Designing Data Repositories to Support Preservation and Publication for the Chemistry Community

Simon Coles
EPSRC National Crystallography Service
School of Chemistry
University of Southampton

Social Networks for Chemists

Google generation: new behaviour and approach

Sharing Rich Media

- Video + Paper = Pubcast

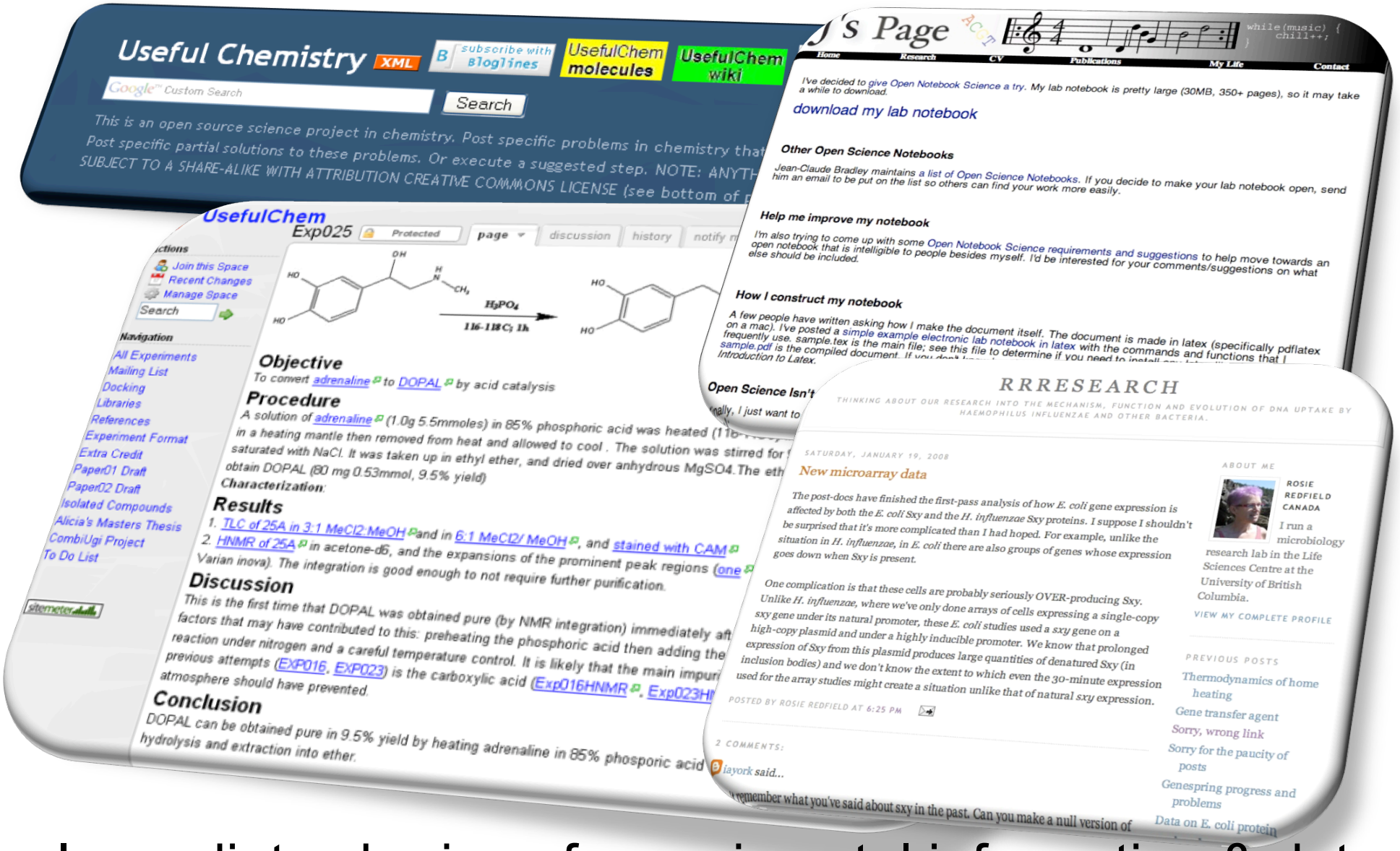
New Approaches to ‘Sharing Experiments’

- Specialised domain-oriented innovations

Formation of Open Communities

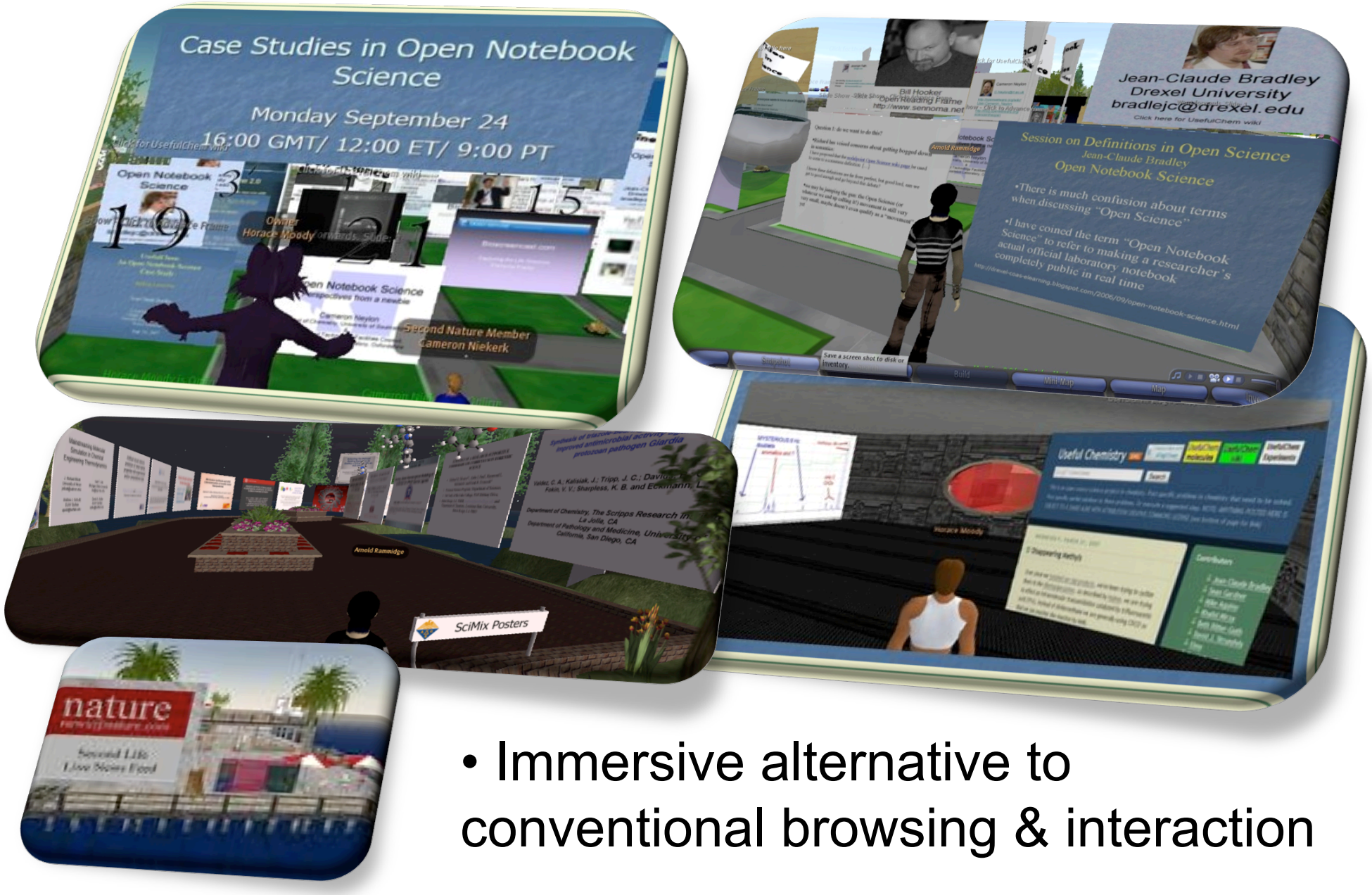
- New approaches surfacing and growing FAST

Open Notebook Science



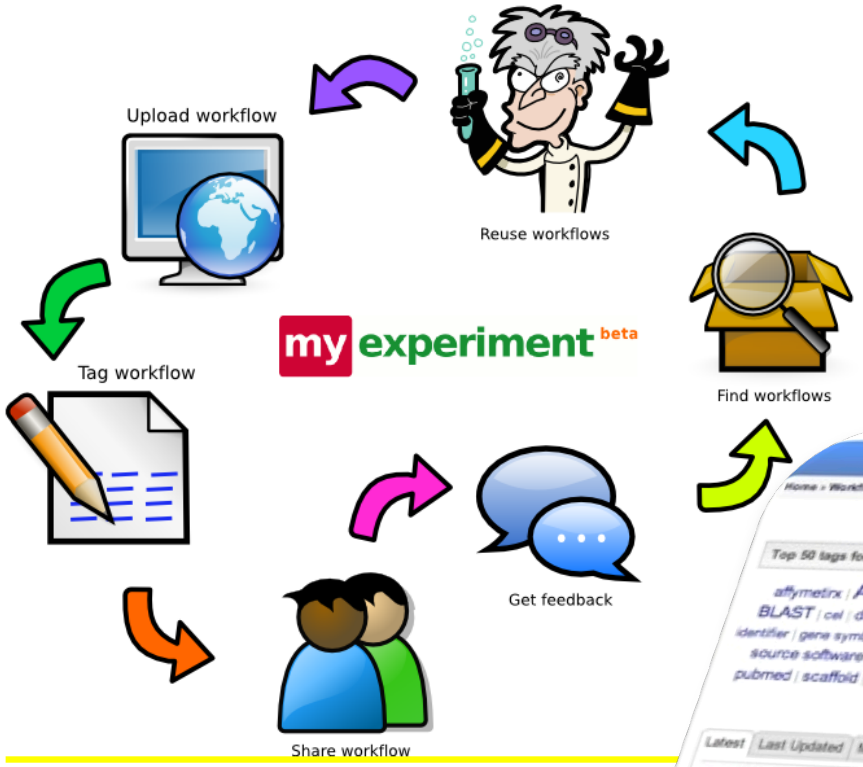
- Immediate sharing of experimental information & data

New Information Exchange Environments



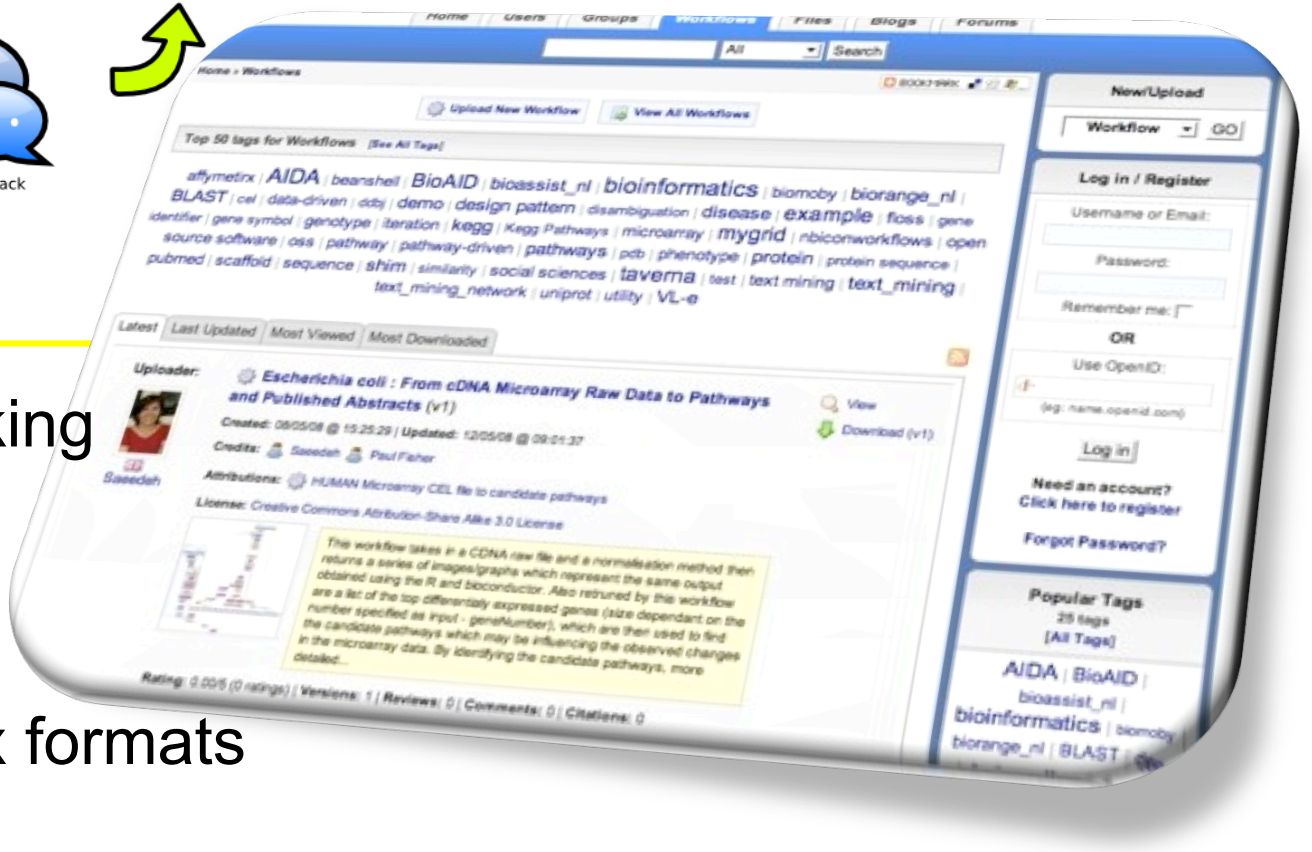
- Immersive alternative to conventional browsing & interaction

A Virtual Research Environment



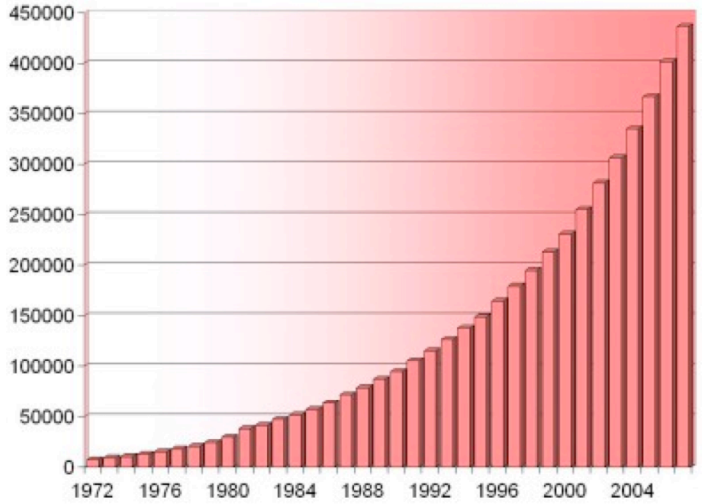
Adoption of "Web2.0" philosophy and methodology to promote exchange of research information: The New eScience?!

- Community networking
- Targeted audiences
- Open or closed
- Discoverable
- Multimedia/complex formats



Data Deluge

• 40 years ago a PhD student would determine about 3 crystal structures during the course of their study – this can now be easily achieved in a day

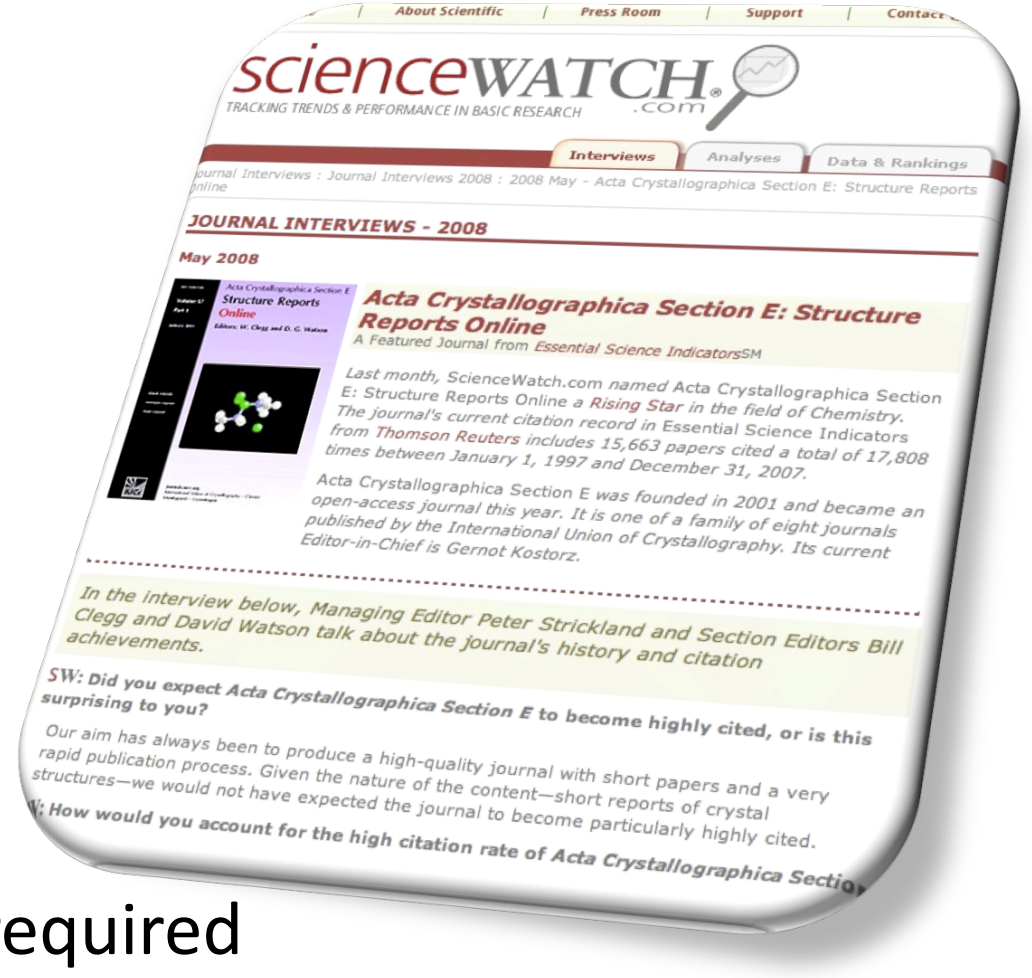


• There are approx. 30 million known chemical compounds
• Approx. 2 million crystal structures have been determined
• There are less than 0.5 million published crystal structures residing in (licensed) curated databases
• There are just a few thousand ‘open’ crystal structures

• The primary cause of this is the current data publication process, which is tied to journal articles and peer review

Data Publishing & Open Access

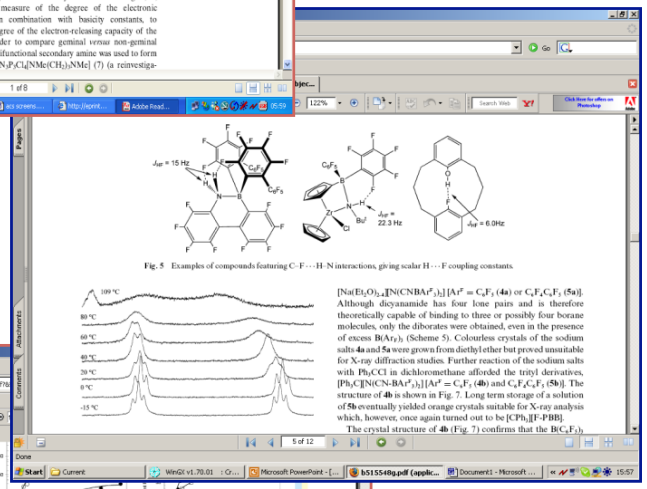
- Short communications
- Electronic only
- Rapid publication
- Open access (01/2008)
- Highly cited
- Written in CIF
- Freely available tools
- Still cannot keep up!
- Journal ‘article’ format required
- Not all crystal structures are of primary importance to the underlying chemistry: by-products / unexpected results / tracking reactions

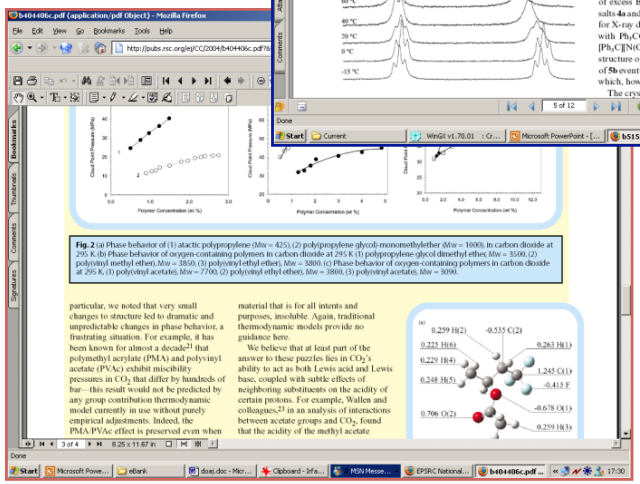


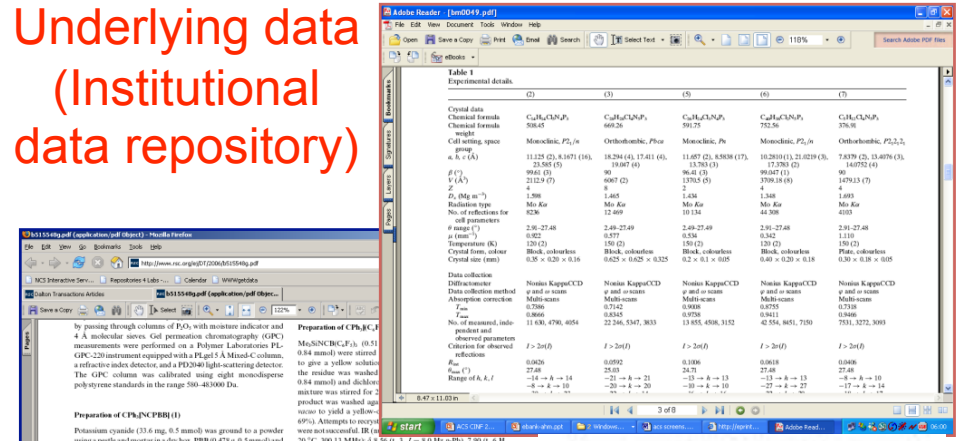
The Solution

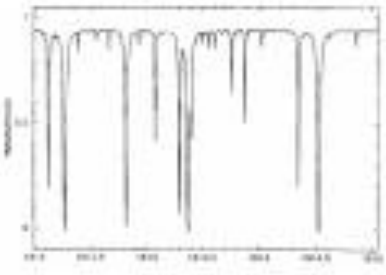
Intellect & Interpretation (Journal article, report, etc)

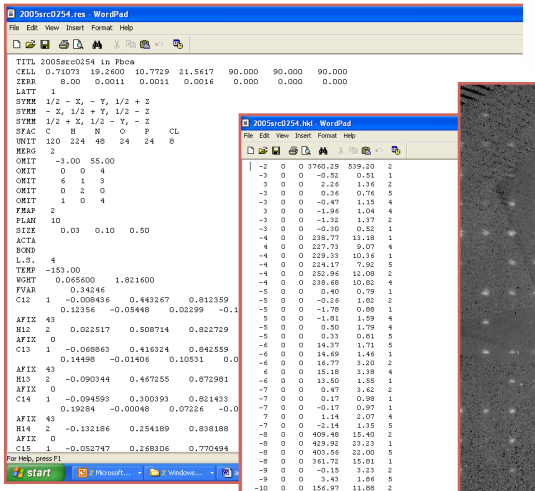
Underlying data (Institutional data repository)

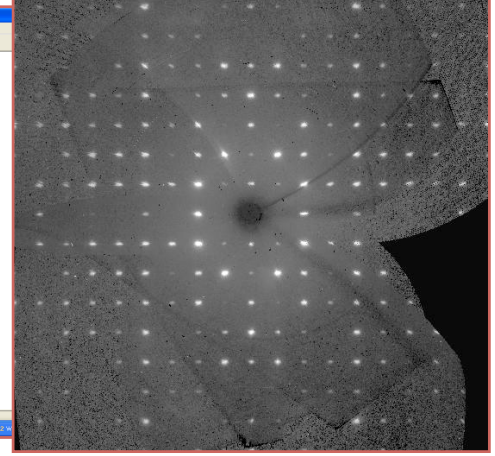




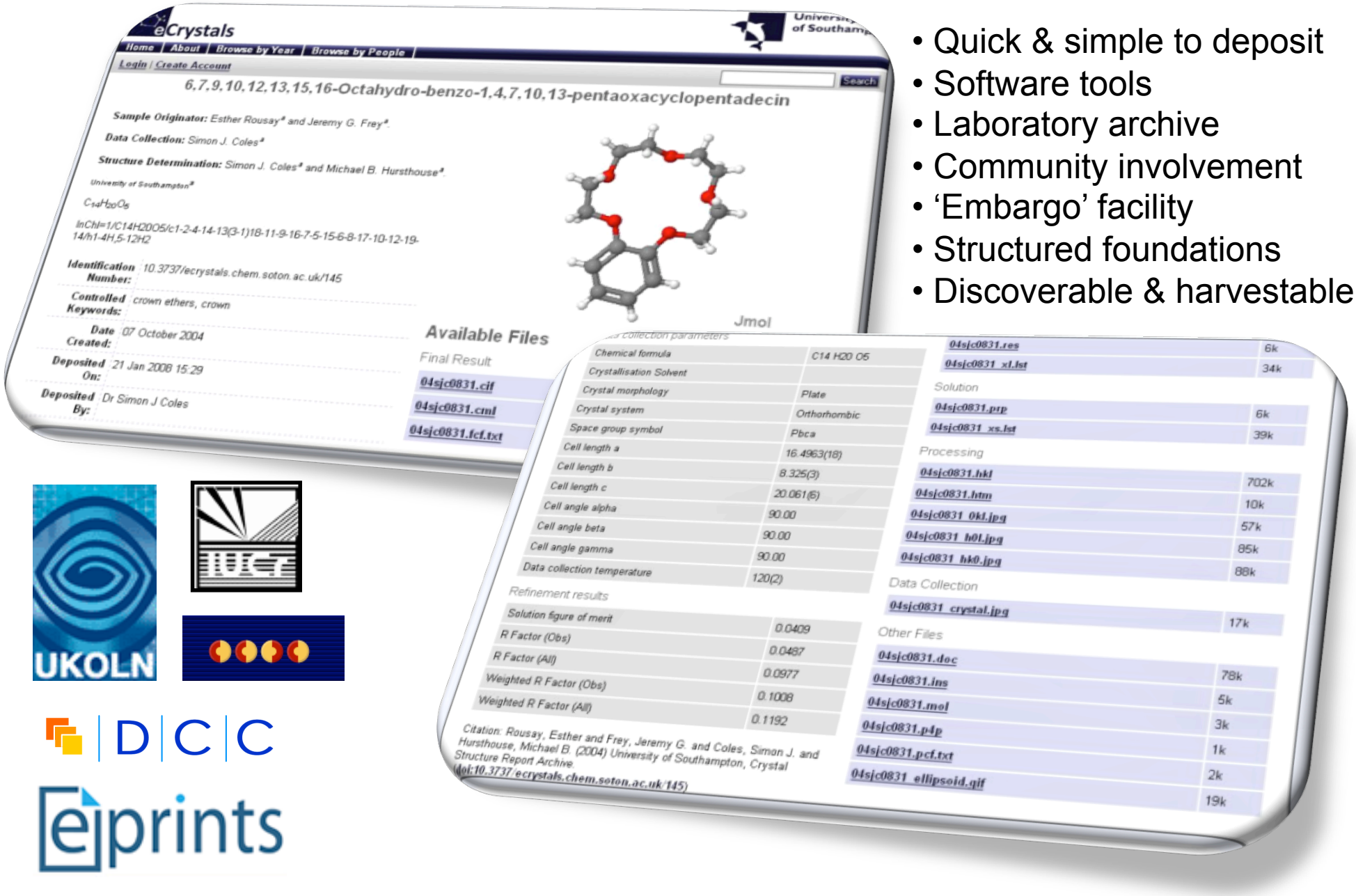








The eCrystals Data Repository



- Quick & simple to deposit
- Software tools
- Laboratory archive
- Community involvement
- ‘Embargo’ facility
- Structured foundations
- Discoverable & harvestable

UKOLN

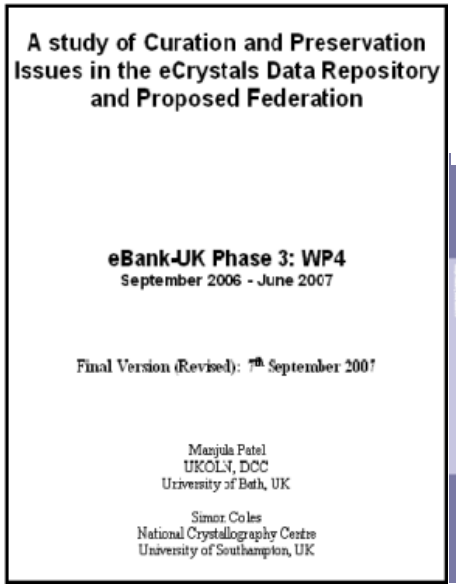
DCC

eprints

A Thorough Approach to Dissemination

- Using simple Dublin Core protocol (OAI-PMH)
    - Crystal structure
    - Title (Systematic IUPAC Name)
    - Authors
    - Affiliation
    - Creation Date
- Additional chemical information through Qualified Dublin Core
    - Empirical formula
    - International Chemical Identifier (InChI)
    - Compound Class & Keywords
- Specifies which ‘datasets’ are present in an entry
- Application Profile *http://www.ukoln.ac.uk/projects/ebank-uk/schemas/*
- DOI links *http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145*
- Rights & Citation *http://ecrystals.chem.soton.ac.uk/rights.html*

A Thorough Approach to Preservation



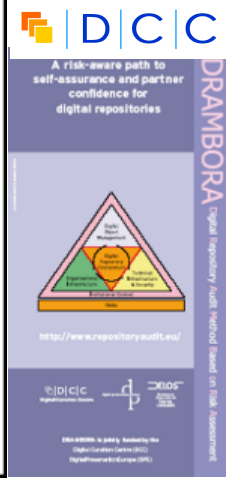
A study of Curation and Preservation Issues in the eCrystals Data Repository and Proposed Federation

eBank-UK Phase 3: WP4
September 2006 - June 2007

Final Version (Revised): 7th September 2007

Manjula Patel
UKOLN, DCC
University of Bath, UK

Simon Coles
National Crystallography Centre
University of Southampton, UK



D|C|C

A risk-aware path to self-assurance and partner confidence for digital repositories

DRAMBORA

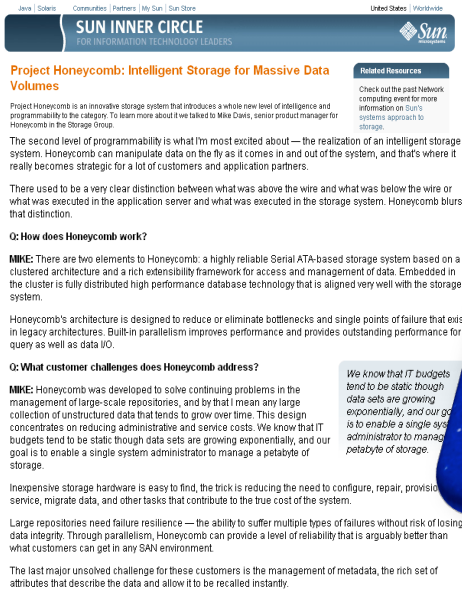
http://www.repositoryaudit.eu/

Keeping research data safe

Author: Neil Beagrie; Julia Chruszcz; and Brian Lavoie
Publication date: 12 May 2008
Publication Type(s): Reports
JISC theme(s): e-Research, Information environment

Executive summary

Research funders are implementing policies requiring researchers to submit data management, preservation or data sharing plans with their applications



SUN INNER CIRCLE
FOR INFORMATION TECHNOLOGY LEADERS

Project Honeycomb: Intelligent Storage for Massive Data Volumes

Related Resources

Project Honeycomb is an innovative storage system that introduces a whole new level of intelligence and programmability to the category. To learn more about it we talked to Mike Davis, senior product manager for Honeycomb in the Storage Group.

The second level of programmability is what I'm most excited about — the realization of an intelligent storage system. Honeycomb can manipulate data on the fly as it comes in and out of the system, and that's where it really becomes strategic for a lot of customers and application partners.

There used to be a very clear distinction between what was above the wire and what was below the wire or what was executed in the application server and what was executed in the storage system. Honeycomb blurs that distinction.

Q: How does Honeycomb work?

MIKE: There are two elements to Honeycomb: a highly reliable Serial ATA-based storage system based on a clustered architecture and a rich extensibility framework for access and management of data. Embedded in the cluster is fully distributed high performance database technology that is aligned very well with the storage system.

Honeycomb's architecture is designed to reduce or eliminate bottlenecks and single points of failure that exist in legacy architectures. Built-in parallelism improves performance and provides outstanding performance for query as well as data I/O.

Q: What customer challenges does Honeycomb address?

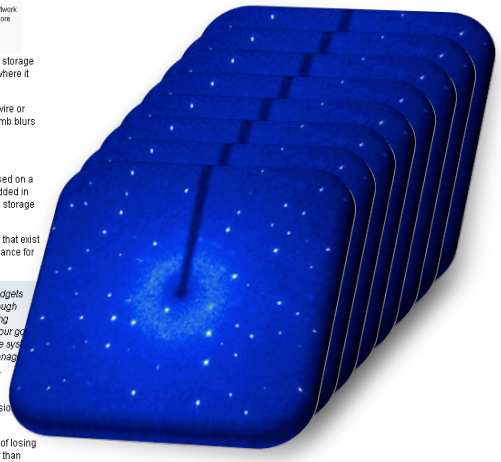
MIKE: Honeycomb was developed to solve continuing problems in the management of large-scale repositories, and by that I mean any large collection of unstructured data that tends to grow over time. This design concentrates on reducing administrative and service costs. We know that IT budgets tend to be static though data sets are growing exponentially, and our goal is to enable a single system administrator to manage a petabyte of storage.

We know that IT budgets tend to be static though data sets are growing exponentially, and our go is to enable a single sys administrator to manag petabyte of storage.

Inexpensive storage hardware is easy to find, the trick is reducing the need to configure, repair, provisio service, migrate data, and other tasks that contribute to the true cost of the system.

Large repositories need failure resilience — the ability to suffer multiple types of failures without risk of losing data integrity. Through parallelism, Honeycomb can provide a level of reliability that is arguably better than what customers can get in any SAN environment.

The last major unsolved challenge for these customers is the management of metadata, the rich set of attributes that describe the data and allow it to be recalled instantly.



KEEPING RESEARCH DATA SAFE

A COST MODEL AND GUIDANCE FOR UK UNIVERSITIES

Neil Beagrie, Julia Chruszcz, and Brian Lavoie

with case studies contributed by the Universities of Cambridge, Southampton, King's College London, and the Archaeology Data Service University of York.

Final Report - April 2008

Prepared by:

Charles Beagrie Limited

www.beagrie.com

A study funded by

JISC

Scaling Up: A Community Solution

Interviews & analysis

Synthesis: IR Policy & Practice, Laboratory Practice & Workflows, Technical Interoperability & Standards, Metadata Schema & Application Profiles, Semantic Interoperability, Data Citation, Identifiers & Linking, Federation Architectures & Third Party Services, Rights & Licensing, Data Quality & Validation, Preservation, Curation & Sustainability

Recommendations, commentary

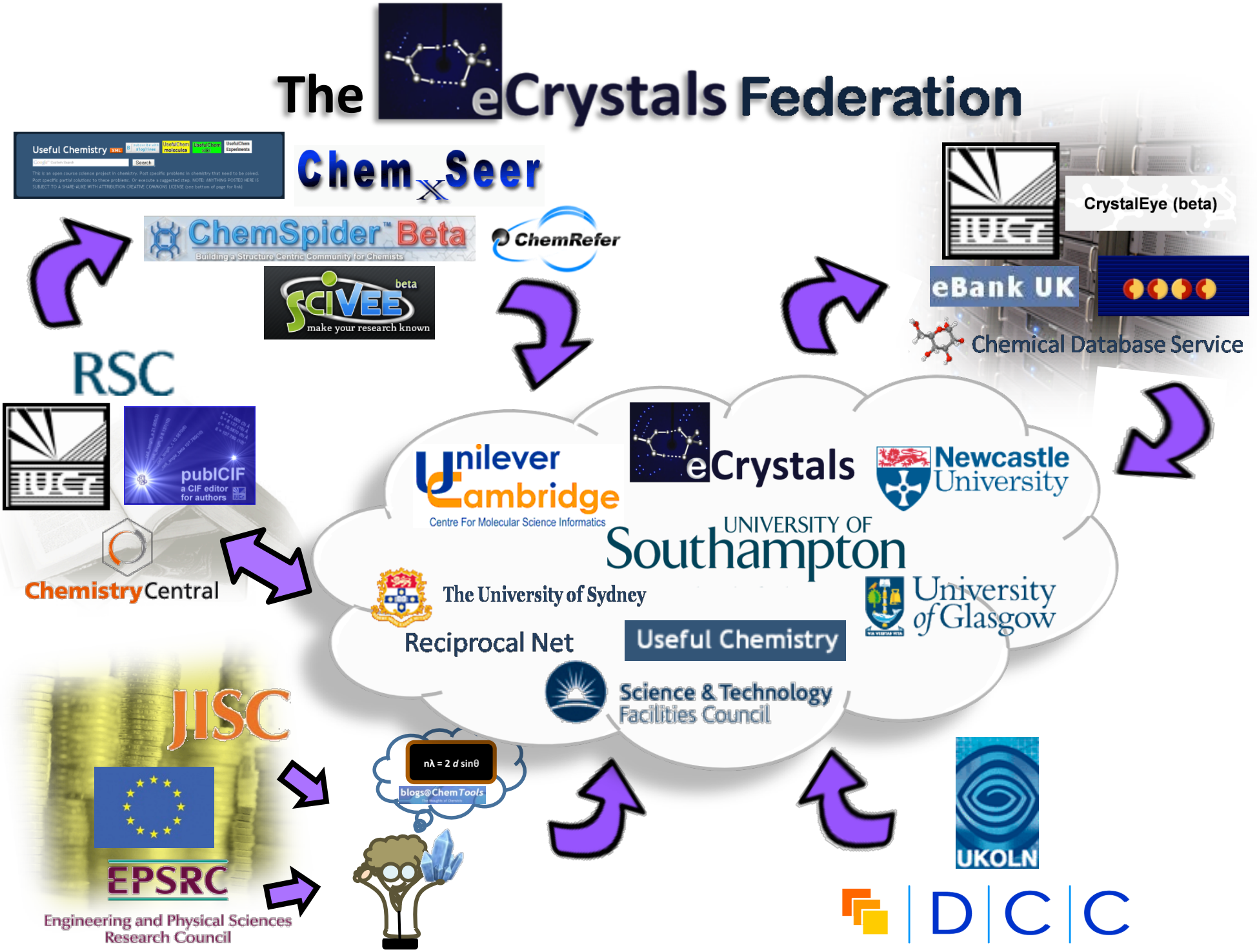
Matters Arising: Diverse lab practice, LIMS and proprietary formats, Data policy should reflect lab practice & institutional model, Data quality criteria/validation, "Prior publication" problem, We need scalable assignment of "terms" for data discovery, No discipline preservation model

Scaling Up: Towards a Federation of Crystallography Data Repositories

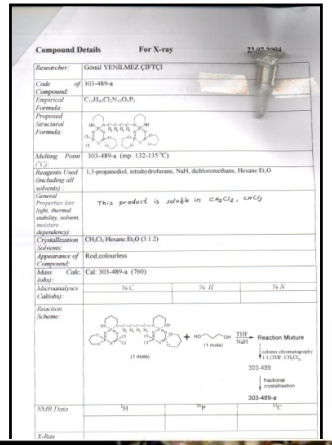
Document details

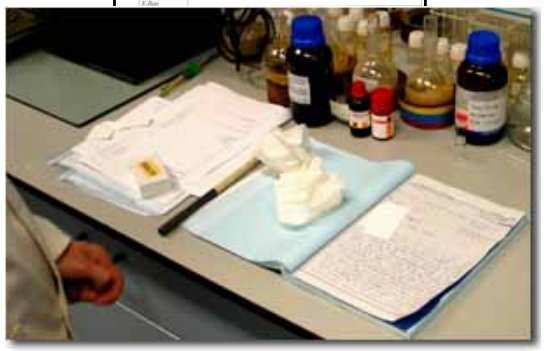
Author:	Liz Lyon, Simon Coles, Monica Duke, Traugott Koch
Date:	12th May 2008
Version:	1.0 Final
Document Name:	ebank-phase3-report-final.doc
Notes:	

http://wiki.ecrystals.chem.soton.ac.uk


The eCrystals Federation
Useful Chemistry
Chem Seer
ChemSpider Beta
Building a Structure Centric Community for Chemists
ChemRefer
SciVEE beta
make your research known
CrystalEye (beta)
eBank UK
Chemical Database Service
RSC
publCIF
a CIF editor for authors
ChemistryCentral
Unilever Cambridge
Centre For Molecular Science Informatics
eCrystals
Newcastle University
UNIVERSITY OF Southampton
The University of Sydney
University of Glasgow
Reciprocal Net
Useful Chemistry
Science & Technology Facilities Council
JISC
EPSRC
Engineering and Physical Sciences Research Council
nλ = 2 d sinθ
blogs@ChemTools
UKOLN
DCC

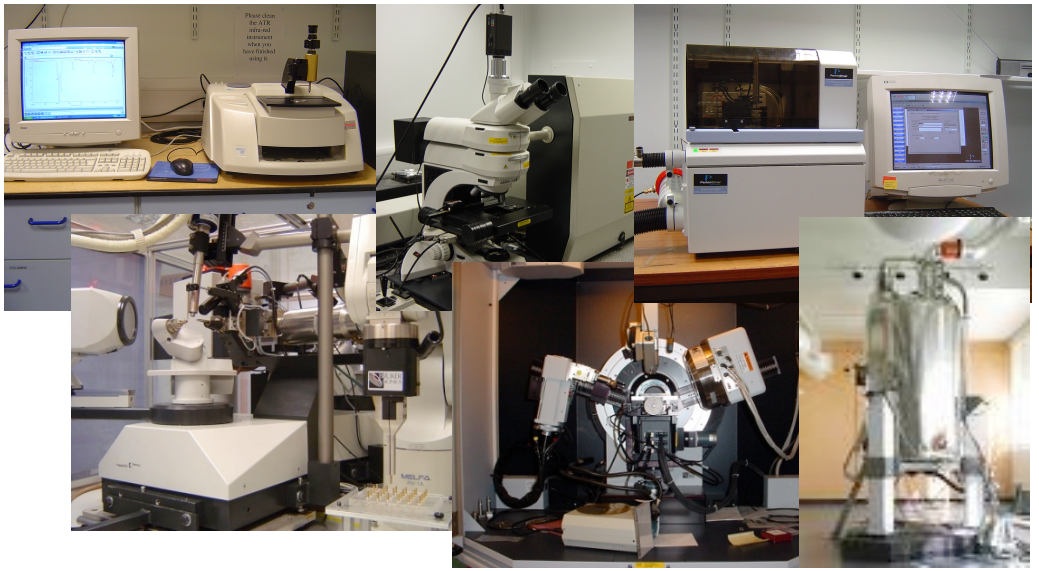
General Chemistry Issues: Data Generation

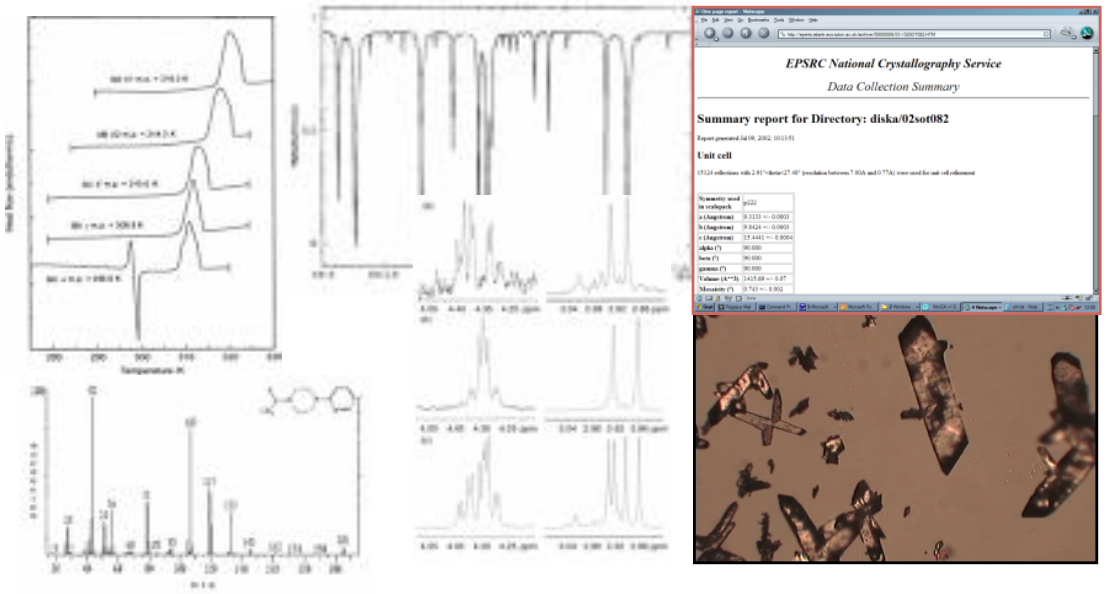




Synthesis

Characterisation





Shortfalls in Data Management

"Data from experiments conducted as recently as six months ago might be suddenly deemed important, but those researchers may never find those numbers – or if they did might not know what those numbers meant"

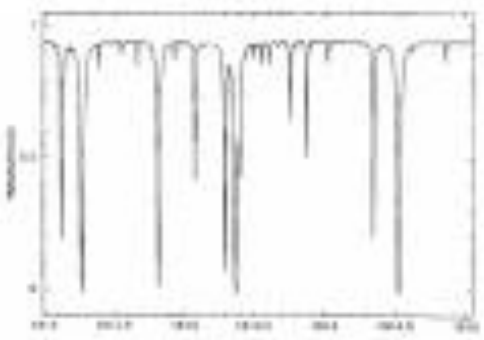
"Lost in some research assistant's computer, the data are often irretrievable or an undecipherable string of digits"

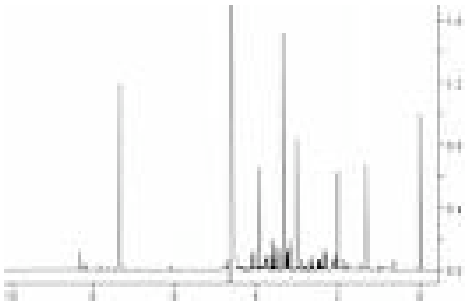
"To vet experiments, correct errors, or find new breakthroughs, scientists desperately need better ways to store and retrieve research data"

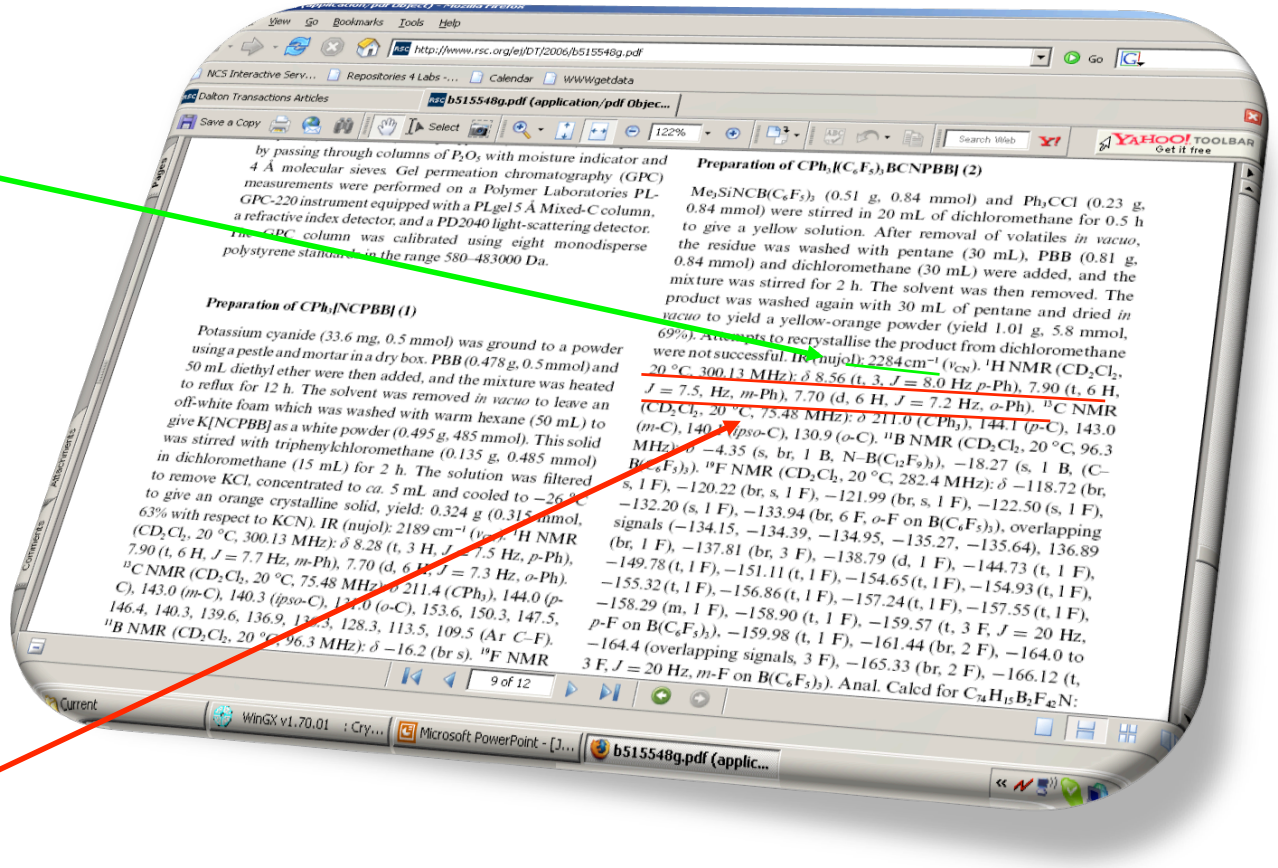
"Data from Big Science is … easier to handle, understand and archive. Small Science is horribly heterogeneous and far more vast. In time Small Science will generate 2-3 times more data than Big Science."

'Lost in a Sea of Science Data' S.Carlson, The Chronicle of Higher Education (23/06/2006)

Analysis Data Publication & Information Loss



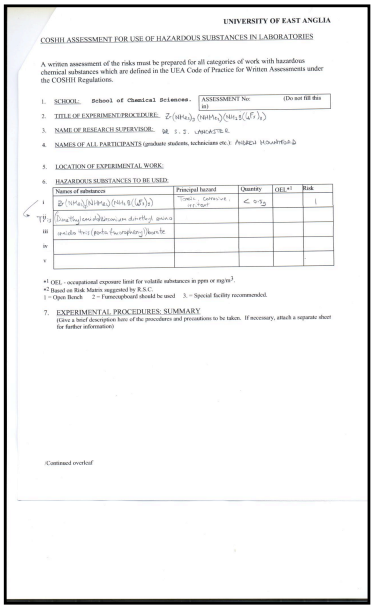




Spectroscopic analysis is often performed to ensure a reaction is proceeding according to plan – as a result <5% are published (via a process with heavy information loss)

A New Kind of Electronic Lab Notebook

the Smart Tea Project



UNIVERSITY OF EAST ANGLIA

COSHH ASSESSMENT FOR USE OF HAZARDOUS SUBSTANCES IN LABORATORIES

A written assessment of the risks must be prepared for all categories of work with hazardous chemical substances which are defined in the UEA Code of Practice for Written Assessments under the COSHH Regulations.

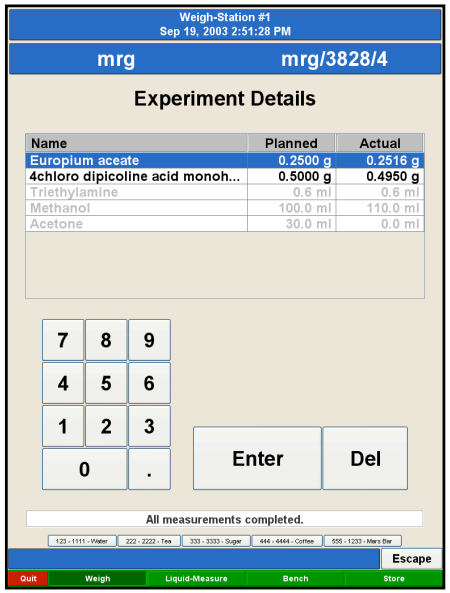
1. SCHOOL: School of Chemical Sciences. ASSESSMENT No: (Do not fill this in)
2. TITLE OF EXPERIMENT/PROCEDURE:
3. NAME OF RESEARCH SUPERVISOR:
4. NAMES OF ALL PARTICIPANTS (graduate students, technicians etc.):
5. LOCATION OF EXPERIMENTAL WORK:
6. HAZARDOUS SUBSTANCES TO BE USED:

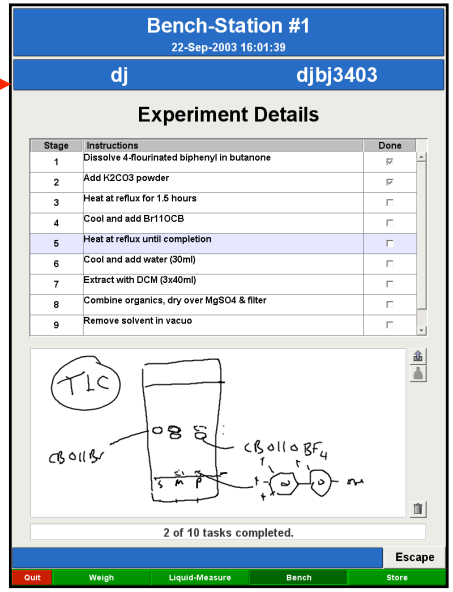
	Names of substances	Principal hazard	Quantity	OEL *1	Risk
i					
ii					
iii					
iv					
v					

*1 OEL - occupational exposure limit for volatile substances in ppm or mg/m³.
*2 Based on Risk Matrix suggested by R.S.C.
1 = Open Bench 2 = Fumecupboard should be used 3. = Special facility recommended.

7. EXPERIMENTAL PROCEDURES: SUMMARY
(Give a brief description here of the procedures and precautions to be taken. If necessary, attach a separate sheet for further information)

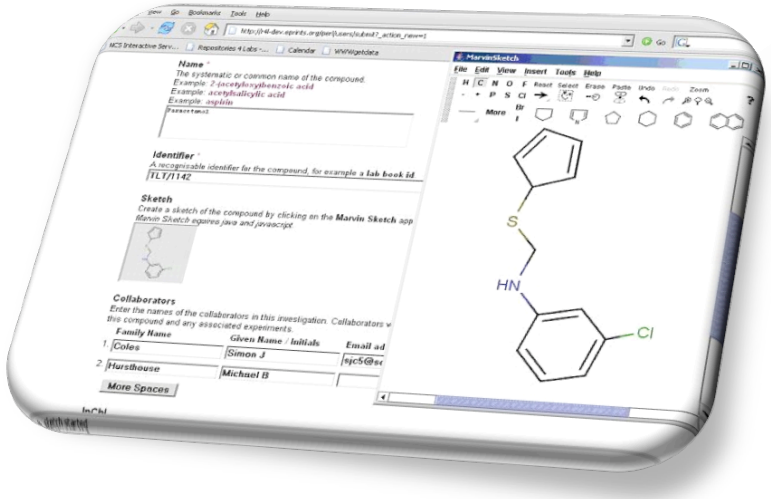
/Continued overleaf





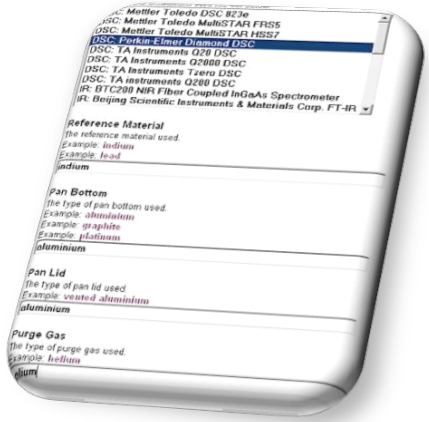
A General Chemistry Laboratory Repository

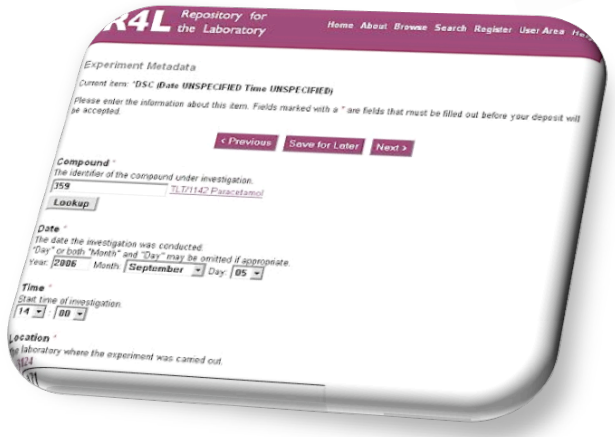
Create new compound
(parent record)



Add new experiment type

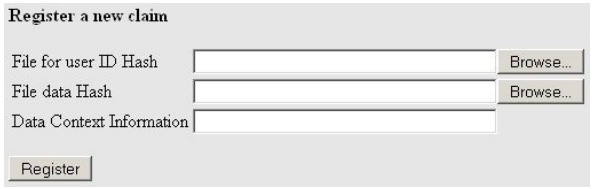
Add metadata and
upload data files

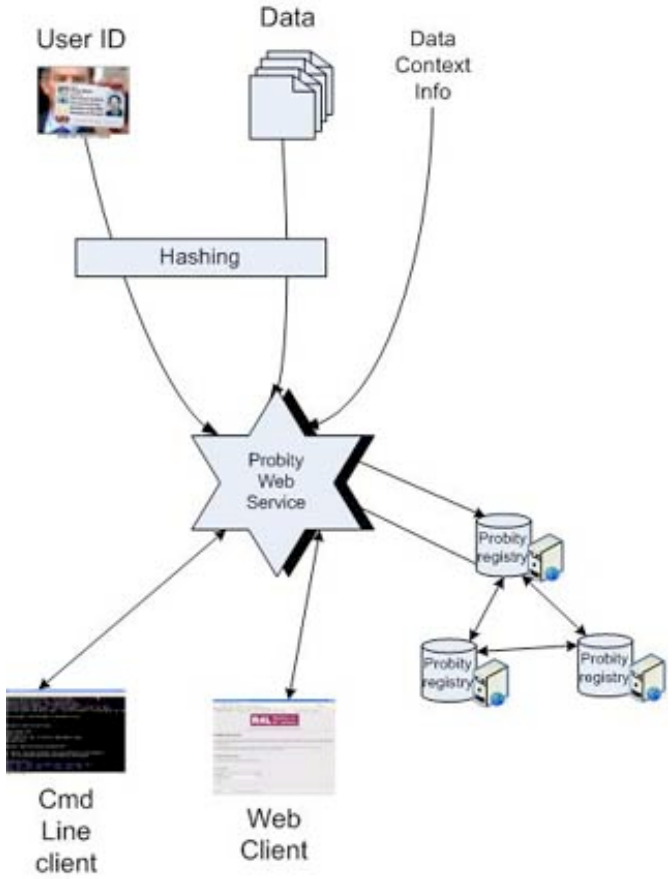




A General Chemistry Laboratory Repository

- Probity: A process to assert originality of a data record

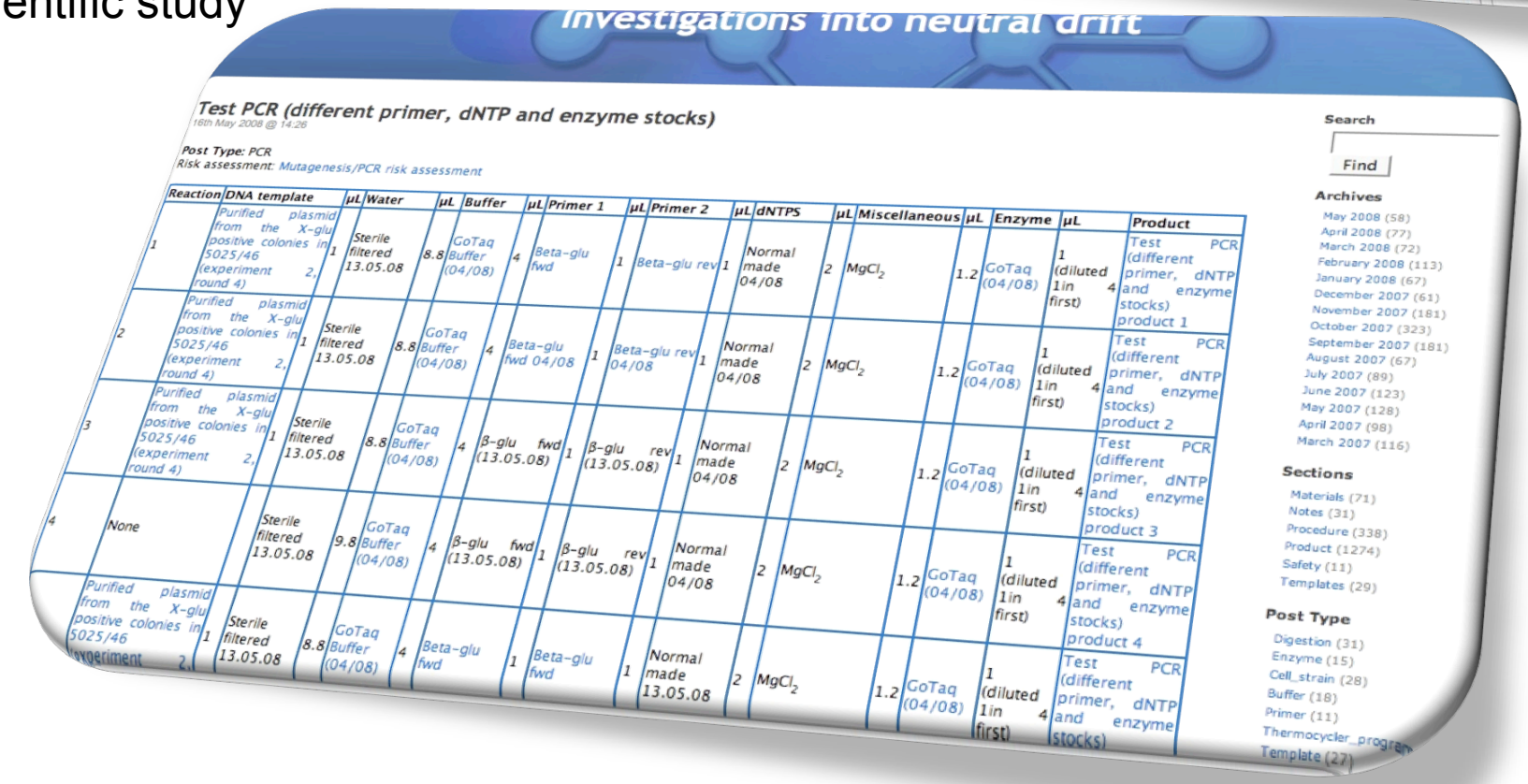




Analysis & Discussion: Blogging Experiments

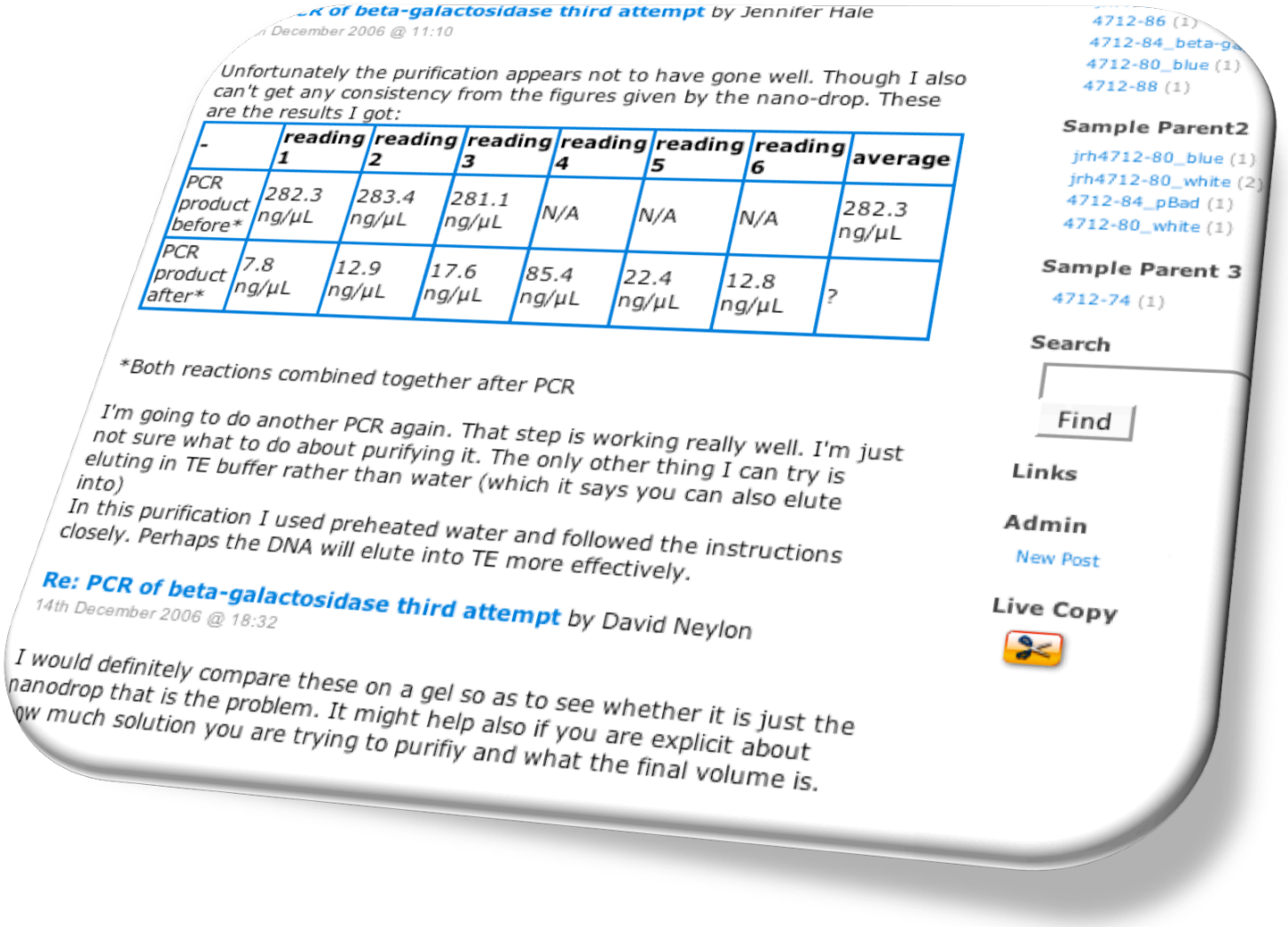
A repository can…

- Allow one to *put*, *store* and *get*
- Provide *search* and *browse* functionality
- NOT provide the *presentation* and *discussion* functions essential to working up a scientific study

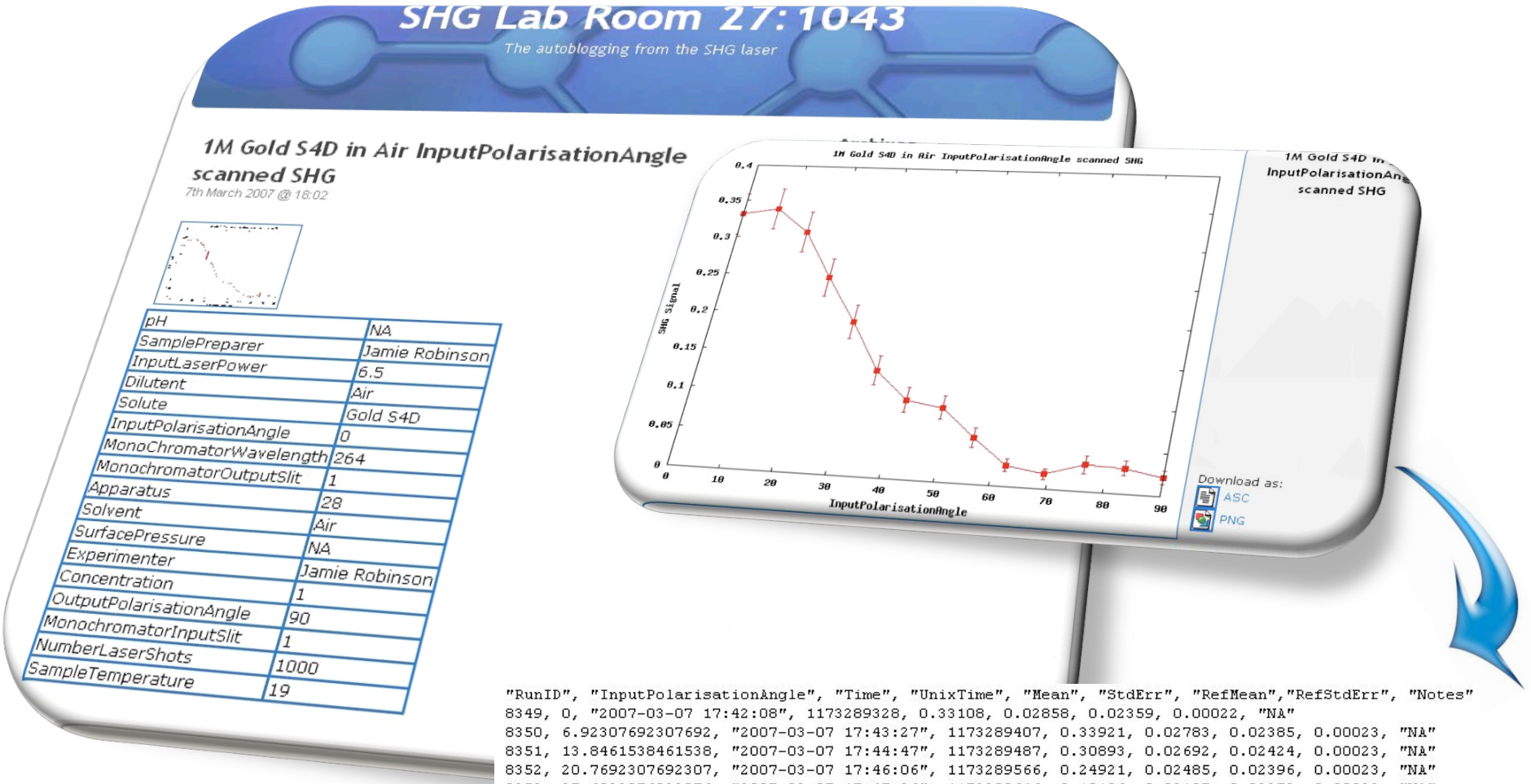


Facilitating Research

- Enables ‘geographically distributed collaborative research’
- Can be open or private
- A useful approach for sharing ‘failed’ experiments?



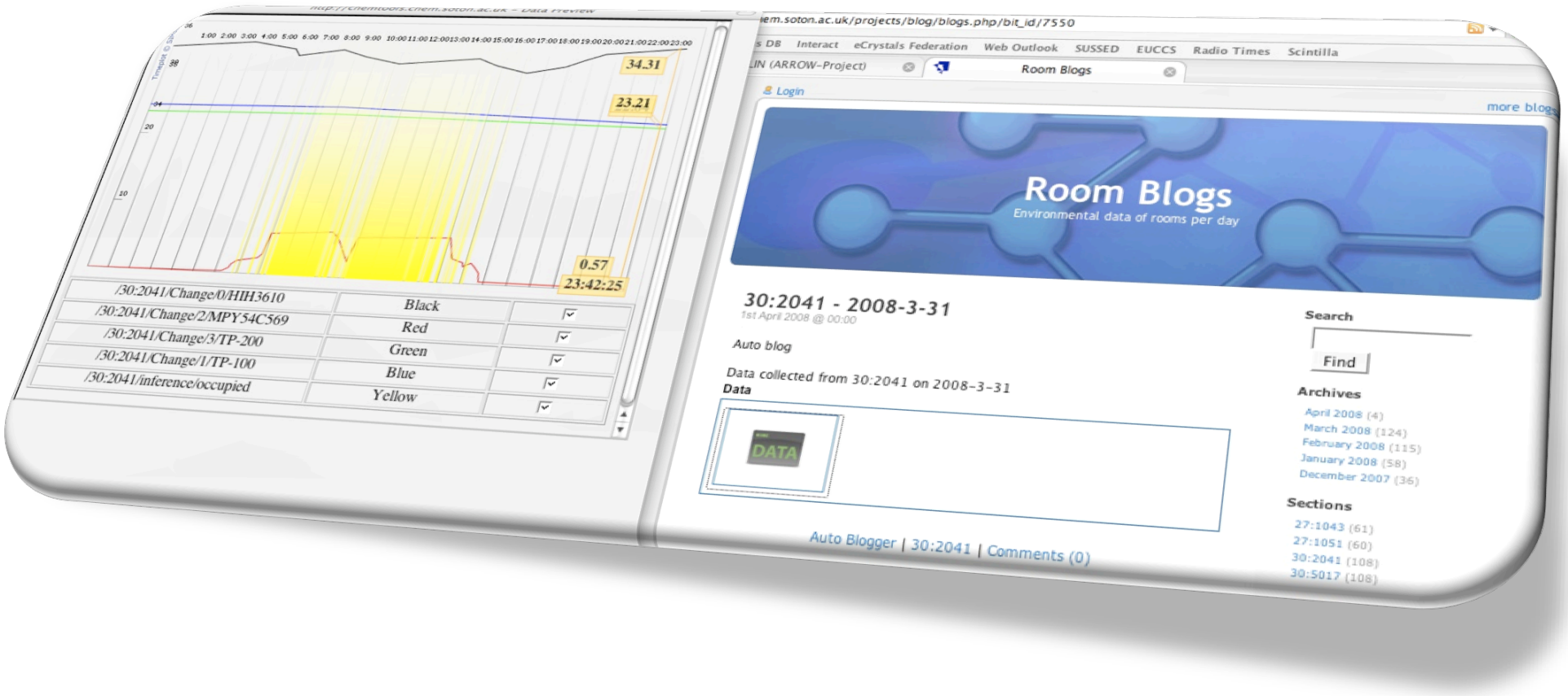
Automatic Blogging by Machines




```
"RunID", "InputPolarisationAngle", "Time", "UnixTime", "Mean", "StdErr", "RefMean","RefStdErr", "Notes"
8349, 0, "2007-03-07 17:42:08", 1173289328, 0.33108, 0.02858, 0.02359, 0.00022, "NA"
8350, 6.92307692307692, "2007-03-07 17:43:27", 1173289407, 0.33921, 0.02783, 0.02385, 0.00023, "NA"
8351, 13.8461538461538, "2007-03-07 17:44:47", 1173289487, 0.30893, 0.02692, 0.02424, 0.00023, "NA"
8352, 20.7692307692307, "2007-03-07 17:46:06", 1173289566, 0.24921, 0.02485, 0.02396, 0.00023, "NA"
8353, 27.6923076923076, "2007-03-07 17:47:26", 1173289646, 0.19186, 0.02197, 0.02372, 0.00022, "NA"
8354, 34.6153846153845, "2007-03-07 17:48:45", 1173289725, 0.13147, 0.01842, 0.02377, 0.00022, "NA"
8355, 41.5384615384614, "2007-03-07 17:50:05", 1173289805, 0.09684, 0.01536, 0.02407, 0.00023, "NA"
8356, 48.4615384615383, "2007-03-07 17:51:24", 1173289884, 0.08984, 0.01470, 0.02406, 0.00023, "NA"
8357, 55.3846153846152, "2007-03-07 17:52:43", 1173289963, 0.05578, 0.01183, 0.02456, 0.00023, "NA"
8358, 62.3076923076921, "2007-03-07 17:54:03", 1173290043, 0.02520, 0.00770, 0.02401, 0.00023, "NA"
8359, 69.230769230769, "2007-03-07 17:55:22", 1173290122, 0.01750, 0.00661, 0.02438, 0.00023, "NA"
8360, 76.1538461538459, "2007-03-07 17:56:41", 1173290201, 0.03136, 0.00998, 0.02397, 0.00023, "NA"
8361, 83.0769230769228, "2007-03-07 17:58:01", 1173290281, 0.03012, 0.00846, 0.02393, 0.00023, "NA"
8362, 89.9999999999997, "2007-03-07 17:59:20", 1173290360, 0.02174, 0.00789, 0.02383, 0.00023, "NA"
```

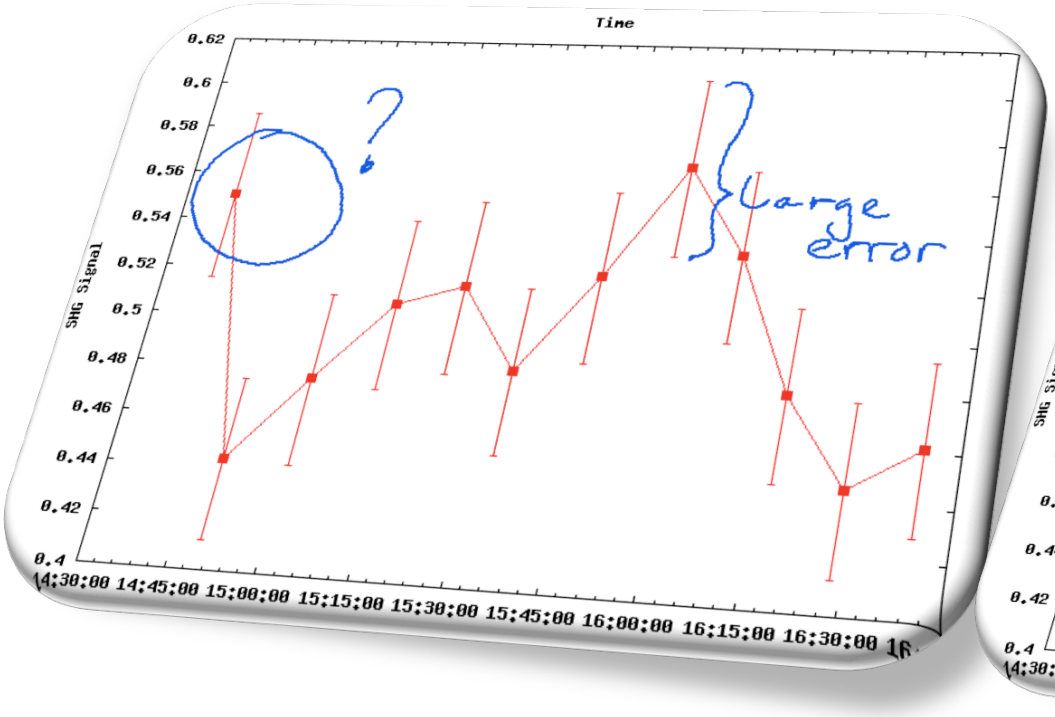
Automatic Blogging by Sensors

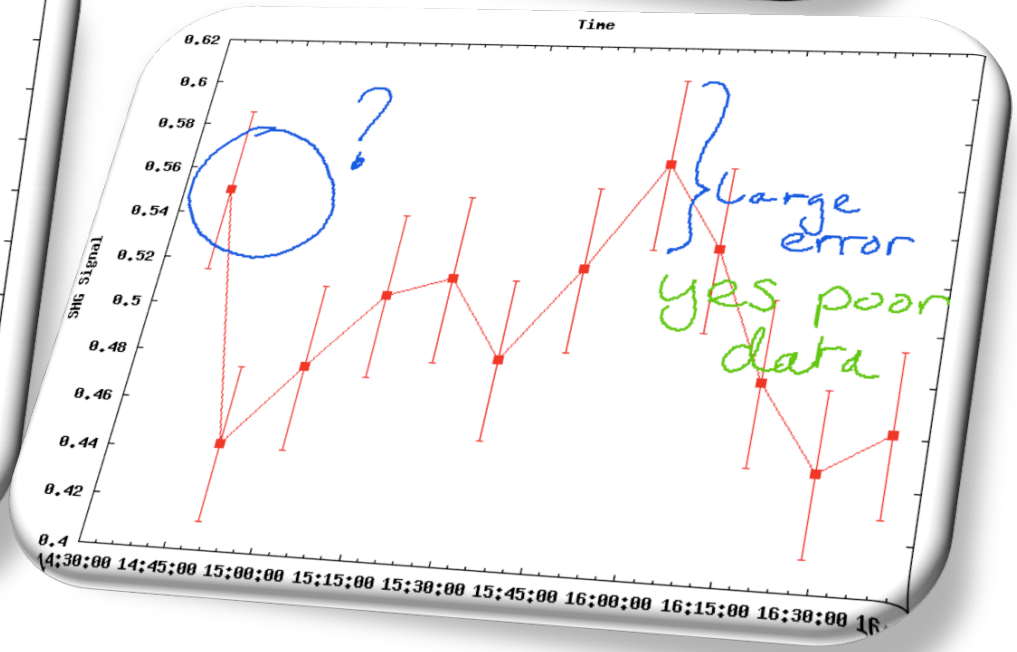
- Continuous log of ‘environmental’ conditions in a laboratory
- Instant detection of erroneous events
- Correlate with inconsistencies in datasets



Comments and Collaborative Tools

- Annotation tools allow comments and foster collaboration and / or communication
- Need for more advanced Blog tools / technology around data





Packaging and Interoperability

- New moves in Digital Libraries community to enable distributed repositories to exchange content
- OAI-ORE (Open Archives Initiative – Object Reuse and Exchange)
- http://www.openarchives.org/ore/
- Describes an aggregation of objects in an exchangeable format
- Microsoft funded eChemistry testbed project

Los Alamos NATIONAL LABORATORY

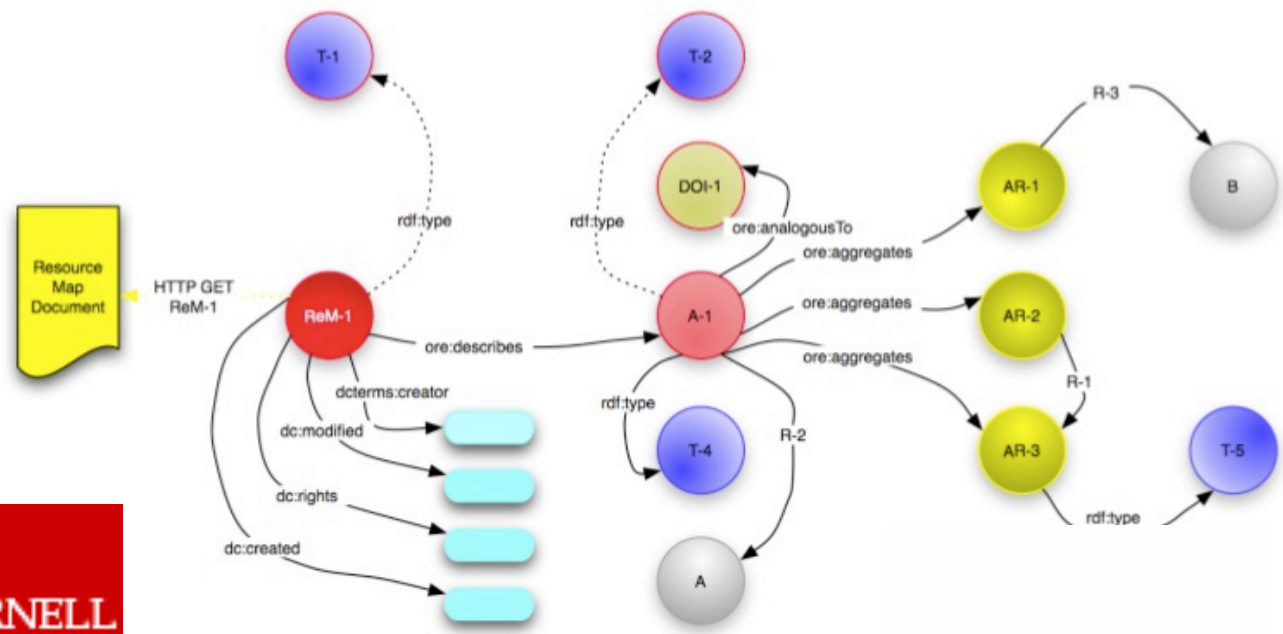
Indiana University School of informatics

ChemXSeer

PubChem

Unilever Cambridge Centre For Molecular Science Informatics

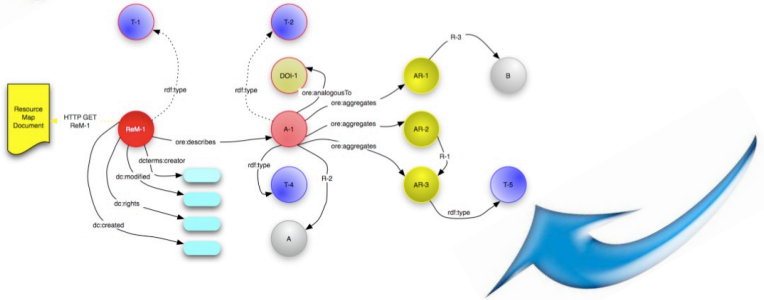
CORNELL

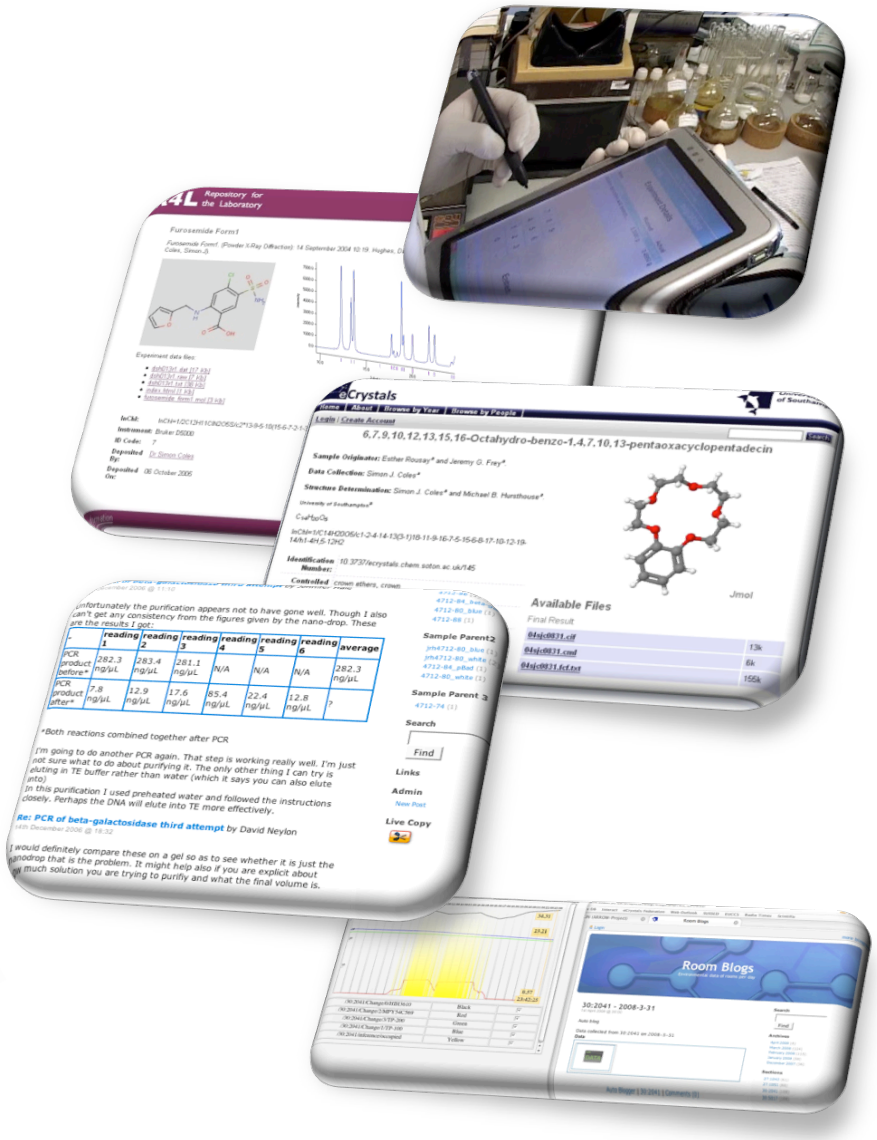


Microsoft

Towards a New Model for Chemical Information Exchange

myexperiment





A solid foundation for Open/Self-Publishing of Chemistry Data???

Thanks to:

• Jeremy Frey, Andrew Milsted, Richard Stephenson, Cameron Neylon, Jamie Robinson, Steven Wilson

• Dave DeRoure, Les Carr, Chris Gutteridge, Tim Myles-Board, Arouna Woukei

• Liz Lyon, Rachel Heery, Monica Duke, Michael Day, Traugott Koch, Manjula Patel