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”
CombeChem Data and Knowledge Cycle
End-to-End Management

Smart Laboratory

Smart HCl

Smart Storage

Smart Dissemination

Digital Model

Analysis

Synthesis

Knowledge

Literature

Goal

ACS 10 Sept 06

Cyberinfrastructure

Jeremy Frey 3
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
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. Need a usable digital lab book. Design by analogy to help Chemists and Computer Scientists work together.

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

Only some equipment is networked

ACS 10 Sept 06

Cyberinfrastructure
COSHH
leverage off things we already have to do

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

<table>
<thead>
<tr>
<th>SUBSTANCE NAME</th>
<th>PHYSICAL FORM</th>
<th>QUANTITY</th>
<th>NATURE OF HAZARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>liquid</td>
<td>1000 ml</td>
<td>None</td>
</tr>
<tr>
<td>Dextrose</td>
<td>solid</td>
<td>&lt;20 g</td>
<td>Possible irritation to eyes and skin</td>
</tr>
<tr>
<td>Caffeine</td>
<td>solid (tea)</td>
<td>&lt;1 g</td>
<td>Harmful &amp; nauseous, induce vomiting.</td>
</tr>
<tr>
<td>Miller</td>
<td>liquid</td>
<td>&lt;100 ml</td>
<td>No particular hazards</td>
</tr>
</tbody>
</table>

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

ACOS 10 Sept 06 Cyberinfrastructure Jeremy Frey 8
Sample of 4-flourinated biphenyl

Add Cool Reflux Butanone

Sample of K2CO3 Powder

Weigh 0.9031 g
Measure 40 ml
Add

Weigh 2.0719 g

Sample of Br11OCB

Reflux

Cool Water

Measure 30 ml

Liquid-liquid extraction DCM

Measure 3 of 40 ml

Dry MgSO4

Filter (Buchner)

Remove Solvent by Rotary Evaporation

Fuse Silica Column Chromatography Ether/Petrol Ratio

Butanone dried via silica column and measured into 100ml RB flask. Used 1ml extra solvent to wash out container.

Started reflux at 13.30. (Had to change heater stirrer) Only reflux for 45min, next step 14:15.

Inorganics dissolve 2 layers. Added brine ~20ml. Organics are yellow solution

Washed MgSO4 with DCM ~ 50ml

Plan

Observation Types

Weight - grammes
Measure - ml, drops
Annotate - text
Temperature - K, °C

Key

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

Ingredient List

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

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

**Sample of 4-flourinated biphenyl**

- Dissolve 4-flourinated biphenyl in butanone
- Add 0.931 g of Fluorinated biphenyl
- Weigh

**Add K2CO3**

- Add K2CO3 powder
- Reflux

**Heat at reflux for 1.5 hours**

- Measure Butanone: 40 ml
- Weigh 2.0719 g of K2CO3

**Notes**

- Started reflux at 13:30. (Had to change heater stir) Only reflux for 45 min, next step 14:15.
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.
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

Data Source

Translator Service

Message Broker

PDA

Mobile phone

Web Client

Archive Client

BLOG
Temperature – room, laser

Air Conditioning failed

SHG Lab environment recall tool
Wed 16 Mar 2005 00:00:18-Thu 17 Mar 2005 00:00:18

Temperature Data

KEY: Big Room, Wiring Box

Door & interlock, Motion Sensors

SHG Lab environment recall tool
Mon 24 Jan 2005 00:00:00-Sun 06 Feb 2005 00:00:00

Temperature Data

KEY: PIR 1 PIR2 PIR3 Light Interlock (High = Disabled) Door -> 27:1081 Door -> 27:1057

Other Sensors

ACS 10 Sept 06

Cyberinfrastructure

Jeremy Frey 15
Members of Dr Frey's non-linear laser spectroscopy group, working as part of the Combechem project, successfully tried out the system at conferences and the university bar.

The next step is to evolve the system so that a button on a mobile, scientists will be able to remotely control the conditions in the lab, like turning down the temperature.

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.

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

Memory leak in testing program!
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 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

National Crystallography Grid Service

No, the computers are up, We’re down
Security and trust for experiments and data

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

“On the Internet, nobody knows you’re a dog.”
National Crystallography Service – Sample Status

Viewing samples for M E Light (light@soton.ac.uk)

<table>
<thead>
<tr>
<th>NCS ID</th>
<th>Customer ID</th>
<th>Received</th>
<th>Collection</th>
<th>Status</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>04MEL0098</td>
<td>Grad test</td>
<td>2004-02-12</td>
<td>001</td>
<td>Successful</td>
<td>HEL file / Report</td>
</tr>
<tr>
<td>04MEL0093</td>
<td>mcl01</td>
<td>2004-02-06</td>
<td>001</td>
<td>Successful</td>
<td>HEL file / Report</td>
</tr>
<tr>
<td>04SRC0104</td>
<td>#13-123</td>
<td>2004-03-08</td>
<td>001</td>
<td>Next</td>
<td>Due at 00:00:00 (est)</td>
</tr>
<tr>
<td>04SRC0103</td>
<td>#12-01</td>
<td>2004-03-08</td>
<td>001</td>
<td>Failed (Referred)</td>
<td>Diffraction too weak</td>
</tr>
<tr>
<td>04SRC0105</td>
<td>HSF-HCI</td>
<td></td>
<td>002</td>
<td>Failed (No Further Action)</td>
<td>Crystals too small</td>
</tr>
</tbody>
</table>

X-Ray Diffractometer Images

Status Log
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

30,000,000

1.5,000,000

450,000

30,000,000
The eCrystals Data Archive

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

C41H40O5

InChI=1/C14H20O5/c1-2,4-14-13(3-1)13-11-9-16-7-5-15-8-17-10-12-19-14/h1-4H5-12H2

DOI: 10.5281/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

Final Result
04sc00831.cif 13k
04sc00831.cml 6k

Validation
04sc00831_checksum.htm 7k

Refinement
04sc00831.res 6k
04sc00831_xl.lst 34k

Solution
04sc00831.psp 5k
04sc00831_xl.lst 39k

Processing
04sc00831_hil 70.2k
04sc00831_hil1 10k
04sc00831_001.jpg 57k
04sc00831_n01.jpg 85k
04sc00831_n00.jpg 88k

Data Collection
04sc00831_crystal.jpg 17 k

Other Files
04sc00831.doc 78k
04sc00831.tcf.txt 155k

http://ecrystals.chem.soton.ac.uk
Access to the underlying data
4.2 Polarisation Dependence

The polarisation dependence of the SHG signal was recorded for bulk phenylalanine concentrations from 0.005 mol dm$^{-2}$ up to 0.8 mol dm$^{-2}$. 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 $I_x$, $I_y$, and $I_z$, where $P$, $S$, and $N$ correspond to the output harmonic polarisation $h=1$ (For Harmonic and $h=90^\circ$ for Short harmonic light). The conditions were optimised to avoid the photon fluorescence present with the SHG signal. A representative sample of five plots at different concentrations are shown in Figures 711.

 SVG “active” graphics

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

Link to simulation, full simulation data archived in BioSimGrid

Paper organized using RDF

R4L

Cyberinfrastructure
Need for a data archive in the laboratory

Not just the published spectra!

Jeremy Frey

ACS 10 Sept 06

Cyberinfrastructure

Preparation of CF₆H₅NC (PFB) (1)

Potassium cyanide (23.5 mg, 0.5 mmol) was ground to a powder using a pestle and mortar in a dry box. PBB (247.8 g, 0.5 mmol) and 20 mL of THF were then added, and the mixture was heated to reflux for 12 h. The solvent was removed in vacuo to leave a white foam which was washed with warm hexanes (3 mL) to give KCN/PBB as a white powder (0.40 g, 2.1%). This solid was stirred with triphenylphosphine (24.3 g, 0.945 mmol) in dichloromethane (15 mL) for 1 h. The solution was filtered to remove KCl, concentrated to ca. 5 mL and cooled to -26 °C to give an orange crystalline solid. Yield: 0.224 g (0.315 mmol, 62%) with mp 90-92°C (KCN, 95%).

1H NMR (CDCl₃, 20°C, 500 MHz): \( \delta = 7.55 (t, J = 8.2 Hz), 7.40 (d, J = 8.2 Hz), 7.25 (m, J = 7.7 Hz) \) ppm.

13C NMR (CDCl₃, 20°C, 125 MHz): \( \delta = 143.0 (m-C), 140.0 (p-C), 130.9 (p-C), 121.3 (q-C) \) ppm.

Preparation of CF₆(CF₃)₂O (PFB) (2)

Me₂SnNCBF₆₂ (0.60 g, 0.36 mmol) and Ph₃CCl (0.25 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 (0.84 mmol) and dichloromethane (30 mL) were added, and the mixture was stirred for 2 h. The solvent was then removed. The residue 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 purify the product from dichloromethane were not successful. IR (KBr): 2960 (m, CH₃), 1707 (m, C=O), 1300 (m, N=O).

1H NMR (CDCl₃, 20°C, 300 MHz): \( \delta = -118.72 \) ppm (CH₃).
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
All I am saying is that now is the time to develop the technology to deflect an asteroid.