ROYAL SOCIETY OF CHEMISTRY | CHEMICAL INFORMATION AND COMPUTER APPLICATIONS GROUP

NEWSLETTER Summer 2015

Diana Leitch with husband David outside Buckingham Palace on Friday 30th January 2015, after collecting her MBE. For the full story see page 3.

CICAG aims to keep its members abreast of the latest activities, services, and developments in all aspects of chemical information, from generation through to archiving, and in the computer applications used in this rapidly changing area through meetings, newsletters and professional networking.

Chemical Information & Computer Applications Group: http://www.rsc.org/CICAG

LinkedIn® http://www.linkedin.com/groups?gid=1989945

MyRSC http://my.rsc.org/groups/cicag

https://twitter.com/RSC_CICAG

QR Code

Contributions to the CICAG Newsletter are welcome from all sources - please send to the Newsletter Editor: Lindsay Battle, email: lindsay.battle@chem.ox.ac.uk

Table of Contents

Chemical Information & Computer Applications Group Chair's Report

Contributed by RSC CICAG Chair Dr Helen Cooke, email: helen.cooke100@gmail.com

The first half of 2015 has been a busy period for the CICAG, and has included some significant changes to the Committee membership. Two of our members, Doug Veal and Hannah Morgan, have resigned from the Committee, and seven new members have joined us.

Doug Veal, our longest serving member with 25 years' service, was an outstanding contributor to the CICAG and its predecessor, the Chemical Information Group. Doug held Chair and Secretary roles in the 1990s and 2000s, organised and presented at scientific meetings, and provided continuity for the Committee through many changes. Doug was, and will continue to be, the CICAG's point of contact for the Tony Kent Strix award (http://www.cilip.org.uk/uk-einformation-group/awards-and-bursaries/tony-kent-strix-award) which the CICAG sponsors. On a personal note, before I joined the CICAG Committee I had known Doug while we both were members of the Institute of Information Scientists Publications Committee in the 1990s, and was delighted to have the opportunity to work with him again. Thank you Doug, you will be missed.

Thanks also to Hannah Morgan, a Committee member for several years, for her contributions, including an article in our last Newsletter: http://www.rsc.org/images/CICAG-Newsletter-Winter-2014_tcm18-244171.pdf. Hannah's career has changed direction, which led to her decision to resign from the Committee.

We undertook a recruitment drive for new Committee members in the spring, and I'm very pleased to welcome seven new members:

Dr Neil Berry MRSC, University of Liverpool
Dr Peter Bladon FRSC, Interprobe Chemical Services
Dr Nathan Brown MRSC, Institute of Cancer Research
Professor Jonathan Goodman CChem FRSC, University of Cambridge
Dr Brian Lynch FRSC, St. Francis Xavier University, Nova Scotia
Dr Michelle Lynch CChem MRSC, HIS
Dr Chris Swain CChem FRSC, Cambridge MedChem Consulting

A full list of RSC CICAG Committee Officers and Members, with their affiliations and contact details, is provided on the CICAG website:
http://www.rsc.org/Membership/Networking/InterestGroups/CICAG/committee.asp

I mentioned in the last Newsletter that one of my missions while in the role of CICAG Chair is to broaden our outreach by forming partnerships with other RSC interest groups, and outside organisations with whom we share common interests. In the early stages of planning the recent meeting on "big data", we were approached by the Dial-a-Molecule Grand Challenge Network (DaM; http://www.dial-a-molecule.org/wp/) team, who were interested in partnering with us as they had been considering running an event on a similar topic. We enthusiastically accepted this proposal, the outcome of which was the "From Big Data to Chemical Information" meeting, held at Burlington House in April. The partnership proved to be very fruitful, as the DaM team helped with identifying speakers, publicity and meeting logistics. The meeting was fully booked, and well received by delegates. A report from this meeting appears later in this Newsletter, and speakers' slides are available - see below for link.

We are also partnering with the Automation and Analytical Management Group (AAMG), to organise a scientific meeting with the provisional title 'Measurement, Information and Innovation: Digital Disruption in the Chemical Sciences' scheduled for 20th October 2015. Following successful past events run jointly with the Historical Group, we are continuing to explore opportunities to partner with them in the future.

To keep in touch with the CICAG, please follow us on Twitter @RSC_CICAG and LinkedIn or join our MyRSC group (links on the front page of this Newsletter). Our Social Media Editor, Keith White, makes frequent posts and looks forward to your comments and interactions. Please also feel free to contact me directly with any questions or suggestions as to how your Committee can support you.

RSC CICAG Annual Report for 2014

The CICAG Annual Report for 2014 can be downloaded from the CICAG website:
http://www.rsc.org/images/Annual-Report_tcm18-245531.pdf

Notice of Future Meeting

In partnership with the RSC Automation and Analytical Management Group (AAMG), CICAG is organising the following meeting, to be held at Burlington House on 20th October 2015:
Measurement, Information and Innovation: Digital Disruption in the Chemical Sciences,
http://www.rsc.org/events/detail/18885/measurement-information-and-innovation-digital-disruption-in-the-chemical-sciences

More details and a registration form will be available nearer the time from the CICAG website: http://www.rsc.org/CICAG

Meeting Report: From Big Data to Chemical Information

CICAG held a very successful meeting "From Big Data to Chemical Information" on 22nd April 2015 at RSC, Burlington House, London. A detailed report is appended to this newsletter, along with contributions from four bursary recipients describing their experiences.

The report and speakers' presentations are available from the CICAG 'Previous Meetings' web page:
http://www.rsc.org/Membership/Networking/InterestGroups/CICAG/meetings.asp

Diana Leitch's MBE

Snow came to Didsbury in south Manchester on the morning of Thursday 29th January 2015 as David and Diana Leitch were making last minute preparations to go to London for Diana to receive her MBE at Buckingham Palace on the morning of Friday 30th January. Would they get there as the A34 was gridlocked, the M60 was at a standstill and Manchester Airport was closed. Taxis were few and far between and taking up to an hour to get to Stockport Station from Didsbury, a journey that would normally take 10 minutes. Would their daughter, Fiona, a surgeon in Glasgow get to London overnight from snowbound Scotland? With a sigh of relief everyone, including son, Andrew, did get to London where it was very cold but snow free and the sun was shining. The Prince of Wales officiated at the ceremony which took place in the great ornate ballroom and was very impressive. Diana's was the 26th out of 27 investitures of people who had been awarded honours in the Queen's Birthday Honours in June 2014. One week later on 6th February and she would have been there as the author, Hilary Mantel, received her Dame's insignia. Quite coincidentally, Diana later discovered that the person who received her Dame's insignia on the same day as Diana, Dame Louise Makin, (for services to the life sciences industry), had attended the same school, the Queen's School in Chester, as Diana, but several years later. Everyone enjoyed themselves at the Palace, lunch followed and then a visit to the Royal Society of Chemistry at Burlington House, off Piccadilly, for afternoon tea courtesy of CEO, Dr Robert Parker and his PA, Mrs Gill McGrath, as Diana, who is a Fellow of the RSC, had received her MBE for 'services to chemistry'. It was the culmination of a long career in the chemical information world at national and international level and since retirement in the promotion of chemistry and the chemical sciences to people of all ages through her Trustee's role at the Catalyst Science Discovery Centre and Museum in Widnes.

See photo on cover, showing Diana and David outside Buckingham Palace that morning.

Tony Kent Strix Award

The **Tony Kent Strix Annual Lecture Series** begins with a talk by Dr Susan T Dumais of Microsoft Research at the Geological Society in London on the afternoon of Friday November 6th. The talk will be followed by a discussion and some time for refreshments and networking. The event is free but register will be required:
http://www.cilip.org.uk/uk-einformation-group/news/tony-kent-strix-annual-lecture-major-free-event-your-diary

The **Call for nominations for 2015 Strix Award** has recently been announced. Nominations should be received by August 28th 2015.
http://www.cilip.org.uk/uk-einformation-group/awards-bursaries/tony-kent-strix-award/2015-call-tony-kent-strix-nominations

CICAG is one of the sponsors of the Strix Award, which is presented in memory of Dr Tony Kent, a past Fellow of the Institute of Information Scientists, who died in 1997. Tony Kent made a major contribution to the development of information science and information services both in the UK and internationally, particularly in the field of chemistry.

The Award is managed by UKeiG, a Special Interest Group of CILIP: the Chartered Institute of Library and Information Professionals. Further details can be found at:
http://www.cilip.org.uk/uk-einformation-group/awards-and-bursaries/tony-kent-strix-award

Other Awards - Calls for nominations

CSA Trust Grants: Applications Invited for 2015: The Grant Program has been created to provide funding for the career development of young researchers who have demonstrated excellence in their education, research or development activities that are related to the systems and methods used to store, process and retrieve information about chemical structures, reactions and compounds. Closing date: 25th March 2016.
http://bulletin.acscinf.org/node/730; http://www.csa-trust.org/?page_id=9

Call for Jason Farradane Award Nominations: UKeiG is now seeking nominations for this award, which is made to an individual or a group of people in recognition of outstanding work in the information profession. Closing date: 11th September 2015.
http://www.cilip.org.uk/uk-einformation-group/awards-bursaries/jason-farradane-award/2015-call-jason-farradane-nominations

CINF Scholarship for Scientific Excellence: Call for Applications.
http://bulletin.acscinf.org/node/729

Other Awards - Recent Recipients

The **2014 CODATA Prize has been awarded to Professor Sydney Hall**, who devised a universal self-defining text archive and retrieval (STAR) file format that evolved into the Crystallographic Information Framework (CIF):
http://www.codata.org/news/13/62/CODATA-Prize-2014-awarded-to-Professor-Sydney-Hall

The **2015 CINF Lifetime Award**: **Helen A. "Bonnie" Lawlor** was selected as the third recipient. The Lifetime Award was established by the ACS Division of Chemical Information in 2006 and recognizes long-term membership and outstanding service and active contributions to the Division over the years.
http://www.acscinf.org/content/cinf-lifetime-award

Lucille Wert Scholarship: Siu Hong Yu has been selected as the 2015 recipient of the Lucille M. Wert Student Scholarship. The award is to "help persons with an interest in the fields of chemistry and information to pursue graduate study in library, information, or computer science."
http://bulletin.acscinf.org/node/728

Student Bursaries

UKeiG Student or Early-Career Professional Conference Grant. Applications are accepted throughout the year. Enables a student or early-career professional to attend any conference relevant to their professional studies or career, either in the UK or overseas.
http://www.cilip.org.uk/uk-einformation-group/awards-and-bursaries/student-or-early-career-professional-conference-grant

2015 John Campbell Trust: call for applications for bursaries for study or conference trips is now open; closing date for applications is 12 June 2015.
http://www.cilip.org.uk/cilip/membership/benefits/advice-and-support/grants-and-bursaries/john-campbell-trust/john-campbel-2

Chemical Information / Cheminformatics and related Books

Exploring Materials through Patent Information
http://pubs.rsc.org/en/content/ebook/9781782621126

First to File: Patents for Today's Scientist and Engineer. M. Henry Heines, Wiley
http://www.wiley.com/WileyCDA/WileyTitle/productCd-111883965X.html

Chemistry Education: Best Practices, Opportunities and Trends. Javier García-Martínez, Elena Serrano-Torregrosa, Peter W. Atkins. Wiley
http://www.wiley.com/WileyCDA/WileyTitle/productCd-3527336052.html

What You Need For the First Job, Besides the PhD in Chemistry. Oxford University Press
http://ukcatalogue.oup.com/product/9780841229624.do
See review in CINF Information Bulletin Summer 2015:
http://bulletin.acscinf.org/node/740

Cambridge Cheminformatics Network

Contributed by Chris Swain, Cambridge MedChem Consulting

The quarterly Cambridge Cheminformatics Network Meeting took place on Wednesday, 27 May 2015 at the Centre for Molecular Informatics (CMI) on Lensfield Road, Cambridge. The first talk "Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis in-house HTS Data" from Shardul Paricharak, Centre for Molecular Informatics and NIBR Basel described a retrospective analysis of HTS data to see if data from a small (2500 compounds) manually curated data set representing compounds with known mode of actions could be used to select similar compounds from the total 1.5M compound collection. Similarity was based on ECFP_4 fingerprints and HTS_FP a fingerprint based on measured screening results. After 10 iterations they had selected and screened 1% of the 1.5M compounds and identified 45% of the active scaffolds. In comparing a range of different targets it seems that protein-protein assays perform particularly poorly and cell-based assays also give a poorer return.

In the second talk John May, NextMove Software gave a talk entitled "Substructure Search Face-off", in which he described their work in benchmarking and comparing different tools and strategies for substructure-search large compound collections. They used the eMolecules dataset (7M structures) and the structure query collection compiled by Andrew Dalke. Whilst there was a considerable range in the speed of execution one of the most striking observations was that relatively simple substructure searches (ortho-xylene) can return between 440,961 and 529,396 hits for the same dataset depending on the tool used. It seems the difference in the number of hits returned can be largely explained by the different aromaticity models used and whether standardisation is supported. Slides are also available here: http://www.slideshare.net/NextMoveSoftware/substructure-search-faceoff

More information about the Cambridge Cheminformatics Network Meetings can be found on the website (http://c-inf.net). You would like to present? All talks from the cheminformatics - and neighbouring - scientific areas are welcome. Talk are usually about 15-20 minutes long, followed by questions. Everyone is welcome to present his or her work - please just contact Andreas Bender (ab454 at cam.ac.uk) if you would like to present.

RSC News

Contributed by David Allen, Library Collections Coordinator, RSC

Digitisation Project: The RSC are happy to announce the launch of the first phase of the Historical Collection (http://pubs.rsc.org/historical-collection) – a new archive that, when complete, will contain nearly 1,000,000 pages of content and over 3,000 unique items. This product is launching in two phases:

- **Society publications and minutes, 1841-2012** (currently live) – covering material published by the Chemical Society, Royal Institute of Chemistry, and Royal Society of Chemistry. This includes back copies of Chemistry in Britain (now Chemistry World), Education in Chemistry, Monographs for Teachers, Lectures, Monographs and Reports, and minutes of Council and Committees.
- **Historical books and papers, 1505-1991** (coming soon) - over 2,000 books and letters from some of the most famous chemists in history

The collection is available through institutional subscription and is available free to all RSC Members

Test your ideas with Reaxys:
Using this chemical information database you can search reactions, formulas and literature to find experimental data from journals. You can also start your search by browsing through properties, methods and transformations. To use Reaxys, please request access from the Library: http://www.rsc.org/Library/Services/LICEnquiryRequestForm.asp. You have automatic access to Reaxys if you already have a login for Elsevier's ScienceDirect through the RSC.

National Chemical Database Service News

ChemSpider now has a new website. A key feature of the new design is to make ChemSpider work on as many devices as possible, from desktops to mobile phones. The changes are explained in the ChemSpider blog: http://www.chemspider.com/blog/

CAS / SciFinder / STN News

Contributed by Dr Anne Jones, CAS Applications Specialist UK & Ireland

CAS News

NCI™ Global, a new regulatory and compliance solution from CAS, delivers the most timely and accurate information for commercial use of chemicals

CAS has introduced a new web-based solution for anyone who needs regulatory information about chemicals in commerce. NCI Global is a timelier, higher value replacement for National Chemical Inventories™ (NCI) on CDROM.

NCI Global launched in January, 2015 and provides:

- A web-based regulatory and compliance solution, with significant usability improvements
- Access to nearly 20% more content than NCI on CDROM, including substances from regulatory lists, non-commenced pre-manufacturing notifications, harmonized tariff codes, classification, labeling and packaging information from more than 50 countries
- The most timely source of regulatory information, with content updated weekly
- Weekly email alerts with substance updates determined by the user
- Content built by the experts at CAS
- Subscription-based pricing, sold by company site; customers are qualified by IP address with no login id or password required. Monthly payment options are available upon request.

SciFinder News

Preserve Your Most Valuable Resource – Time – With PatentPak™

PatentPak saves users up to half the time they spend scrutinizing patents by providing instant access to hard-to-find chemistry in languages they know.

Key Features:

- Full-text patent documents from 11 major patent offices
- Patent family coverage in multiple languages, including English, German, Chinese, Japanese, French, Korean and Russian
- Patent page numbers for key indexed substances
- Unique, interactive patent document viewer (coming soon)

All SciFinder users receive five free samples of PatentPak content to give them a glimpse into the benefits afforded with PatentPak and links to product information on the PatentPak web page: http://www.cas.org/patentpak

STN News

STN Search Service Value Pricing Now Available for Information Brokers

STN Search Service Value Pricing (SSVP) is now available to support intellectual property and related search services of all sizes in managing costs and maximizing the value they can provide to their clients using STN.

The key benefits of SSVP pricing include:

- Budget predictability
- Simplified administration
- The ability of information brokers to focus on search quality rather than cost

Annual MEDLINE® Reload on STNFeatures Enhanced Clinical Trial Information and the 2015 MeSH Thesaurus

The 2015 MEDLINE reload on STN was launched in January, 2015 and featured the 2015 MeSH (Medical Subject Headings) thesaurus, which introduced 310 new controlled indexing terms. MEDLINE backfile records which included index terms deleted or changed for 2015 were updated to reflect 2015 MeSH terminology. Key improvements introduced with the reload include:

- More clinical trial information. The number of authorities for which abbreviated clinical trial authority names and clinical trial numbers are indexed, in the /FS and /NCT fields, respectively, has been expanded greatly

- One-click access to source documents, including full text in many instances, via complete Digital Object Identifier hyperlinks in the /DOI field. Similarly, the /AUID field now also provides complete hyperlinks allowing expanded author access information from ORCID
- Multiple addresses per author, when provided
- Information on Associated Datasets and Associated Publications in the /CM (Comment) custom display field, along with MEDLINE on STN Accession Numbers, as available

CAS Training in the UK 2015

In addition to the e-learning materials, CAS continues to offer instructor-led training for both STN and SciFinder in the UK.

We conduct 'in-house' or WebEx training sessions on all aspects of STN or SciFinder searching. Also, if you wish to know more about any of the CAS products or would like further information or help with STN or SciFinder, then please contact annejones@acsi.info

---

InfoChem News

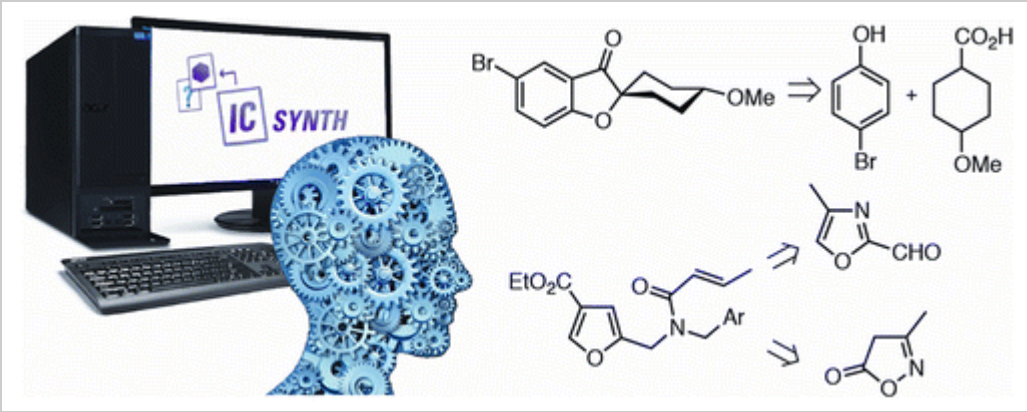
Contributed by Stephanie North, Allyl Consulting Ltd, representing InfoChem in the UK

InfoChem, AstraZeneca and ChemNotia have recently published a paper in Organic Process Research and Development on computer aided synthesis design:

Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction
Anders Bøgevig, Hans-Jürgen Federsel , Fernando Huerta, Michael G. Hutchings, Hans Kraut, Thomas Langer, Peter Löw, Christoph Oppawsky, Tobias Rein, and Heinz Saller
Org. Process Res. Dev., 2015, 19 (2), pp 357–368.
DOI: 10.1021/op500373e
Publication Date (Web): January 22, 2015

E-mail: Hans-Jurgen.Federsel@astrazeneca.com, mghutchings@infochem.de

Abstract:



The new computer-aided synthesis design tool ICSYNTH has been evaluated by comparing its performance in predicting new ideas for route design to that of historical brainstorm results on a series of commercial pharmaceutical targets, as well as literature data. Examples of its output as an idea generator are described, and the conclusion is that it adds appreciable value to the performance of the professional drug research and development chemist team.

The article is available on open access as follows:
http://pubs.acs.org/doi/full/10.1021/op500373e

For more information about InfoChem please email Dr Stephanie North: sn@infochem.de.

Forthcoming Meetings/Conferences

2015

Jul 1: **CILIP Library Information Research Group Member's Day & AGM**, Liverpool
http://www.cilip.org.uk/library-information-research-group/events/library-information-research-group-member-s-day-agm

Jul 2-3: **CILIP Conference 2015: Bringing the information world together**, Liverpool
http://www.cilip.org.uk/conference2015

Jul 7: **Open Information Science: exploring new landscapes,** iFutures conference for doctoral students in Information Science, Sheffield
http://ifutures.group.shef.ac.uk

Jul 13: **Get it sorted...positive approaches to non-standard access**, JIBS Workshop, London
http://www.jibs.ac.uk/events.html

Jul 13-14: **Knowledge Organization - Making a difference**. ISKO UK Conference, London
http://www.iskouk.org/content/knowledge-organization-making-difference

Jul 15: **NoWAL Conference: Where is the Library?** Manchester
http://www.nowal.ac.uk/conference/

Jul 20: **Subject Librarians: time for a fresh look?** University of Hertfordshire, Hatfield
https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1505&L=lis-link&F=&S=&P=82413

Jul 20-22: **11th Northumbria International Conference on Performance Measurement in Libraries and Information Services**, Edinburgh
http://www.york.ac.uk/about/departments/support-and-admin/information-directorate/northumbria-conference/

Jul 29: **Wikipedia Editathon**, London
https://en.wikipedia.org/wiki/Wikipedia:GLAM/Royal_Society_of_Chemistry/BH-2015
Book here: http://www.rsc.org/events/detail/18874/wikipedia-editathon

Aug 8: **Wikipedia Editathon**, Catalyst, Widnes
Book here: http://www.rsc.org/events/detail/18911/wikipedia-editathon

Aug 16-20: **CINF symposia at ACS 2015 Fall National Meeting**, Boston
http://bulletin.acscinf.org/node/732

Aug 26: **Cambridge Cheminformatics Network Meeting**
http://c-inf.net

Aug 31-Sep 3: **10th European Conference on Computational Chemistry**, Fulda, Germany
http://www.euco-cc-2015.org

Sep 14-15: **With Power Comes Great Responsibility - How librarians can Harness the Power of Social Media for the Benefit of Users**. Cilip Multimedia, Information and Technology Group annual conference, Sheffield http://lanyrd.com/2015/mmit2015/

Sep 15-17: **7th Conference on Open Access Scholarly Publishing** (COASP), Amsterdam
http://oaspa.org/conference/

Sep 25: **Getting to grips with developing and managing e-book collections: an introduction**, UKeiG Training, London
http://www.cilip.org.uk/uk-einformation-group/events/ukeig-training-getting-grips-developing-managing-e-book-collections

Oct 7-8: **2:AM Altmetrics conference**, Amsterdam
http://www.altmetricsconference.com/registration/

Oct 19-20: **ICIC 2015 27th ICIC International Conference for the Information Community**, Nice, France
http://www.haxel.com/icic

Oct 19-21: **Dynamic disruption: transforming the library**, Internet Librarian International, London
http://www.internet-librarian.com/2015/

Oct 20: **Measurement, Information and Innovation: Digital Disruption in the Chemical Sciences**, CICAG/AAMG meeting, RSC London
http://www.rsc.org/events/detail/18885/measurement-information-and-innovation-digital-disruption-in-the-chemical-sciences

Oct 29-30: **Classification and Authority Control: Expanding Resource Discovery**, Lisbon
http://seminar.udcc.org/2015/

Nov 6: **Tony Kent Strix Annual Lecture Series**: Dr Susan T Dumais of Microsoft Research, Geological Society, London
http://www.cilip.org.uk/uk-einformation-group/news/tony-kent-strix-annual-lecture-major-free-event-your-diary

Nov 8-10: **11th German Conference on Chemoinformatics**, Fulda, Germany
https://www.gdch.de/index.php?id=2348

Nov 18: **UKSG 2015 Forum**, London
http://www.uksg.org/event/FORUM2015

Nov 19-20: **1st Relationship Management for Academic Libraries Conference**, Stirling
https://relationshipmanagementgroup.wordpress.com/2015/06/03/1st-relationship-management-for-academic-libraries-conference/

Nov 26: **Cambridge Cheminformatics Network Meeting**
http://c-inf.net

2016

Jan 12-13: **International Conference on Data and Information Management**, Loughborough
http://idimc.org/programme/

Apr 11-13: **UKSG Annual Conference and Exhibition**, Bournemouth
http://www.uksg.org/events/annualconference

Jul 4-6: **Seventh Joint Sheffield Conference on Chemoinformatics**. Molecular Graphics and Modelling Society and the Chemical Structure Association Trust, Sheffield
http://cisrg.shef.ac.uk/shef2016/

Recent meetings

which you may have missed, but can follow up online

Jun 1-2: **Advocating Libraries: Innovate and Thrive**, CILIPS Conference, Dundee
https://cilips.formstack.com/forms/cilipsconference2015

Jun 5: **Hard to reach? Resources, people and promotions**, University Science and Technology Librarians' Group (USTLG) meeting, Bath http://www.ustlg.org/?page_id=950

Jun 23-26: **5th i³ Conference**, Aberdeen
http://www.rgu.ac.uk/research/conferences/i-2015

People News

Knighthoods for two chemists:
Prof Martyn Poliakoff - for services to the chemical sciences
Dr Simon Fraser Campbell, Past President of RSC plus Pfizer and Astex - for services to chemistry
http://www.rsc.org/news-events/rsc-news/articles/2015/jan/new-year-honours-2015/

Dana Roth was awarded the title of Fellow of the Royal Society of Chemistry (FRSC).
http://library.caltech.edu/news/index.php/archives/1999

Ms. Sharon Todd is appointed as Executive Director of the SCI.
http://www.soci.org/news/sci/sharon-todd

Appointment of Jisc chair
Professor David Maguire, vice-chancellor at the University of Greenwich, has been appointed to the role of Jisc chair.
http://www.jisc.ac.uk/news/appointment-of-jisc-chair-14-apr-2015

Robert J. Massie, 66, who led Chemical Abstracts Service for nearly 22 years until his retirement in March 2014, died on June 7 at his home in Columbus, Ohio.
http://cen.acs.org/articles/93/web/2015/06/Robert-Massie-Dies-66.html

Professor Paul von Ragué Schleyer passed away on November 21, 2014. Paul was a major force in physical organic and computational organic chemistry.
http://comporgchem.com/blog/?p=3403

Other News Items

Forgotten synthetic PhD theses set to be given new lease of life, Chemistry World March 2015
http://www.rsc.org/chemistryworld/2015/03/forgotten-synthetic-phd-theses-set-be-given-new-lease-life

The RSC's **Virtual Library** is described in **Chemical science data at your fingertips** in RSC News, Jan 2015, p. 6. Download PDF at:
http://www.rsc.org/images/rsc-news-january-2015_tcm18-244235.PDF

Notes is a free quarterly newsletter for librarians & information specialists from the RSC. Find past issues and subscribe at: http://pubs.rsc.org/en/content/data/librarian-newsletters

National Compound Collection: a pilot scheme funded by the RSC in partnership with the University of Bristol: http://www.rsc.org/blogs/escience/national-compound-collection

RCUK publishes first independent review of its open access policy
http://www.rcuk.ac.uk/media/news/openaccess/
see also **Jisc Response**: Where are we on the open access roadmap?
http://www.jisc.ac.uk/news/where-are-we-on-the-open-access-roadmap-26-mar-2015 and
Jisc report: **How publishers might help universities implement OA**:
http://scholarlycommunications.jiscinvolve.org/wp/2015/03/26/how-publishers-might-help-universities-implement-oa/

Nature promotes read-only sharing by subscribers
http://dx.doi.org/10.1038/nature.2014.16460
See also comments:
http://www.michaeleisen.org/blog/?p=1668
http://del-fi.org/post/104125242971/natures-shareware-moment

Confusion over publisher's pioneering open-data rules. Nature Nov 2014
http://dx.doi.org/10.1038/515478a

Scientists losing data at a rapid rate. Nature Dec 2013
http://dx.doi.org/10.1038/nature.2013.14416

Instant translation – no longer sci-fi
BBC News Dec 2014
http://www.bbc.co.uk/news/technology-30539198
See also Skype Translator: http://www.skype.com/en/translator-preview/

100 Years of Industrial Chemistry. ChemViews Dec 2014
http://www.chemistryviews.org/details/ezine/7086821/100_Years_of_Industrial_Chemistry.html

Academic libraries from the UK's Northern Collaboration to use OCLC QuestionPoint for out-of-hours enquiry services. OCLC, Dec 2014
http://www.oclc.org/en-UK/news/releases/2014/201409emea.html

Notes on notation - Chemistry World Flashback, Jan 2015
The UK chapter of the Chemical Notation Association met for the first time - reported in Chemistry in Britain Feb 1970
http://www.rsc.org/chemistryworld/2015/01/45-years-ago-notes-notation

Chemical Information and Computation 2014, Number Two. 248th ACS National Meeting and Exposition, San Francisco, August 2014
Wendy Warr & Associates' 44th ACS report covers papers on Drug Discovery, Chemistry Text Mining in Patents and Other Documents, Herman Skolnik Award Symposium 2014 Honoring Engelbert Zass, plus a news section covering people, awards, and 65 organisations. Contents list and order forms at http://www.warr.com.

Report of CINF Technical Program at 249th ACS National Meeting, March 2015, Denver
http://bulletin.acscinf.org/node/732

350 years of publishing from the world's oldest science journal. The Guardian Feb 2015
http://www.theguardian.com/higher-education-network/gallery/2015/feb/12/philosophical-transactions-of-the-royal-society-350-years-of-science-publishing-in-pictures

50 Years of the Cambridge Structural Database: Some Personal Perspectives, CINF Information Bulletin Summer 2015
http://bulletin.acscinf.org/node/738

Survey of Academic Library Use of Online and Other Survey Tools, Report from Primary Research Group Inc.
http://www.primaryresearch.com/view_product.php?report_id=536

Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction, Organic Process Research & Development **2015**, *19* (2), pp 357–368
http://pubs.acs.org/doi/abs/10.1021/op500373e

Bibliometrics: The Leiden Manifesto for research metrics. Nature April 2015
http://www.nature.com/news/bibliometrics-the-leiden-manifesto-for-research-metrics-1.17351

Fast-track peer review trial ends following resignations. Chemistry World April 2015
http://www.rsc.org/chemistryworld/2015/04/scientific-reports-fast-track-peer-review-trial-ends

Assent: UK researchers set to benefit from easier access to digital services, Jisc March 2015
http://www.jisc.ac.uk/news/uk-researchers-set-to-benefit-from-easier-access-to-digital-services-27-mar-2015

ChemTrove: Enabling a Generic ELN To Support Chemistry through the Use of Transferable Plug-ins and Online Data Sources. *J. Chem. Inf. Model.*, **2015**, *55* (3), pp 501–509
http://pubs.acs.org/doi/abs/10.1021/ci5005948

Protein-Ligand Interaction Profiler: a web service for analysis of non-covalent interactions in protein-ligand complexes from PDB files, simply enter a PDB code and the result page lists all detected non-covalent interactions (hydrogen bonds, water bridges, salt bridges, halogen bonds, hydrophobic interactions, π-stacking, π-cation interactions) in atom-level detail. The results can also be rendered using the embedded JSmol.
https://projects.biotec.tu-dresden.de/plip-web/plip/index

Replacing Photoshop With NSString
One for everyone's inner geek, a really clever way to create icons for applications using ascii art.
http://cocoamine.net/blog/2015/03/20/replacing-photoshop-with-nsstring/

ReseachKit: an OpenSource framework from Apple intended to help developers build applications for medical research
http://researchkit.github.io
How does a scientist's h-index change over time?
https://jeffollerton.wordpress.com/2015/05/10/how-does-a-scientists-h-index-change-over-time/

Review of Medicinal Chemistry Toolkit - this app is a growing suite of resources to support the work of medicinal chemists; created in conjunction with The Handbook of Medicinal Chemistry: Principles and Practice, a book based on the Medicinal Chemistry summer school run by the RSC.
http://www.macinchem.org/reviews/mctk/medchemtoolkit.php

Unleashing Technical Talent: Leveraging knowledge frameworks that combine complex data with advanced research capabilities and sophisticated analytics. White paper from IHS Solutions, 2014.
http://cdn.ihs.com/www/Tridion 2013/PDF/Unleashing-Technical-Talent-WhitePaper-FINAL.pdf

Old papers find new life online. Researchers on social media buzzed about an article that showed a growing citation rate for older papers.
Nature 26 Nov 2014
http://dx.doi.org/10.1038/nature.2014.16399

Chemistry Papers Rank High Among Once-Obscure Studies That Recently Racked Up Citations, C&E News June 2015
http://cen.acs.org/articles/93/i22/Chemistry-Papers-Rank-High-Among.html

Public attitudes to chemistry: results of a study on how the UK public thinks and feels about chemistry, chemists and chemicals.
http://www.rsc.org/campaigning-outreach/campaigning/public-attitudes-chemistry See also:
A first look at the findings:
http://www.rsc.org/news-events/rsc-news/articles/2015/jun/understanding-public-perceptions/
and commentaries in Chemistry World June 2015:
http://www.rsc.org/chemistryworld/2015/05/getting-know-you
and Chemical & Engineering News June 2015:
http://cen.acs.org/articles/93/web/2015/06/UK-Positive-View-Chemistry.html

The Oligopoly of Academic Publishers in the Digital Era
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0127502
See also Chemistry World June 2015: **Chemical sciences literature dominated by five publishing houses**:
http://www.rsc.org/chemistryworld/2015/06/chemical-sciences-literature-dominated-big-five-publishing-houses

Thomson Reuters 2015 State of Innovation Report: A number of UK Academic Institutions have been included for their influential and prolific research contributions.
http://stateofinnovation.thomsonreuters.com

And Finally…….

All set for chemistry - relive memories of your first chemical experiments with Philip Ball's article in Chemistry World on chemistry sets through the years:
http://www.rsc.org/chemistryworld/2015/05/chemistry-sets

Please send future items for the CICAG Newsletter to the newsletter editor:
Lindsay Battle, email: lindsay.battle@chem.ox.ac.uk

From Big Data to Chemical Information - Meeting Report

This meeting, organised by the RSC Chemical Information and Computer Applications Group (CICAG) and Dial-a-Molecule Grand Challenge Network took place on 22nd April 2015 at RSC, Burlington House, London.

The following report is provided by Colin Bird, University of Southampton

Introduction

The event was co-sponsored by RSC Chemical Information and Computer Applications Group (CICAG) and the Dial-a-Molecule Grand Challenge Network, which is funded by the EPSRC. The meeting brought together a diverse group of attendees interested in the challenges presented by "big data" and whether the chemistry situation might be different in any way. The morning session was devoted to talks assessing the scope and effect of "big data" from a chemistry perspective, while the afternoon session comprised talks about various approaches to managing "big data" and exploiting the opportunities that it presents. The day concluded with a keynote by Tony Williams, covering a notably wide range of topics related to RSC activities with large datasets.

Emergent themes

While it was to be expected that our speakers would offer different perspectives on the definition of "big data", it was perhaps less obvious that several of them would suggest that chemical data is not necessarily "big" data. Issues related to data integration arose in most of the talks and the need to capture metadata at source was one of the themes that arose more than once. A number of our speakers urged us to look first at what was already available before creating a new resource such as an ontology. While "big data" does present chemistry with diverse challenges, the tone of the meeting was optimistic: there are opportunities to meet those challenges.

Rise and Impact of Big Data

Big Data and the Dial-a-Molecule Grand Challenge
Richard Whitby

As the Principal Investigator for Dial-a-Molecule, Richard presented a range of big numbers: quantities that underpin the challenges associated with making novel molecules quickly. The real challenge for the synthetic organic chemist lies in deciding how to plan a synthesis so that you know it will work. Consequently, organic synthesis has to change from its current compound-driven approach to being a data-driven discipline. It is the data that has the lasting value.

Richard suggested that the predictability of a reaction sequence was analogous to a chess problem, except that we do not know enough about the weightings, owing to a lack of data. To improve that situation, we need to capture data at source, especially for reactions that we deem to have failed. Big numbers are involved, the largest being the estimated upper limit of the reaction space in terms of connections between molecules (10^18 to 10^200). Although we can to some extent reduce the reaction space with techniques such as functional group approximation, the space remains huge; it is difficult to say where we are, as the amount of information is still restricted.

Richard reviewed four computer-aided synthesis design programs that are currently in use, but concluded that they are essentially ideas generators. He contended that it should be possible to use data more effectively, illustrating his point with an example having limited information in the reaction database, more in the publication, but with far more information not there. More data will become available, particularly from ELNs, but researchers have to be willing to share and provide data in a format that can be exchanged. Automated capture of experimental conditions can help considerably, as experimenters tend to record only a fraction of what they actually observe. How we make effective use of the flood of data that will be produced will be a real big data challenge: should we keep it all or could we throw some away? Ultimately, the challenge is to make the most effective use of the existing data to predict reaction outcomes.

Big, broad and blighted data
Jeremy Frey

Chemical data is diverse and heterogeneous, which breadth differentiates it from the big data produced by, for example, the Large Hadron Collider. However, chemical data is sometimes not what we think it is, so might as a result be 'blighted'. These characteristics arise in part because chemical data comes from a lot of

sources of different sizes: the distribution has a long tail, as it does when chemical data is analysed by country of origin. There is a huge amount of heterogeneous information now available from a lot of countries around the world. Moreover, since 1980, chemical information publications have given way to chemical informatics.

The use of social networking has increased the amount of user-generated content, but in a form that is potentially unprocessable. Such content might even include information about failed reactions, albeit emerging by unconventional routes. The community could also become involved in carrying out work, although questions of control inevitably arise.

Echoing Richard's message, Jeremy advanced the need to automate data capture, emphasising the importance of metadata, which people are known to be reluctant to assign. Metadata has to be captured at source; there are risks with adding it later. Insight and further information are essential for climbing the Data-Information-Knowledge-Wisdom pyramid, as Jeremy illustrated by reference to black bananas.

Semantic web technologies offer hope, with the caveat that human understanding of machine-machine interactions is important; otherwise we will not trust the findings. Typically, the further information we need will represent context, as Jeremy illustrated by reference to keto-enol tautomerism and to water. Work is going on with chemical schema and ontologies, including building links to other disciplines.

Jeremy then introduced this theme of *reducing uncertainty* (which we might relate to increasing understanding), illustrating his point with images of the Amazon and Okavanga deltas. The analogy is with data streams coming together to form a data river that then spreads out into a delta. However, the Okavanga spreads into a desert, rather than an ocean, which would not be useful in the case of data. Exploiting the integrated data relies not only on knowing its provenance, but also being confident about its correctness. If an assertion is subsequently shown to be untrue, it is almost impossible to remove that defect, because the information necessary to do so is not there. The challenge is to get people to do the work to ensure their data is reusable.

Digital disruption in the laboratory: joined-up science?
John Trigg

After teasing us with a "spot the difference" between pictures of labs old and new - his answer being that it is no longer necessary to wear a tie when working in a lab - John embarked on a comprehensive overview of the transformation wrought by the evolution of digital technologies. The result has been a fundamental change in the way we communicate, arising from the implementation of digital technologies, causing disruption. We no longer rely on a third party; we can do things ourselves. We have to manage IP differently; to adopt different business models, doing more with less; and to adjust our scientific method to be conscious of data curation, provenance, integrity, and preservation.

Against a background of knowledge and expertise being dispersed geographically, and chemistry becoming more complex and less certain, John used the Snowden and Stanbridge landscape of management diagram to reason that the current drive is towards process engineering (rules and order), because social complexity (un-order and heuristics) is difficult to manage.

John believes that the nature of laboratory work will change, creating a need for more education (for understanding) as opposed to training (for doing). Currently we have too much of the latter, raising the question: what happens if/when no longer need cognitive input? The Internet of Things is increasing the number of devices with machine-machine protocols, giving us unprecedented opportunities to exploit new technologies, provided we increase our understanding and do not rely on black-box technology without cognitive input.

John concluded with Charles Darwin's quote to the effect that the species that survives is the one most responsive to change.

Big data chemistry
Jonathan Goodman

Jonathan began by asking whether chemistry has big data, inviting comparison with astronomy, which can generate more data than all the CCTV cameras in the UK. He cited the Wikipedia view that big data is characterised by being difficult to process with traditional techniques, noting that chemistry has few reactions that we really understand and many more that we would like to understand.

He depicted a machine for making molecules as a box, which in its Mark III incarnation would take sunlight plus raw ingredients such as O_2, N_2, and CO_2 and deploy a library of processes to make the best molecules

with specific properties. He asked us to identify the simplest molecule that we could not make: his answer was tri-t-butyl isopropyl methane.

Jonathan then suggested that we need different ways of looking at molecules, while acknowledging that there would be some resistance to change. To build his Mark III machine, we will also need new models for 'understanding' chemical data. Not all of the data that we would need is openly available, yet if life depends on it, we will want to know that a structure is correct, giving the debate about the structure of maitotoxin as an example. Even when we have a lot of data, do we understand it?

How long might it take from an invention to a derived product being available from Tesco, for example? Jonathan gave several examples: Teflon to non-stick pans; lasers to CD players; and (somewhat tongue-in-cheek) molecule-making machines to ready meals.

Discussion
Panel comprising the first session speakers

Asked whether chemists would be happy publishing failed reactions, Richard Whitby replied that we need a culture change: we know there are errors, so should not be afraid to disclose them. John Leonard queried what a failed reaction might be, as a 1% yield might still be regarded as a success.

In response to a question about the scientific method, Jeremy Frey said that it should be taught in schools, adding that ethics has to be dealt with early, a point reinforced by John Trigg.

The observation that, under US Patent law, machines cannot make inventions elicited a range of responses. Jonathan Goodman saw it as a challenge for lawyers; Jeremy Frey argued that encouragements for creativity deserve rewards; John Trigg pointed out that inventions are kicked off ideas, which prompted Jeremy to point out that cognitive computing (such as IBM's Watson) could suggest new ideas.

Asked why chemistry lags behind in depositing data to a repository, as required for open access, Jonathan Goodman observed that it was intrinsically difficult, then Donna Blackmond said that funding agencies require a plan. Jeremy Frey asserted that intelligently accessible data is the lifeblood of future progress. Depositing data needs to to become good, standard practice.

Approaches to Managing Big Data and Maximising Opportunities

Managing and searching large chemical structure data resources
Mark Forster

Mark's perspective was that chemical data is not necessarily big data, but computations could be big. As part of its portfolio, Syngenta is interested in developing new pesticides, which, with *in vivo* testing, can go from hypothesis to bioactivity testing in a few weeks, so the model for compound discovery is quite dissimilar to pharmaceuticals.

Mark's first example of the large structure datasets that Syngenta manages was ChEMBL: there are 28,000 compounds in the pesticide literature that are not in ChEMBL, but are now being added as a result of a collaboration between Syngenta and the EBI ChEMBL group.

Syngenta are investigating new search processes to find candidate compounds, surveying both corporate and vendor compounds: for example, Openbabel Open Source Chemistry Toolbox and Chemfp which uses fingerprints. They also use a script to find 'new' compounds using the eMolecules public chemical data set, extracting those not seen previously and performing property calculations and analyses. Pesticide physical property scoring produces HFI similarity scores: H(erbicide), F(ungicide), I(nsecticide) likeness. They also calculate a compound's novelty relative to Syngenta corporate compounds. Mark illustrated the use of Knime to filter and visualise structure, with the data flow set by search criteria. He showed a scatter plot in which size represented H score and colour the Novelty. Mark noted that they use InChI keys for linking data, observing that one can sometimes can put an InChI into Google and find a web page about the compound.

Data-rich organic chemistry: enabling and innovating the study of chemical reactions
Donna Blackmond

Donna's presentation was based on a two-day NSF-sponsored workshop held in Washington, DC, in September 2014. She brought copies of the reports for us to take away. The motivation for the workshop came mainly from the pharmaceutical industry, their collective interest being in: enabling technologies for capturing process information; precompetitive collaborations; and promoting the use of tools for reaction monitoring.

One aim of the workshop was to find new ways to fund academic research and to train the next generation of 'workers'. There were five talks about models for collaboration. The Caltech model is that of a central catalysis facility where different research groups can access both sophisticated instrumentation and expertise. The collaboration with Merck, for example, operates by Merck offering postdocs short assignments in the company, working in ways that will not harm competitive aspects.

The workshop also covered recent progress with pre-competitive collaboration models. The Pfizer approach relies on data being transportable, which requires compatibility of software, enabling the integration of data into a searchable architecture. For Merck, data-rich tools should be able to be to run without headaches, so that they can make use of the data.

Donna then talked about the need for transformative solutions: obtaining quality in a way that can accelerate development with fewer people. Pfizer are developing the concept of the Lab of the Future, which would require new skills and education as well as training.

Among the challenges is the development of a common data framework, which the Allotrope Foundation is working towards, developing standards and aiming to: improve integrity, reduce waste, realize the full value of the data, bridge the gap between ideas and execution.

The IQ Consortium aims to share ideas without hurting commercialisation, so the Allotrope Foundation and IQ are dealing with IP issues and seeking agreement about what collaborators are looking for. Donna envisaged the possibility of a new heyday of physical organic chemistry with new tools, asking what might have been if the pioneers of the discipline had had our tools. Achieving the aims requires education to enhance critical skills in data-rich science.

The workshop identified four ideas that would be important for the future, of which were given priority: developing new educational models and the development of a Caltech-like centre for data-rich experimentation. Donna expects these topics to arise at the next CCR Meeting, in May 2015, in a session entitled "Disruption in Biotechnology and Process Chemistry".

Use of data standards and metadata in information exchange
Rachel Uphill

Rachel began by identifying categories of big data, such as: gene expression profiles; interactions; reactions in our bodies; and citations. Pharmaceutical companies have a lot of data, which is increasingly complex and of higher dimensionality. They also get data from other sources, albeit often as PDFs, which might contain structures, but do not really use it.

Metadata is the answer. To integrate substance, result, experiment, and project data, we have to rely on metadata; there is no point in storing big data and not doing anything with it. However, questions do arise about the integrity of data, sometimes epitomized as Garbage In Garbage Out (GIGO). Without the right data, and the right metadata, we are not going to get correct answers.

For GSK, the need is to be able to use the data, adding standards and then embedding the tools according to an information blueprint. This process involves stewardship and governance, to find, understand, and use the data, and to integrate with the information blueprint. There is a range of requirements and measures to give trust in the data and to enable its use. Master Data Management (MDM) provides one reference point: one view of the information.

With regard to standards, that is where the Allotrope Foundation comes in. Data held in Allotrope format does not lose context, so we can look back at its provenance. Rachel advised against creating a new standard without looking first at what is already out there; Allotrope is looking at the gaps, aiming for an open document standard and open metadata repository. Allotrope is also integrating the regulatory aspects, which lead to more requests for information.

Rachel concluded with an example from GSK, joining datasets together in different ways, integrating from external as well as internal sources.

100 million compounds, 100K protein structures, 2 million reactions, 4 million journal articles, 20 million patents and 15 billion substructures: Is 20TB really big data?
Noel O'Boyle

Citing the Wikipedia article, Noel suggested that any dataset could be considered big data if we lack the means to process it, giving examples of large numbers of 'things', adding some wry comments.

He went on to talk about searches for matched pairs [2] and matched series [>=3] in the ChEMBL dataset, which identified 391,000 matched series. In contrast substructure searches are relatively slow, especially when compared with a typical Google search, which can be very fast, owing to its look-ahead feature. Ideally, a sketch search should be underway, using a similar feature. A fingerprinting screen would be fast, although it would produce false positives, and could be followed by slow matching. However, in some situations, such as a structure containing benzene, a fingerprinting screen is not very effective. As the worst-case behaviour arises from slow matching, Noel went on described attempts to speed that up.

The approach is to pre-process the database, matching the rarer atoms first, which Noel showed to be significantly faster, with or without fingerprinting. An alternative approach is to pre-process all substructures, using NextMove's SmallWorld technology, which takes a lot of time and requires a lot of space. Maximum common subgraph techniques are computationally expensive but can be implemented efficiently using SmallWorld.

NextMove also have text mining technologies, which extract chemical names from text, and can find ~90% of the structures (131,000) in all the open access papers from PubChem. Noel ended with his view that many classic cheminformatics problems can be handled with today's techniques.

Dealing with the wealth of open source data
John Holliday

John began with overviews of the Sheffield research areas and the open source data available circa 1999/2000. In comparison, he showed a "scatter plot" of the open resources now available, comprising drug databases and compound databases, offering breadth as well as depth.

Sheffield will be using new as well as old techniques to investigate approaches such as hyperstructures, virtual screening, and data fusion; they are using CASREACT for reaction schemes. They are also exploring cross-database integration issues, for example, multiple formats, with various databases distributed in various formats; consistency (e.g., gaps) is a problem that they cannot do much about. If a test has not been done, they cannot use the data. However, they can report the issue back to EBI, for example, asserting that a particular assay is wrong.

With unstructured data, the question arises whether one can be confident that the data is right. There are now more data types and chemical mime types, such as XML formats, including CML: essentially there are too many formats from too many different sources. Translation is feasible, but can get some loss of data in the process. Looking ahead, we might evolve standard format(s) by virtue of the way we use the data. John thought the situation could settle down with time, as everyone starts to use the same formats.

John then considered data management issues, such as: how to back up databases holding many terabytes; and the need for the right metadata, with the right metadata vocabulary, so they have to enforce the use of metadata. The overall need is for a proper management system. At Sheffield all databases are on MySQL, with a "nice" new front-end.

Using car design in 60s and 90s as his illustration of "soul", John argued for more human input into the design of decision support systems: include some "soul". We have to make sure our output is communicating to everyone. Sheffield uses benchmarking: they now have several benchmark sets for screening databases, where they used to use one dataset. People are now becoming more data aware, but have to pick and choose what is best for them. Although we are all becoming data scientists, programming skills are going down, despite an increase in the use of tools.

Discussion

Tom Hawkins noted that the examples used had been around the paradigm that a molecule has a single structure and a reaction a single result, then asked what was available in cheminformatics to support polymers and mixtures of products. Mark Forster replied that we can register a chemical without a structure; Rachel Uphill that we can register different structures with relationships between them.

Asked how today's tools would scale and what might happen in the future, Noel O'Boyle replied that NextMove tools all scale well and many are parallelisable. Rachel Uphill added that tools now operate on the databases: they are no longer downloaded.

A question about the lack of incentive and credit for publishing good, reproducible, data elicited several responses. Donna Blackmond noted that the "incentive to publish" system is working pretty well; John Holliday pointed out that there is a lot of data available now, so making the data fit the hypothesis might become an issue. Jeremy Frey then queried the possibility of safe "exchanges", to which Donna responded

that there is so much incentive now, although there is concern about other groups knowing; however, she and her fellow moderators are trusted. Jeremy asked whether we could create blueprints, Donna replying that the NSF case studies were going to lead to some form of blueprint. Rachel Uphill proposed a hosting centre, so that each company doesn't have to go through the process each time.

Speakers taking part in a discussion at the end of the afternoon session of CICAG's Big Data meeting. From left to right: Prof Donna Blackmond (Scripps Research Institute), Dr John Holliday (University of Sheffield), Dr Mark Forster (Syngenta), Rachel Uphill (GlaxoSmithKline), Dr Noel O'Boyle (NextMove Software, Cambridge).

Keynote: ***Activities at the Royal Society of Chemistry to gather, extract and analyze big datasets in chemistry***
Tony Williams

Using a colourful slide, Tony illustrated big data in terms of the number of things going onto the web in 60 seconds, then showed a count of 95,736,025 substances in the CAS Registry at the time of capture in the afternoon of 22nd April 2015. Tony then traversed more chemistry-related numbers: the prophetic compounds in patents; the compounds in PubChem, including "similar" compounds that require manual curation; proliferation of InChIs, enables Google to access over 50 million records; and the chemicals held by ChemSpider. However, the reality is that relative to brontobytes (10^27), the numbers are not "big data".

The RSC has taken up open access and also open data, although it leads to some problematic conversations: funders tell you to make data available, but not how to do it; do you put data in a repository that might not be supported tomorrow? There is not as much open chemistry data as there should be. Some teams will want open access but be reluctant to release their own data, saying that it is "really important".

Turning to the scientific literature, Tony cited ContentMine, which claims to "liberate 100 million facts", but queried whether they were really facts or actually assertions about a particular measurement. The RSC has published more than 36,000 articles in 2015, posing several "how many" questions. However, much information is lost, particularly relationships, as publications are only a summary of work. Trying to find work from years ago is a problematic area, especially as much of the data in our hands still lies in PDF files. Data in publications should be available; it should not be locked up. Tony posed the question: how much data might be lost to pruning? Nobody will *rush* to publish to the Journal of Failed Reactions, so how much data is thrown away? How much data resides in ELNs? It would be great to sit at an ELN and make requests. How many compounds are made that are never reported? Tony thought he had probably published less than 5% of the work he did; the rest is mostly lost. There are data management systems in most institutions, so it should be feasible to share more data.

Tony then talked about his experiences with computer-assisted structure elucidation (CASE) and in associating structures with NMR spectra and selecting the highest ranked. One of the challenges of data analysis is access to raw data. For example, if 3 NMR peaks lie close together, they cannot be resolved from

an image; you need the data in a CSV file. We publish into document formats, from which we have to extract data, whereas they could conform to a community norm. Currently, there isn't even a reference standard. Tony argues that we can solve these problems. In ChemSpider, supplementary spectra information is in JCAMP format: analytical data should be produced in standard rather than proprietary formats.

Mandates do not offer data deposition solutions: we have to build them. The RSC will offer embargoing, collaborative sharing, and links to ELNs. There are standards, although students are not taught about them. There are also ontologies I use, so we should not create new ones.

Tony illustrated the issue of data quality with several examples, such as: detecting corrupted JCAMP files that have got flipped; an allegedly "high quality dataset" giving Mn++ as the symbol for a selenium oxide cadmium salt; a database with only 34 correct structures out of 149; and several other telling examples. His final example was that of domoic acid, for which C&E News had taken the (wrong) structure from Wikipedia rather than from SciFinder, because Wikipedia was free!

Tony then gave the Open PHACTS project as an example of ODOSOS (Open Data, Open Source, Open Standards) before moving on to open source validation with CVSP (the Chemical Validation and Standardisation Platform), asking whether publishers could use it before submission if all rules were available. He noted that when he started his work, 8% of structures on Wikipedia were wrong: checking and correcting took 3 years.

After mentioning the RSC Archive, Tony gave an example of a reaction description, comprising a diagram and a method: the description would be useful, but we still do not know the context: that's in the publication.

With regard to modelling "big data", Tony talked about melting point models, showing a relatively narrow distribution. He went on discuss building a database of NMR spectra, noting the problems that can occur with names, especially those with brackets. Overall, there are issues with textual descriptions, such as erroneous and incomplete information.

In conclusion, Tony remarked that we are sitting on big data: what it takes is to apply the techniques and standards.

Discussion

Asked why, in the context of open data and open publishing, OpenArchive was not popular in the chemistry community, Tony replied that there was so much value in chemicals, so a reluctance to put data up.

With regard to data quality, it was suggested that a button to report errors would be very useful and should be encouraged. Tony said that ChemSpider publishes changes, but no one wants to take the feed. Responding to a question about publishing data, Tony said the need was to publish data without extracting it.

He was then asked about big data links to other sources and about appropriate curation of keys in other people's databases. Tony said that was what Open PHACTS was about: linking with biological data. It had a very specific focus; one would have to deal with appropriate organisations to deal with other areas.

Posters

Towards statistical descriptions of crystalline compounds
Philip Adler *(University of Southampton)*

This poster presented research demonstrating the use of statistical methods to address relationships between molecular and crystallographic structure. It illustrated example problem domains, and discussed issues with the methodology, both in terms of difficulty with statistical methods, and problems with gathering data in a standardised fashion from the published literature. In particular, sparse and uneven coverage of the chemical space by the literature, especially with respect to 'failed' experiments, has proven to be a large hindrance.

Multiparameter optimization of pharmaceuticals: What 'BigData' can tell us about small groups that make a big difference
Al Dossetter *(Med Chemica)*

Matched Molecular Pair Analysis (MMPA) shows promise for assessing the pharmacology of new biological targets but the process requires many matched pairs to achieve statistical significance and thus new design rules. Combining and analyzing data from many pharmaceutical companies is both cheaper and faster than

making and screening large numbers of compounds. To enable the contributing companies to share knowledge without exposing their intellectual property or critical data, datasets are encoded using only *changes* in structure and property. This poster illustrated the process that created and analyses the 'big data' involved, illustrated with examples of rules found within ChEMBL toxicity data.

Physical chemists' attitudes towards management of laboratory data
Isobel Hogg *(Royal Society of Chemistry)*

This poster presented research into the data management needs of physical chemists, who often face difficulties owing to the quantity of data that they generate. The research also discovered issues with recording the narrative that goes alongside experiments and provides context to the data recorded in the lab, leading to an investigation of how physical chemists currently manage their data and the extent to which their needs would be served by electronic laboratory notebooks (ELNs). The conclusion was that improved note taking and data organisation within a comprehensive system could improve working practices but data sharing would not be a strong driver for the adoption of an ELN.

Batch correction without QCs
Martin Rusilowicz *(University of York)*

Quality Control (QC) samples are often used to assess and correct the variation between batch acquisitions of Liquid Chromatography – Mass Spectrometry (LC-MS) spectra for metabolomics studies. This poster showed how the use of QC samples could lead to certain problems. As an alternative, "background correction" methods use all the experimental data to estimate the variation over time, rather than relying on the QC samples alone. The poster reported comparisons of non-QC correction methods with standard QC correction.

From Big Data to Chemical Information - Student Reports

Four students received a bursary to attend, and provided the following reports of their own experience of the meeting.

Report from Ramatoulie Camara *(University of Hertfordshire):*

I would like to express my heartfelt gratitude for your generosity in providing me with this opportunity to attend the recently concluded RSC-CICAG / Dial-a-Molecule "From Big Data to Chemical Information" in London. The meeting was a great insight into the relationship between big data and the chemical sciences.

The program lined up interesting topics and speakers that I found very interesting. The talk by Dr Mark J Forster from Syngenta R&D on Managing and searching large chemical structure data resources did resonate with me. As part of my PhD I have searched and profiled thousands of compounds to compile a list of suitable candidate compounds for my research. It was reassuring that some of the techniques used within this team at Syngenta had a great deal of similarity to the methods used in my own research. His talk also offered some other techniques that could be of use in my future work. I have also learned about some freely available resources that I had no prior knowledge of that could be beneficial to my work.

The keynote speech delivered by Dr Tony Williams to end the meeting was worth waiting for. His talk on the Activities at the Royal Society of Chemistry to gather, extract and analyse big datasets in chemistry highlighted the challenges faced by the RSC in trying to preserve chemical information for generations to come. His excellent talk also delved into the problems his team at the RSC has encountered and some of the solutions they have come up with when dealing with the semantic accuracy and representation chemical structures in publically available data sets. His departure from the RSC will be missed and it is hoped that the good work started will continue to prosper.

Report from Mudasser Rafiq *(University of Liverpool):*

From Big Data to Chemical Information – my experience

The "From Big Data to Chemical Information" meeting was the first time I attended a CICAG meeting. It gave me the opportunity to meet a range of people from both academic and industrial backgrounds and discuss the projects they are working on. It also allowed to gain advice about my project and learn about some of the issues in cheminformatics.

Attending the meeting has helped my research as it gave me a clearer understating of how data associated with different chemicals is collected and stored. The talks enhanced my understating of the amount of data which is available and the potential for it to increase as people move towards electronic lab books and automated laboratories. This meeting also introduced me to some tools developed by next move software such as LeadMine and CaffeineFix.

I saw how chemical data can be extracted from existing literature by using these tools and how the data can be used to build models which can predict some of the properties of compounds. These predictions can then be compared literature to recognise errors in existing literature and help identify mistakes in research before it is published.

The talks also emphasised the importance of sharing data both published and unpublished to help create databases which can help other to design reactions and verify their results.

I would like to thank the CICAG for providing the bursary and I hope to attend future events where I hope to have an opportunity to present my own research and learn from others.

Report from Martin Rusilowicz *(University of York):*

Thank you for inviting me to the From Big Data to Chemical Information meeting on the 22nd March. It was a great opportunity to be able to put my research into context with other researchers' work in the field and there were some interesting discussions. I particularly enjoyed the "dial a molecule talk" and "big data chemistry" talks by Professors Whitby and Goodman. I am currently researching the effects of drought and disease in crops at the metabolite level and since I am often involved in identifying the potential pathways of resistance marker molecules it was especially interesting to see similar problems being tackled in reverse – finding out how to use data to propose new syntheses. I will be looking up many of the novel ways of combining data discussed to see if I can put them to use. There were also some great opportunities to network and I'd like to thank everyone for creating such a welcoming and friendly environment. I received some great feedback on my research on batch correction with some thought provoking suggestions on novel uses.

Report from Matt Swain *(University of Cambridge):*

The RSC CICAG "From Big Data to Chemical Information" meeting was a great opportunity to meet a diverse group of people with a common interest in tackling the challenges of dealing with large and complex chemical datasets. It was very interesting to hear many speakers reflect on what we really mean by "big data" and whether we actually have it in chemistry, given that the actual quantity of data may not be as large as in other fields. In these talks and the subsequent discussions, many people made the compelling argument that, regardless of quantity, the complexity, variety and frequently unstructured nature of chemical data is more than enough to justify the title "big data".

Noel O'Boyle, Mark Forster and many others gave highly informative talks on their current tools and methods for tackling "big data" challenges in chemistry, providing useful ideas that are already informing my current research. Meanwhile, Richard Whitby and Jonathan Goodman inspired and entertained with more forward-looking talks that prompted great discussions around the biggest challenges and opportunities for the future.

A particular highlight of the day was the keynote talk by Antony Williams, in which he presented the activities of the RSC in managing large chemistry datasets. As well as highlighting the role of ChemSpider as a repository for depositing structured chemical information, he shared some fascinating insights into text-mining projects focused on exploiting the huge quantity of unstructured data present in the archive of RSC publications. It was particularly useful to hear about his approach to automatically extracting spectroscopic data and physicochemical properties from the literature, and using these to build predictive models, as I use similar methods in my own research.

Many thanks to the CICAG and Dial-a-Molecule teams for organising.