Subcommittee on Nomenclature for Properties and Units (NPU)

In laboratory medicine one of the most basic challenges is to ensure that there is a common understanding of what is being measured in a biological system, as well as how the results will be expressed and in what units. To address this issue, the subcommittee has partnered with the International Federation of Clinical Chemistry (IFCC) and the Danish National e-Health Authority (DeHA) to develop, test, and refine an intuitive and comprehensive NPU terminology. This is essential to providing quality assurance and to unequivocally interpreting the results of clinical laboratory analysis. In 2014, a formal agreement between the three partners was developed to provide a template for greater international promotion of the NPU terminology as an aid to harmonized practice and better patient safety.

Scientists interested in participating in activities related to Chemistry and Human Health are invited to contact the Division President, Tom Perun <tjperun@aol.com>.
www.iupac.org/body/700

From Big Data to Chemical Information

by Colin L. Bird and Jeremy G. Frey
Chemistry, University of Southampton

A meeting on Big Data [1] was jointly organized by the RSC Special Interest group on Chemical Information and Computer applications (CICAG) [2] and the UK Engineering and Physical Sciences Research Council Grand Challenge of Dial a Molecule (DaM) [3] and held 22 April 2015 at the Royal Society of Chemistry, Burlington House, London, UK.

"Big data" is very much a current term, for chemistry no less than for other disciplines. While there is an understandable tendency to interpret "big" as "voluminous", scope is an equally important yardstick for chemical data. Meeuwis van Arkel summed up the situation in a letter to *Chemistry & Engineering News*, [4] "Chemists need information from a multitude of different sources, each with its own origins. But there's a huge gap between volume and relevance that needs to be bridged. … Big data must be focused on breaking huge blocks of information down to the smallest particles. Only when we can ensure that our tools enable confident decision making at every stage of chemical research will we realize big data's value rather than feel as if we are drowning in the chaos of too much."

In April 2015, the RSC Chemical Information and Computer Applications Group (CICAG) and the EPSRC-funded Dial-a-Molecule Grand Challenge Network co-sponsored a meeting: "From Big Data to Chemical Information" (programme available from the RSC web site). [5] The morning session addressed the "Rise and Impact of Big Data"; the afternoon session considered "Approaches to Managing Big Data and Maximizing Opportunities", then concluded with a keynote by Tony Williams, "Activities at the Royal Society of Chemistry to gather, extract and analyze big datasets in chemistry". While it was to be expected that the speakers would offer different perspectives on "big data", it was perhaps less obvious that several of them would suggest that chemical data is not necessarily "big" data. Nevertheless, consistent aspects were the heterogeneity, high dimensionality, and complexity of chemical data, the utilisation of which is often complicated by uncertainty.

Challenges

Richard Whitby (University of Southampton) began by presenting the Dial-a-Molecule challenges associated with making novel molecules quickly, the main issue for the synthetic organic chemist being in deciding how to plan a synthesis such that we know it will work. Consequently, organic synthesis will have to change to being a data-driven discipline. At present, we do not know enough about reaction outcomes, and so need to capture data at the source, especially for reactions that we deem to have failed. The reaction space is huge, so it is difficult to say where we are, as the amount of information is still restricted. Current computer-aided synthesis design programs are essentially idea generators. Richard contended that it should be possible to use data more effectively, particularly by getting more information into reaction databases rather than in publications.

Jeremy Frey (University of Southampton) introduced issues that can arise from the diversity and heterogeneity of chemical data, noting that it comes from a lot of sources of different sizes, so some data might not be what we think it is. The use of social networking has increased the amount of user-generated content, but in a form that is potentially not processable. Such content might even include information about failed reactions, albeit emerging by unconventional routes. Echoing Richard's message, Jeremy advanced the need to automate data capture, emphasising the importance of metadata, which researchers are known to be reluctant to assign. Metadata has to be captured at the source; there are real risks with adding it later. Semantic Web technologies offer hope, with the caveat that human understanding of machine-machine interactions is important; otherwise we will not trust the findings. The objective must be to reduce uncertainty.

Digital Transformation

John Trigg (RSC) embarked on a comprehensive overview of the transformation wrought by the evolution of digital technologies, resulting in a fundamental change in the way we communicate, which in turn causes disruption. John believes that the nature of laboratory work will change, creating a need for more education (for understanding) as opposed to training (for doing). The "Internet of Things" is increasing the number of devices with machine-machine protocols, giving us unprecedented opportunities to exploit new technologies, provided that we ensure that we retain our cognitive input.

Jonathan Goodman (University of Cambridge) began by comparing chemistry with astronomy, which generates a large amount of data. 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. Using his model of a machine for making molecules, depicted as a box that takes sunlight plus raw ingredients as input, Jonathan suggested that we need different ways of looking at molecules, while acknowledging that there would be some resistance to change. 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.

Tony Williams (RSC) [6] introduced his keynote by illustrating big data in terms of the number of things going onto the Web in a 60 second period, then showed a count of substances in the CAS Registry that was over 95 million on the day of the meeting. [7] Traversing a range of chemistry-related numbers, Tony contended that these were not, in reality, "big data". The RSC has taken up both Open Access and Open Data, but there is still not as much open chemistry data as there should be. Some teams will want open access but nevertheless be reluctant to release their own data, arguing that it is "really important". However, much information is lost, particularly relationships, as publications are only a summary of work. Such data should be available, not locked up. Tony posed the questions: "How much data might be lost to pruning? Nobody rushes to publish in the Journal of Failed Reactions, so how much data is thrown away? How much data resides in Electronic Laboratory Notebooks (ELNs)?" 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 reviewed his experiences in a variety of areas: computer-assisted structure elucidation, associating structures with NMR spectra; data deposition; data quality, including the detection of corrupted files; the Open PHACTS project as an example of ODOSOS (Open Data, Open Source, Open Standards); [8] reaction description, noting that we rarely know the context, because it is in the publication. In conclusion, Tony remarked that we are sitting on big data: what it takes is to apply the techniques and standards.

Chemistry data small and large

The afternoon session opened with Mark Forster (Syngenta AG) offering the perspective that chemical data is not necessarily big data, but computations could be big. As part of its portfolio, Syngenta develops new pesticides. Unlike pharmaceuticals, these pesticides, with in vivo testing, can go from hypothesis to bioactivity testing in a few weeks. To facilitate their candidate compounds searches, Syngenta are adding to ChEMBL [9] the 28,000 compounds in the pesticide literature that are not in that database, and are also investigating new search processes, surveying both corporate and vendor compounds. Searches produce a pesticide physical property score based on HFL similarity scores: H(erbicide), F(ungicide), L(ikeness) and they also calculate a compound's novelty relative to Syngenta corporate compounds.

Training

Donna Blackmond (The Scripps Research Institute, California) then presented the outcomes of a two-day NSF-sponsored workshop held in Washington, DC, in September 2014. The motivation for the workshop came mainly from the pharmaceutical industry, one aim being to find new ways to fund academic research and to train the next generation of workers. The workshop also covered recent progress with pre-competitive collaboration models, which require the integration of data into a searchable architecture. Donna then talked about the need for transformative solutions: obtaining quality in a way that can accelerate development with fewer people. 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, realise the full value of the data, and bridge the gap between ideas and execution.

Standards

Rachel Uphill (GlaxoSmithKline [GSK]) took up the theme of data standards and metadata in information exchange, initially 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. To integrate substance, result, experiment, and project data, we have to rely on metadata, although questions can and do arise about the integrity of data. Without the right data, and the right metadata, we are not going to get correct answers. Stewardship and governance are important, so GSK uses Master Data Management (MDM), [10] with a range of requirements and measures to instill trust in the data and to enable its use. With regard to standards, GSK is a member of the Allotrope Foundation. [11] Data held in Allotrope format does not lose context, so we can look back at its provenance. Allotrope is also integrating the regulatory aspects.

Speed

Noel O'Boyle (NextMove Software) [12] suggested that any dataset could be considered big data if we lack the means to process it, going on to give examples of searches for matched pairs (2) and matched series (>=3) in the ChEMBL dataset, which identified 391,000 matched series. Such searches are relatively slow, especially when compared with a typical Google search, so Noel described NextMove's attempts to speed up matching. Their approach is to pre-process the database, matching the rarer atoms first, which Noel showed to be significantly faster. 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. [13] Noel ended with his view that many classic cheminformatics problems can be handled with today's techniques.

Open Source

John Holliday (University of Sheffield) addressed the management of open source data, comparing the resources now available with those obtainable circa 1999-2000. 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. [14] They are also exploring cross-database integration issues, for example, multiple formats, with various databases distributed in various formats. Consistency can be a problem: can we be confident that the data is right? There are now more data types and chemical mime types, such as XML formats, including CML. [15] Essentially there are too many formats from too many different sources. Looking ahead, we might evolve standard formatting 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.

Four posters were also on display at the meeting, covering the following topics: analysing matched molecular pairs for assessing the pharmacology of new biological targets, attitudes to laboratory data management among physical chemists, correction of variations in LC-MS data for metabolomics studies without using quality control samples, and statistical methods to address relationships between molecular and crystallographic structure.

Lively discussions followed each session and the keynote address and it was widely agreed that the meeting had offered broad coverage of the issues relating to big data and chemical information and that the talks and discussions had been interesting and stimulating.

References

1. www.rsc.org/Membership/Networking/InterestGroups/CICAG/meetings.asp
2. CICAG www.rsc.org/Membership/Networking/InterestGroups/CICAG/
3. DaM http://generic.wordpress.soton.ac.uk/dial-a-molecule/
4. C&E News, 92, 2, http://cen.acs.org/articles/92/i2/Deal-Big-Data.html
5. www.rsc.org/events/download/Document/479d45aa-bc86-44c0-8391-a773904a11f6
6. Tony Williams is now at the EPA
7. www.cas.org/content/chemical-substances
8. www.openphacts.org/
9. www.ebi.ac.uk/chembl/
10. www.gartner.com/it-glossary/master-data-management-mdm
11. www.allotrope.org/
12. www.nextmovesoftware.com/
13. https://pubchem.ncbi.nlm.nih.gov/
14. www.cas.org/content/reactions
15. www.xml-cml.org/

This article is based on the meeting report prepared by Colin Bird, which can be found at http://www.rsc.org/images/BigData-Meeting-Report_tcm18-246660.doc.