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

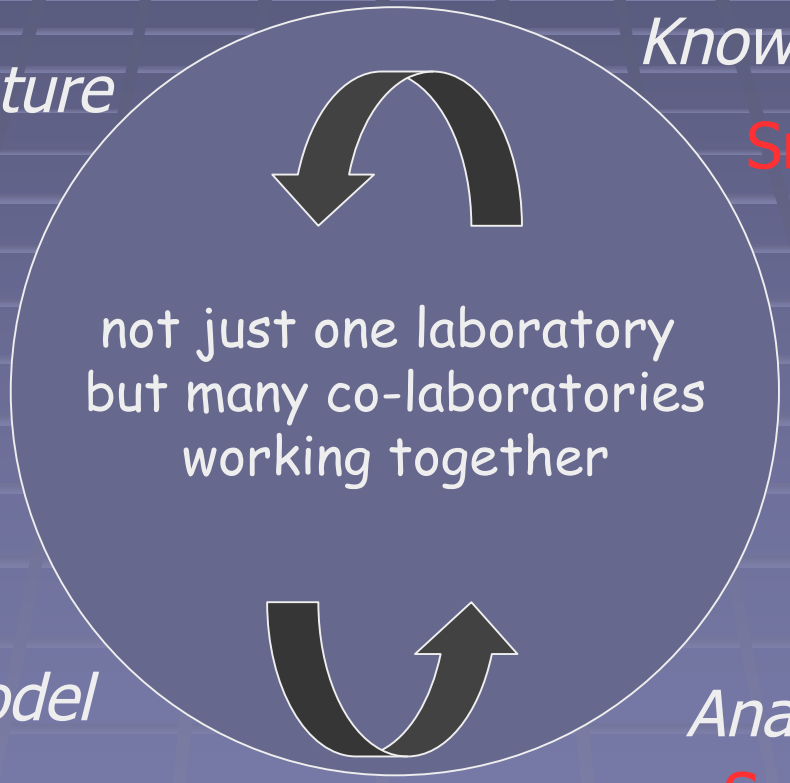
Instruments on the Grid

e-Bank

Statistics

BioSimGrid

The concept of Publication @ Source



Goal

Knowledge

Smart Dissemination

Report

Information Integration

Analysis

Smart Workflow

Synthesis

Smart Storage

Digital Model

Smart HCI

Plan & COSHH

Smart Laboratory

Literature

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 on-line search."

First, they do an online search

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

The details should be in her thesis……

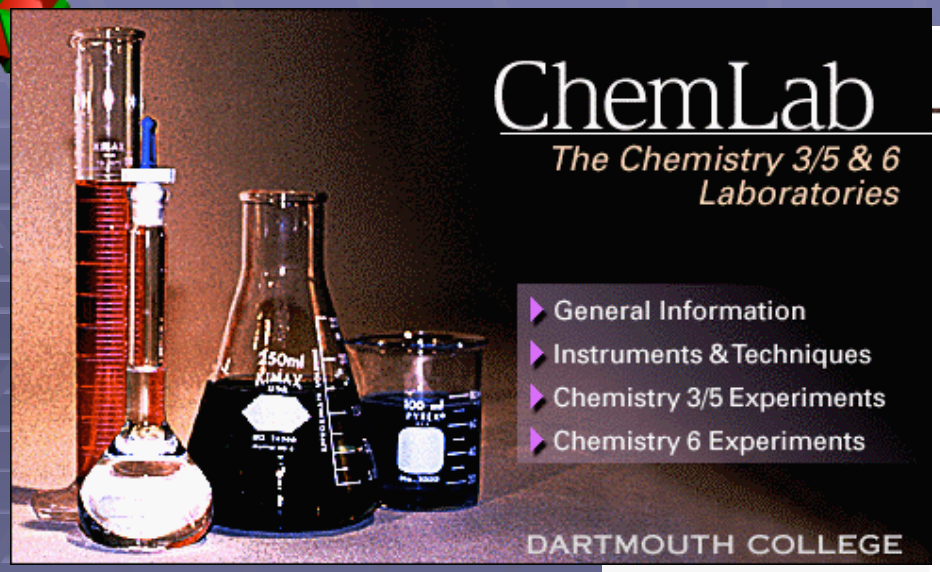
Can you read what he says here….?

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



Permanent, documented and primary record of laboratory ***observations***

21 Nov 2006

Safety
General Rules
Safety Equipment
Safety Hazards
[illegible] Procedures

Resources
Applets
General FAQ
Uncertainty

ChemLab Home

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.

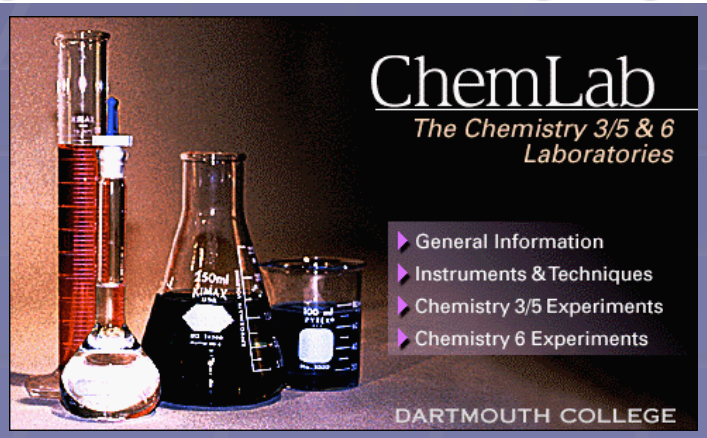
The laboratory notebook is a permanent, documented, and primary record of laboratory observations. Therefore, your notebook will be a bound journal with pages that should be numbered in advance and never torn out. A notebook will be supplied to you before the first laboratory period. Write your name, 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 t… your prelab problems from the manual for grading. P… 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?

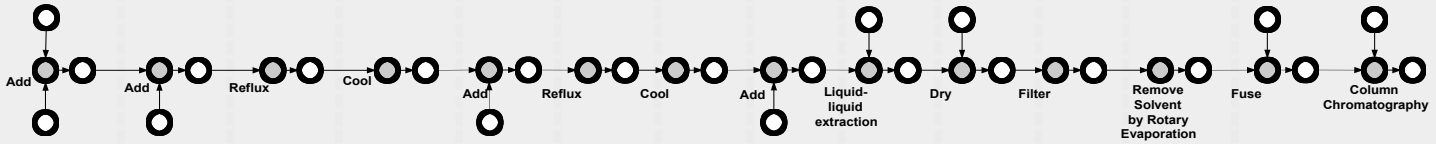
CONTROL MEASURES REQUIRED No specific measure required
(Local exhaust ventilation, personal protection, etc.)

TO DO LIST

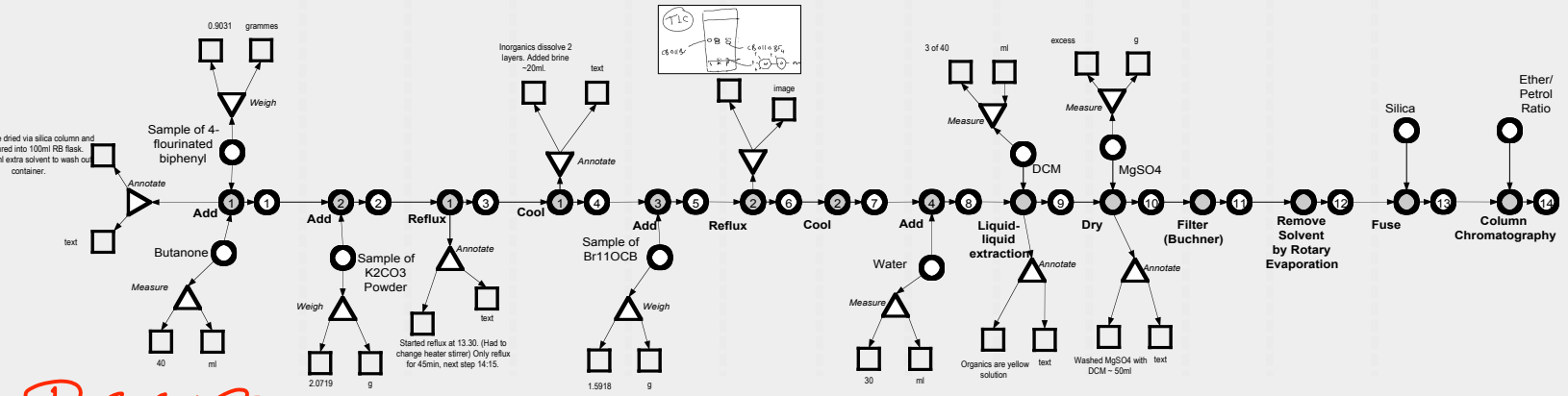
Ingredient List

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



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

Dissolve 4-flourinated biphenyl in butanone

Add K2CO3 powder

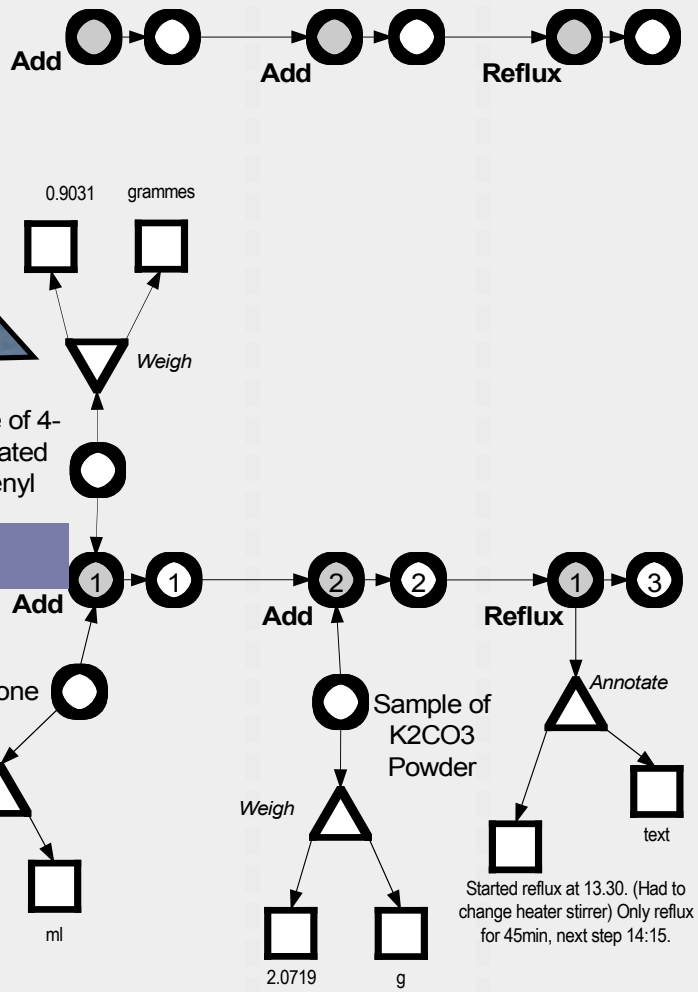
Heat at reflux for 1.5 hours

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

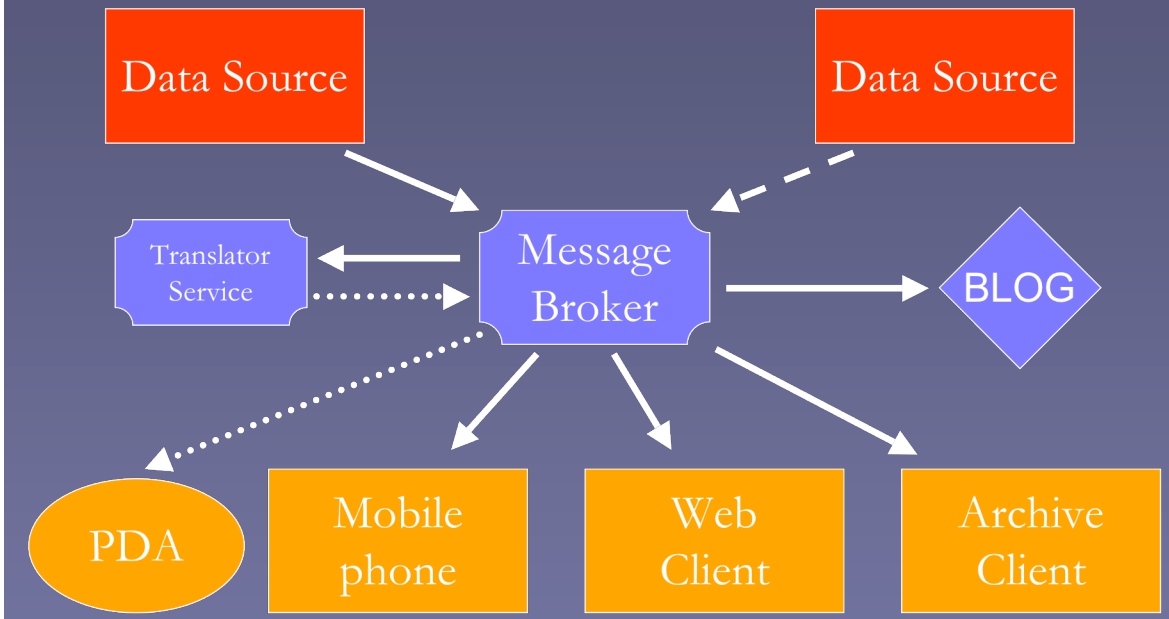
7	8	9
4	5	6
1	2	3
0		.

Enter | Del

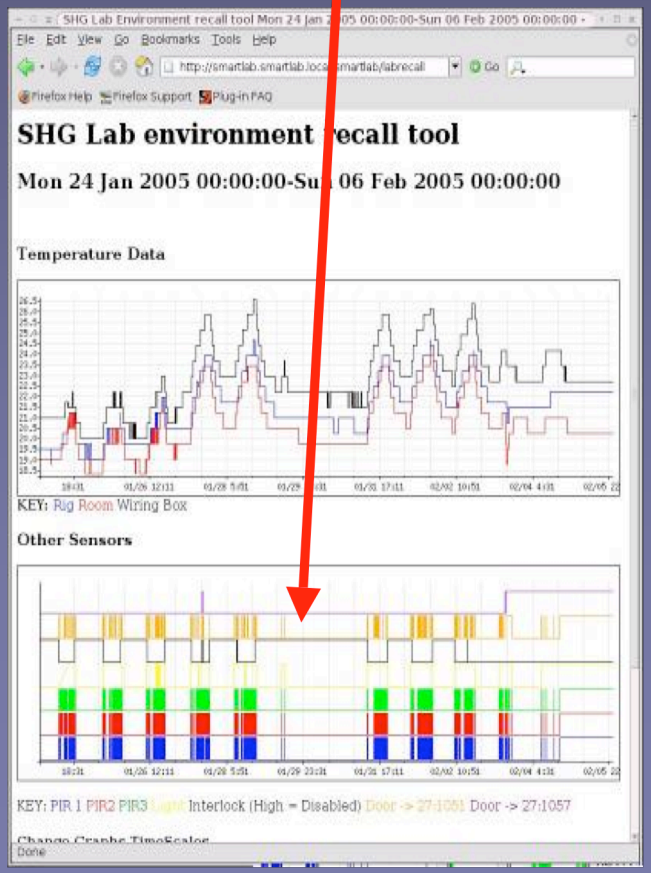


21 Nov 2006

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



Smart Laboratory Spaces



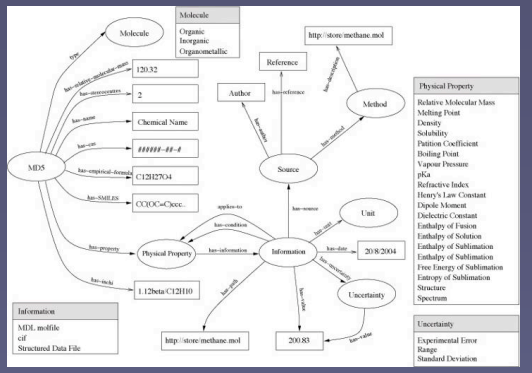
But what about the laboratory environment?

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

From *The New Yorker Book of Technology Cartoons.*

Semantic DataGrid

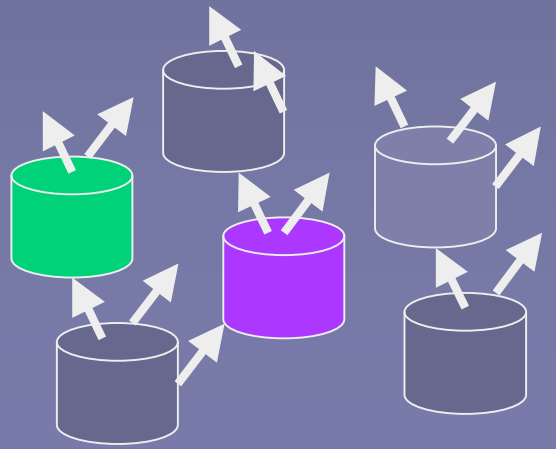
- CombeChem used, tested & strained the Semantic Web for
  - Enhanced (annotated) DataGrid over multiple diverse stores
  - Storage of Provenance Information
  - Some Data Storage
  - Annotated multimedia streams
  - Units & Propoerties Ontology
  - Multiple Triple Stores



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

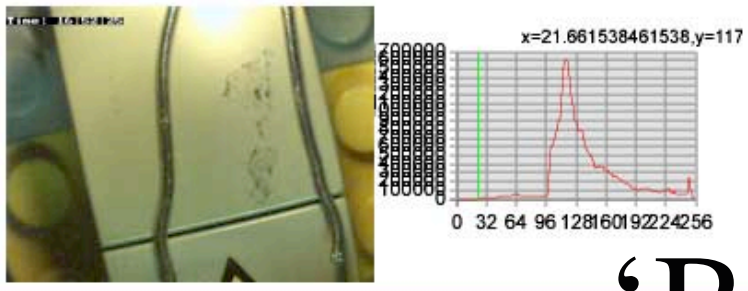
A highly advanced remote control microscope.

Data Collection

15th September 2006 @ 16:52

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

Data



ajm3 | Data Collection | Comments (0)

Archives

September 2006 (2)

Sections

Data Collection (2)

Search

Find

Links

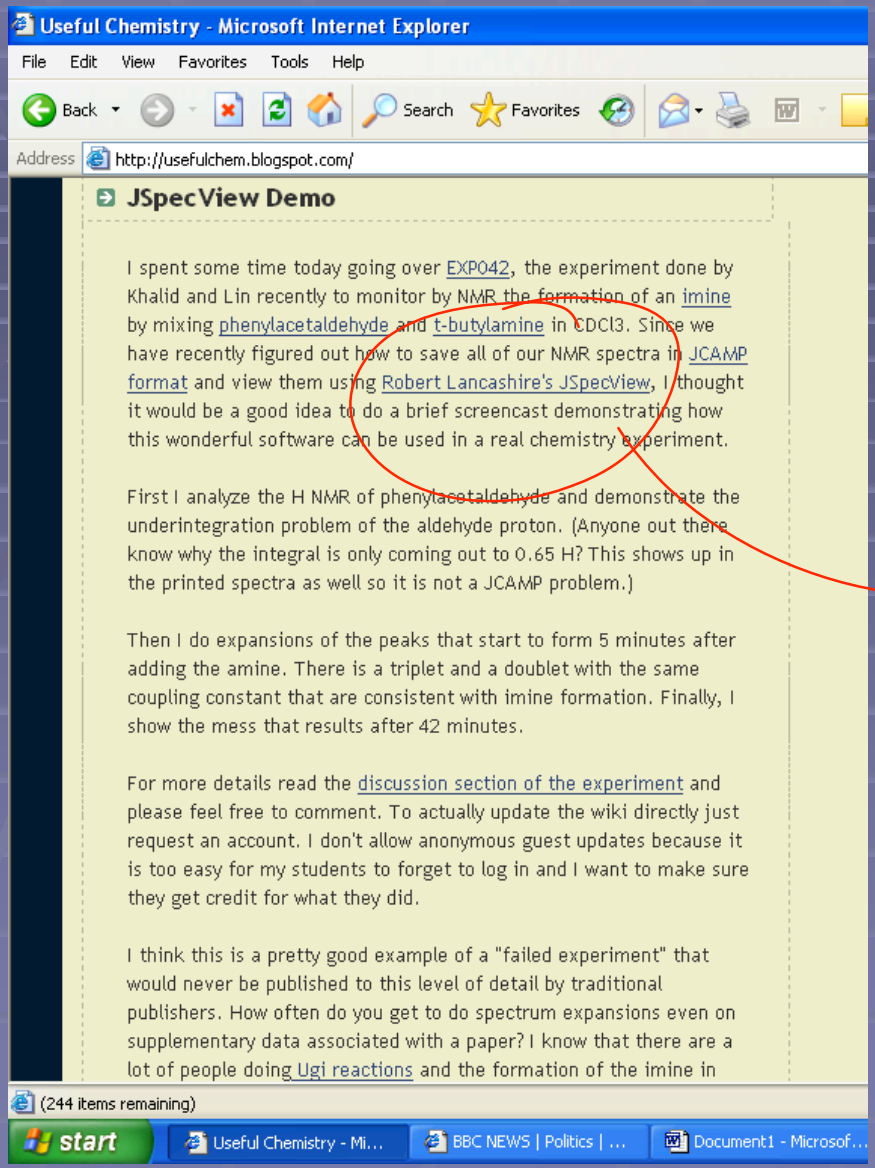
‘Blog-jects’

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

Public

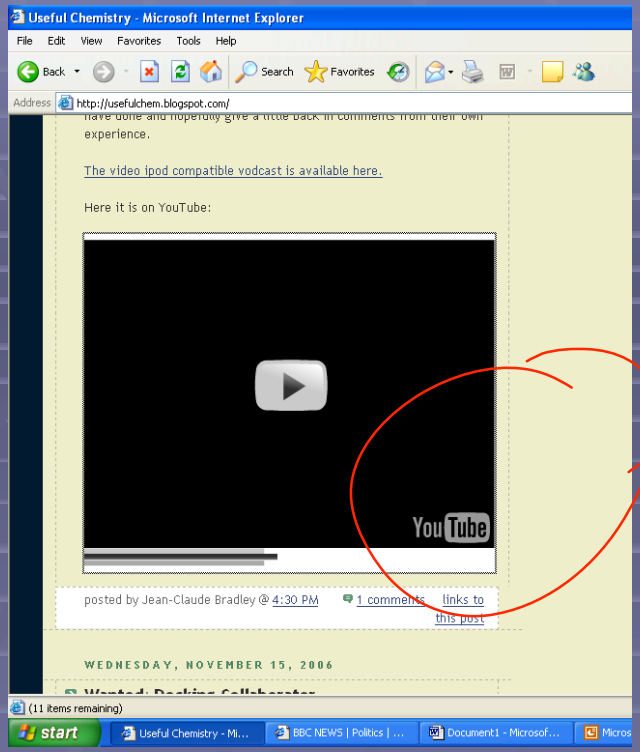
Attribution

Immediate

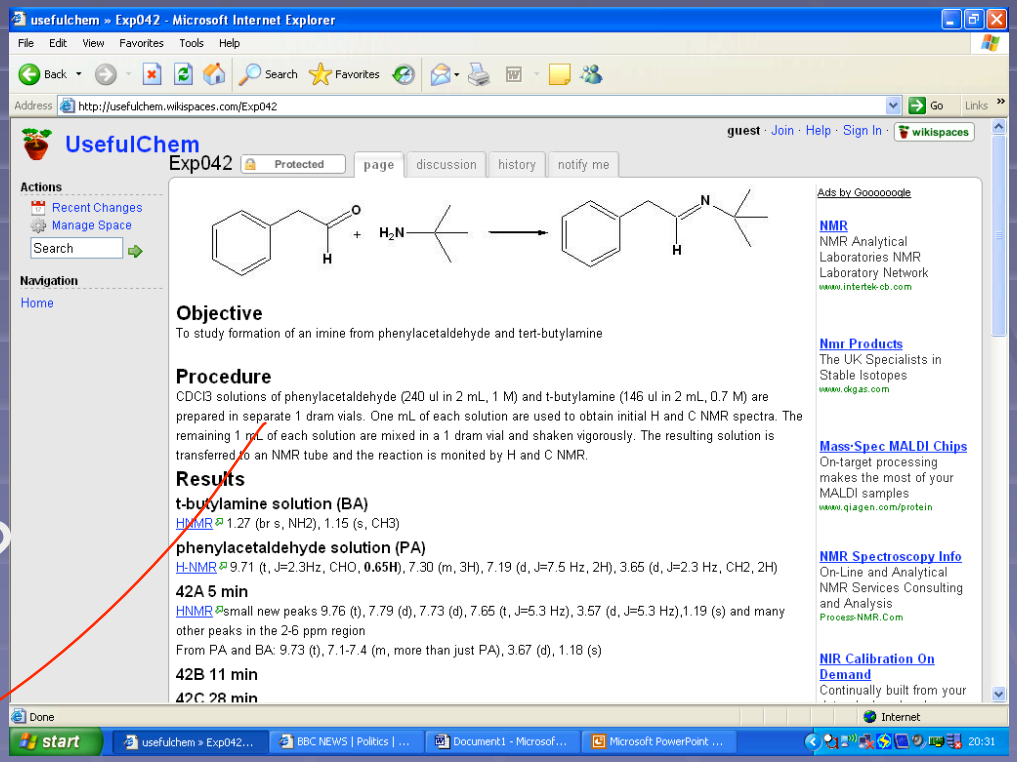


Format Issues - everyday and for the long term

Jeremy G. Frey
University of Southampton



Note the use of "YouTube"



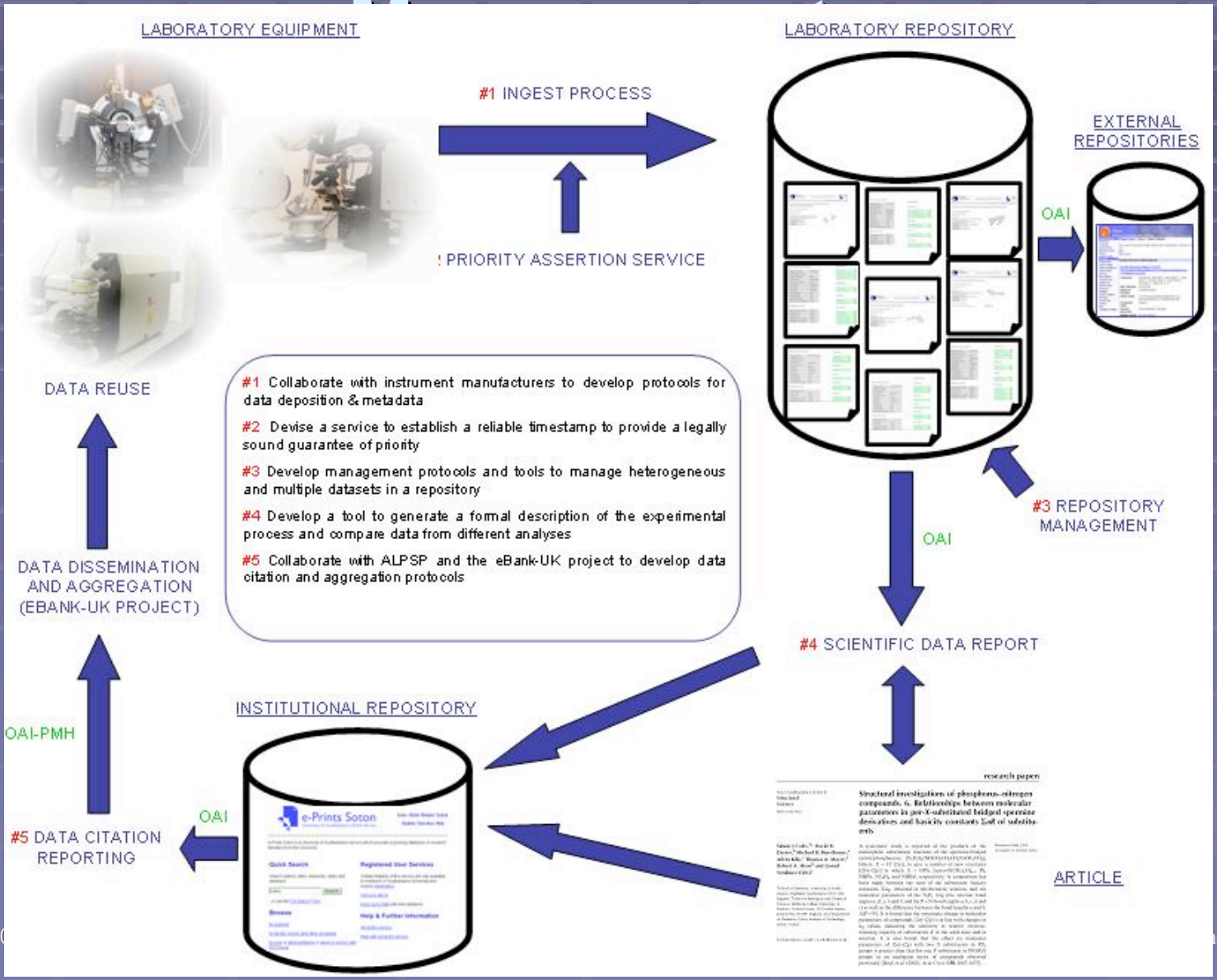
An experiment that failed… Publishable? Useful?

CoAKTing

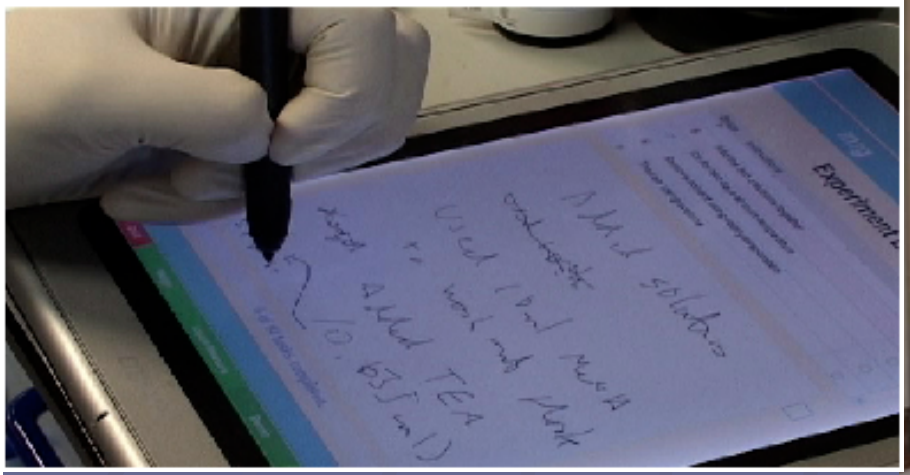
Memetic

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

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."

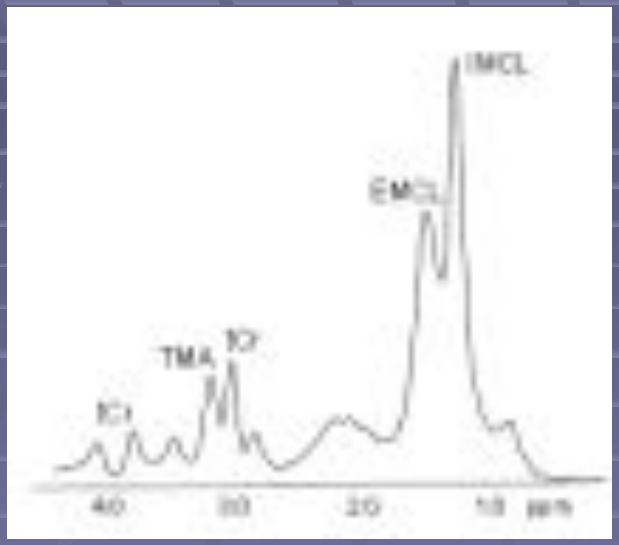
- Chris,
a real chemist, on using Smart Tea instead of a paper lab book.

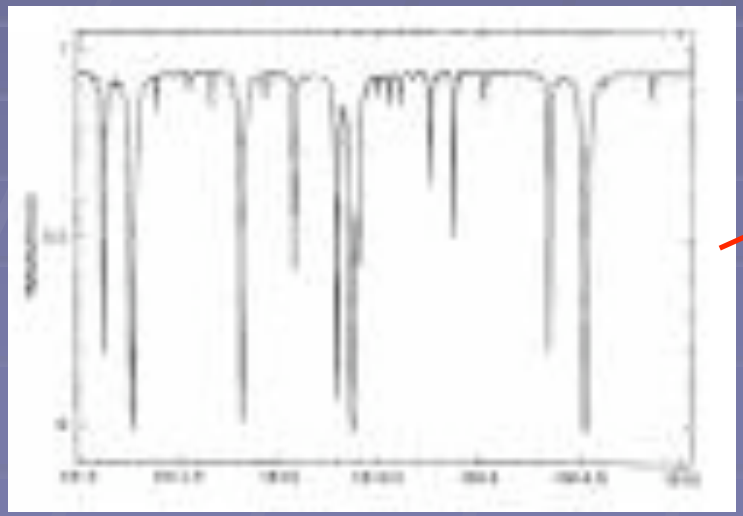
Smart Tea is about improving the information environment for chemists doing chemistry - within and beyond the lab. Smart Tea is about supporting chemists in the preparation, execution, analysis and dissemination of their experimental work.

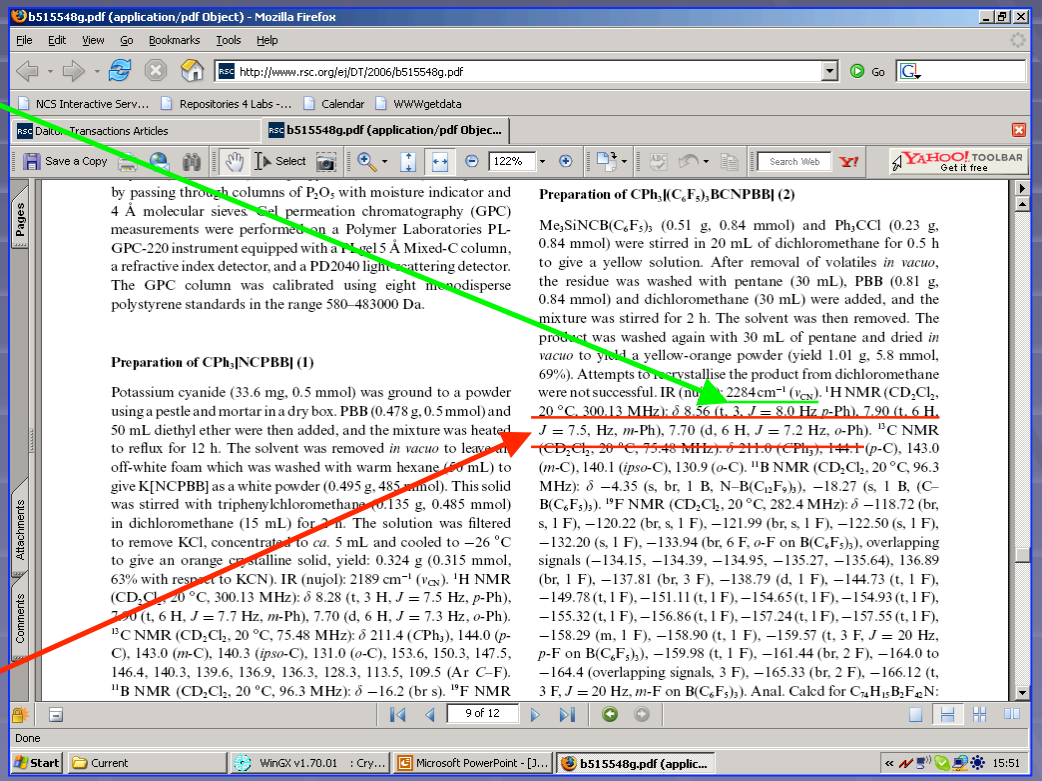
Validation

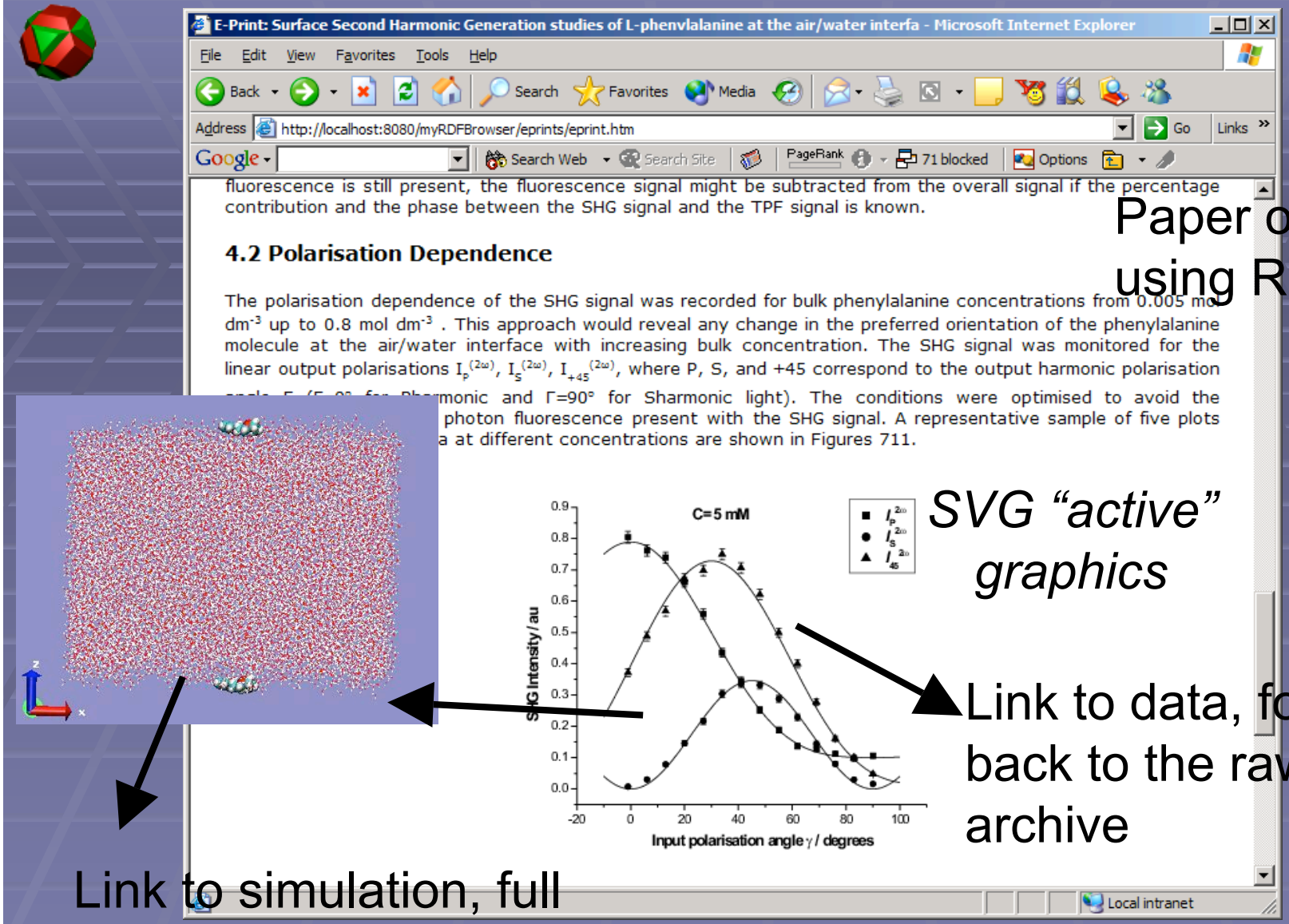
- Increasing the value of data
- How to bring all the necessary information together to enable appropriate validation
- Increasingly difficult & expensive to achieve
- Need provenance and context
- Essential step otherwise just a collection of items

Why?
Publishing Data and Information Loss









Paper organized using RDF

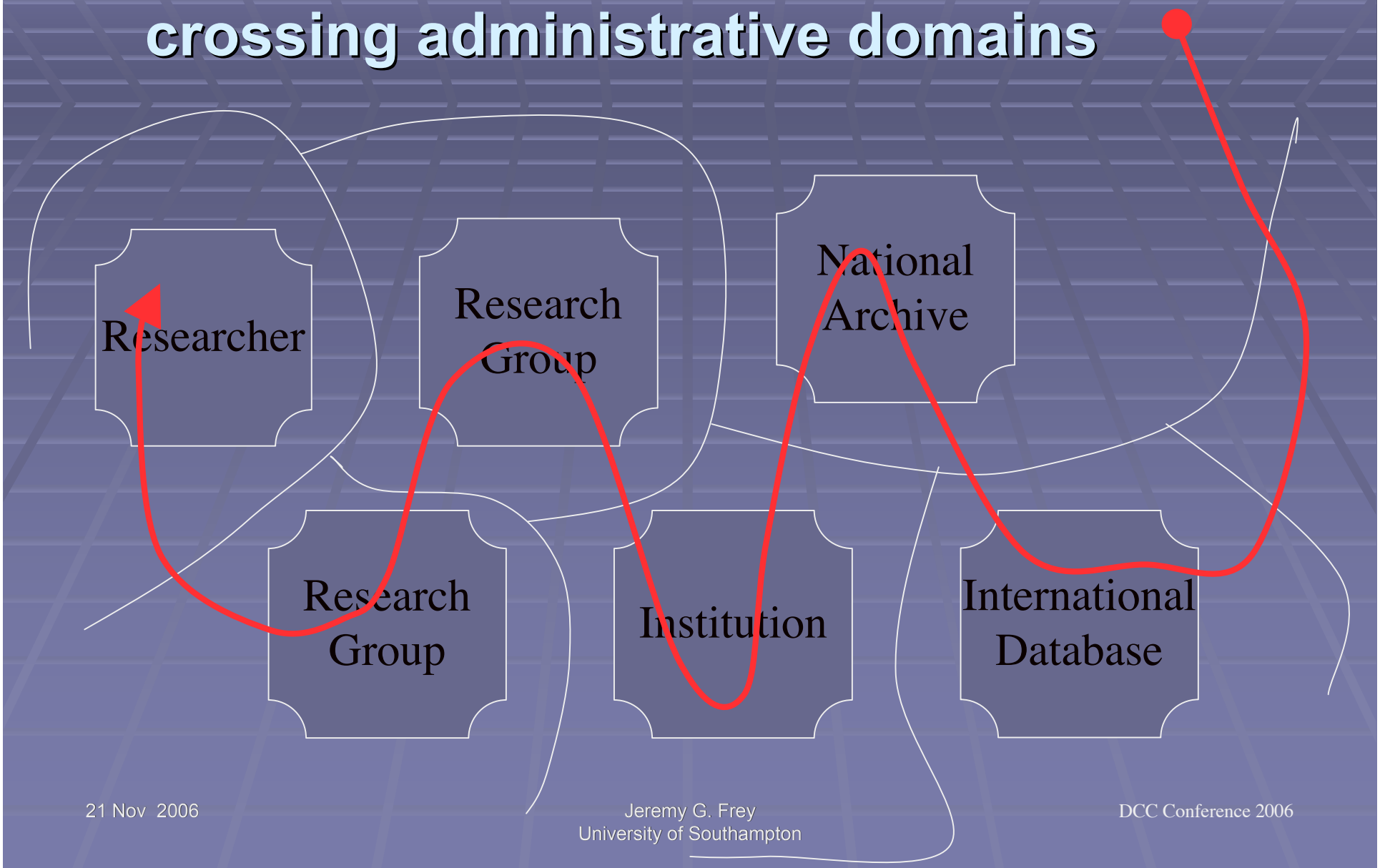
SVG "active" graphics

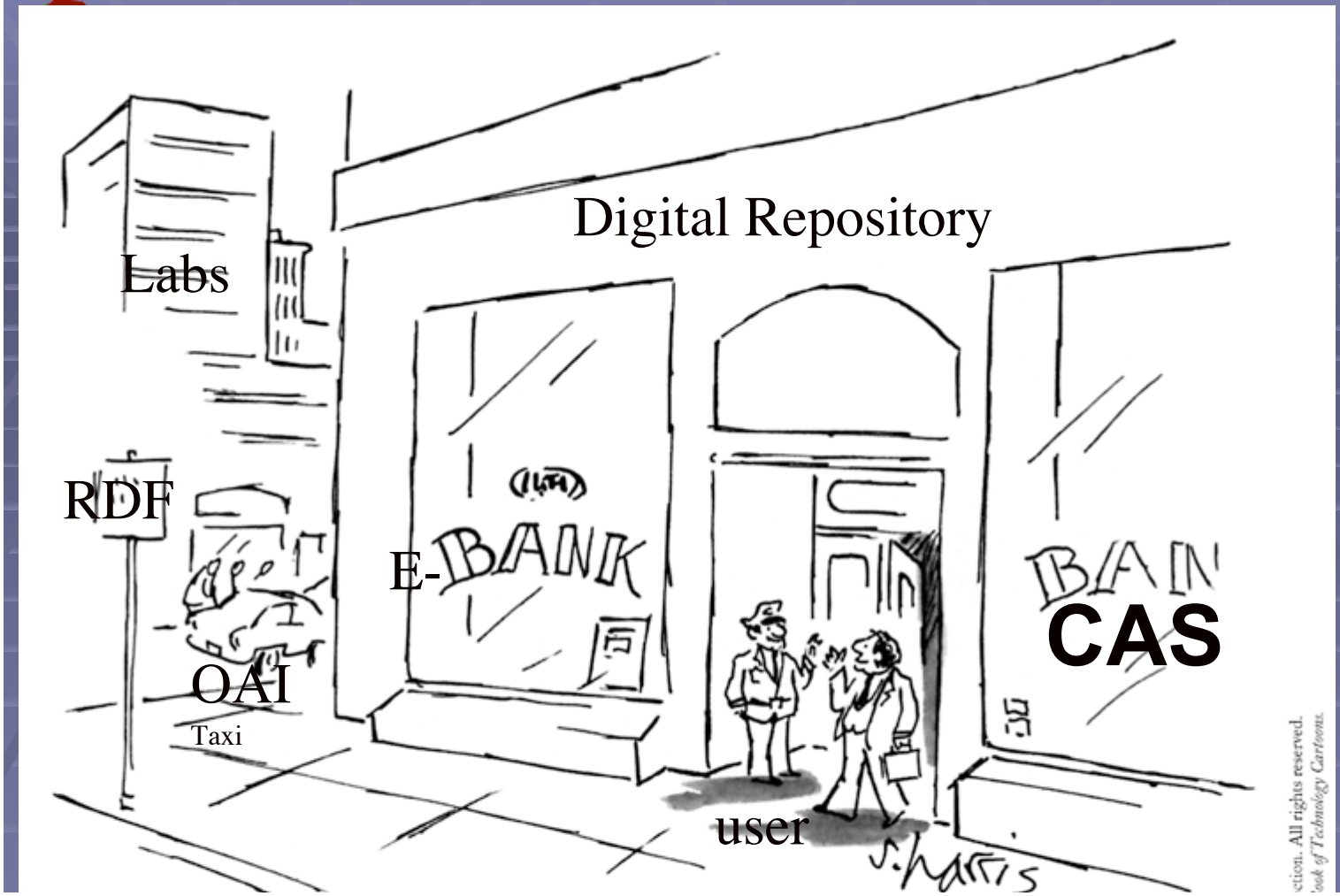
Link to data, follow links back to the raw data archive

Link to simulation, full simulation data archived in BioSimGrid

R4L

Access to information requires crossing administrative domains

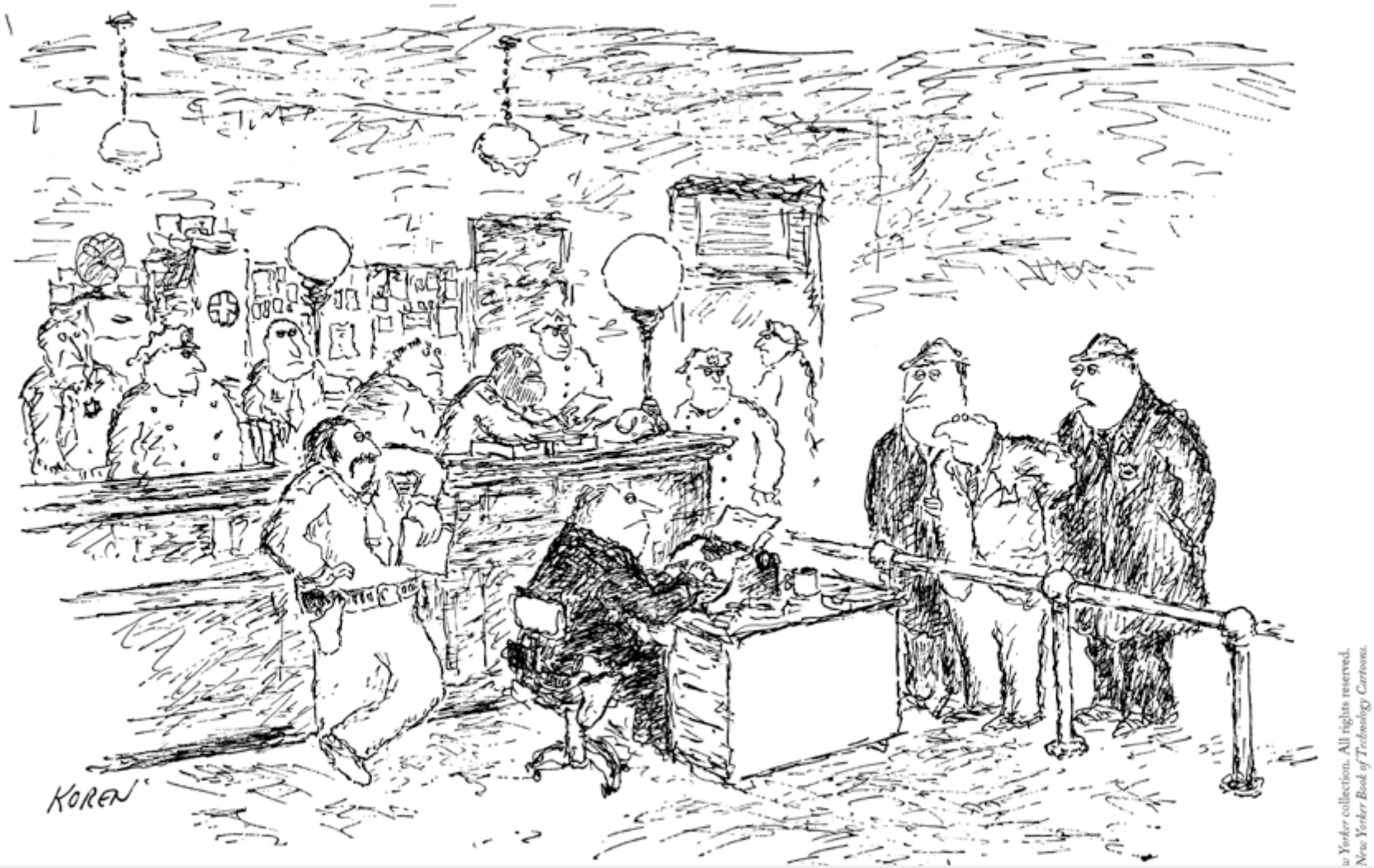




Subversive and furtive sharing & exploitation of data in virtual space

PERPETRATOR OF A DARING DAYLIGHT ILLEGAL ELECTRONIC TRANSFER OF FUNDS FLEEING THE SCENE OF THE CRIME

Data



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

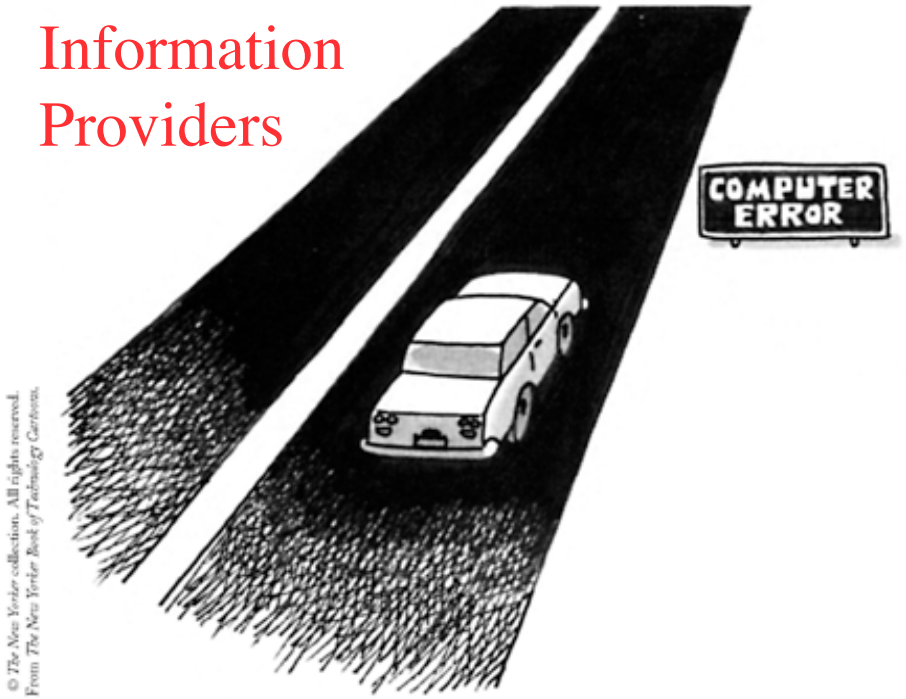
Plans

- 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

ZIEGLER

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

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