UNIVERSITY OF Southampton
School of Chemistry

eCrystals

The eCrystals Model for Gathering and 'Publishing' Crystallographic Data

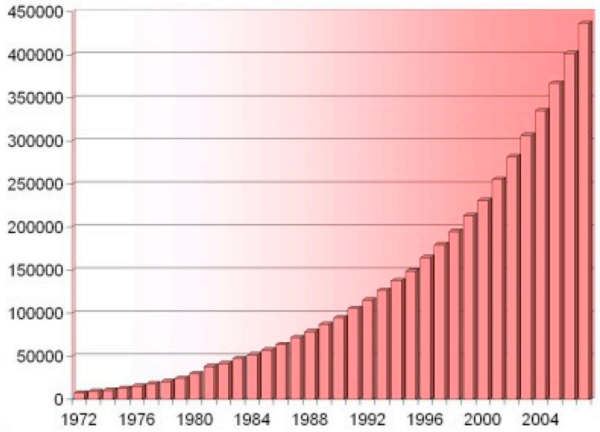
http://wiki.ecrystals.chem.soton.ac.uk

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

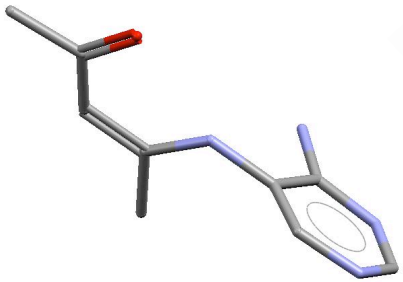
New Approaches to ‘Publishing’, ‘Sharing’ & ‘Gathering’ Data

Data Deluge

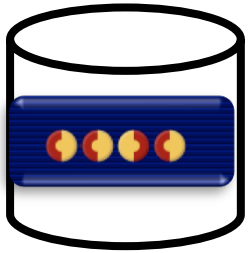
• 40 years ago a PhD student would determine about 3 crystal structures for their thesis – this can now be easily achieved in a day



35 million



2.5 million

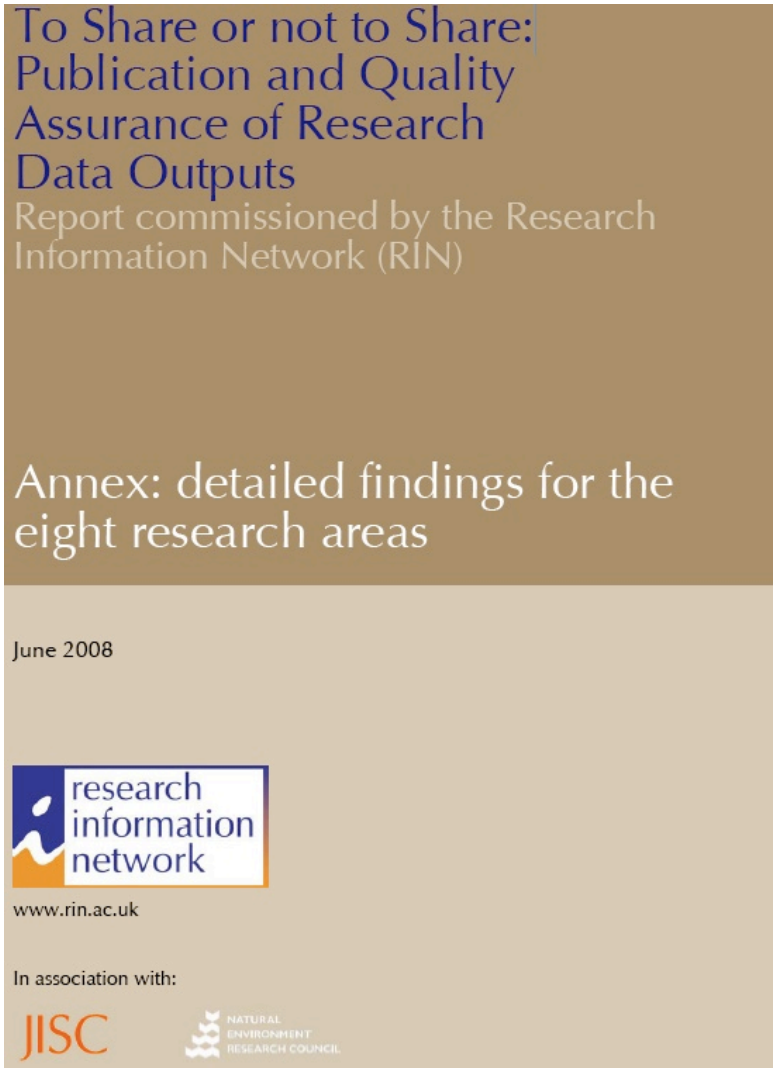


0.5 million

'Few thousand'

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

Current (Open?) Practice in Crystallography



CHEMICAL CRYSTALLOGRAPHY

Overview

Crystallography is the most-used unambiguous technique for identifying the structures of chemicals. This structural characterisation is done either for basic research purposes or as a service for other chemists who do not have their own crystallographic facility.

The chemical crystallography community has long had a relatively organised approach to data. Crystallographic data are highly-structured. Any heterogeneity resides in the instrumentation and the software this uses rather than in the data outputs. Outputs are in 3 or 4 possible formats but there is a *de facto* standard adopted by the bulk of the community, the CIF (Crystallographic Information File), and a standard protocol for making data available to others. Probably 95+% of crystallography data is in the form of CIF files.

A CIF file represents derived data rather than raw data and chemists may derive further data from the CIF. Various players have made software available for generating and manipulating CIF files, including the Cambridge Crystallography Data Centre (CCDC) (see below for more on this organisation). Examples of such software are Mercury (from the CCDC)[16], WinGX from Glasgow University[17] and Olex, developed at Durham University[18]. All of them are free to the community. Additionally, researchers often write their own software to perform particular tasks. There have been people capable of writing software for crystallographic purposes in every major crystallography research group since the 1960s and frequently in smaller outfits everyone is a one-man-band, able to perform all roles in the data production process.

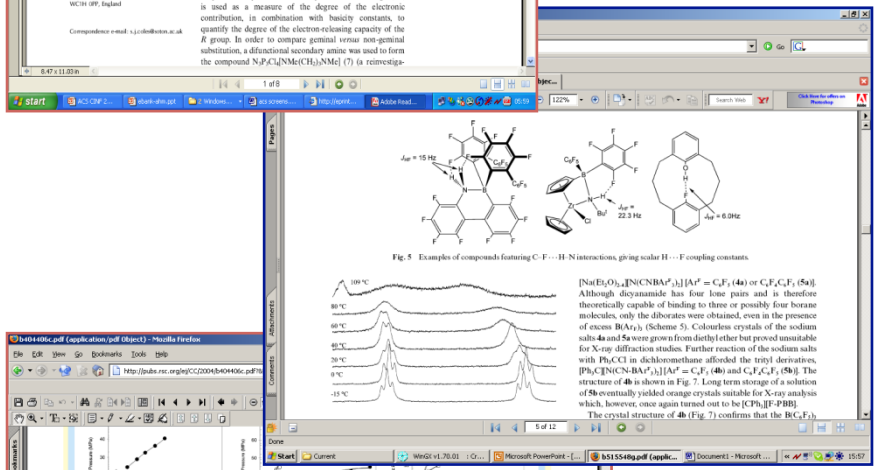
The major public funder of chemical crystallographic research in the UK is the Engineering and Physical Sciences Research Council (EPSRC), but a considerable amount of other crystallography work carried out in universities is paid for by pharmaceutical or chemical companies. This of course has some repercussions on what can be made publicly-available for sharing, but in general if a project is a purely academic one, even if funded by industry, then the funder places no barriers to public revelation of the data outputs. At the other end of the scale, some industry funded research is completely client-confidential, with the university crystallographic centres providing a bespoke service accordingly.

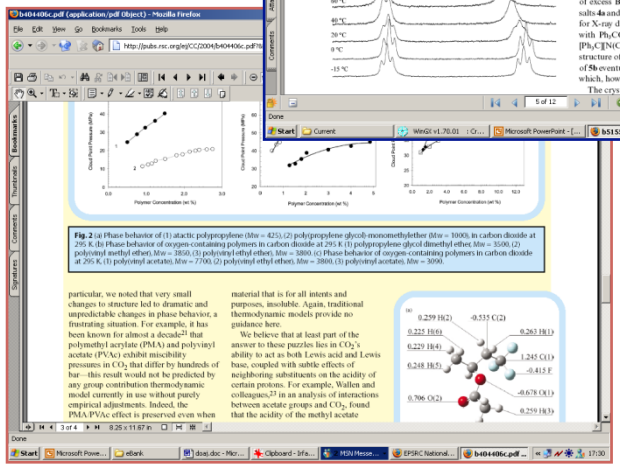
http://www.rin.ac.uk/data-publication

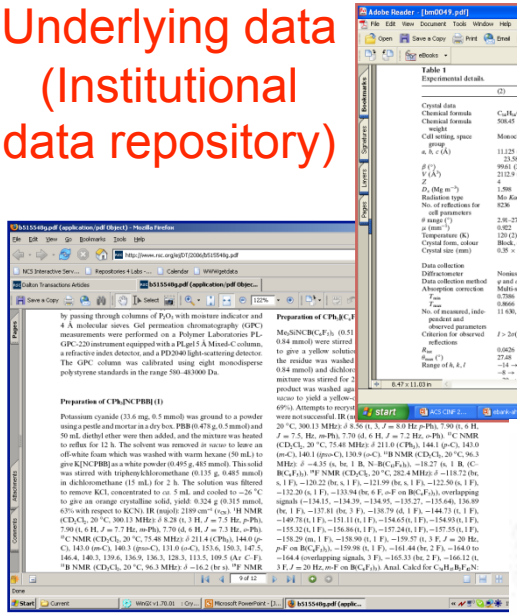
The Solution

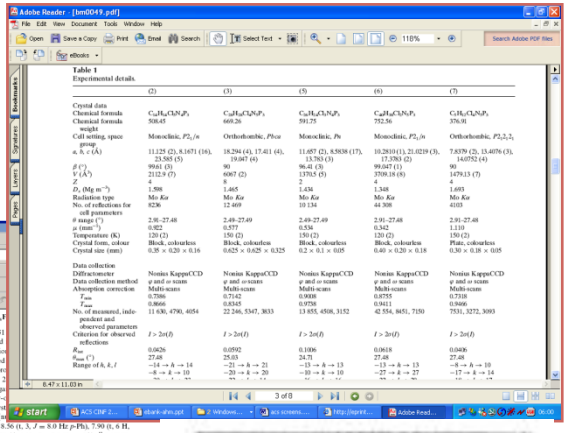
Intellect & Interpretation (Journal article, report, etc)

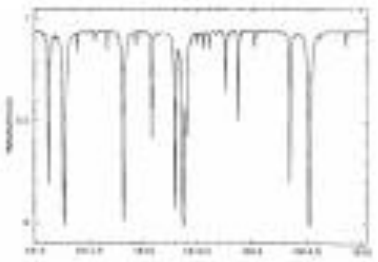
Underlying data (Institutional data repository)

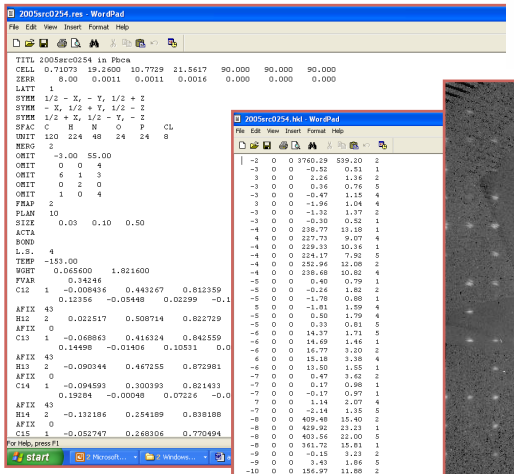


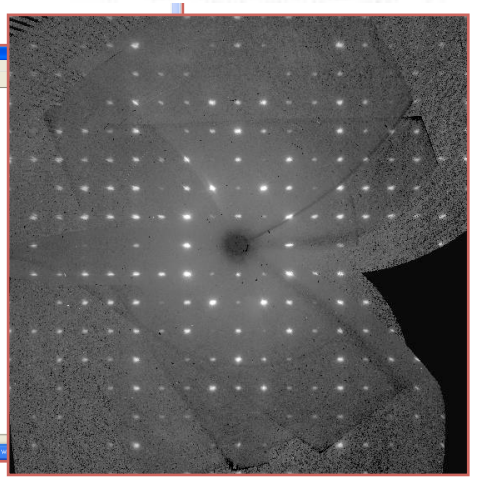




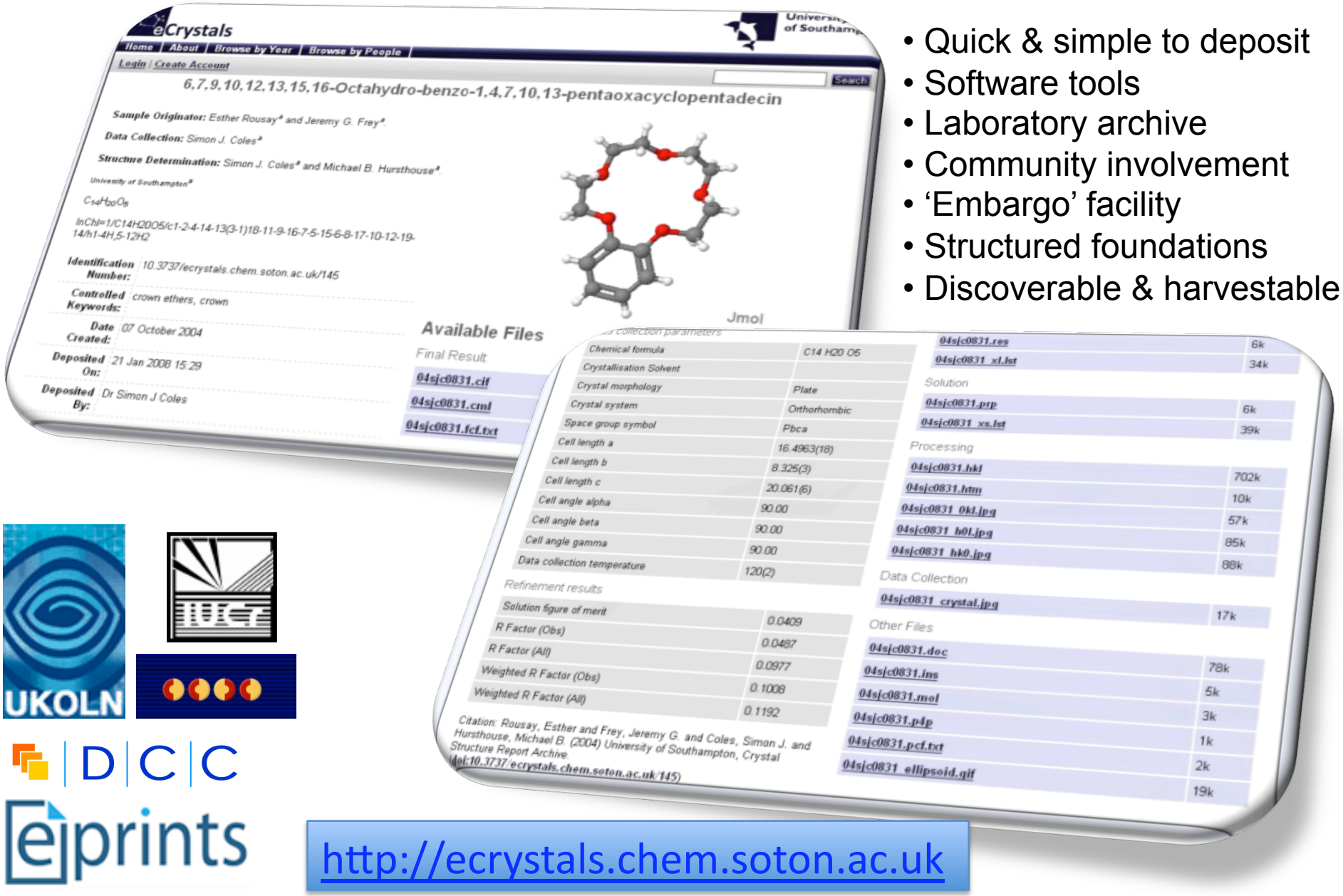








The eCrystals Data Repository



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

UKOLN

DCC

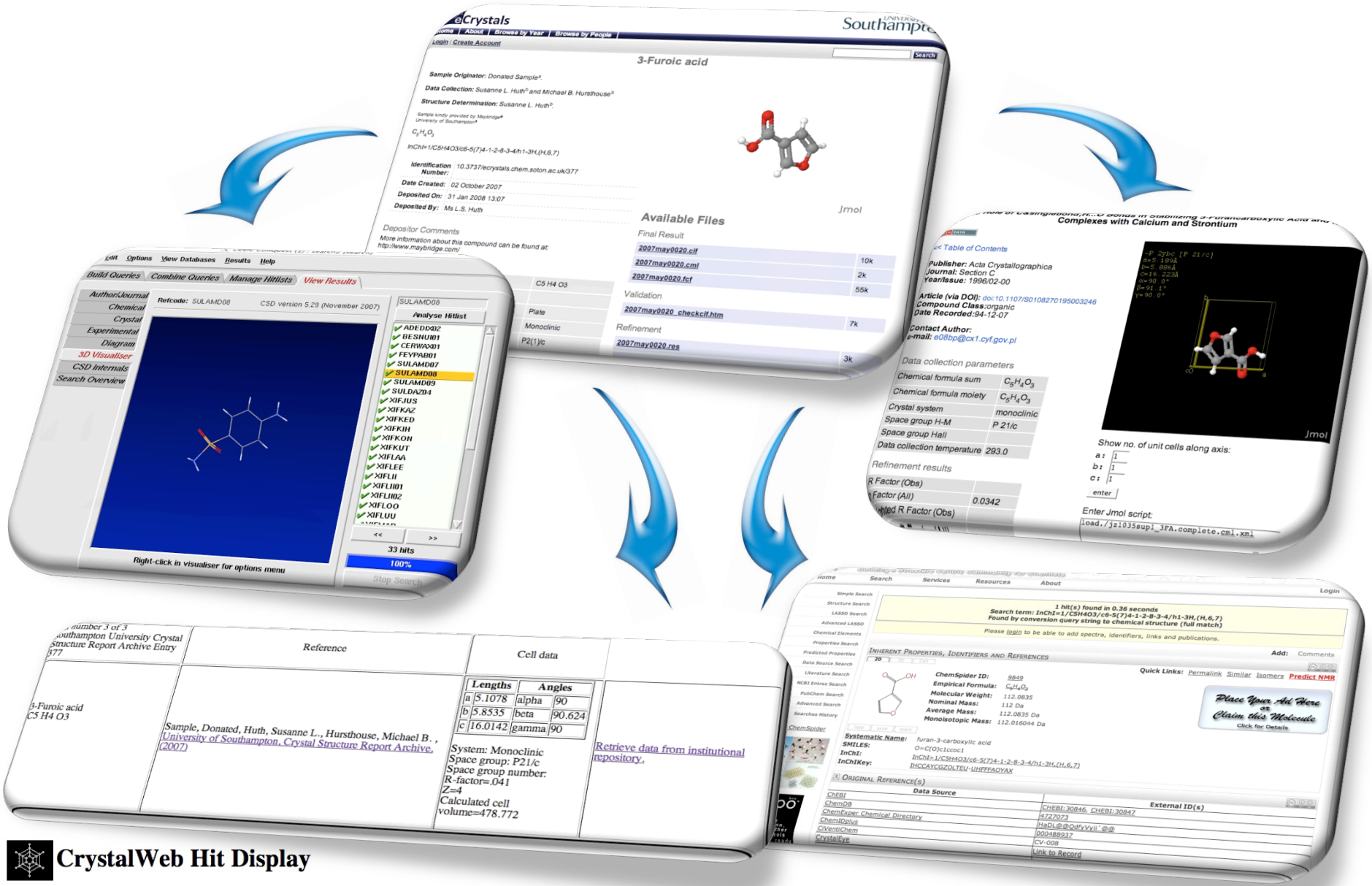
eprints

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

A Thorough Approach to Dissemination

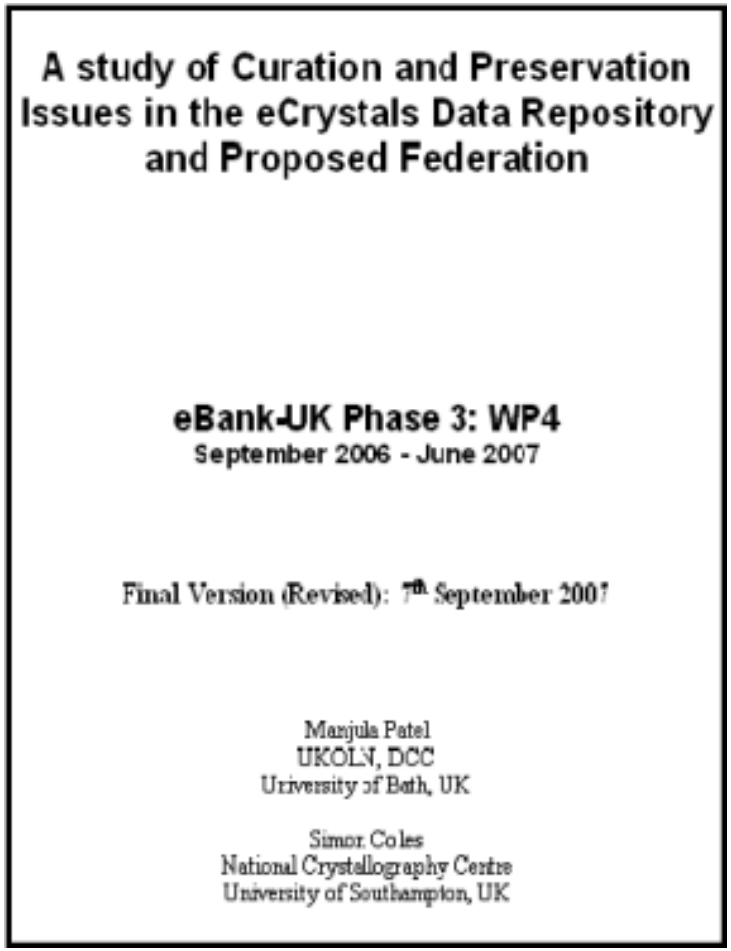
- Using simple Dublin Core protocol (OAI-PMH)
    - Crystal structure
    - Title (Systematic IUPAC Name)
    - Authors
    - Affiliation
    - Creation Date
- Additional chemical information through Qualified Dublin Core
    - Empirical formula
    - International Chemical Identifier (InChI)
    - Compound 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*

Data Harvesting



CrystalWeb Hit Display

A Thorough Approach to Preservation



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

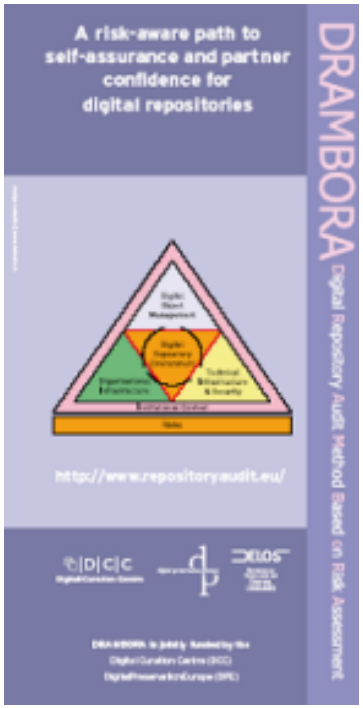
eBank-UK Phase 3: WP4
September 2006 - June 2007

Final Version (Revised): 7th September 2007

Manjula Patel
UKOLN, DCC
University of Bath, UK

Simon Coles
National Crystallography Centre
University of Southampton, UK

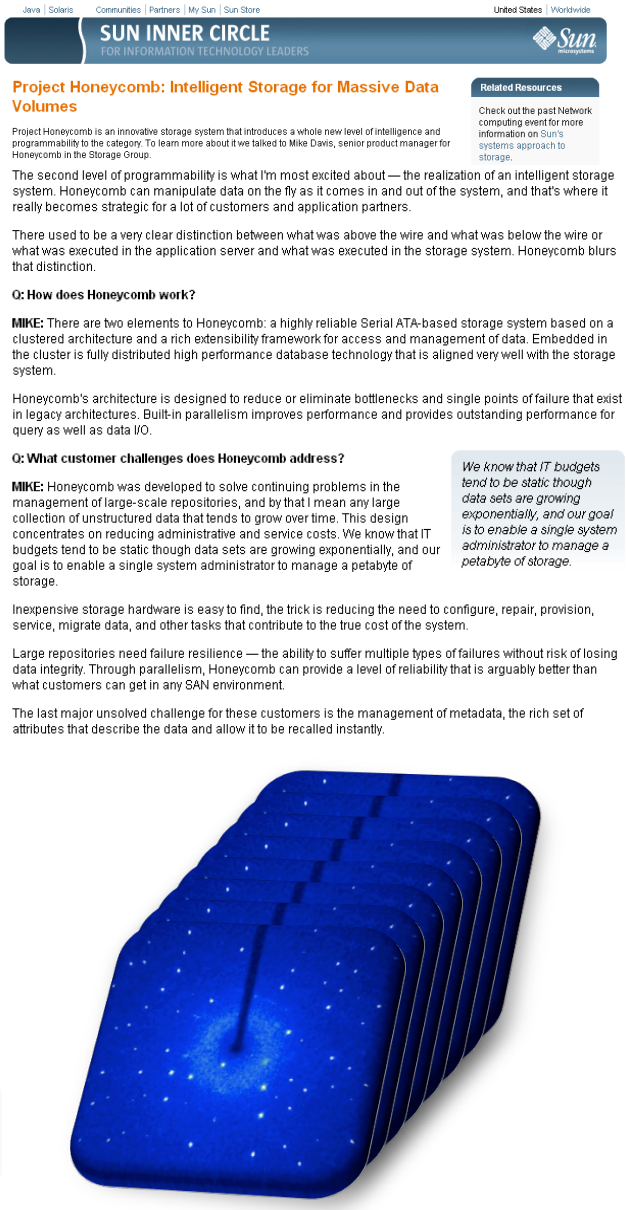
DCC



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

DRAMBORA Digital Repository Audit Method Based on Risk Assessment

http://www.repositoryaudit.eu/



Java | Solaris Communities | Partners | My Sun | Sun Store United States | Worldwide

SUN INNER CIRCLE
FOR INFORMATION TECHNOLOGY LEADERS

Sun Microsystems

Project Honeycomb: Intelligent Storage for Massive Data Volumes

Related Resources

Check out the past Network computing event for more information on Sun's systems approach to storage.

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 goal is to enable a single system administrator to manage a petabyte of storage.

Inexpensive storage hardware is easy to find, the trick is reducing the need to configure, repair, provision, 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.

http://www.ukoln.ac.uk/projects/ebank-uk/curation/

So How Much Does it Cost?

Appendix 4- Case Study : The University of Southampton

Background

Information to be considered as a case study for the cost model will be collected from the UK National Crystallography Service (NCS), The School of Chemistry and the National Oceanography Centre, Southampton (NOCS).

The National Crystallography Service

History

The National Crystallography Service has been operation since 1981, firstly at Queen Mary College London, then at the University of Wales College of Cardiff and since 1998 at the University of Southampton. For the whole period of its existence the NCS has always been housed within the respective Schools or Departments of Chemistry at these institutions and has been funded by a succession of research grants under both the rolling grant and responsive mode funding schemes of the EPSRC. The NCS provides an analytical service for UK chemists, based on state of the art experimental data collection facilities. This service includes the provision of raw data for those 'skilled in the art', who wish to work up a crystal structure themselves but don't have experimental facilities available to them or the provision of fully analysed crystal structures for chemists who do not have the necessary training or facilities to conduct these experiments.

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

http://www.jisc.ac.uk/publications/publications/keepingresearchdatasafe.aspx

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

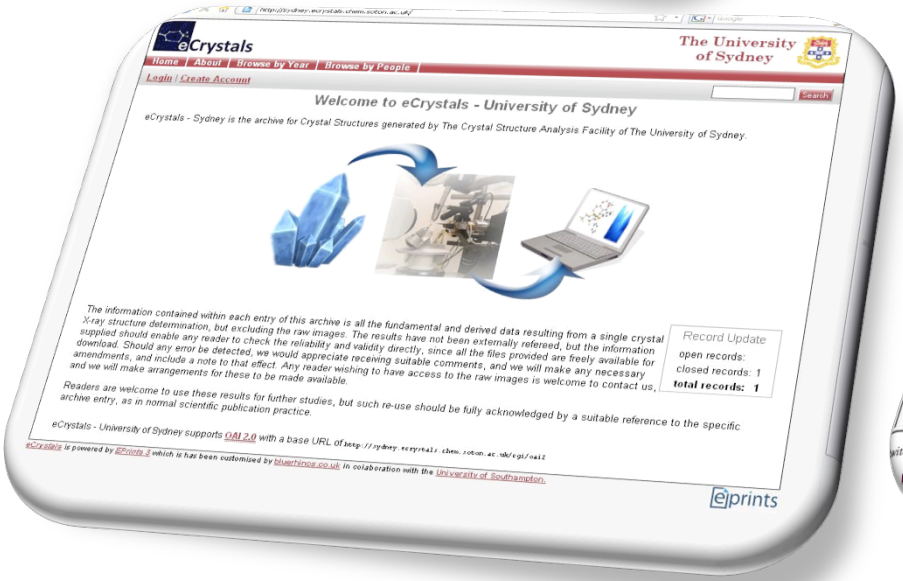
Scaling Up: Towards a Federation of Crystallography Data Repositories

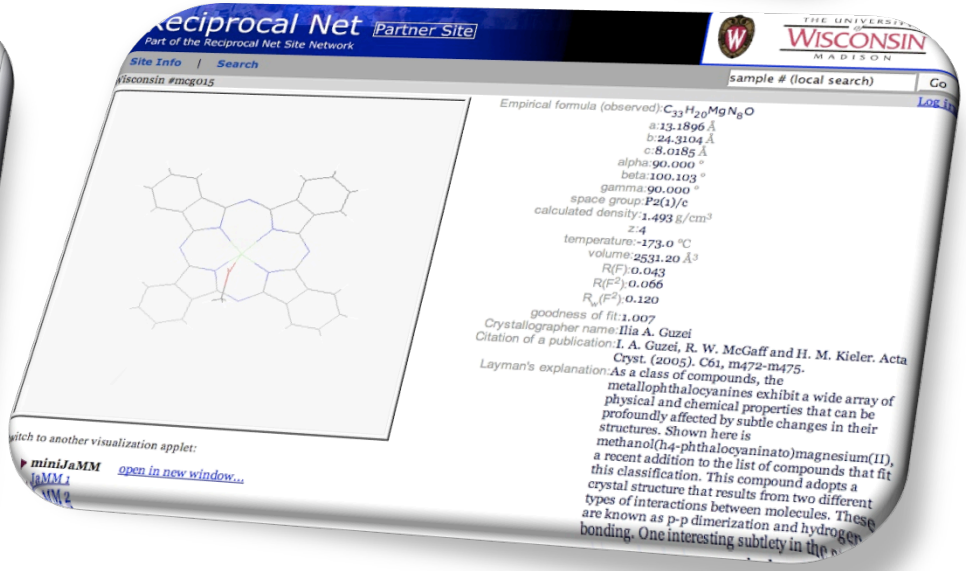
Document details

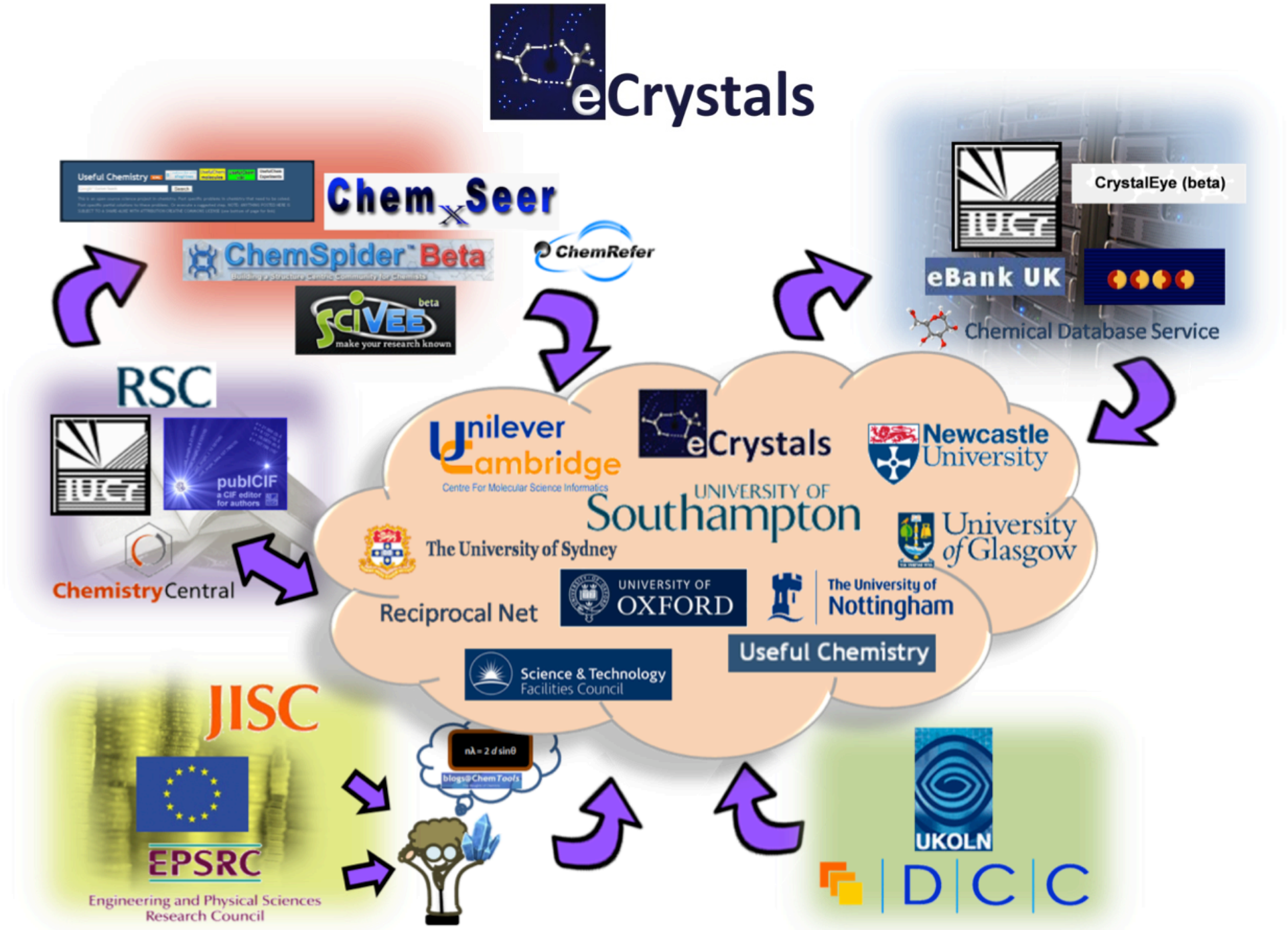
Author:	Liz Lyon, Simon Coles, Monica Duke, Traugott Koch
Date:	12th May 2008
Version:	1.0 Final
Document Name:	ebank-phase3-report-final.doc
Notes:	

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

Building a Federation of Repositories

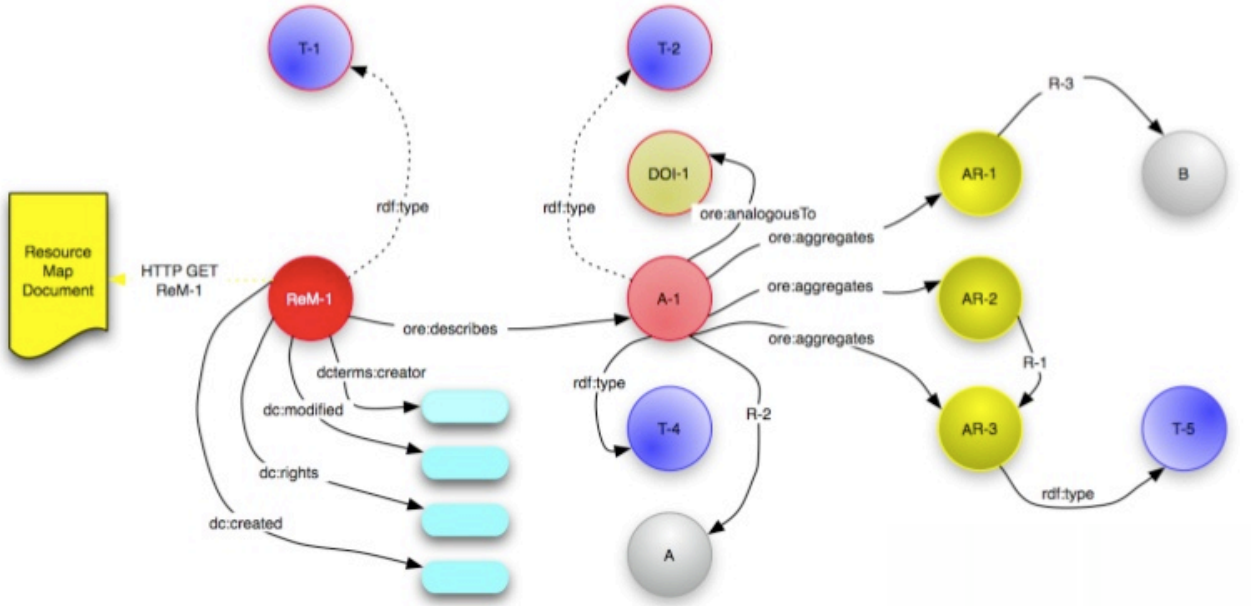





eCrystals
Useful Chemistry
Chem Seer
ChemSpider Beta
ChemRefer
SciVEE beta
make your research known
CrystalEye (beta)
eBank UK
Chemical Database Service
RSC
publCIF
a CIF editor for authors
ChemistryCentral
Unilever Cambridge
Centre For Molecular Science Informatics
eCrystals
Newcastle University
UNIVERSITY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

Packaging and Interoperability

- New moves in Digital Libraries community to enable distributed repositories to fully describe and 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
- OREChem testbed project



Big Business

Google to Host Terabytes of Open-Source Science Data

By Alexis Madrigal ✉ January 18, 2008 | 2:23:21 PM Categories: Dataset, Research

Sources at Google have disclosed that the humble domain, http://research.google.com, will soon provide a home for terabytes of open-source scientific datasets. The storage will be free to scientists and access to the data will be free for all. The project, known as Palimpsest and ~~first~~ previewed to the scientific community at the Science Foo camp at the Googleplex last August, missed its original launch date this week, but will debut soon.

Building on the company's acquisition of the data visualization technology, Trendalyzer, from the oft-lauded, TED presenting Gapminder team, Google will also be offering algorithms for the examination and probing of the information. The new site will have YouTube-style annotating and commenting features.

The storage would fill a major need for scientists who want to openly share their data, and would allow citizen scientists access to an unprecedented amount of data to explore. For example, two planned datasets are all 120 terabytes of Hubble Space Telescope data and the images from the Archimedes Palimpsest, the 10th century manuscript that inspired the Google dataset storage project.

UPDATE (12:01pm): Attila Csordas of Pimm has a lot more details on the project, including a set of slides that Jon Trowbridge of Google gave at a presentation in Paris last year. WIRED's own Thomas Goetz also mentioned the project in his fantastic piece of freeing dark data.

One major issue with science's huge datasets is how to get them to Google. In this post by a SciFoo attendee over at business|bytes|genes|molecules, the collection plan was described:

> (Google people) are providing a 3TB drive array (Linux RAID5). The array is provided in "suitcase" and shipped to anyone who wants to send they data to Google. Anyone interested gives Google the file tree, and they SLURP the data off the drive. I believe they can extend this to a larger array (my memory says 20TB).

You can check out more details on why hard drives are the preferred distribution method at Pimm. And we hear that Google is hunting for cool datasets, so if you have one, it might pay to get in touch with them.

Image: flickr/dannysullivan

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

Thanks to:

• Mike Hursthouse, Jeremy Frey, Andrew Milsted, Richard Stephenson, Mark Borkum

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

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