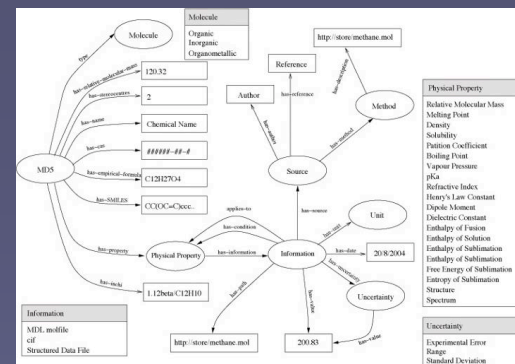


**"I just realized, Howard, that everything
in this apartment is more sophisticated
than we are"**

© 2006 New Yorker collection. All rights reserved.
From *The New Yorker Book of Technology Cartoons*.

Semantic DataGrid

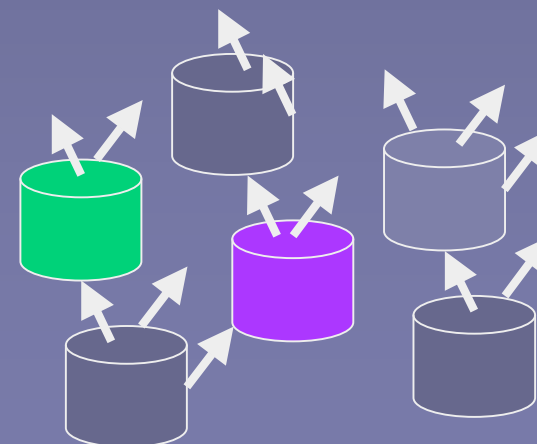
- CombeChem used, tested & strained the Semantic Web for
 - Enhanced (annotated) DataGrid over multiple diverse stores
 - Storage of Provenance Information
 - Some Data Storage
 - Annotated multimedia streams
 - Units & Properties Ontology
 - Multiple Triple Stores



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

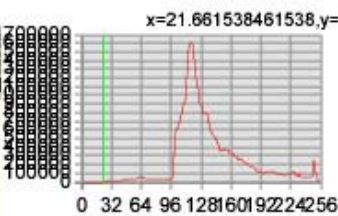

A highly advanced remote control microscope.

Data Collection

15th September 2006 @ 16:52

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

Data



x=21.661538461538,y=117

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

Archives

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

Sections

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

Search

Links

‘Blog-jects’



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

Useful Chemistry - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites RSS Print Mail

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

Blogger SEARCH THIS BLOG SEARCH ALL BLOGS BlogThis! GET YOUR OWN BLOG FLAG? NEXT BLOG»

Useful Chemistry

UsefulChem Experiments

An attempt at open source science in chemistry. Post specific problems in chemistry that need to be solved. Post specific partial solutions to these problems. Or execute a suggested step. NOTE: ANYTHING POSTED HERE IS MADE PUBLIC IMMEDIATELY AND DONATED TO THE PUBLIC DOMAIN . ANYONE MAY USE, EVEN FOR COMMERCIAL PURPOSES, AS LONG AS ATTRIBUTION IS MADE TO THE RELEVANT POSTS IN THIS BLOG

THURSDAY, NOVEMBER 16, 2006

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

Done

start Useful Chemistry - Mi... BBC NEWS | Politics | ... Document1 - Microsof... Microsoft PowerPoint ... Internet 20:28

Public

Attribution

Immediate

21 Nov 2006

DCC Conference 2006



Useful Chemistry - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites

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

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 JSpeView](#), 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.

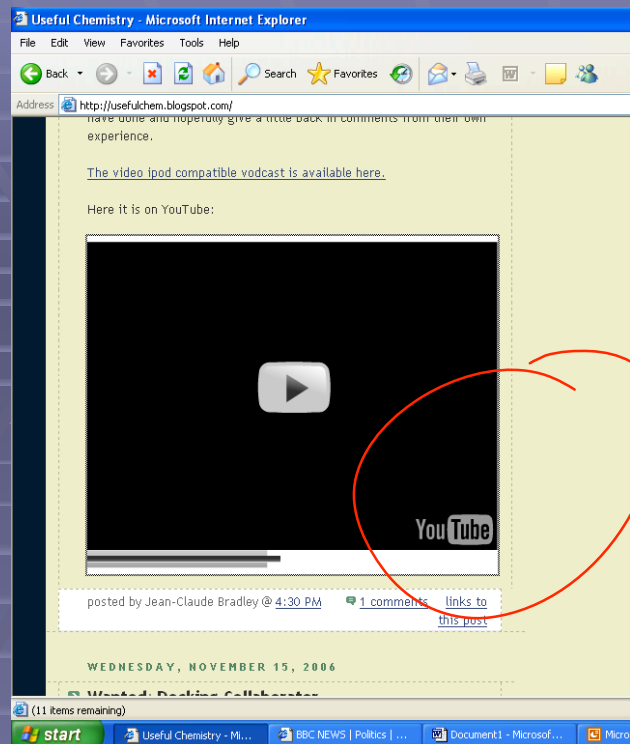
For more details read the [discussion section of the experiment](#) and please feel free to comment. To actually update the wiki directly just request an account. I don't allow anonymous guest updates because it is too easy for my students to forget to log in and I want to make sure they get credit for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

start Useful Chemistry - Mi... BBC NEWS | Politics | ... Document1 - Microsof...

Format Issues
- everyday and
for the long
term



Note the use of
"YouTube"

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

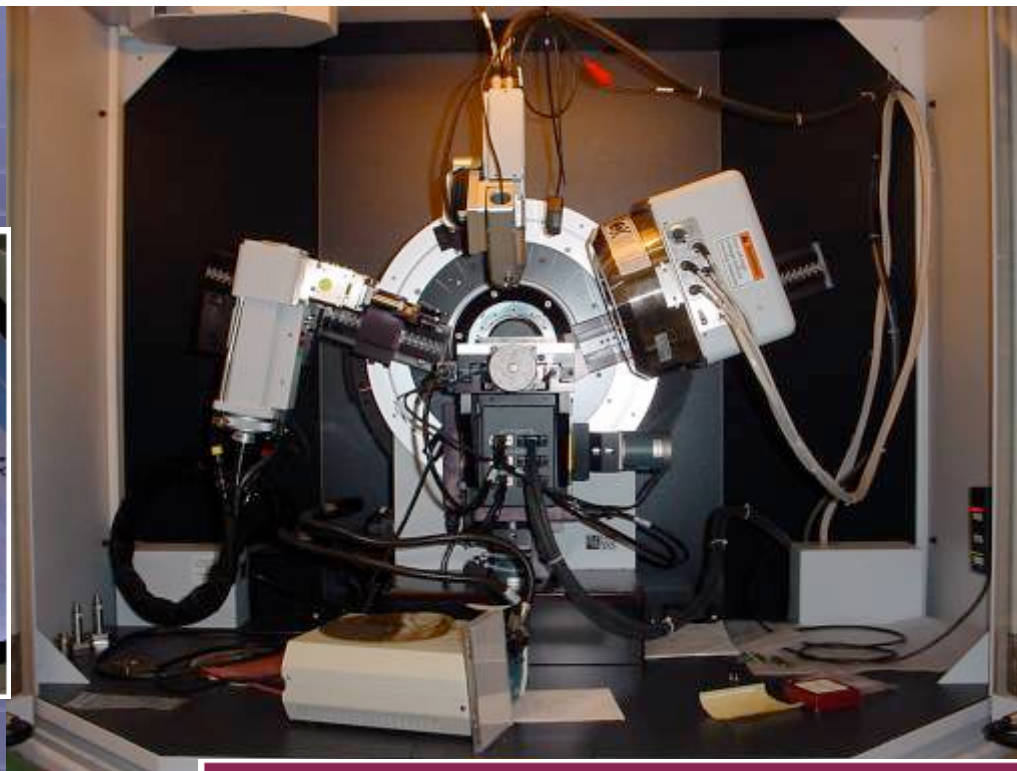
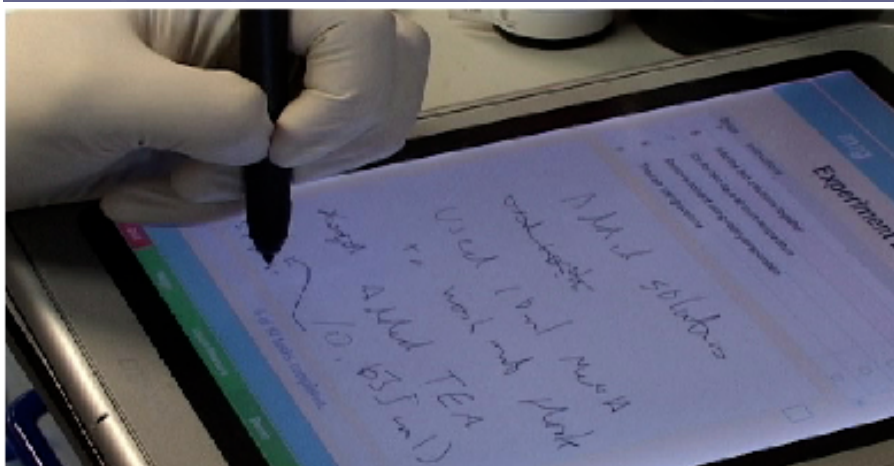
Jeremy G. Frey
University of Southampton

DCC Conference 2006





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.

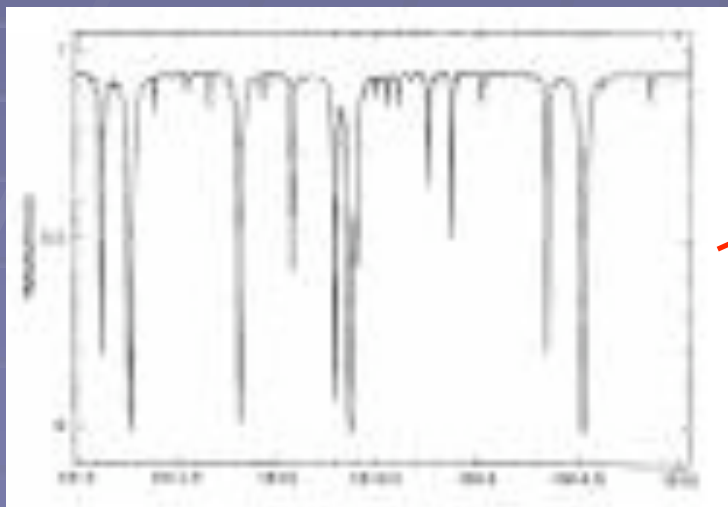
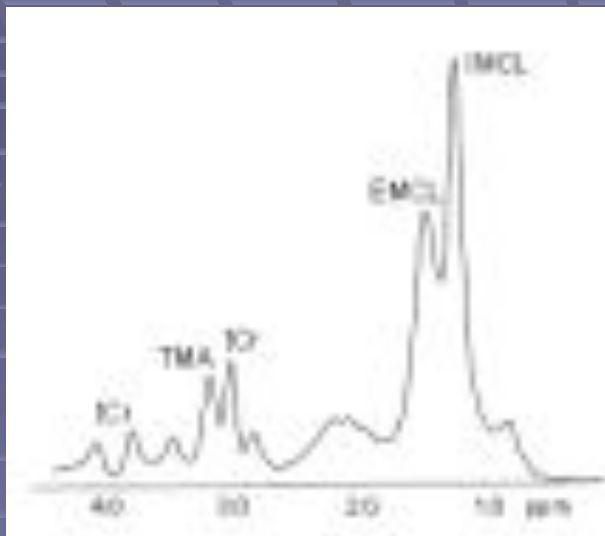


Validation

- 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



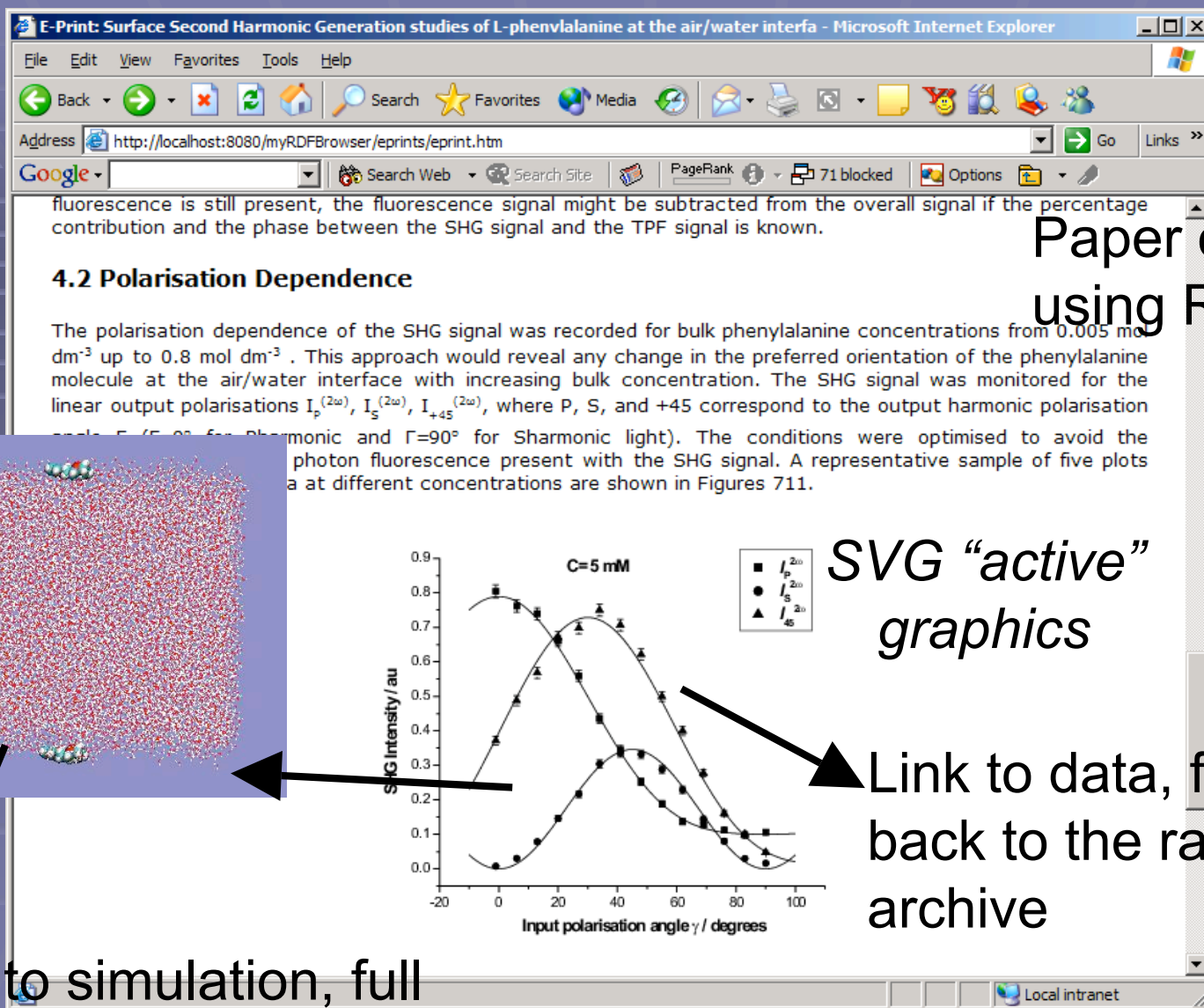
by passing through columns of P_2O_5 with moisture indicator and 4 Å molecular sieves and permeation chromatography (GPC) measurements were performed on a Polymer Laboratories PL-GPC-220 instrument equipped with a PL-gel 5 Å Mixed-C column, a refractive index detector, and a PD2040 light scattering detector. The GPC column was calibrated using eight monodisperse polystyrene standards in the range 580–48300 Da.

Preparation of $CPh_3[NCPBB] (1)$

Potassium cyanide (33.6 mg, 0.5 mmol) was ground to a powder using a pestle and mortar in a dry box. PBB (0.478 g, 0.5 mmol) and 50 mL diethyl ether were then added, and the mixture was heated to reflux for 12 h. The solvent was removed *in vacuo* to leave an off-white foam which was washed with warm hexane (50 mL) to give $K[NCPBB]$ as a white powder (0.495 g, 485 μmol). This solid was stirred with triphenylchloromethane (0.135 g, 0.485 mmol) in dichloromethane (15 mL) for 2 h. The solution was filtered to remove KCl, concentrated to ca. 5 mL and cooled to $-26^\circ C$ to give an orange crystalline solid, yield: 0.324 g (0.315 mmol, 63% with respect to KCN). IR (nujol): 2189 cm^{-1} (ν_{CN}). 1H NMR (CD_2Cl_2 , $20^\circ C$, 300.13 MHz): δ 8.28 (t, 3 H, $J = 7.5$ Hz, *p*-Ph), 7.90 (t, 6 H, $J = 7.7$ Hz, *m*-Ph), 7.70 (d, 6 H, $J = 7.3$ Hz, *o*-Ph). ^{13}C NMR (CD_2Cl_2 , $20^\circ C$, 75.48 MHz): δ 211.4 (CPh_3), 144.0 (*p*-C), 143.0 (*m*-C), 140.3 (*ipso*-C), 131.0 (*o*-C), 153.6, 150.3, 147.5, 146.4, 140.3, 139.6, 136.9, 136.3, 128.3, 113.5, 109.5 (Ar C-F). ^{19}F NMR (CD_2Cl_2 , $20^\circ C$, 96.3 MHz): δ -16.2 (br s).

Preparation of $CPh_3[(C_6F_5)_2BCNPBB] (2)$

$Me_2SiNCB(C_6F_5)_2$ (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 recrystallise the product from dichloromethane were not successful. IR (nujol): 2284 cm^{-1} (ν_{CN}). 1H NMR (CD_2Cl_2 , $20^\circ C$, 300.13 MHz): δ 8.56 (t, 3, $J = 8.0$ Hz *p*-Ph), 7.90 (t, 6 H, $J = 7.5$ Hz, *m*-Ph), 7.70 (d, 6 H, $J = 7.2$ Hz, *o*-Ph). ^{13}C NMR (CD_2Cl_2 , $20^\circ C$, 75.48 MHz): δ 211.0 (CPh_3), 144.1 (*p*-C), 143.0 (*m*-C), 140.1 (*ipso*-C), 130.9 (*o*-C). ^{19}F NMR (CD_2Cl_2 , $20^\circ C$, 96.3 MHz): δ -4.35 (s, br, 1 B, N- $B(C_6F_5)_2$), -18.27 (s, 1 B, $C-B(C_6F_5)_2$). ^{19}F NMR (CD_2Cl_2 , $20^\circ C$, 282.4 MHz): δ -118.72 (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 $B(C_6F_5)_3$), 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, $J = 20$ Hz, *p*-F on $B(C_6F_5)_3$), -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, $J = 20$ Hz, *m*-F on $B(C_6F_5)_3$). Anal. Calcd for $C_{24}H_{15}B_2F_{12}N$:



Paper organized
using RDF

SVG "active"
graphics

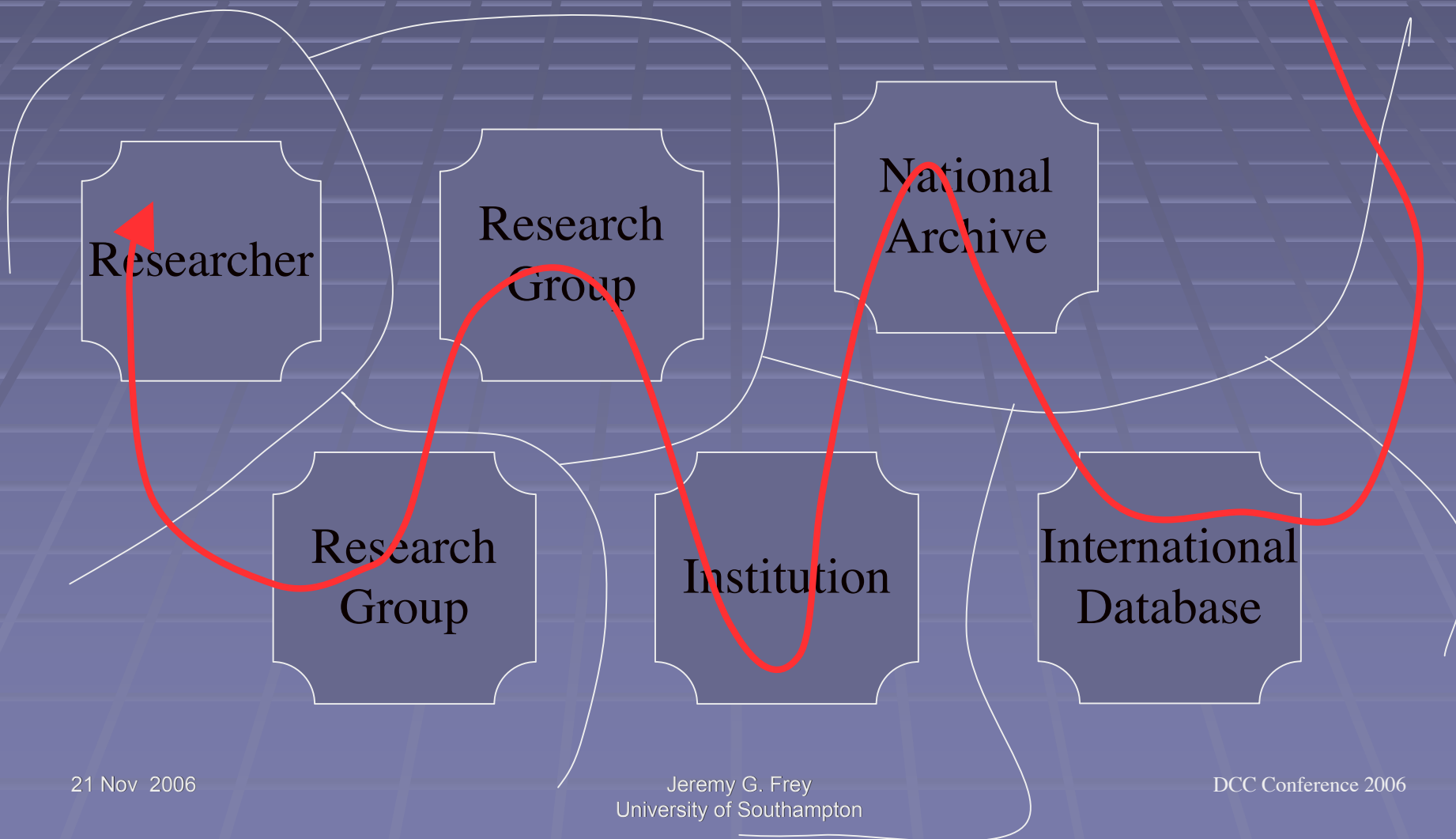
Link to data, follow links
back to the raw data
archive

Link to simulation, full
simulation data archived
in BioSimGrid

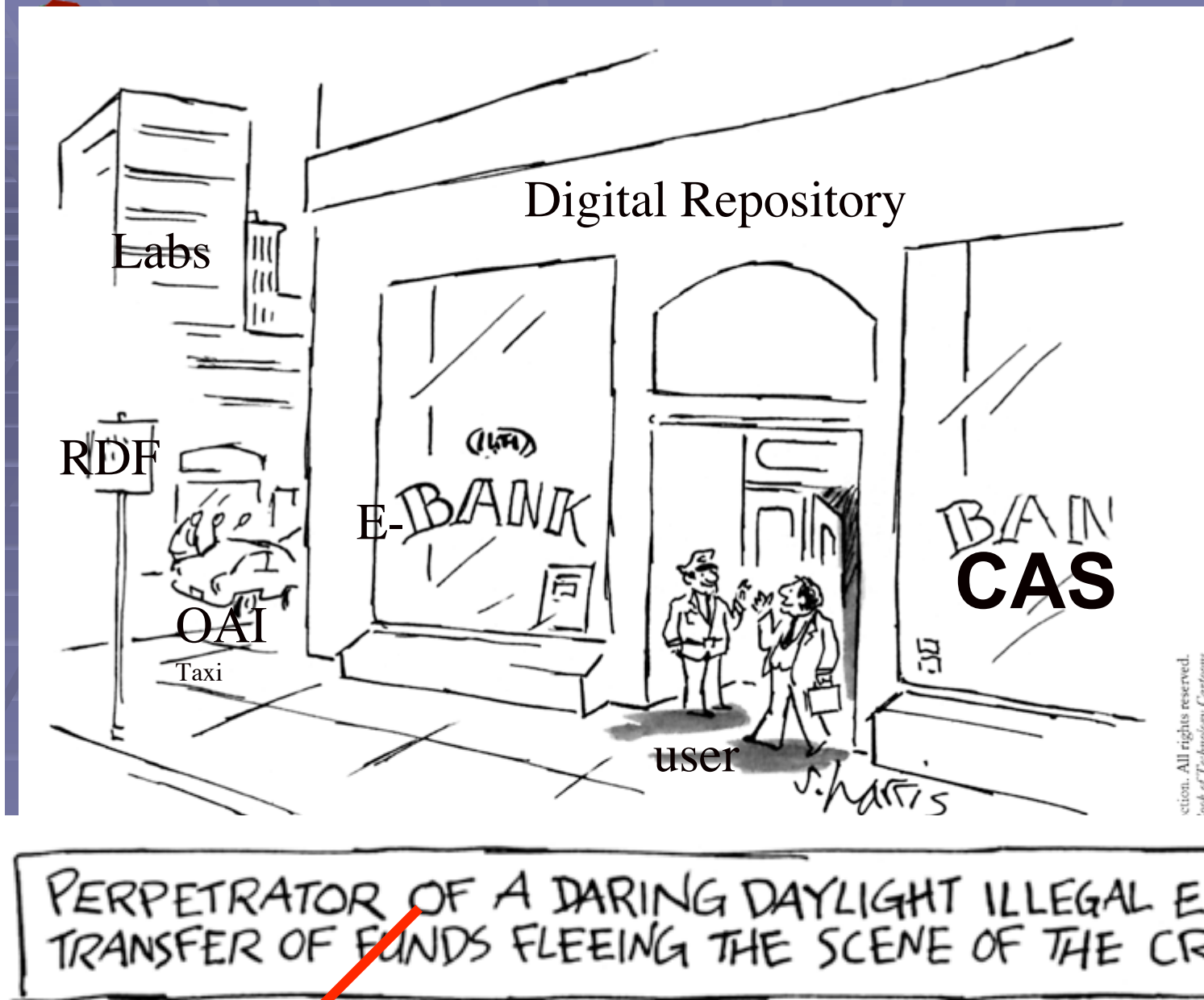
R4L



Access to information requires crossing administrative domains



Subversive
and furtive
sharing &
exploitation
of data in
virtual
space

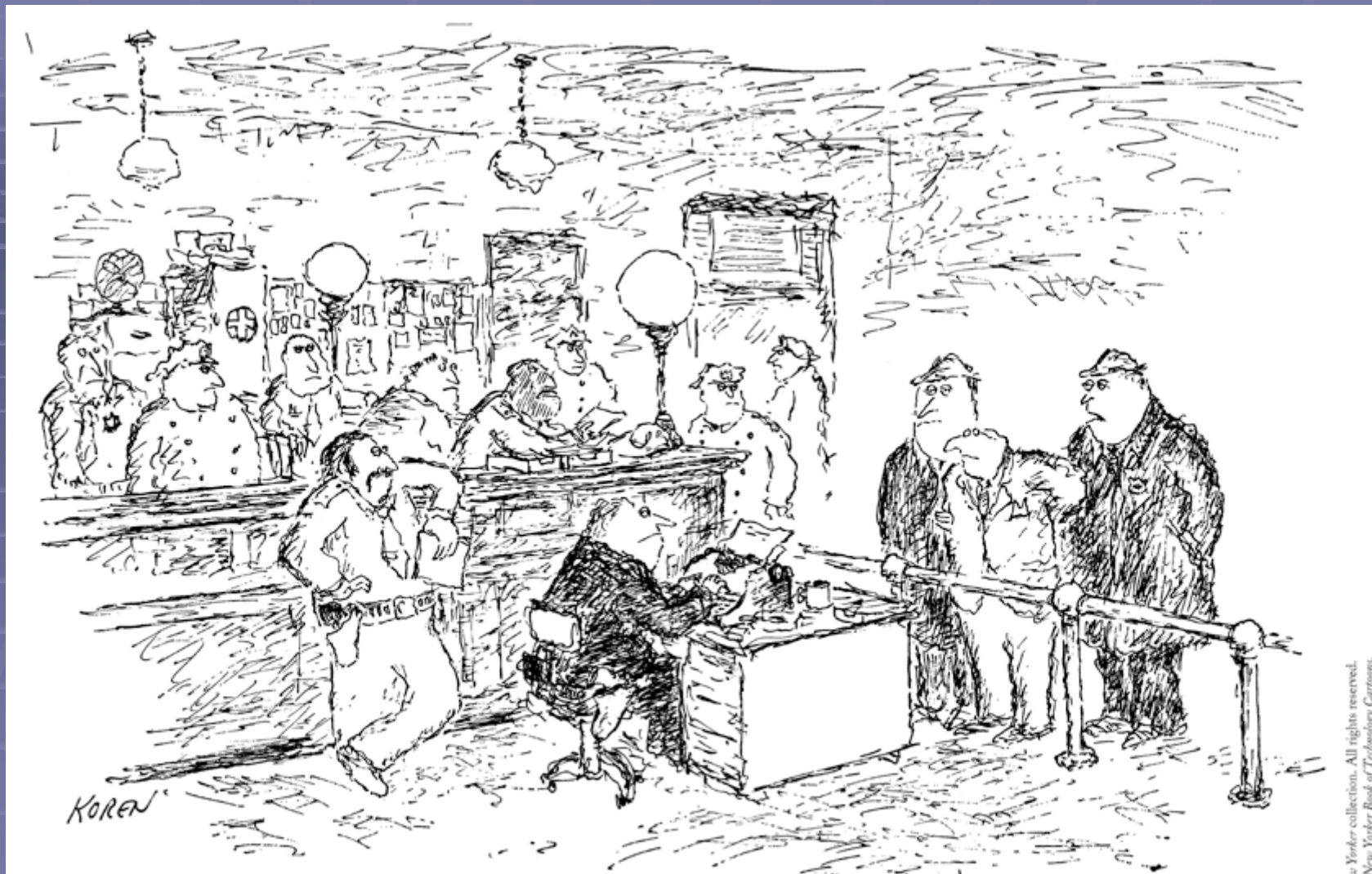


21 Nov 2006

Data

Jeremy G. Frey
University of Southampton

DCC Conference 2006



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?

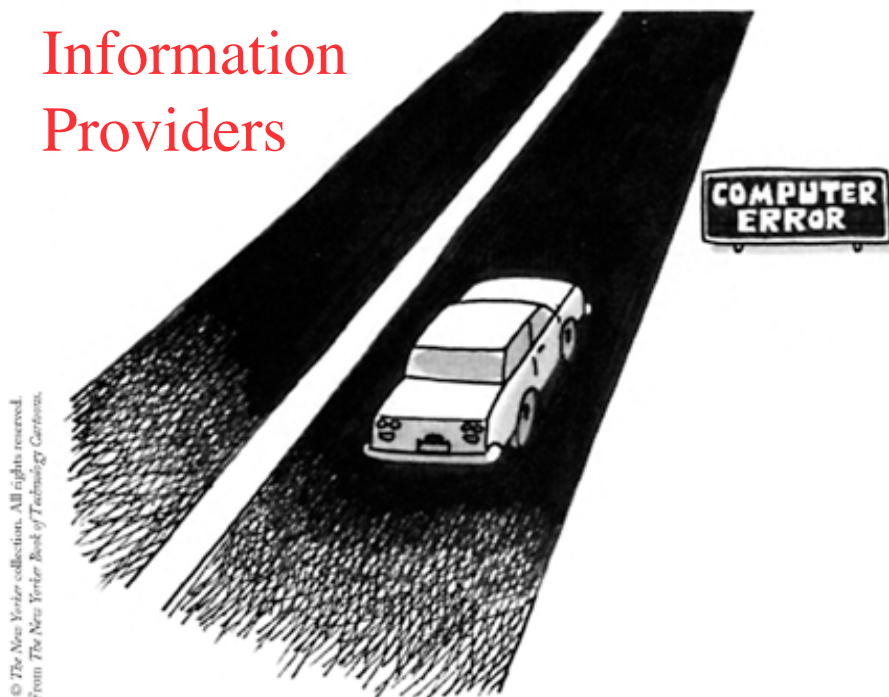


Who is responsible

- 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



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



We must speed up the knowledge discovery process



*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
- EPSRC e-Science & Chemistry Programmes
- JISC e-Infrastructure
- DTI
- See web site for full details and links
- www.combechem.org