SUSTAINABILITY & DIVERSITY THROUGH CHEMISTRY

Special Symposium on Research Data, Big Data, and Chemistry

Date: Thursday, 13 July 2017

Symposium Organizers: David Martinsen, ACS (Retired), currently David Martinsen Consulting, Rockville, MD, USA; Martinsen.david@gmail.com; Leah McEwen, Chemistry Librarian, Cornell University, Ithaca, NY USA, lrml@cornell.edu.

Contents
CHEMISTRY International July-September 2017 Volume 39 No. 3

Special Issue on Big Data

Secretary General’s Column
Research Data, Big Data, and Chemistry by Richard Hartshorn 2

Guest Editors’ Introduction
The Rise of Primary Research Data by Leah McEwen and David Martinsen 3

Features
Connecting Chemistry with Global Challenges through Data Standards by Jan Bruno and Jeremy G. Frey 5
The Future of Chemical Information Is Now by Antony J. Williams and Harry E. Pence 9
Are We Nearly There Yet? A Perspective on Data Sharing in (Chemical) Crystallography by Simon Coles 15
Managing Standards and Critical Evaluation in a World of Big Data by D. Brynn Hibbert*, David Shaw and M. Clara F. Magalhães 22
The IUPAC Gold Book: An Exemplar for IUPAC Asset Digitization by Stuart Chalk* and Leah McEwen 25
Chemical Health and Safety Data Management: Supporting Prudent Practices in Research Laboratories by Leah McEwen* 31
Primary Research Data and Scholarly Communication by David Martinsen* 35
Leveraging the Web: Production, Management, and Dissemination of Chemical Information by William Fyson, Simon Coles, and Jeremy Frey* 39
From Experiments to Knowledge: Reproducibility, Validation, and Reuse of Crystal Structure Data by Ian Bruno* 41
Big Data in Chemical Industry by Dr. B. Saha* 42
Supporting Information Review and Data Analysis at Organic Letters by Angela M. Hunter* 43
Update on ThermoML by Kenneth Kroenlein 44
Beware the Hype of Digital Publishing: Creating Useful Information Requires Work by Martin G. Hicks* and Carsten Kettner 45
The Status of the IUPAC InChI Chemical Structure Standard: Today and the Future by Ray Boucher*, Stephen Heller, and Alan McNaught 47

IUPAC Wire
Election of IUPAC Officers and Bureau Members 48
IUPAC Announces the Winners of the 2017 IUPAC-Solvay International Award for Young Chemists 49

Mark Your Calendar 50
Stamps International 52

* Meet the authors and learn about this topic at the special Symposium organized on Thursday, 13 July 2017
The IUPAC centenary in 2019 is fast approaching, and this will naturally lead people to look back at the significant achievements of the organisation and its dedicated volunteers over the past one hundred years. Equally important, however, will be the need to look forward to the roles for IUPAC in its second century. This special issue of Chemistry International (CI) could well feature in that assessment, as technology in the digital age, and particularly the data that technology produces, will clearly be an essential tool for the future of chemistry as a discipline.

The IUPAC vision, as espoused in our new strategic plan, is to be an indispensable resource for chemistry through the development of tools for the application and communication of chemical knowledge. In this issue of CI, you will find examples of the ways that data analysis can assist chemists and lead to the evolution of new chemical knowledge, and also of the ways that the effective utilization of data can assist in the communication of that knowledge.

Throughout its nearly-completed first century, IUPAC has been recognised particularly for its contributions to nomenclature, terminology, and the symbols of chemistry; for its standardisation of chemical methods; and for its critical evaluation of data and the development of standards for data exchange. The colour books and the curation of the periodic table, along with the atomic weight data within it, are particularly well-known, widely-used, and appreciated by students and researchers alike—even if they may sometimes appear to be a necessary evil. The periodic table and atomic weight data will always be essential to the discipline, and some of the uses to which it has been put are both fascinating and educational. [1] By contrast, many people have commented on the reduced importance of conventional nomenclature (and by implication, the colour books), as the quality of structure drawings and the ease with which such drawings can be incorporated into documents, websites, and other media has improved. Indeed, the rise of the graphical representation of molecules in documents has created challenges for database manipulation and searchability, and it is within this context that the IUPAC International Chemical Identifier (InChI) was invented, implemented, and developed. [2] The InChI identifier is now globally embraced and is being used in a wide variety of applications. In fact, in this issue of CI you will find InChI mentioned numerous times under a variety of topics. This issue of CI will also address a wide-range of issues in data management and data usage across the entire discipline.

From a personal perspective, my involvement with IUPAC mirrors, at least in a small way, the evolution of IUPAC activity over recent times. I began as part of the team producing a new version of the "Red Book", Nomenclature of Inorganic Chemistry--IUPAC Recommendations 2005 [3] and then took on leadership roles in the Division of Chemical Nomenclature and Structure Representation (Division VIII). In those roles I was involved in the development of standards for graphical representation, [4,5] which collectively were guides to drawing chemical structure diagrams that are as unambiguous and informative as possible. I also began to learn more about InChI, particularly about its use in database management/merging. It gradually became clear to me that there was potential for significant application of InChI's beyond databases, and I have become involved in InChI development, at least in a small way, through projects on the development of InChI QR codes [6] and InChI for mixtures. [7]

Now, in my role as IUPAC Secretary General, one of my major responsibilities is to help identify and encourage the development of new IUPAC activities and projects, particularly those that have strategic importance: those that will shape future IUPAC activities and enhance IUPAC's relevance in its second century. One of the key steps in doing this is to collaborate with other organisations and groups that have similar interests. I have been very pleased to see the development of collaborations between the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS) and the Chemistry Interest Group of the Research Data Alliance (RDA) and those individuals and organisations who are involved with it.

This special issue of CI describes many of the recent activities that I believe will have future significance, given the likely importance of "Big Data," the potential of data mining, and the benefits that will derive from being able to properly search, access, and mine all of the research data that scientists around the globe are busily accumulating.

(continued on page 4)
The Rise of Primary Research Data

by Leah McEwen and David Martinsen

As the scale of global commerce and opportunities for multidisciplinary collaboration increase, there is greater pressure on basic research to supply a quick return on investment (ROI). The emergence and development of digital information technologies in the new millennium have inspired a new look at how research outputs are managed and disseminated. The driving question in the minds of many research funders is this—will lowering the barriers for access increase the value of research for the greater society? This is a particularly interesting question to consider for measurement data, the greater amount of which are scattered across millions of separate, fixed publications (not to mention those never published and lingering in file drawers and on hard drives). Can the advent of cloud technologies, exchange standards, and provenance tracking facilitate improved access, evaluation, and use of data for both research and commerce? Can new value and discovery be realized through the greater aggregation of measured scientific data as “Big Data”?

The past five years has seen practical conversations among stakeholders increasingly focused on the publication of primary research data associated with journal articles. Data publication advocates have lobbied for the availability of data, funding agencies have issued mandates requiring funded researchers to publish their data, and repositories have been created to support researchers in fulfilling these requirements. The arguments put forth are many: it is important that science be as transparent as possible so that the community can properly assess the integrity of the research being published; it is valuable for interested scientists to have access to machine-readable data to more deeply examine and interact with the data described in a journal article; it is important that editors and reviewers have access to all of the available material to better understand the validity of the conclusions being presented, or consider whether the data themselves exhibit evidence of manipulation in a fraudulent manner.

This interest in the publication of research data, among other scholarly communication challenges, has spawned a number of new organizations (for example, FORCE11, [1] the Research Data Alliance), [2] which augment long-standing organizations (such as CODATA [3] and ICSU [4]). In addition, repositories for depositing research datasets, such as Data Dryad, [5] figshare, [6] and Mendeley Data, [7] have appeared. In chemistry, these new services may, in some sense, augment traditional curated data collections, such as the former Beilstein and Gmelin Handbooks, the Cambridge Structural Database, [8] the Protein Data Bank, [9] the Powder Diffraction File, [10] the Spectral Database for Organic Compounds (SDBS), [11] Wiley and NIST’s Mass Spectral Databases, [12,13] BioRad’s Spectroscopy Databases, [14] and others.

As a result of the emerging expectations for researchers to publish data, scientific publishers and research libraries are beginning to offer support services to their communities in navigating this evolving landscape. Balancing both sides of the time-cost equation for data generators and consumers will be key to how well new practices are established.

Taking a look at how the movement to publish research data more accessibly intersects the practice of research data dissemination in chemistry is the impetus behind a Special Symposium on Research Data, Big Data, and Chemistry at the 46th IUPAC World Congress, and the basis for this special issue of Chemistry International. The perspectives represented here examine a range of issues from coordinating global initiatives to workflows for publication, review, and evaluation to education to applications in industry and
The Rise of Primary Research Data

society. Also considered are some IUPAC digital initiatives for supporting chemistry data publication, including the International Chemical Identifier (InChI) [15] and the online Gold Book Compendium of Chemical Terminology. [16]

We hope you enjoy the reading, and look forward to meeting you at the Congress in São Paulo, Brazil, 9-14 July and the Special Symposium on 13 July 2017. [17]

References
1. www.force11.org
2. www.rd-alliance.org
3. www.codata.org
4. www.icsu.org
6. https://figshare.com
7. https://data.mendeley.com
8. www.ccdc.cam.ac.uk
9. www.rcsb.org/pdb/home/home.do
10. www.icdd.com
11. http://sdfs.db.aist.go.jp

(continued from page 2)

References
1. www.isotopesmatter.com
2. www.inchi-trust.org
7. https://iupac.org/project/2015-025-4-800

Richard Hartshorn <richard.hartshorn@canterbury.ac.nz> is a member of the chemistry faculty at the University of Canterbury, in Christchurch, New Zealand. So far, his involvement with IUPAC has been largely based in nomenclature, and dates from the late 1990s, when he was persuaded to join the group preparing a revision of the Red Book (“Nomenclature of Inorganic Chemistry, IUPAC Recommendations 2005”, ISBN 0-85404-438-8). Since then he has been involved in numerous projects and has been a member of the Committee on Chemistry Education (since 2006). He was elected to positions of responsibility in the Division of Chemical Nomenclature and Structure Representation (Titular member 2003-07, Vice President 2008-09, President 2010-13) and the Bureau (2014-17), and took over as as IUPAC Secretary General in January 2016.
Connecting Chemistry with Global Challenges through Data Standards

by Ian Bruno and Jeremy G. Frey

The new millennium, now almost 20 years old, has been characterised by a recognition within the research community of the importance of the free flow of research data; not simply in the ability to access the data, but also in the understanding that this valuable resource needs to be reused and built upon. We believe there have been at least two main drivers for this. First, those who pay for the research want to know it is leading to useful outcomes with impact—the transparency and accountability agenda. Second is an appreciation that the major global concerns (food, health, climate, economy) are extraordinarily complex (‘wicked’) problems, [1] whose solution requires interdisciplinary teams able to exchange data, information, and knowledge across domains. Moreover, ensuring data are understandable by other researchers, a hard-enough proposition in its own right, is no longer sufficient. The scale of modern data-intensive research is now only possible using computational techniques that require data to also be understandable by machines. There is a broad consensus across expert groups and scientific organisations that mutually-agreed data standards are essential to achieving these aims. [2-4]

The required investment in standards is significant and it is important that it is spread across traditional silos. This effort needs to work for both academic and commercial interests: for many, a well-constructed but expensive commercial solution would simply be inaccessible and thus work against the degree of open data sharing needed to effectively address global concerns. A focus on the adoption of standards, which in many instances revolves around the definition of appropriate metadata, has created the need for whole new level of discussion in the global community. Further, the increasing opportunity for computer-based access to data (via the Internet and the Web) increases the need for computationally tractable definitions of metadata.

Global Data Standards

Historically, international and national standards bodies have attempted to meet the challenges of agreeing on data. Organisations such as the USA National Institute of Standards (NIST) [5] have created standard test samples, standardised measurement, and assessed the accuracy of much data of interest to chemists (for example thermodynamic and kinetic data). The Committee on Data for Science and Technology (CODATA), [6] an interdisciplinary committee of the International Council for Science (ICSU), [7] established in 1966 and probably most widely known for its task group on Fundamental Constants, is a prime example of international collaboration providing agreed data standards. CODATA, though, has a much wider remit and exists to promote global collaboration to improve the availability and usability of data for all areas of research, by promoting the necessary cultural and technical environment for sharing research data. In 2008, ICSU created a parallel activity to promote long-term stewardship of and access to quality-assured scientific data, the World Data Systems (WDS). [8] The activities of WDS extend to promoting the adoption of recognised standards to help ensure trust in scientific data services and the organisations providing them.

Several of the international unions have been involved in setting data standards, alongside their established roles in defining nomenclature. The

Image drawn by Natalia Talkowska (www.natalkadesign.com) at a workshop run by the EDISON Data Science project, http://edison-project.eu
International Union of Crystallography (IUCr) is a leader in this area with the successful evolution of the universally adopted Crystallographic Information Framework (CIF). [9] IUPAC recently redefined its committee on publications, the Committee on Publications and Cheminformatics Data Standards (CPCDS), [10] to better position the union to play a role in the exchange of data between chemists (academic and industrial, pure and applied) and the wider community. CPCDS itself has set up a specific sub-committee to look at the data standards currently available and identify priorities for their development. [11]

Whilst the ICSU-based organisations (e.g. CODATA, WDS, IUPAC, IUCr, and others) have led the way in building up from data to metadata within disciplines, there has also been a need to create new organisations with a much wider cross-disciplinary outlook that can bring disciplines together to identify, and address, common challenges. One such organisation in which Chemistry is set to play a major role is the Research Data Alliance (RDA). [12]

The Research Data Alliance
The role of the Research Data Alliance (RDA) is not only to look across scientific areas, something ICSU and CODATA certainly do, but also to look across the different professions involved in data, metadata, and standards. Launched in 2013 by the European Commission, NSF and NIST in the US, and the Australian Government's Department of Innovation, the aim of the RDA is to build the social and technical bridges needed to enable open sharing of data to help address grand challenges of society. It currently boasts over 4900 volunteer members, from 118 countries, including researchers as well as information and data professionals from across academia, government, and industry. [13]

Central to the RDA's activities are working groups that aim to deliver outputs such as policy, specifications, standards, or recommended practices that will help eliminate roadblocks to data sharing. RDA outputs produced thus far address fundamental building blocks, such as data terminology, data type registries, and metadata standards; offer recommendations relating to data publishing, data citation, and repository certification; and define frameworks for delivering training in and around data management and data science. [14] Complementing the more time-limited and focused RDA working groups are RDA interest groups that take a broader and longer-term view of a particular dimension of research data management. There are interest groups that focus on general challenges or concepts, such as provenance, reproducibility, vocabularies, or metadata. Others aim to represent the perspective of a particular discipline.

Some RDA groups are run jointly in conjunction with other organisations, including CODATA and WDS. One of these, the RDA/WDS Publishing Data Workflows Working Group, has looked in detail at workflows across a diverse set of current data publishing paradigms to identify common components and standard practices. The group identified reference models that could be adopted by those venturing into data publishing for the first time and highlighted important gaps and challenges worthy of further consideration. [15] One of these gaps is the issue of identifying and exposing links between articles and datasets. This has been the focus of another RDA/WDS working group that has defined a high-level interoperability framework for scholarly link exchange called Scholix, [16] as well as a proof of concept Data-Literature Interlinking Service. [17] These ideas continue to be developed within the RDA working group ecosystem. [18]

Whilst some RDA working groups are largely discipline-agnostic, others have a much tighter focus on issues pertinent to a particular domain. For example, the RDA Wheat Data Interoperability Working Group has identified guidelines, vocabularies, and ontologies for creating, managing, and sharing the various data types relevant to wheat research, covering areas such as genetics, genomics, and physiology. [19] Whilst the development of these resources was implemented under the RDA umbrella, ongoing maintenance will be done within a pre-existing framework outside of the RDA.

Chemistry and the Research Data Alliance
Since 2016 there has been an RDA interest group focused on chemistry data. The idea for this group emerged from discussions taking place in and around meetings of national chemistry societies, in particular within the ACS Division of Chemical Information (CINF). The possible role that such a group might play was the focus of a “Birds of a Feather” session at the 6th RDA Plenary meeting in France in 2015 and a discussion at Pacifichem later the same year: together these provided opportunities for researchers from around the globe, inside and outside of chemistry, to help define the role that a chemistry group within RDA might play.

The central aim of the RDA Chemistry Research Data Interest Group (CRDIG) [20] is to provide a bridge between the data needs and activities of the chemistry community and opportunities emerging from the
Connecting Chemistry with Global Challenges through Data Standards

RDA. As well as reflecting on how RDA outputs could be adopted to benefit chemistry research, it also aims to engage with ongoing RDA discussions and projects and contribute use cases inspired by the data needs of chemists.

The seeds of synergy between the RDA and the chemistry community were nurtured at a workshop jointly supported by the RDA and IUPAC, held at the Environmental Protection Agency in North Carolina in 2016. [21] Entitled Prioritizing Digital Data Challenges in Chemistry, the workshop focused on the two main “languages” of chemistry: chemical terminologies and chemical structures. Out of this workshop came several project proposals, some of which have subsequently progressed, each taking a slightly different path.

Discussion on chemistry terminologies were continued at a CRDIG session at the 8th RDA Plenary meeting held during International Data Week 2016 in Denver [22] and at a VoCamp event in Washington DC. [23] Both events provided an opportunity to seek input from terminology experts in the wider data science community. Additionally, an IUPAC project has been established to overhaul the current online manifestation of the IUPAC Gold Book; once realised, this could form a foundation for future chemistry terminology projects. [24]

Efforts to address challenges in structure representation more naturally sit within chemistry communities. There are workshops and symposia planned for 2017 to engage stakeholders with the necessary domain expertise. One such example is an EMBL-EBI Industry Programme Workshop which was held in Cambridge, UK in March and focused on IUPAC standards for information, including InChI. [25] This workshop continued conversations begun in North Carolina and sought input from industrial stakeholders to help shape priorities to enable industrial application as well as academic pursuit. The InChI Trust are organising a similar workshop that will take place at the NIH in August, just prior to the Fall 2017 ACS National Meeting in Washington, DC. [26]

Chemistry and Global Data Challenges

At the 2016 CODATA General Assembly, held alongside the International Data Week, the issues of the exchange of data was very much the major underlying theme. Importantly this has support not only from the research community (bottom up), and those funders looking for impact, but also from those charged with dealing with global human conditions (humanitarian support in crisis). The need to be able to get access to the right data quickly, efficiently, and in a way that can be combined with other information has been highlighted in the many medical emergencies of recent times.

A task group set up by a CODATA commission on data standards for science is the right venue for exactly the work needed to facilitate data exchanges within and between communities. This group is planning to undertake a mapping of activity related to standards by scientific unions and other organisations, with the aim of raising awareness of standards in development or use. It intends to create resources that will support adoption of these standards and determine models that provide a guide to their maturity and fitness for use. It anticipates providing good practice guides for the development, application, maintenance, and governance of standards and will link to the activities of other key groups, such as the RDA.

During 2017, the CODATA commission will invite the organisations involved, including IUPAC, to meet to progress the ideals of data standards in a digital world. It is hoped that with ICSU support, funding to enhance these activities and provide a global repository of standards will become a reality. As part of this activity, CODATA is arranging a workshop this summer to promote the activities of the task group on exchanging data.

ICSU, in its UN role, has highlighted that access to data underpinning disciplines such as Chemistry, Physics, and Biology is one of the major factors that will make a difference in the global response to emergencies and emerging disasters. This should remind the Chemistry community, if a reminder is necessary, of the importance of access to chemical data. As a community, we need to make a deliberate effort to channel our many years of experience with data into global activities taking place today to ensure that the results of our research are available and useful across domains and not just hidden or kept to ourselves. This is the challenge of the next decade!
Connecting Chemistry with Global Challenges through Data Standards

References


Ian Bruno <bruno@ccdc.cam.ac.uk> is Director, Strategic Partnerships at the Cambridge Crystallographic Data Centre, Cambridge, UK. ORCID.org/0000-0003-4901-9936
Jeremy G. Frey <j.g.frey@soton.ac.uk> is Professor of Physical Chemistry at the University of Southampton, Southampton, SO17 1BJ, UK. He is a member of the IUPAC CPCDS committee and Commission on Physicochemical Symbols, Terminology, and Units (commission I.1 responsible for the IUPAC Green Book). ORCID.org/0000-0003-0842-4302
Search and retrieval of chemical information has been dramatically changed by the application of “Big Data” techniques. This development continues to be driven by the massive growth of chemical scientific literature and of online data and databases. Not only is there an expansion of the traditional avenues of publication, but many new contributing resources, such as open access journals, MOOCs (Massive Open Online Courses), Wikis, and blogs have arisen. Powerful tools, like APIs (application programming interfaces) and Big Data interrogation are providing innovative ways to retrieve and analyze data and connect different databases. Materials, pharmaceutical, and environmental research, to name just a few, are especially challenged by the need to organize and access vast amounts of data. What skill-sets will need to be developed in order to get the greatest value out of the available data? Will it be coding and information technology skills, or awareness and better delivery of the data by the available systems? We believe that, in the short term, efforts are needed to expand awareness and training.

Exponential Growth of the Amount of Chemical Information

One of the earliest efforts to quantitatively measure the growth of the scientific literature was made by Derek J. de Solla Price over five decades ago. He determined that the number of scientific journals was increasing by about 5.6% per year, with a doubling time of 13 years, and that the number of abstracts in Chemical Abstracts was also growing exponentially. More recently, Larsen and von Ins have reported a similar rapid growth for scientific articles, with a slightly longer doubling time than Price, and they report that the Science Citation Index is covering a decreasing portion of the traditional scientific literature. These authors also point out that publishing has expanded into new channels, such as open access archives and web pages, especially in the form of blog posts and an increasing number of alternative distribution channels available to any scientist (e.g. LinkedIn, GoogleDocs, Slideshare). In late 2014, there were about 28,100 active scholarly peer-reviewed English-language journals (plus a further 6450 non-English-language journals), collectively publishing about 2.5 million articles a year. The number of publishers and journals continues to increase and is being boosted by the Open Access publishing movement. Faizul and Hilal surveyed the number of chemistry journals listed in the Directory of Open Access Journals (DOAJ) and found 164 journals as of 2014.

In parallel to the ongoing growth in the number of journals and articles and the availability of a number of for-fee databases, there has also been an enormous growth in the number of free online chemistry databases. The premier for-fee chemistry resources are CAS’ SciFinder, with over 127 million chemical substances in its registry at the time of writing, and Elsevier’s Reaxys database, which contains over 100 million organic, inorganic, and organometallic compounds. Both SciFinder and Reaxys index tens of millions of chemical reactions and hundreds of millions of experimental facts. While there are a number of other commercial databases that can be acknowledged, it is the explosion in web-based data that continues to feed the Big Data revolution in access to chemistry. Wikipedia alone lists over forty chemical databases. The majority of these are freely available resources on the Internet. They represent an increasing thrust in terms of public data dissemination and release as encouraged by funding agencies.

Despite the name, Big Data in chemistry is not just the amount of chemical information. Often the size of the dataset is compounded by the complexity of the information. For example, chemical information is reported in a bewildering variety of structured and unstructured formats including closed, instrument-specific format data. Currently the majority of these data are not being made available in standardized open formats. For example, NMR spectra generated in research labs, if published, are normally reduced to an image, despite there being digital exchange standards for such data. In our experience, the vast majority of scientists have never heard of the standard, do not know how to generate this form of the data and, in any case, would not know how to share it. Some publishers are starting to consider this need and efforts are underway to put a spectral database online. However, this situation will not change until there is a general increase in awareness.
of the problem and potential advantages that will result if the community collaborates in the delivery of data in a form that can be aggregated for better access. One approach to address this issue is an educational effort that includes professional chemists, as well as undergraduate majors.

Even though the data from one individual’s sequenced DNA is only about 750 MB, analyzing the genome for a single person in order to find the best treatment for a disease represents a massive computing problem. It is estimated that just storing the genomic data for the entire U.S. population would represent 222 petabytes. [12] Approaches are being developed by the bioinformatics community to consume and digest omics data (i.e. genomics and metabolomics) to support personalized medicine. [13] Despite the increasing complexity of chemical information, it is likely only through analytical data dissemination that the masses of data will move towards the scale offered by the biomedical community. There is some movement towards sharing big data to support discovery, certainly in terms of file sizes, in the analytical sciences.

Any analytical chemistry laboratory running spectral analyses currently produces multi-megabyte file sizes. Multidimensional NMR spectroscopy data files can measure from 10s to 100s of megabytes, but tandem mass spectrometry files, especially in proteomics, can consume terabytes of space. The online database, the Center for Computational Mass Spectrometry at UC San Diego, already has datasets over 1 terabyte in size. The largest of these is almost 13 terabytes in size. [14] Despite the sheer size of these datasets, and their applications in proteomics, can such data be used for discovery purposes? A recent report regarding how such approaches can be used for the identification of new antibiotics suggests that there really are needles in the haystack that can be extracted from mass spectrometry data. [15,16]

The Global Natural Product Social (GNPS) molecular network, launched in 2015, [17] is also pursuing a Big Data approach to the discovery of novel drug candidates from enormous quantities of mass spectral data assembled from the work of over a hundred laboratories. As has been the case for the analytical sciences for many years, data generation is rarely the main problem. The analysis algorithms associated with data interpretation remain the bottleneck, accompanied by a lack of data interchange standards to migrate data into alternative data processing platforms. Analysis, however, can be assisted by the availability of online data resources, providing access to tens of millions of chemical structures and tens of thousands of analytical spectra. These resources represent tens of millions of dollars of investment in informatics architecture, but also in the collection, curation, and annotation of the data.

Similar to the work that GNPS is doing, scientists can today access online “big data collections” that can be used in the identification of chemical compounds. With nothing more than a monoisotopic mass or molecular formula derived by mass spectrometry, and access to an online database, structure candidates can be identified using a search for “known unknowns”, searching for well-known chemicals that are held in public databases, though the scientists themselves do not know the candidate structure. This approach has been demonstrated previously. [18,19] Chemists can combine this information with an online search of NMR data for tens of thousands of chemicals [20] and even use Robien’s Spectral Robot Referee [21] to help confirm structural hypotheses. Mobile applications on phones and tablets, hosting over 700 000 chemicals (with masses and predicted

<table>
<thead>
<tr>
<th>Unit</th>
<th>Symbol</th>
<th>Size in bytes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>kilobyte</td>
<td>KB</td>
<td>1 000</td>
</tr>
<tr>
<td>megabyte</td>
<td>MB</td>
<td>1 000 000</td>
</tr>
<tr>
<td>gigabyte</td>
<td>GB</td>
<td>1 000 000 000</td>
</tr>
<tr>
<td>terabyte</td>
<td>TB</td>
<td>1 000 000 000 000</td>
</tr>
<tr>
<td>petabyte</td>
<td>PB</td>
<td>1 000 000 000 000 000</td>
</tr>
<tr>
<td>exabyte</td>
<td>EB</td>
<td>1 000 000 000 000 000 000</td>
</tr>
</tbody>
</table>

* a byte is a group of binary digits or bits (usually eight)
Big Data Tools for Chemical Information

Most chemists have been using Big Data tools, even though they might not have been aware of it, since search engines, like Google or Bing, use a combination of MapReduce and Hadoop to distribute a search among multiple servers and then analyze the huge amounts of information that result. [23] Hadoop facilitates the distributed processing of massive unstructured data sets across large computer clusters, while MapReduce distributes work to various nodes within the cluster (or ‘Map’), organizes it, and then ‘Reduces’ the results into a coherent answer for a particular query. [24] Big Data tools are also being used for data analytics in many areas of chemistry. Large arrays of inexpensive sensors connected through a computer cloud may generate very large datasets in environmental chemistry. [25] In addition to size, environmental data may be complicated because it consists of records in differing formats. Data collected in recent years may be distributed across a large number of databases, each with a different data model, potentially using different data ontologies (if any), with the potential for enormous migration and data integration efforts to mesh together valuable data. Historical data, while valuable, may not even be available in digital form, other than as scanned documents where optical character recognition software will produce only some limited level of retrieval, so these are only available for consumption into Big Data analysis tools using the limited metadata associated with the documents.

Researchers are being overwhelmed by so much data that experimental results can be overlooked or repeated unnecessarily. This is increasingly important at a time where reproducibility overall is being called into question. [26] Mullin reports, according to one estimate, "... 40% of all R&D experiments are repeat runs necessitated by inefficient experimental design or inadequate IT." [27] Another arena where Big Data tools are useful is predicting protein structures. Ovchinnikov and co-workers have used Big Data techniques to better predict 3D protein structures. [28] Big Data tools are also valuable in chemical toxicology, where the use of high-throughput screening produces both structured and unstructured information that is so large and complex that it is difficult to analyze using traditional methods. [29] Just managing the chemistry data (in terms of chemical compounds and challenges of chemical structure detail vs. the myriad associated identifiers) is enough of a problem. [30] Working to blend the chemistry data with the associated bioassay screening data into a form consumable by scientists as openly accessible data, and for consumption by Big Data algorithmic approaches, is a significant and often underestimated difficulty.

Big Data, artificial intelligence, and machine learning are today commonly aggregated into the same conversation. This is certainly true when it comes to the promise of these approaches combined through the IBM Watson project. [31] In 2013, MD Anderson partnered with IBM to pursue a cure for cancer, starting with leukemia, [32] and medical centers are already implementing Watson to help oncologists make data-driven decisions. [33] While promise remains, a little of the luster has rubbed off recently with the announcement that the Anderson-IBM collaboration has been halted after a scathing report from auditors at the University of Texas says the project cost MD Anderson more than $62 million and yet did not meet its goals. [34] This example indicates how these approaches might not solve all of our current challenges, but it does not mean we should not attempt these efforts as, after all, this is research.

Big Data Tools to Search the Web for Chemical Information

Text-mining of chemistry data in patents and documents has been underway for many years [35] and IBM text-analytics, and Watson specifically, have been applied to problems in the life sciences. By analyzing millions of pages of text in the medical literature, patents, genomics, and chemical and pharmacological data, Watson made novel connections. Early results suggest that Watson can indeed accelerate the identification of novel drug candidates and targets by harnessing Big Data. [36] Similar technology could be applied to the mass extraction of chemical reactions from literature articles and patents, as demonstrated by Lowe, [37] and then used as the basis of retrosynthetic reaction synthesis algorithms such as those underpinning Wiley’s ChemPlanner. [38]

While there are certainly some naysayers regarding the potential of these supercomputers, i.e., “neither could compete with a toddler at some of the most basic forms of human cognition,” [39] this level of negativity has been pointed at various technologies at some point in their development, whether it be the potential contributions of solar power, self-driving cars, or even putting a man on the moon. All of these, clearly, have proven to be possible.

While the breakthrough technologies are on the bleeding edge and newsworthy, there are many capabilities already available for every Internet consumer. More and more information is available in online databases, mined from the literature in ever increasing large scale
data and then made available as downloadable datasets (e.g., ca. 300 000 melting points extracted from patent literature, [40] connected by appropriate Application Program Interfaces (APIs) and increasingly available via the semantic web). Wikipedia both delivers data for consumption and, increasingly, is being served by the developing Wikidata project. [41] It is getting easier to access and integrate data with components, add-ins, and widgets.

For example, the PubChem project provides access to chemistry data for about 94 million chemical substances and about 1.2 million bioassay measurements. [42] The data are not only available by browsing the PubChem website data, but their widgets can be directly integrated with other websites so that they are accessible to different audiences interested in the data in different contexts. [43] For example, environmental chemists surfing the U.S. Environmental Protection Agency (EPA) CompTox Chemistry Dashboard have direct access to the PubChem bioassay data via the use of PubChem widgets. [44]

Increasingly, chemical information is stored on the Internet in the form of videos. The Internet hosts many chemistry videos (for example the Periodic Table of Videos [45] in addition to the Journal of Video Experiments), [46] and chemistry videos are also be found in many chemistry-related Massive Open Online Courses (MOOCs). [47] If the use of virtual and augmented reality environments develops as expected, even more valuable data will be made available. At present, the ability to search these resources is limited by the associated metadata. The challenge becomes how to search and get the greatest benefit out of these complex datasets and environments.

How should Chemists be Trained to Use Big Data?

Does the increased importance of Big Data mean that every chemist needs to learn how to use Big Data tools? In-depth knowledge of how such tools and algorithms work is unlikely to be useful to all chemists, since it is more probable that most chemists will simply use the results of Big Data algorithms and searches. It is, and will be, more important for chemists to recognize the strengths and weaknesses of a Big Data approach, rather than to be able perform a direct analysis themselves. A complex computer algorithm, such as that used to analyze Big Data, is for many simply a black box, and for most it is human nature to assume that the algorithm is performing its function in an ideal manner rather than questioning the results. Consider the parallels with how many individuals interact with a general online Internet search, accepting partial results rather than insisting upon completeness.

However, we think training is required at two different levels. First, there is a need for some chemists to have in-depth training in data analysis. These individuals, who would have a combined background in Computer Science and in Chemistry, will have skills needed to ensure that chemical-specific information is used appropriately in Big Data analyses. Second, there is a need for practicing chemists to have a background in Big Data analytics sufficient to recognize the potential uses of these techniques, as well as some of the potential pitfalls. Some training is already available: a Google search will turn up a number of courses that may be appropriate. For example, there is an Online Learning Cheminformatics Course sponsored by the Committee on Computers in Chemical Education of the ACS Division of Chemical Education. [48] David Wild, director of the Indiana University Cheminformatics Program, maintains a Cheminformatics Education Portal (ICEP), a repository of freely accessible cheminformatics educational materials. [49] as well as an online course, Introducing Cheminformatics: Navigating the World of Chemical Data. [50]

What does the Future Hold for Searching Chemical Information?

The Internet has catalyzed a shift in expectations for many stakeholders in terms of chemical information. The primary consumers of chemical information today are those driving the search at their desktop, on their tablet, or on their phone using one of the common browsers. The majority of chemists likely think they know enough in terms of Internet searching that they can find what they are looking for themselves and do not need training. They are satisfied with a simple search box. While such results can be found with the majority of Internet searches, it is a fallacy to consider that such approaches are not without issues. Chemical information professionals, and librarians specifically, commonly bring significant experience to the array of commercial platforms available and are able to answer questions, supply training in “search strategies,” and teach users how to best utilize Internet resources. In academia especially, basic training in cheminformatics and chemical information resources is increasingly encouraged. It is hoped that this will expand both in depth and in general availability across institutions. Certainly there are no signs that the rate of publishing is decreasing, that the amount of data coming online is slowing, or that the expectations for high productivity and more innovation are waning. To the contrary, it is likely that chemists will contribute directly to the growth in Big Data, using both personal publishing platforms (e.g., blogs), community collaboration tools for
information dissemination (e.g. wikis), and data sharing platforms. In fact, many funding agencies are now demanding the release of data associated with scientific research: the relevant skills to do so need to be developed and the supporting tools to make it happen need to be continually improved.

Conclusion
It has been two decades since Carla Hesse predicted that the future of the book might consist of “paths of inquiry, modes of integration, and moments of encounter.” [51] That may yet serve as a good description of the future of chemical information, since data, information, and knowledge is hardly static—it is changing moment by moment and accessible via a web search. Chemical researchers may integrate the results of multiple search modes using a variety of paths to the data (i.e., literature, online data, analytical tools). The resulting vast amount of information may allow for less than the optimal time for careful examination of all but the most obviously essential resources. Despite the vast amounts of information to be surveyed, both online and print, it will still be necessary to insure that as little as possible of the most important sources are not lost. This will require new levels of sophistication and ingenuity from the researchers of the future.

References
6. CAS REGISTRY. www.cas.org/content/chemical-substances (accessed 2 March 2017).
25. Vitolo, C., et. al., Web Technologies for...


Antony J. Williams <tony27587@gmail.com> is a cheminformatics at the National Center for Computational Toxicology, U.S. Environmental Protection Agency in Research Triangle Park, USA. ORCID.org/0000-0002-2668-4821

Harry E. Pence <Harry.Pence@oneonta.edu> is SUNY Distinguished Teaching Professor Emeritus in the Department of Chemistry and Biochemistry, SUNY at Oneonta, USA. ORCID.org/0000-0002-0412-9018
Are We Nearly There Yet?

A Perspective on Data Sharing in (Chemical) Crystallography

by Simon Coles

When contemplating the subject of sharing knowledge in academia, I am particularly drawn to the following statements in a monologue on the “The Social Function of Science”: The present method of publication...has the following deficiencies:

- Research results cannot be reported promptly.
- Research results cannot be reported completely, with all necessary data, illustrations, discussion, background, and other pertinent details.
- Waste is widespread, in that only a small number of subscribers are interested in any particular research report as now published.
- Increasing volume of the scientific literature is increasing the financial burden upon individual scientific workers and upon institutions.
- The multiplication of journals...Failure of libraries to subscribe to journals due to measures of economy will result in serious hindrances to the availability of requisite scientific literature.
- Editing and business management of journals is usually undertaken on an unpaid volunteer basis by scientists.
- A scientific worker can only read a small fraction of the papers.

I am not stating anything here that most readers don’t already know—these issues have been raised by various arms of academia for at least the last decade and are the subject of much concern and debate. What is particularly striking, however, is that “The Social Function of Science” [1] was written by J.D. Bernal in 1939!

Visionary Beginnings...

Bernal was a crystallographer and an academic visionary, and he set out to address these (and many more interdependent) matters. Accordingly, the crystallographic community has gone on from Bernal’s basis to lead the way in terms of the organisation of science. However, in some respects we actually find ourselves back in the same position despite moving forwards radically in the last 75 years. This article will reflect on how the crystallographic community, and in particular the chemical arm of the discipline, has structured its communication of results, leading up to the issues of the current times and how they are being addressed.

Back in 1939, Bernal also made two other remarks that are of particular significance to the topic of this article. Firstly, that “Science has moved from direct observation of nature and is more and more dependent on the previous observations of other workers and on their methods of interpretation.” It became apparent over the following decades that crystallography is fundamentally a subject that is about the data, i.e., a crystal structure, and that discussions, interpretations, and scientific insights are all secondary derived material based on that data. Secondly, he discussed at length the need for the reorganisation of science: that research should be carried out for the benefit of society as a whole. This requires the effective coordination of individuals: “The general problem of the organisation of scientific research can be simplified by dividing it into two problems, those of the inner and outer organisation. The first is an internal problem of how a laboratory should be run and the second of how the work of the different laboratories should be co-ordinated into a coherent structure of scientific research.”

Many of Bernal’s 1939 observations were based on the distribution problem for scientific publications and the lack of coordination of research, however it was during the next 30 years in collaboration with his protégé, Olga Kennard, that a route to addressing these problems was realised. In 1965, Kennard founded the Cambridge Structural Database (CSD). [2] What this did was:

- make the data the primary unit of scientific endeavour
- instigate a centralised model, whereby all data were contributed to, and held in, a single place
- provide a single resource for data sharing
- ensure that all data were held in the same form

These achievements, in a single approach, clearly address the majority of concerns raised many years earlier. However, their thinking had subsequently moved beyond these concerns, and in fact the CSD was also conceived in order to generate a collection of data that could facilitate new science in itself. This goal is captured in a comment by Kennard, “We had a passionate belief that the collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments.” [3]
In 1965, there were only around 1500 published structures, which were collated manually into printed volumes, named ‘Molecular Structures and Dimensions’. However, this number began to rise quite rapidly and it soon became clear that a database approach would be required in the longer term. The CSD evolved as one of the world’s first scientific databases, essentially containing bibliographic information and associated crystallographic numerical data.

The success of the CSD grew over time, but the largest changes occurred through the implementation of tools to search the database, which in turn engaged an audience that went far beyond the core crystallographic community. Software that was developed to enable the processing of structural data for incorporation into the CSD quickly evolved into applications for searching the data. The real key to success was developing the software beyond providing search functionality for crystallographers (essentially recall based on the data fields entered). Chemical searching was enabled by indexing the data in new ways. Ultimately, this enabled 2D structure searching and analysis of 3D structural data, both molecular conformation and, more recently, interactions between molecules. This functionality clearly went beyond simply questioning whether a structure had been done before to pioneer new areas of structure-based chemistry research. It drew in the chemistry community, not only by providing usable, intuitive tools that could help them with their research, but also by providing new insights into chemical phenomena. It is worth noting here that other, related crystallography databases have followed suit—the Protein Data Bank (PDB) and the Inorganic Crystal Structure Database (ICSD) are now equally successful in serving their respective communities, albeit with slightly different organisational structures.

Realising the Need for a Standard
By the late 1980’s, it was clear that this data-based discipline was moving fast; the sheer scale at which the science was operating clearly required the community to be well organised in order to manage the situation. At this time the CSD had reached somewhere in the region of 100000 records, each of which had to be re-encoded from journal articles and their associated hard copy supplementary information. With this rate rapidly increasing, the manual approach was clearly unsustainable—an obvious fact when we look back from today’s ‘digital era’ at how the CSD has grown. Figure 2 on the following page shows that the 1980’s are the lead up to the inflection point.
With computing becoming more ubiquitous in the laboratory, as well as being available for all to use on demand and in a timely manner, it was not long before the community turned to finding a way in which crystal structures could be more ‘computable’, i.e., processed by a computer. Accordingly, the Crystallographic Information File (CIF) [4] was adopted as a standard in 1991 by the International Union of Crystallography (IUCr). Although CIF pre-dates the prevalent use of the worldwide web, its ethos and design maps onto it very well and could easily be considered as ‘semantic web compliant’ today. Critically, it continues to be compatible with evolving standards.

CIF is a file structure developed for the archiving and interchange of crystallographic data and is now ubiquitously used in all aspects of the discipline, whilst enabling integration with other disciplines. CIF can capture the results of a diffraction experiment and take them all the way through to the publication of results, alongside journal articles and/or in data repositories. In fact, the IUCr use CIF as the basis for the submission of crystal structures to their journals and it can therefore be considered as a publishing format in its own right. The dictionary structure of CIF empowers much of this multidisciplinary work, as there is an overarching ‘core’ into which different dictionaries can be added, such that the same framework can be adopted for macromolecular crystallography as for chemical crystallography or for powder diffraction, for example. Its adoption success is due to its strengths as a general, flexible, and easily extensible free-format file that is both human and machine readable and can be edited by a simple text editor.

However, its success is not just down to providing a great technical solution; community acceptance and coherence is crucial. The IUCr proposed the format in 1991: it was well on the way to being universally adopted within five years and rapidly thereafter became ubiquitous in all aspects of the field. This is also due to exceptional and rigorous governance—the Committee for the Maintenance of the CIF Standard (COM-CIFS) [5] was established in 1993 under the auspices of the IUCr and comprises a range of working groups overseen by a management group.

CIF, therefore, has evolved in many ways and provides the underpinning for the whole discipline, transcending being merely a file format. CIF is now known as the Crystallographic Information Framework, and the breadth of support that the CIF provides for the discipline is illustrated in Figure 3. It has been adopted by the CSD (and other crystallographic databases) as the vehicle for rapid and automated processing of crystal structures, and has fuelled their accelerating growth. There are now also many graphics packages that can import and render a CIF, making it a format that supports a whole range of visualisation techniques.
Perhaps the greatest example of its power is the fact that software can automatically check and validate it, not just for syntactic correctness, but also against a set of rules using algorithms. This is embodied in the checkCIF service, [6] which checks for completeness, quality, and consistency in chemical structures and reports back to the user. This unique capability enables all practitioners, irrespective of their level of expertise, to rapidly assess a result and is a key factor in maintaining the rate of data growth in the field, not just according to sheer numbers, but also to quality.

The Rate of Change—New Models and New Science

A unified file format and a centralised model don’t in themselves mean that a database will grow on its own—there are other stakeholders involved and incentives necessary. In 2015, 90 % of structures released into the public domain as part of an academic publication were deposited into the CSD prior to acceptance by the journal. This is a remarkable statistic, due mainly to coordination in the community and engagement with formal publication routes, and this new model of ‘push’ rather than ‘pull’ has certainly eased the path for data into the database.

This is the generally accepted route for the publication of crystallographic data and thus is coupled with, and often governed by, the underlying science and the peer review process. In some senses this can be considered a curse as much as a blessing, for a crystal structure will reach the database only if it is associated with a formal publication. Straw polls indicate that only around 20 % of all crystal structure analyses performed are “published”, largely a result of being tied to this process. Moreover, while electronic publication is now the route of choice for the dissemination and discovery of scientific works, this remains simply a mechanism for the process, and the structure and content of an electronic article is largely the same as that of a paper version. The data contained within articles, which is often as important to other scientists as the commentary, remains a second-class citizen in this model. The current rate at which data may be generated and captured, therefore, far outweighs the rate of dissemination.

In the last 15 years the data repository concept has been much explored and, as ever, the crystallographic community has been at the forefront. Institutional, or research group, projects have been undertaken [8] and the Crystallography Open Database [9] has gained significant traction. These approaches enable orphaned data to be readily deposited and disseminated without the need for a parent publication. The CSD has had a similar, ‘Private Communication’ submission route in place for around 25 years, which has recently been rebranded as ‘CSD Communications’ and has grown in
popularity. It is very interesting to consider that in 2016 this route had a greater number of submissions than any one journal. These innovations raise interesting opportunities for a distributed, somehow federated, network of repositories that could open up the flood gates from the laboratory into the public domain. It is important to note, however, that this doesn’t mean the ‘centralised CSD’ model should be considered outmoded. Rather, it takes on the role of the authoritative resource, which a diverse set of systems feed into. These new models therefore also require the consideration of a range of factors that have hitherto not been of primary concern, such as socio-cultural issues, sustainability, advocacy, organisational factors (both academic/research/labs and library/IT support), the context of the digital scholar, the digital research cycle and data curation, different research practices and workflows, technical interoperability, and open standards.

Moving to these data centric approaches is not just about efficiency and maximising the size of databases—most importantly, this move has enabled new science. Data mining in crystallography in order to provide new chemical insight stems from the work of pioneers, such as Dunitz in the 70’s and 80’s. However, the first time the power of systematically analysing such data collections struck me was in the heroic work that resulted in reams of tables of bond lengths for particular chemical environments in organic and organometallic compounds, [10,11] which became not only the go-to reference for every crystallographer, but also the basis for a wealth of other work. It is the inspiration for a knowledge base of molecular geometry, [12] essentially a database of these values, but one which has been extended to include angles and torsions and can therefore be used to assess both geometry and conformation in an automated manner. Such data is also the basis for many approaches in molecular mechanics, where knowledge of all the observed geometries for a particular environment is encoded into a force field that is used for predictive purposes—most theoretical studies these days begin with a quick geometry optimisation of the proposed molecule/system that is based on such data. This approach has been extended to intermolecular interactions, but was in fact initiated before the molecular geometry knowledgebase. The wealth of information in a crystal structure goes beyond the boundaries of an individual molecule! This in turn has resulted in a knowledge base of molecular interactions [13] that encodes hydrogen bonding and other intermolecular interaction information, which opens up a whole new area. One can not only query, then make use of, the propensity of a particular interaction, as in crystal engineering, but also evaluate the range of possible interactions between two entities (e.g. see Figure 4). This function has extensive applications in fields such as crystal/materials engineering.

Figure 4. The use of a molecular interactions knowledge base in the design of pharmaceutically acceptable co-crystals.
and drug docking. The complementarity of chemical and macromolecular crystallography in using these approaches in the field of drug discovery is particularly well covered in a recent article by Colin Groom, the current Director of the Cambridge Crystallographic Data Centre. [14]

What Now?
It is clear that every crystal structure is important—it contributes to a collection which can then be exploited in a range of new scientific areas. As alluded to above, the challenges that persist are the requirement for quality assurance, custodianship, and organisation of the data and also the need to enable routes that maximise the volume of data. When considering all the possible, hypothetical ‘chemical space’ that has or could be explored and comparing that with coverage in the CSD, there are large, significant differences—many classes of compounds don’t have any crystal structures present. There are several significant likely reasons for this. While some of these are perhaps insurmountable, others are not. Examples of the latter include: for one reason or another structures were never published in the scientific literature; crystals were never made, as they were not deemed necessary; compounds were too difficult to crystallise and no significant efforts were made to overcome this; a compound simply won’t crystallise; only a small number of structures were deemed necessary to characterise a whole class of compounds; or synthetic chemists have not tried to make a particular class of compounds (and may never do so if there are not research ‘drivers’ to do so!). A significant culture change is still required to ensure that the considerable secondary benefits of databases continue to evolve and grow in scope.

However, we now live in a data-driven world. The application of algorithms and data to address chemical problems, Cheminformatics, is very much on the rise. Crystallographic data is now being used in ever different ways, which weren’t envisaged, either in Bernal’s time or more recently! A single example that illustrates this point is the notion of crystallisability, which is of considerable importance in manufacturing many products in the chemical industries. Efforts are being made by mining databases of reported syntheses and comparing them to those of crystal structures in order to understand whether a compound is likely to crystallise. There is also a need to be able to optimise crystallisation based on this work. There are obvious big pitfalls in these studies and, at the very least, the utility of these approaches would be much improved if more laboratory data were available. For this kind of derivative research to be really useful, we would need to know the outcome of every crystallisation trial, not just the ‘1 in a 100’ that work.

This leads into my final point—it is primarily the results of science that get published. Often, very little about the research journey that leads to these results is made known. Yet surely science can move forward at a much greater rate if we can run computer algorithms on bodies of data that include information on what a researcher intended to do (and why) and what they actually did, tied to what resulted—preferably for everything that they do! This is still lacking in crystallography—the databases are full of results, not raw data. ‘Data’ can generally be considered to be raw data, processed data, and derived data. In the crystallographic context, these are diffraction images, structure factors, and crystal structures, respectively. Recently, some progress has been made, in that software will include derived data (structure factors) in the CIF result: validation processes and the CSD will make use of and curate this data. But we can go further still—not only would raw data improve validation processes and provide valuable training sets for software developers to improve algorithms, etc., but there is a more interesting issue—a diffraction experiment records the average signal from the whole sample, which includes defects, impurities, etc., yet often only the data that gets a perfect result is extracted. For materials engineering, it can be crucial to be able to understand these additional effects, yet it is never made public that they have been observed!

Raw data availability, therefore, can be very important, but there are often counter arguments related to costs. The diffraction experiment is relatively quick and cheap, so why not just do it again? The real cost of doing a structure again was assessed by the UK National Crystallography Service as part of the ‘Keeping Research Data Safe Project’. [15] There are many nuances to such a cost calculation, but if one has to factor in that the research expertise/group/laboratory that originally generated the material may no longer exist or still be set up (people, apparatus, etc.) to make such materials, then the cost rapidly escalates. The replacement cost of the CSD is therefore almost immeasurably large!

Data transfer and storage are problems that are now being overcome. For around 15 years there has been an ‘extension’ to CIF (imgCIF) that can cater for raw data, although its uptake has been very slow. So why aren’t we amassing more of our valuable raw data for the community to widely exploit? For the last five years, a group known as the Diffraction Data Deposition Working Group has been looking into the issues surrounding this topic. The outcome from the activity of this
A Perspective on Data Sharing in (Chemical) Crystallography

group is that the International Union of Crystallography has recently convened a Committee on Data, ‘CommDat’, [16] as an advisory committee to the Executive. CommDat will work across the organisation, subsuming data-related interests of Journals, the now discontinued Committee on Crystallographic Databases, and of the Committee on Electronic, Publishing, Dissemination and Storage of Information, and will also have a formal relationship with COMCIFS.

To conclude, the answer to the question in the title, “Are we nearly there yet?”, is that I doubt we ever will be! The scientific endeavour will continue to stretch us, and data is the foundation of our research. However, the lessons we have learned along the route so far are applicable to many disciplines outside of crystallography and provide us with a very strong basis for the future. As a coherent community that considers data matters very seriously indeed, we continue to look forward and encourage others to follow.

References
5. www.iucr.org/resources/cif/comcifs
8. http://ecrystals.chem.soton.ac.uk
9. www.crystallography.net/cod/
16. IUCr Committee on Data (CommDat) www.iucr.org/iucr/governance/advisory-committees/committee-on-data.

Simon Coles <S.J.Coles@soton.ac.uk> is Professor of Structural Chemistry at the University of Southampton and Director of the UK National Crystallography Service. ORCID.org/0000-0001-8414-9272
Managing Standards and Critical Evaluation in a World of Big Data

by D. Brynn Hibbert, David Shaw and M. Clara F. Magalhães

IUPAC is very interested in data, big or small. Its website opens with the statement, “The International Union of Pure and Applied Chemistry (IUPAC) is the world authority on ... many other critically-evaluated data.” While the ‘...’ covers compelling and widely popular topics, such as naming new elements, the mission of IUPAC to give its imprimatur for chemical data is of great importance to health, security, and trade in the world. In this article, after a review of present activities, we will contemplate how a comprehensive approach might be structured under IUPAC project rules and then look to the future in a world of ‘big data’ and ‘smart instruments’.

History
IUPAC does not perform or sponsor experimental work that produces chemical data. Rather, it encourages the formation of international teams of qualified experts to compile and critically evaluate data gathered by others. Teams have formed within various Divisions to critically evaluate data ranging from fundamental chemical data (e.g., atomic weights) to a wide range of chemical data used by non-chemists as well as chemists in efforts to satisfy human needs and help resolve industrial, environmental, health-related, and other problems, both local and global. Prior to 2001, when IUPAC activities were organized around Commissions within the Divisions, a number of Commissions and Subcommittees within those Commissions focused on critical evaluation, constituting an ongoing pool of expertise in critical evaluation of chemical data. In 2001, [1] groups with ongoing interests in critical evaluation of data included the Subcommittee on Thermodynamic Data (within the Commission on Thermodynamics, I.2), the Subcommittee on Gas Kinetic Data Evaluation for Atmospheric Chemistry (within the Commission on Chemical Kinetics, I.4), the Commission on Atomic Weights and Isotopic abundances, II.1, the Commission on Equilibrium Data, V.6, and the Commission on Solubility Data, V.8.

Since the reorganization of IUPAC to a project-driven system, only three bodies have existed with a continuing focus on the compilation and critical evaluation of data—the Commission on Isotopic Abundances and Atomic Weights (II.1), within the Inorganic Chemistry Division; the Subcommittee on Modeling of Polymerization Kinetics and Processes, within the Polymer Division; and the Subcommittee on Solubility and Equilibrium Data (SSED), within the Analytical Chemistry Division. Based on a review of the titles of active projects listed on the IUPAC website in November 2016, it appears that three projects involving the critical evaluation of data are underway in the Inorganic Chemistry Division, five in the Polymer Division, and 16 in the Analytical Chemistry Division.

A project, “Interdivisional Discussion of Critical Evaluation of Chemical Data” (project 2016-043-1-500), has recently been started to bring a ‘whole-of-IUPAC’ approach to critical evaluation of data. The Task Group carrying out this project is composed of the authors of this article. We believe that the current organization of IUPAC, in which critical evaluation is organized along disciplinary lines, is appropriate and should be maintained. We also believe that an exchange of information and experience across disciplinary lines can be valuable and should be encouraged. The focus of this project is to convene a working discussion, open to all interested persons, on Tuesday, 11 July, during the 2017 IUPAC General Assembly. To begin this discussion, the organizers have posed four questions for consideration:

• How can IUPAC produce critical evaluations that are more useful to chemists and to non-chemist users of chemical data?
• How can IUPAC adjust presentation formats and dissemination channels to make critically evaluated data more accessible to potential users?
• How can groups of critical evaluators within IUPAC better learn from one another’s experience?
• How can IUPAC identify overlooked data categories of high societal value for critical evaluation and organize efforts in response?

As an example of how critical evaluation operates at present, we will describe the workings of the Subcommittee on Solubility and Equilibrium Data (SSED).

Figure 1: ‘Like dissolves like’, the logo of the Subcommittee on Solubility and Equilibrium Data
The Subcommittee on Solubility and Equilibrium Data (SSED)

The SSED, chaired by Clara Magalhães, coordinates projects whose objectives are:

- the comprehensive compilation and critical evaluation of published experimental data on the chemical solubility and other related thermodynamic data of well-defined substances (solubilities of gases, liquid and solids in liquids and solids) and other equilibrium systems (stability constants for homogeneous reactions); and
- the dissemination of the evaluated data on solubility and stability constants for homogeneous reactions through traditional (journal) and electronic (internet-accessible database) means.

At present the outcomes from the projects directly related to solubility are published in the *Journal of Physical and Chemical Reference Data* and the outcomes of the projects related to the Critical Evaluation of homogeneous systems are published in *Pure and Applied Chemistry*, usually as an IUPAC Technical Report.

The first annual meeting of the Commission on Solubility Data was held at McGill University in Montreal in 1974. Various topics were discussed, including guidelines for the creation of data sheets and evaluations; methods and the preparation of useful evaluations; and the process of recruitment of compilers, evaluators, and editors. By 2016, more than 100 volumes were published on the solubility of well defined systems. In 2001, after the restructuring of IUPAC, the Commission on Equilibrium Data merged with the Commission on Solubility Data, originating the present SSED. Experts in the fields under analysis have done this work, and the recruitment of compilers and other useful evaluators has been a constant effort for more than forty years. The SSED has also taken up work on the Stability Constants Database, a long-standing effort initiated by the Equilibrium Data Commission, and begun work on additional projects in critical evaluation.

To discuss the progress on the various projects, the SSED organizes annual meetings, either in conjunction with the IUPAC General Assembly and the World Chemistry Congress, or at other related international conferences. The SSED is also responsible, through the organization of biannual scientific conferences (International Symposia on Solubility Phenomena), for promoting research in all areas related to solubility and other equilibrium phenomena.

Approaches to Data Evaluation

Whether data are big or small, their usefulness is in proportion to their accuracy. To a metrologist the problem is easy: data that can be demonstrated to be metrologically traceable to a reference of some standing (such as a realization of a unit in the SI), with an appropriate estimate of measurement uncertainty, are evaluated as fit for some intended use if the measurement uncertainty is within a target range. Even now, when we have a better understanding of the requirements for metrological traceability, [2] published data do not always have clear-cut evidence of traceability or measurement uncertainty estimated by a method compatible with the *Guide to the Evaluation of Uncertainty of Measurement* (GUM). [3] Real-world evaluation is complex and very dependent on both the field of interest and the data themselves.

From its beginnings in the mid-1970s to the end of 2016, the IUPAC-NIST Solubility Data Series has published 103 volumes of compiled and evaluated solubility data and wrestled with many details of the critical evaluation process. Present thinking about best practices in the evaluation of solubility data are described in a guide for preparers and users of evaluated solubility data. [4] According to this guide, “Where two or more compiled independent measurements of solubility for a given system at similar conditions of temperature and pressure exist, an evaluation is prepared. ... The evaluator checks that the compiled data are correct, assesses their reliability and quality, estimates errors where necessary, and recommends numerical values based on all the published data (including theses, reports, and patents) for each given system. Thus, the evaluator reviews the merits or shortcomings of the various data. Only published data are considered.” While the Solubility Data Series uses a consistent format to present compiled and evaluated data, evaluators are given considerable latitude to use expert professional judgment to suit their evaluations to the systems with which they are dealing, provided that the process of reasoning leading to an evaluation is made clear. Fitting and smoothing equations as well as graphical representations of the data may be used, but the focus is always on experimentally obtained data rather than modeled values. Evaluators must be aware that their work may be used for purposes with a range of needs for data quality and by individuals with varying levels of knowledge of chemistry. Stability constants are evaluated in a similar manner.

If the knowledge is undigested or simply wrong, more is not better.
Data Evaluation in a World of Big Data

‘Big data’, by its nature, is not amenable to individual scrutiny, and only when condensed, analyzed, and transformed in some way can it be evaluated by traditional methods. Although chemistry is not immediately seen as a producer of big data, we can produce gigabytes of raw signal from NMR or LC-MS/MS, particularly in the biological –omics. Such data can be processed and is often visualized in the form of colorful connection maps expressing a maximum amount of information. It is interesting to speculate how such output might be evaluated if it is to be offered as definitive information. Back to the instrument—which will have been validated and calibrated for the task at hand at each stage of data collection, storage and processing—some trail must be generated that can give a green light to the evaluator. Taking the steps outlined above at each up-scaling of the process, there will need to be evidence that the software and algorithms have done what they purport to do. Coverage intervals on inferences will be necessary to understand how confident we might be in drawing conclusions.

A problem with evaluating big data will be the reliance on the generator and processor of those data to adequately describe and provide evidence for the evaluator. We cannot rely on the ‘wisdom of crowds’, especially when those crowds are rather small and self-supporting. With small data, we believe we know quite a lot about analytical techniques, for example pH titrations to obtain acid dissociation constants, their limitations and when results are likely to be reliable. This is unlikely to be the case with big data.

Even with modern approaches that might challenge traditional methods (for example “The death of the Job plot,…” [5]) the attention to careful assessment of uncertainties arising both from the measurements and the modelling of data leads to results in which we can have greater confidence. To conclude, we recall the words of A. S. Kertes in the forwards of volumes 1 to 53 of the Solubility Data Series (1979 – 1993), “If the knowledge is undigested or simply wrong, more is not better.” [6]

The authors hope, through the Interdivisional project, that IUPAC will continue to lead in the critical evaluation of chemical data – no matter how ‘big’!

References

Brynn Hibbert <b.hibbert@unsw.edu.au> is Emeritus Professor of Analytical Chemistry, UNSW Sydney. He is Past-President of Division V and editor of the revision of the Orange Book (Compendium of Terminology in Analytical Chemistry). ORCID.org/0000-0001-9210-2941

M. Clara F. Magalhães <mclara@ua.pt> is a professor in the Department of Chemistry and CICECO-Aveiro Institute of Materials, at the University of Aveiro, Aveiro, Portugal; Visiting Fellow in the School of Biological, Earth and Environmental Sciences, Faculty of Science, at UNSW Sydney, Australia, and Chair of the Subcommittee on Solubility and Equilibrium Data of the IUPAC Analytical Chemistry Division. ORCID.org/0000-0002-3359-850X

David Shaw <davidshaw@post.harvard.edu> is Emeritus Professor of Chemistry and Marine Science, University of Alaska, Fairbanks, USA. He is a Titular Member of the Division V and a former Chair of its Subcommittee on Solubility and Equilibrium Data. ORCID.org/0000-0001-8914-2862
As IUPAC approaches its 100th anniversary, it is important to re-evaluate the mode(s) in which it operates to sustainably support the standards that are vital to the chemical enterprise—chemical naming, chemical identification, and essential terms for commerce and the community. In addition, the importance of IUPAC’s underpinning of the digital representation of chemistry in the current ‘big data’ environment cannot be underestimated. Sustainable global support for the digital use of IUPAC assets is vital to the chemical sciences.

In 2014, Jeremy Frey promoted the idea of a “Digital IUPAC,” or “iUPAC” to highlight the need to support the computer readability of chemical information in addition to that of humans. [1] Specifically, he stated that,

The comprehensive conversion of IUPAC’s knowledge base of standards and definitions from human-readable to computer-readable form is essential. It is vital that this conversion be done now, as a matter of extreme urgency, if IUPAC is to maintain its role as the international authority for the chemical sciences.” [2]

We cannot agree more with this statement. Three years later, we argue that it is not only important for chemistry, but for all related sciences, as there are significant gaps in the development of semantic representation of chemical and biological entities and chemical concepts.

In this article, we present a discussion of the IUPAC Gold Book [3] as an exemplar of an important asset that needs to be digitally represented for preservation, maintenance, dissemination, programmatic access, and semantic application. This is an immensely important activity for IUPAC when we consider the number of volunteer-hours invested over the years in the development of the color books upon which the Gold Book is based. This effort continues previous work to formalize a future-looking digital management plan that could be applied to other IUPAC assets. [4] Issues that fall out of the development process are likely to be important in other digitization efforts and a proactive stance can inform appropriate policies and procedures for implementing and sustaining future projects.

The Gold Book is a compendium of authoritative chemical terminology originally compiled from IUPAC recommendations published by the scientific divisions of the Union. Since standardized naming of compounds became important, IUPAC has been involved in defining standards for chemistry. This effort has resulted in a series of ‘color’ books that define concepts in chemistry, many of which are integrated into the Gold Book. Scientific divisions in IUPAC are responsible for updating terms and drafting new definitions, which are then ratified by the IUPAC Interdivisional Committee on Terminology, Nomenclature and Symbols (ICTNS) and published in the Color Books or in IUPAC’s premier journal, Pure and Applied Chemistry (PAC). The Gold Book provides a unified portal into these definitions. The compendium was initiated by Victor Gold in the early 1980s [5] and has undergone several revisions to add and modify terms [6]. It was compiled online in 1997 in Portable Document Format (PDF) by Alan D. McNaught and Andrew Wilkinson of the Royal Society of Chemistry [7] and has since evolved into a web-based form to facilitate global access and use of these important terms.

The first digital version of the Gold Book was originally conceived in 2002 as part of an initiative to translate existing IUPAC standard terminologies into electronic “data dictionaries” in XML format. [8] This was an early vision of a “Digital IUPAC” and the first digital rendition was released in 2006 at the now familiar namespace, goldbook.iupac.org. Further enhancements to the content, markup, and site functionality were implemented in a project from 2007-2009, including functionality to promote linking and citing of the term definitions. [9] At this time, the Gold Book was registered in the CrossRef system of scholarly publishers, and each term record was assigned a Digital Object Identifier (DOI) for persistent referral back to IUPAC authority name-space. A snapshot of the interface is shown in Figure 1.

At the time of this writing, the current project has begun a move of the Gold Book to a new website built on a relational database, from which the pages can be dynamically created. This change will make the website more manageable through the ease of updating the scripts that are used to generate pages, more secure through frequent software updates, and more robust via integrity checking of the data in the database. Fear not though, the changes that will be enacted will
not change the content of terms and relationships between them.

**Lessons Learned for Sustainability**

In the Fall of 2016, it was reported that a large portion of the Gold Book website (terms A thru G) was not working. An analysis of website snapshots in archive.org’s WayBack Machine [10] shows that, somewhere between June and October, the number of pages declined significantly. As it turns out, over 8000 of the HTML files that contain the text of the terms had just disappeared—to this day there is no explanation for the loss. According to a poster mailed to chemistry departments by the Chemical Abstracts Service (CAS) last year, 50 % of data loss is because of hacking, theft, and loss. If that is the case, then the other 50 % of the loss is due to... what? Servers going down? Data getting corrupted? Formats not being readable? If nothing is safe when it’s on the web, how do we move forward with digital assets securely and sustainably?

It is thanks to the WayBack Machine that the Gold Book website content has been restored. With a little scripting, the missing files could be retrieved from previous snapshots and added back to the website. Based on this event, IUPAC migrated the pages from the Gold Book website to a new (modern) server, where the site could be better managed, actively maintained, and backed up. This event should be taken as the wake-up call that it is. While there was ultimately no loss of data (as far as we know), it could have been much worse. There was little information regarding the complete content and structure of the site as implemented over time and a backup was not known to exist previously.

**Lesson learned: Web versions of important IUPAC assets need to be continually managed and documented; backups should never be an afterthought.**

Taking a critical look at the Gold Book site in light of this event, we realize not only how valuable it is, but also how complex are its contents. Previous projects put
much effort into building a community resource that is not only comprehensive, but also interactive, and of course scientifically rigorous. Looking at the pages with more than a glance reveals a wealth of information and integration. Take for example the encoding of symbols, mathematical equations, and chemical reactions. Each of these components was developed using scaled vector graphics (svg) files—XML based image files—that were then converted into portable network graphic (png) files so that the presentation was preserved on all browsers. This involved much effort, automated scripts, and a naming convention for the files.

Another very important feature of each term is the mapping that shows how the current term relates to other terms in the Gold Book. The implementation of this on the page is in the form of three linked pages that contain image maps with links out to the terms related to a term at the first, second, and third levels (see Figure 2). This visual navigation of the content of the Gold Book can only be done via the web and highlights additional important information about the categorization of a Gold Book term, useful context for those less familiar with the Gold Book’s content. In today’s linked data perspective, image-based navigation is excellent for humans, but is not a true digital representation of the context that computers can interpret and can be difficult to maintain over time.

Lesson learned: Websites need ongoing technology updates and planned migration; IUPAC must support the value of digital assets as well as physical ones.

Given these recent concerns, it is clear that there needs to be a strategy for backing up the website. Going forward, this will involve mirroring the scripts that create the webpages on the new site at GitHub, [11] a popular online code repository, which will allow multiple authorized administrators and developers ready and secure access to the code. In addition to the code, backups of the database behind the new website will be uploaded to GitHub, as well as a mirror of the database on a server maintained by the current project lead at the University of North Florida. GitHub provides tools to document code and features, with complete logs of the changes that have been made and by whom. Let us know if you are interested in helping with this strategy to support the Gold Book website.

Anticipating Usage of the Gold Book for Humans and Computers

Two forward-thinking digital enhancements to the new site are eagerly anticipated: the development of the Gold Book terms as a formal, machine-readable controlled vocabulary (available in different formats), and the on-demand digital publication of Gold Book terms via a web-based application programming interface (API). These enhancements will allow IUPAC to sustainably support digital functionality of the Gold Book as well as access to content.

The Gold Book serves as a basis from which to reference specific concepts using approved formal terminology. Information scientists are clamoring for authoritative controlled vocabularies in machine-readable
format to reference in databases, computational models, and other applications. Distributed scientific information systems need to be able to exchange and connect data and controlled vocabularies can facilitate linking across these systems. Of course, controlled vocabularies are not a new thing (your librarian will remind you of this), but the digital application of highly specialized scientific vocabularies is a hot topic right now as we start to understand the ramifications of ‘big data’ for accelerating science.

Machine readability of the Gold Book terms in a controlled vocabulary is only the first step in the logical progression toward the eventual development of an ontology. Formal digital ontologies allow the expression of the meaning of a concept as well as the relationships among concepts, written in a formal language. [12] Ontologies can be referenced by databases to incorporate greater context around data. Figure 3 shows how an ontology entry for the term ‘absorbance’ might be formulated in the Web Ontology Language (OWL) [13] using the Protégé software package. [14] The representation is based on the content of the current Gold Book page, including the source and related terms. This approach could be used to digitally represent the term link maps illustrated in Figure 2. To formalize such structures, IUPAC will need to formulate an approved specification for defining the metadata necessary to correctly characterize and contextualize each term.

Computer-based access to the Gold Book terms can be provided from the new site through an API service. Initially implemented as a proof-of-concept, the API will allow non-HTML versions of the information on each term to be delivered via a documented, standard format. APIs are very common these days on many public sites (e.g. Facebook, Wikipedia, PubChem) and provide a way for web developers to integrate their content accurately with other websites.

As an example, the term absorbance might be available to view in HTML at:

http://goldbook.iupac.org/view/term/absorbance
A more transportable output format could be made available at:

http://goldbook.iupac.org/view/term/absorbance/json

JavaScript Object Notation (JSON) is consistent and succinct and can be read by many programming languages. Offering output in JSON could open up many opportunities for the integration of authoritative IUPAC terms into other websites and would allow the promotion of the service. Note the reference back to the IUPAC-based DOI for the authoritative ‘copy of record’ in the JSON example in Figure 4.

The Goldify feature implemented in the previous version of the Gold Book website that allowed automatic addition of term links to other electronic documents could be accomplished via the API in such a way that the usage of the feature could be tracked over time. This in turn could identify major users of the Gold Book and catalyze the development of collaborative projects that make the Gold Book more useful, usable, and relevant. In the long term, there is much potential to develop and manage a multitude of services for Gold Book terms, leveraging their online nature through future projects.

Towards Digital IUPAC

It is clear that to make the Gold Book and other valuable IUPAC resources as widely available as possible, they must be implemented digitally on the web in addition to traditional print copies. This vision has been promulgated over time in several projects addressing the IUPAC Color Book corpus. [15,16] Given the current climate, it must be approached in a way that supports computer as well as human use and honors the rigorous scientific definition process that exists within IUPAC. Many questions arise as we consider what an appropriate web version ‘means’ and how to develop a strategic plan to formulate and sustain digital assets.

- Where are the data? the backup? the documentation?
- In what format(s) are the data and how do we maintain their integrity?
- What processes need to be in place to support the metadata? the web presence? DOI? the formal vocabulary/ontology structures?
- What policies address the digital integration of information managed by IUPAC with information managed by other entities?

Figure 4. Sample JSON representation of Gold Book term ‘absorbance’

```json
{
    "id": "https://dx.doi.org/10.1351/A00028",
    "term": "absorbance",
    "definition": "Logarithm of the ratio of incident to transmitted radiant power through a sample...",
    "relatedTerms": [
        "https://dx.doi.org/10.1351/goldbook.A00013",
        "https://dx.doi.org/10.1351/goldbook.R05040"
    ],
    "history": [
        {
            "revision": 2,
            "file": "2006-11-02T14:43:22",
            "version": "A00028.xml"
        },
        {
            "revision": 1,
            "date": "2006-09-29",
            "activity": "updated",
            "comment": "Taken from the original Gold Book, published in version 1.0.0 of the XML Gold Book.",
            "editor": ""
        }
    ],
    "source": "Green Book. 2nd ed : IUPAC Quantities, Units and Symbols in Physical Chemistry. ...",
    "citation": "IUPAC Compendium of Chemical Terminology, 2nd ed. (the 'Gold Book')... DOI: 10.1351/goldbook.",
    "seealso": [
        "https://dx.doi.org/10.1351/pac19966812Z2Z3",
        "https://dx.doi.org/10.1351/pac19906211Z67"
    ],
    "status": "approved",
    "version": "2.3.3",
    "updated": "2014-02-24"
}
```
IUPAC joins many organizations globally going through a transition from physical to digital resources. The process is inevitably iterative and lessons learned along the way uncover many best practices, example use cases, and technology solutions with which to navigate. It is important that IUPAC defines what is important in this process and continues to work toward solutions that fill the needs of members and of the chemistry community. IUPAC Divisions need to plan for digital migration and consider consistent, sustainable, and cost effective solutions to the critical issues of security, integrity, provenance, and management of digital assets. Members at all levels need to engage and actively recruit new members to help shape the Union as it approaches its centennial.

Conclusion
The digital footprint of IUPAC is an important global resource for the Union as it approaches its 100th anniversary. Managing the corpus of assets that IUPAC has developed in this time is an ongoing and monumental task. We all understand how important it is in the long term to curate and promote these contributions so that the chemical community and the Union continue to flourish in the next 100 years.

References
8. IUPAC project “Standard XML data dictionaries for chemistry” https://iupac.org/project/2002-022-1-024
13. OWL 2 Web Ontology Language. Available at https://www.w3.org/TR/owl-overview/
15. IUPAC project “Software framework for transformation of IUPAC Color Books to XML” https://iupac.org/project/2007-014-1-024

Stuart Chalk <schalk@unf.edu> is an associate professor of chemistry at the University of North Florida, USA. He is a member of the CPCDS Subcommittee on Cheminformatics Data Standards. ORCID.org/0000-0002-0703-7776

Leah McEwen <lrm1@cornell.edu> is chemistry librarian at Cornell University, USA. She is a member of the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS), and co-chair of the CPCDS Subcommittee on Cheminformatics Data Standards. ORCID.org/0000-0003-2968-1674
Supporting Prudent Practices in Research Laboratories

by Leah McEwen

In response to a sequence of dramatic accidents in research laboratories, [1-3] over the last five years the academic chemistry sector in the United States has focused on improving its safety culture and adopting more prudent laboratory practices. [4] Crucial to this effort is improved accessibility and management of chemical health and safety data and information tools that support laboratory process risk assessment. To this end, chemical safety, information, and education professionals are collaborating to facilitate and improve the risk assessment process for chemistry researchers and research students, as reported previously in Chemistry International. [5]

The goal is to connect existing chemical safety and education communities and information technologies to improve Findability, Accessibility, Interoperability and Re-use of chemical safety data for both researchers and information systems. These are referred to as the FAIR principles for sharing research data. [6] There are a number of collaborative efforts applying this concept in order to facilitate the use of existing data, as well as to identify data gaps in the support of hazard analysis and risk assessment for academic research scenarios.

These efforts have identified a number of critical challenges to supporting safety data management at the scale and breadth needed for institutions researching novel compounds. For example:

- Pertinent data for assessing and managing risks are scattered across many agency and industry resources, nationally and internationally, and in many different formats.
- Reporting requirements vary by sector and region, and cause difficulties for exchanging and evaluating data.
- Diverse schemes are employed to identify chemicals and mixed substances are often only indexed by the primary component.
- No formal practice exists for reporting incidents to provide valuable “lessons learned” data.
- Few sources are designed for machine handling of the data to support local applications.

Addressing these challenges engages technologies and expertise from several national and international efforts in chemical data management. These include the IUPAC International Chemical Identifier (InChI); [7] the National Center for Biotechnology Information (NCBI) in the US; [8] the Pistoia Alliance; [9] the Research Data Alliance (RDA); [10] and the technical Divisions of Chemical Health and Safety (CHAS), [11] Information (CINF), [12] and Education (CHED) [13] of the American Chemical Society (ACS). [14] These tools are relevant to supporting the Recognize, Assess, Minimize, Prepare (RAMP) paradigm recently popularized by the American Chemical Society. [15]

Preliminary analysis of the results of a recent survey of the academic chemistry community in the US points to a specific example of these problems. Of those surveyed, Safety Data Sheets (SDS) from chemical vendors were overwhelmingly the go-to source for chemical safety information. [16] There is concern, however, in the professional chemical safety community about the efficacy of using the SDS out of context to support laboratory risk assessment. While the SDS provides a general safety profile for an individual known chemical, chemistry research involves reactions among multiple compounds and substances combined in various proportions under diverse conditions. The chemical safety context is further influenced by processes within the experimental procedure, the management of substances from inventory through disposal, and setup and maintenance of the laboratory environment. Very little information relevant to this type of analysis has been incorporated systematically into the laboratory safety process.

There is a body of research dedicated to analyzing the operations and conditions of large scale chemical processes in industrial settings, where these processes are well-defined and carefully specified as part of the planning process. However, such analysis is rarely conducted for chemical procedures developed iteratively at the laboratory level as defined by OSHA regulations in the United States. [17] Very often, the most important source of relevant information is from reports of incidents where safe control was exceeded and the influence of reactivity and process factors can be considered in retrospect. However, these reports are not the focus of regular research reporting and tend to be brief mentions found sporadically in letters to editors of journals, [18] or as news items, [19] or occasionally rephrased as caution statements in vetted procedures. [20]

The value of compiling this data across many compounds is realized in the Bretherick’s Handbook of
Reactive Chemical Hazards. [21] Not only does this handbook serve as a reference for individual chemicals, but similar reactivity hazards are grouped by chemical classes or atom groupings, suggesting other possible hazardous combinations. This tool is a step toward filling the significant need to project potential hazards beyond well-characterized chemicals when designing novel experiments. However, these reports are still locked in text and not easily parsed for incorporation into digital information workflows.

The current focus on chemical safety in research has inspired a number of additional efforts to fill this gap in experiential information. The Pistoia Alliance recently launched an open tool prototype for collecting pre-competitive reaction hazard data that can be downloaded for use in local inventory and planning systems. [22] The ACS is implementing a requirement that authors publishing in their journals include descriptions of potential hazards during their research processes that require safety measures beyond normal precautions. [23] This policy was prompted by a study of the ACS Committee on Chemical Safety last year that found very little mention of chemical safety reporting in the author guidelines of hundreds of scientific journals. [24] If other scientific societies and publishers embrace similar initiatives, reporting on chemical procedure hazards may begin to permeate the literature. The ability to make this information discoverable at the time of need will depend to some extent on the systematic description of these hazard scenarios, which will further depend on chemical safety terminology. Benefits can be realized by both traditional indexing and semantic methods.

Reframing these challenges and opportunities from a data management perspective, where machine processing is a critical requirement, raises the following questions:

1. How can chemical safety information be collated for further application?
2. How can chemical specific data be mapped to real-world chemical work?
3. How can incident data be described systematically in a laboratory context?

Collating data across sources adds value for analysis and enables researchers to more readily assess risk, especially that of novel compounds and evolving procedures. In the United States, the NCBI at the National Library of Medicine hosts the publically available PubChem database, [25] which collates public data on over 90 million compounds, semantically represented and programmatically accessible for reuse in local applications. Chemical safety information for over 100 thousand chemicals, from authoritative agency sources worldwide, is dynamically presented as Laboratory Chemical Safety Summaries (LCSS, see Figure 1 [26]), [27] based on the format described by the US National Research Council (NRC) in Prudent Practices in the Laboratory. [4] This data augments the information generally provided by Safety Data Sheets, and includes many chemical and physical properties, GHS Hazard Classifications, and specific examples of incompatible reactions, as sourced from the NLM Hazardous Substances DataBank, Sigma Aldrich Safety Center notes, the US National Fire Protection Agency (NFPA) documents, Sax’s Dangerous Properties of Industrial Materials, and Bretherick’s Handbook of Reactive Chemical Hazards, among others. Data are reported directly from multiple sources with full provenance, allowing users to make comparisons and determine their applicability to local contexts. The data are also programmatically accessible and can be provided as a stream to support integration into local systems and applications.

In the safety context, hazards can be presented by any component in a chemical system, such as solvents. In practice, no compound is pure and composition impacts chemical reactivity, unintentionally or by design. Characterization of components provides better experimental results, in addition to safer laboratory functions. Managing information about mixed substances becomes a significant challenge in a world of 100+ million characterized compounds. In the real world, many articles of trade exist as defined or partially defined mixtures, yet often only the components of interest are noted. The identification of chemical entities involved in multi-component systems could enable the connection of safety data, predominately ordered.
Supporting Prudent Practices in Research Laboratories

by individual chemicals, to commercial products and real-world chemical work. For these reasons, there is a project underway to leverage the IUPAC InChI algorithm to articulate the defined chemical aspects in a mixture system (see Figure 2). [28-29]

There are many factors at play in conducting a laboratory procedure that may contribute to the potential risk of a given situation. The safety community is looking to practices in industry to formalize useful approaches to risk assessment in the research laboratory context. [30-32] Analyzing procedures and coupling these with incident data can potentially bring to light incompatible combinations and problematic operations, as well as aid in planning for adjustments to experimental parameters. This opportunity can be greatly informed by tapping into collective information across many documented processes. Domain terminology that describes key factors can enable the systematic analysis of relationships, such as combinations of chemicals, or substances under different conditions. This approach has been used for single analysis of M/SDS documents, [33] and chemical procedures. [34] Melding these vocabularies to consider data from incidents more systematically and across compiled reports is the forefront of another community effort. [35]

The safety data ecosystem is exceedingly complex and a classic management challenge, involving many different uses, applications, stakeholders, expectations, and requirements. The scope is dynamic and exceeds the information space of well-documented chemicals. Several hundreds of thousands of different substances, involving hundreds of sources and thousands of users, may be present at a large research campus. Turning this data into improved safety in practice is a community challenge, engaging the expertise and dedication of a broad range of professionals in the research, education, and service sectors. The good news is that this challenge is increasingly recognized by the chemistry community and emerging information sources and tools show important potential in helping to address this opportunity for improvement.

References
6. www.force11.org/group/fairgroup/fairprinciples

Figure 2. InChI Extension for Mixture Composition (Draft Specification)
Chemical Health and Safety Data Management

7. www.inchi-trust.org
10. www.rd-alliance.org/groups/chemistry-research-data-interest-group.html
11. http://dchas.org
13. www.divched.org
14. www.acs.org/safety
19. https://dchas.org/the-dchas-l-list
22. www.pistoiaalliance.org/projects/chemical-safety-library
31. www.acs.org/hazardassessment
33. www.ilpi.com/msds/ref/demystify.html
34. http://chemicaltagger.ch.cam.ac.uk

Leah McEwen <lrm1@cornell.edu> is chemistry librarian at Cornell University, USA. She is a member of the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDs), and co-chair of the CPCDs Subcommittee on Cheminformatics Data Standards. ORCID.org/0000-0003-2968-1674

Don’t miss new PAC content. Sign up for eTOCs at www.degruyter.com/pac
One of the questions that scholarly publishers have faced as a result of the increasing interest in research data is the relationship of journal articles to the research data behind those articles. From the 1970s forward, journals in the physical sciences published research data in the form of supplemental materials. Due to economic considerations and the sheer volume of data, this supplemental material was generally distributed in microform. In the late 1990s, some publishers began experimenting with digital distribution of research data. By the early 2000s, the volume of research data being submitted was causing problems for editors and reviewers. This tension was captured well in an editorial by Emilie Marcus in Cell in 2009. [1]

The importance of research data has been considered for some time. An interesting article by Müller [2] envisioned a somewhat primitive connected world in the late 1950s:

“It is manifestly impossible to embark upon a program to print prodigious detailed tables of all the data known to physical science. It would exceed our supply of paper, few could afford it, and storage would be a major problem. It would seem eminently feasible, however, to punch program cards for tens of thousands of cards for as many empirical or fundamental equations and the data to which they apply, in a suitable computer center or agency these could be interrogated when necessary and the detailed data sent to a subscriber by teletype or more leisurely by mail. Individual dialing of long-distance ‘phone calls will be with us all very soon. It is not too great an extrapolation to imagine the time when one will be able to dial a call to a computer agency and, upon requesting a certain code designation, have an answer returned in a matter of minutes. There are no existing technical impediments to this scheme. The problem is one requiring vision, organization, and dedicated policy.”

Two comments are interesting to note. First, even today, if one were to store all the raw data being collected through scientific research, from laboratory instruments to space telescopes, the global storage capacity would be strained. [3] Second, in a later part of the paper, Müller bemoans the likelihood that telecommunications might be overwhelmed by entertainment and advertising, to the detriment of science. It seems he was prescient in both the benefits and pitfalls of a large, distributed network.

It is also interesting to note that in the chemistry domain, a long tradition of aggregation of published data has grown without the additional efforts of individual researchers. These include various spectroscopic databases, such as the NIST, Wiley, and BioRad collections of spectral data of various types, as well as structural information in the CCDC, ICSD, and PDB collections. [4] The latter system, with the support of the US government, has become a repository where researchers are expected to deposit structural data concurrent with article submission.

As the cost of storing and disseminating digital content dropped, the economic burden for storing and distributing research data decreased. Because of its importance in describing and understanding research, the publication of research data has been advanced as a mechanism to bring greater transparency to science, to improve reproducibility, to reduce duplication of research, and to better detect fraud. [5] In fact, some advocates have proposed the publication of research data as an end in itself. These proposals, that data be considered a first class object, [6] or that scientific dissemination in the form of journal articles be replaced by data that speak for themselves, is making many in the scholarly publishing world rethink the relationship of the data to the textual description of the research. By extension, this also affects the ways in which scientific reputations and careers are assessed (i.e., metrics).

Given all the positive rationales for researchers to publish and share their data, one might question why more scientists are not actively doing so. A number of reasons can be given. It should be said that, pre-Internet, supplementary material generally took the form of printed pages, even when the data themselves were collected digitally. Even though the transmission of digital data can be accomplished more easily, many researchers continue to provide their research data as word processing documents or PDF files. While these are adequate for human consumption, they are not necessarily useful for importing data into a software package for visualization or for reanalysis. Spreadsheets are also a common way for research data to be shared. These are easier to visualize and re-use than PDF files. However, at issue in all cases is the extra effort involved in providing enough descriptive metadata for readers or software processors to understand the experiments completely. It is simply easier to provide research data in the traditional formats. It takes time and effort to prepare data for publication. It takes more effort to prepare the raw data for

---

*Primary Research Data and Scholarly Communication*  
by David Martinsen

---
In addition, researchers feel that they have already shared their data when an article is published, with the data presented as tables or images in the article itself or in the supplemental material. This is one conclusion that can be inferred from a Wiley survey of 2250 authors, where 50 to 75% of researchers claimed to be sharing their research data as supplemental material. [7] A more recent survey, from Leiden University and Elsevier, [8] confirms this conclusion. Researchers who do publish their data do so almost exclusively within the current journal publishing environment, and over a third of researchers do not publish any research data. This study makes a distinction between disciplines that are intensively data sharing, where sharing of datasets is necessary for analysis, and those disciplines where there is restricted data sharing. In chemistry, computational and cheminformatics studies may fall under the intensively data sharing scenarios, while the experimental disciplines, perhaps the bulk of chemistry research, fall under the restricted sharing scenario. The study presents some of the challenges that need to be overcome to enable better data publication and sharing in those restricted sharing environments.

The scholarly publishing industry has, arguably, been proactive in the research data discussion. The International Association of Scientific, Technical, & Medical Publishers (STM, http://www.stm-assoc.org/) has been involved in a growing number of research data initiatives. STM provided input to the Opportunities for Data Exchange Project, which resulted in a report on integrating data and publications. [9] A data publication pyramid, as shown in Figure 1, has been used to describe some of the issues associated with research data. At the top of the pyramid, the smallest two segments represent the bulk of the data that has been prepared for publication, either within the article or in the supplemental materials. The bulk of the data, though, at the bottom of the pyramid, is never published but rather contained in disks or file drawers. While the pyramid is instructive, it seems to me that an iceberg might be a better metaphor. In the iceberg, as shown in Figure 2, [10] it is much more apparent that the hidden data below the surface is not nice and neat. However, neither is the portion above the surface, the published part, very nice and neat.

Each year, technology leaders in the STM publishing industry gather to review technology issues of the past year, and identify trends that will be important over the next five years. At the most recent meeting, in December 2016, the group produced a graphic that envisioned the scholarly communication ecosystem as a pinball machine. Through a combination of skill and chance, researchers, funding agencies, academic institutions, and publishers navigate a series of obstacles and opportunities, trying to gain points to enhance their reputations. In this representation, shown in Figure 3, [11] the center of the image is that of trust and integrity. With the rise of fake news and the “debunking” of science, the scientific enterprise must pursue the truth with vigor and with transparency. The publication of research data is a significant part of that effort.

One reason given for publishing data is to improve the reproducibility of science. Started in 1955, perhaps as a joke, the Journal of Irreproducible Results [12] now finds itself in the vanguard of studies in reproducibility, including reproducibility guidelines, academic departments devoted to reproducibility studies, [13] and most recently a manifesto. [14] Recent studies have cast doubt on the reproducibility of science, leading to the...
perception that perhaps scientific research itself, or at the very least the peer review process, is flawed. [15] Reports in the popular press of recommendations that contradict earlier recommendations add to this perception. For example, the US National Institute of Allergy and Infectious Diseases recently recommended that the rise in peanut allergies may be due to lack of exposure to peanuts in young children. [16] Thus, rather than avoiding peanuts in the early years of life, exposing children to peanuts may actually be a better choice. There is certainly much that can be done to improve the quality of experiments and the quality of data collected. However, there are a couple of issues related to reproducibility that are relevant as we consider the realm of Big Data. First, there is a difference between correlation and causation. It is tempting to journalists hunting for the next attention grabbing headline to jump from correlation to causation, and it is tempting for researchers to do the same. It is important to ask, though, whether in any given study there were enough samples or a large enough range of samples studied to even propose a general correlation, and only then ask whether the correlation actually means causation. In the life sciences, where it is very difficult to control all of the possible variables in an organism, this challenge is great. However, even in the physical sciences, where it is perhaps easier to control the experimental conditions, the distinction between correlation and causation remains difficult. It is often the outliers that really lead to new understandings of science. For example, the announcement of the 2011 Nobel Prize in Chemistry, awarded to Dan Shechtman, notes that Schectman “had to fight a fierce battle against established science”. [17] Could a machine have made such a remarkable discovery? Big data analytics are designed to look for correlations, with humans providing the judgement that the correlation is related to causation. It will be interesting to see if the tools and techniques related to cognitive computing and artificial intelligence can eventually lead to confidence in a conclusion that a specific correlation is indicative of causation.

A second aspect of reproducibility is related to a fundamental randomness in nature at a granular level and, as we develop more sensitive instrumentation, the limits of detection of a particular method. In fact, one indication of manipulated data is whether a match is too good. There is an inherent irreproducibility in nature; that is one of the reasons that, when conducting experiments, one should always conduct more than a single instance and make several measurements to compare.

Questions of reproducibility should also be considered within the context of the idea that negative data should be published in addition to positive data. [18] This idea perhaps has different significance in the life sciences than in the physical sciences. In the life sciences, it is useful to know whether a particular drug has an adverse affect on a significant percentage of the population, or in fact has any medicinal affect at all (keeping in mind the difference between correlation and causation). In the physical sciences, though, one might imagine that there are no negative results in that same sense. There are only results. It is up to the researcher to interpret the results. If the results don’t match the preconceived notion of the researcher, or the research community, then it is up to the researcher to explain the results. In this way, unexpected results become novel positive results. It is said that publishing negative results will prevent much duplication and waste of research dollars. However, if the negative results of the past were not questioned, would we have had publications on arsenic-based life [19] (subsequently proved to be an invalid conclusion), or on compounds of helium? [20]

On the other hand, despite these cautions, the benefits of access to both primary research data and big data analytics will be great. We are moving beyond the use of technology to simply collect, store, process, and preserve data to using it to analyze and interpret...
data. These cognitive technologies are still in their early stages, and we can expect big failures along the way. The expectation that technology companies need only turn their attention to these problems to quickly solve what researchers have been struggling with for years and years is not a proven assumption, at least while we are in this transition period. Looking back, we may think that the change was remarkably rapid. But for now, the perspectives of Francis Collins and Craig Venter on the 10th anniversary of the human genome publication are still relevant: [21]

“But for all the intellectual ferment of the past decade, has human health truly benefited from the sequencing of the human genome? A startlingly honest response can be found on pages 674 and 676, where the leaders of the public and private efforts, Francis Collins and Craig Venter, both say ‘not much’.”

References:
7. Wiley Data Sharing Survey: https://figshare.com/articles/Data_Sharing_Infographic/3555993
13. John Ioannidis, Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford 94304, California, USA.
15. John P. A. Ioannidis, Why Most Published Research Findings Are False, PLOS Medicine, Published 30 August 2005, https://doi.org/10.1371/journal.pmed.0020124
21. Editorial: The Human Genome at Ten, Nature 464:649-650 https://doi.org/10.1038/464649a; Published online 31 March 2010

David Martinsen <dmartinsen@consultdpm.com>, formerly a Senior Scientist at ACS Publications, consults in scholarly publishing at David Martinsen Consulting in Rockville, MD, USA. He is co-chair of the CPCDS Subcommittee on Cheminformatics Data Standards, and is also a co-chair of the Research Data Alliance Chemistry Research Data Interest Group. ORCID.org/0000-0002-8667-5855
Leveraging the Web

Production, Management, and Dissemination of Chemical Information

by William Fyson, Simon Coles, and Jeremy Frey

As an open, inexpensive, collaborative platform, the Web is ideal for facilitating communication among scholars and enabling nearly free access to knowledge. The Web’s potential allows researchers to utilize the digital, networked medium to publish more of their research data in comparison to paper-based journal articles and to publish them in context-sensitive formats, enabling wider access and visibility along with increased discoverability and the potential for further use, including through the deployment of e-science techniques. [1] However, this potential is often not fully achieved. Academic publishing has created a market, formed over hundreds of years, that fulfills a number of processes, such as peer review, impact assessment, and allocation of funding. Numerous stakeholders, ranging from publishers to academics to research funders, all have their own agendas and conflicting interests, which can inhibit the adoption of innovative approaches to leveraging the Web. Thus, in order to better enable the potential of e-science, an approach that merely leverages the Web to develop new tools and platforms for dissemination is insufficient; the socio-economic forces that influence stakeholders’ behaviour must also be considered. [2]

To better understand how these interactions affect scholarly discourse and to encourage a more open approach to publishing that enables data reuse for a range of stakeholders, a prototype Web service was developed to facilitate a process which we have termed disaggregation—the breaking down of conventional publications, such as journal articles or theses, so that their constituent elements may be disseminated and reused more freely. The Web enables a vast range of research outputs to be communicated and presents new mediums through which a researcher’s contribution to their academic community may be measured. Yet the Web is primarily used to distribute journal articles, [3] a format that has remained largely unchanged over hundreds of years. By presenting new approaches to conventional publications, disaggregation disrupts many of the interactions between stakeholders, offering new opportunities for sharing content and facilitating peer review, and also new metrics for measuring impact.

This paper presents a thorough evaluation of the effect disaggregation would have on the entire process, from research planning through data generation to disseminating and reusing the outputs. This is achieved through the medium of interviews with expert stakeholders, including perspectives and insight from academics, publishers, and librarians. During the interviews, participants were first asked to describe their role in academia before considering a hypothetical world of scholarly discourse where researchers distribute research outputs freely on the Web without the aid of academic publishers, with the aim of encouraging participants to adopt an open mind when exploring the topic of academic publishing. The interviewees were then asked to evaluate how disaggregation may coexist alongside the conventional approaches to scholarly discourse and how it may impact upon their role. All interviews were conducted in a semi-structured fashion to allow the expert interviewees to elaborate on their specialist expertise where appropriate. Through such an evaluation, a number of themes begin to emerge concerning the nature of researchers’ attitudes towards the dissemination of research, how current practices may be adapted, and what needs to be developed and standardized, with the themes having implications for both an understanding of the scholarly process and how disaggregation may be implemented.

The first two themes are interlinked, concerning the nature of cultural norms and behaviours within academia and the roles of stakeholders. The participants’ comments highlight a conflict in the relationship between a long-term shift in cultural norms and a short-term shift in behaviours and decision making. Peer review provides an example of this conflict. There are rarely explicit incentives for reviewing, yet researchers continue to review to fulfil their professional duty. When discussing stakeholder roles, participants focused on the notion of value adding services. For example, researchers emphasised their passion for doing research, with this being the activity that they most want to focus on and the writing of articles being an inconvenience. Thus, publishers add value to scholarly discourse, providing a venue for researchers to disseminate in and prescribing what the final output that the research is aiming towards should look like—the presence of publishers thus ensures the smooth running of all the other functions that stem from it.

These two themes together lead on to the third, the role of reputation in academia and the need for work
to be recognized by the rest of the research community, an established driving mechanism in academia. [4] Researchers offer their contributions to the rest of the community, not for direct financial profit, but to enhance their reputation amongst their peers through recognition for contributing novel and useful research. Researchers, institutions, and funders need to be able to quickly identify what work is reputable, a task that journal publications achieve through the power of their brands and impact factors, even if such metrics may be flawed. This need for the means of dissemination to be recognized and understood can prevent new approaches from being successful, with dissemination platforms requiring a significant mass of users. Publications provide a convenient shorthand for busy academics who need to make quick value judgements concerning the quality of work and hiring decisions, rendering them too convenient for the system to be deprived of. Any substitute approach to research dissemination offered to researchers can only be regarded as a trade off in comparison.

From the above themes, it becomes apparent that the act of disaggregation as carried out by researchers on their publications is flawed. Initially, the notion of disaggregation is sound—publications can be broken down into constituent parts and these can be used to help different stakeholders achieve their goals. It was also suggested that the process may be reversed to allow for the re-aggregation of content to create new outputs. However, the implementation of disaggregation proposed is ultimately flawed. It is both time consuming for the researcher and, as an unrecognised form of dissemination, lacks reputational impact.

The final themes therefore focus on how disaggregation may be implemented instead, with a large focus placed on the importance of the researcher’s day-to-day workflows. It is here that the researcher’s cultural norms may be challenged and innovative new approaches to scholarly communication and knowledge management have the potential to add value. Rather than extract content from existing publications, the processes of disaggregation need to be built into the researcher’s workflow, embedded through the whole research lifecycle. By doing so, it provides researchers with the resources they need to help structure their work and produce the conventional outputs that drive their careers, as well as satisfy many other purposes, from data sharing within the lab, through engaging the general public. The key to motivating researchers to engage with new approaches that may be leveraged to encourage open and structured sharing of knowledge begins with demonstrating how adopting such techniques may help achieve the goals they are already focused on. New approaches need to complement existing methods rather than act as poor substitutes.

Therefore, if disaggregation is to be successful, it needs to be a flexible and robust process that can be applied within any lab context, but also go on to help produce research outputs recognised by the whole academic community. This in turn requires standards for structuring and querying the extracted data. These standards must be flexibly defined and adaptable, allowing the stakeholders across academic publishing and research contexts to each define their own disaggregation requirements, which should ultimately be interoperable with standards defined by other individuals who stand to benefit from the disaggregation process. Such a means for organizing, structuring, and querying data presents new challenges for organizations such as IUPAC, concerning both technological issues as well as governance complexities. If such challenges can be overcome, then chemistry researchers can become better at what they are already doing by adopting techniques that inherently promote the open and organised use of the Web for the production, management, and dissemination of chemical information.

References

William Fyson <rwf1v07@soton.ac.uk> is a Doctor of Web Science at the University of Southampton and currently works on the EPrints open access repository software. Simon Coles <S.L.Coles@soton.ac.uk> is Professor of Structural Chemistry and Director of the UK National Crystallography Service at the University of Southampton. Jeremy Frey <J.G.Frey@soton.ac.uk> is Professor of Physical Chemistry at the University of Southampton, Southampton, UK. ORCID.org/0000-0002-7227-8293
Reproducibility, Validation and Reuse of Crystal Structure Data

by Ian Bruno

The Cambridge Structural Database (CSD) provides a platform for sharing data generated from X-ray and neutron diffraction experiments. [1] It contains experimental determinations of the 3D structures of over 850 000 small organic and metal-organic compounds. In recent years, the rate of growth of the CSD has shown an upward trend, with over 80 000 datasets deposited in 2016. Collectively, these datasets constitute a rich resource of information about 3D molecular structure. Knowledge derived from the CSD is used by academic and industrial scientists worldwide to help address scientific challenges across a range of domains.

Crystallography as a discipline has established standards and best practices that support the reliable exchange of experimental data. [2] Central to these is the Crystallographic Information File (CIF), [3] which enables the semantic representation of metadata pertaining to the experiment, derived and processed data, and the methods used to determine the structure. [4] Publication of both processed and derived data has been common within the crystallographic community for many years, with conversations now turning to the publication and preservation of raw data. [5]

CSD data deposition services aim to make it easy for researchers to comply with recommendations around best practice and policies for publishing crystal structure data. This includes integration with the community-supported checkCIF validation service that assesses the consistency and completeness of the data. [6] checkCIF issues alerts of varying degrees of severity and journal policies often require that severe alerts are addressed or explained prior to publication.

Datasets published in the CSD are uniquely identified by a Digital Object Identifier (DOI) [7] and can be independently cited. CSD deposition services also encourage depositors to supply an ORCID identifier [8] to enable the reliable and unambiguous association of a researcher with their research output. Whilst many structures are associated with a journal article, an increasing number are separately published as CSD Communications. [9]

A CSD entry includes a chemical representation of the substance studied by the diffraction experiment. This is vital for enabling the effective discovery and reuse of the data, particularly in domains beyond crystallography. Generating this representation uses a combination of automated processes and validation by expert scientists. Automated processes generate diagnostic information indicating probable points of error and a reliability score that helps prioritize manual validation activities. [10]

The combination of chemistry and crystallography in the CSD provides a foundation for software solutions that enable knowledge about molecular shape and interactions to be applied to the design of new molecules and materials in areas such as drug discovery [11] and solid form optimization. [12] A reliable chemical representation makes it possible to generate standard InChIs [13] that can be used to establish interoperability between the CSD and other chemical and biological resources. Links have thus far been established between ChemSpider, [14] PubChem, [15] and the Protein Data Bank. [16]

The CSD is part of a wider ecosystem comprising the technical and social components needed to make crystal structure data available in support of published research and for reuse in the pursuit of new discoveries. Achieving this requires the commitment of researchers, publishers, and repositories alike and is greatly aided by community-based standards and recommendations. The experiences of the crystallographic community demonstrate what is required to provide joined-up systems that support the stewardship of data from instrument through publication and subsequent application across domains. The challenges encountered and lessons learnt in the field of crystallography are potentially applicable to initiatives aimed at achieving similar ends for other types of data relevant to chemistry.

References
5. L.M.J. Kroon-Batenburg, J.R. Helliwell, B. McMahon,
From Experiments to Knowledge


Ian Bruno <bruno@ccdc.cam.ac.uk> is Director, Strategic Partnerships at The Cambridge Crystallographic Data Centre, Cambridge, UK. ORCID.org/0000-0003-4901-9956

Big Data in Chemical Industry

by Dr. B. Saha

Big data with advanced analytical software is now being used by many chemical companies to improve business performance.

There are several areas where “Big Data” has started playing an important role in chemical industry. This includes Manufacturing, Supply Chain Management, Marketing, Innovation, and Human Resource Management.

In manufacturing, use of Big Data is helping companies to improve productivity and efficiency. In Supply Chain Management, cost optimization is being done through data analysis of freight, raw material prices, procurement schedules, and storage costs, etc. Big data with advanced analytical software is providing Marketing departments insight into product trends, demand, future customer requirements, and so on. Another important area in which Big Data usage is changing industry is in pricing. Pricing decisions depend on a multitude of variable factors, such as the cost of raw materials, exchange rates, utility costs, competitors’ prices, market demand, etc. Big Data Analytics has started playing an important role in these decisions, as well as in demand forecasting, which can have a major impact on the success of businesses.

Big data is helping innovation management by allowing the analysis of the vast data available on chemicals and their properties, including biological activities. In Human Resource Management, the analysis of data helps to identify parameters impacting job satisfaction, employee retention, and many other issues.

There are more than 15 software companies in India who have specialized in Big Data Analytics. In the Indian chemical industry, it is now felt that one must invest in Big Data and Analytics, which will bring immense benefit to the organization.

Dr. B. Saha <drbsaha@rediffmail.com> is Director, R&D Nagarjuna Agrichem Limited in Hyderabad, India.
In an era where research data is readily accessible during the peer review process, several obstacles persist in verifying the data required for publication. While Supporting Information (SI) for all American Chemical Society (ACS) journals is freely available, there is no standardized review process, and errors frequently linger, despite peer review. The SI for submitted manuscripts routinely contains missing or incorrect data, as well as inconsistencies between the submitted manuscript and the SI. While rare, issues of data manipulation, including digitally altered spectra, and plagiarism have been identified. [1] For the past five years at the ACS journal Organic Letters, the SI for all manuscripts has been reviewed by a data analyst to ensure compliance with author guidelines and to allow authors to address missing data and/or data inconsistencies prior to publication.

Organic Letters works with authors to provide education and resources to improve manuscript and SI preparation. In collaborating with authors, our office has developed SI checklists that can be used prior to submission to ensure complete and accurate SI. [2] Included in the checklists are the data requirements, necessary synthetic details, and guidance to properly format data. The SI review process is continually evolving as new standards and criteria are incorporated. For example, more rigorous safety standards are being enforced, as ACS publications will require authors to report “experimental details to address and emphasize any unexpected, new, and/or significant hazards or risks associated with the reported work.”[3] Furthermore, to increase reproducibility, the Organic Letters guidelines were updated in 2017 to require that one-step organic transformations include a detailed synthetic method at the 1 mmol scale for at least one representative example. These standards ensure the SI is complete, improve reproducibility, maintain a consistent presentation of data, and minimize the likelihood of incorrect and/or manipulated data.

While instances of manipulated data are rare, they are still found in 2-3% of all submitted manuscripts at Organic Letters, although the severity varies. If the data is found to be questionable or unusual during SI review, the Corresponding Author is contacted for an explanation and asked to provide original data for review. Often, authors are unable to provide the original data for their own submission. Explanations vary as to why the original data is not available and have included the following: improper storage of data, deleted data with no back-up, misplaced data, or inaccessible data (i.e., former students have the data and no contact information is available). In dealing with requesting and reviewing data for original submissions, there are several continuing issues: How long should a publisher wait to receive the data from an author? What happens if an author cannot provide the original data? Is newly generated data acceptable to replace questionable data? For an Addition/Correction to a publication, if new data is submitted, how will the data be verified?

One recommendation to enhance the quality of the manuscript and SI is to have additional guidance for peer reviewers. This may be beneficial to the review process and could lead to a higher quality publication. A thorough review requires careful examination and a systematic approach to determine if the experiments, data, and conclusions provided by the authors are worthy of publication. Since the SI is typically longer than the manuscript, it can be overwhelming for reviewers to determine how best to give useful feedback to an author. Ideally, a reviewer should follow these steps:

- Before starting an evaluation of the manuscript and SI, review the journal guidelines, focusing on the sections pertaining to the SI. Each journal has different criteria and it is helpful to understand what data is required before SI review begins.
- Determine that all necessary information is included in the SI and detail any data omissions, inconsistencies between the reported methods and the data presented, unusual data, or errors in data reporting in the review.
- Moreover, include any potential ethical concerns detected in the review, such as possible data manipulation or duplicate submission by the...
Supporting Information Review and Data Analysis at Organic Letters

Authors. This information is invaluable to the Editor in determining a manuscript’s outcome.

A systematic SI review prior to publication improves the quality of the published SI by having a consistent presentation of data that benefits both authors and readers. Organic Letters’ continually examines best practices for reviewing the SI, resolving issues encountered when requesting data from authors, and advising authors on submitting data for publication.

References:

Angela M. Hunter <angela.hunter@orglett.acs.org>, is Data analyst, Organic Letters in the Department of Chemistry at the University of Pennsylvania, Philadelphia, USA. ORCId.org/0000-0002-3474-525X

Update on ThermoML

by Kenneth Kroenlein

ThermoML is the XML-based IUPAC Standard for Thermodynamic Property Data, first developed as an IUPAC project in 2006, and extended in 2001. [1] At present, ThermoML is being used by a number of process design packages as a data input format, as well as more broadly on the Internet as a data dissemination format. The freely-accessible ThermoML archive maintained by the Thermodynamics Research Center (TRC) is the primary source of these data and has new address https://www.nist.gov/mml/acmd/trc/thermoml-archive. We fully expect this archive to continue to grow as new materials as published.

The most recent development impacting ThermoML is the expansion of TRC’s data collection activities into metal-based systems. Several elements of metals-based systems reporting requirements, around concepts such as phase description and differing expectations for sample characterization, made substantial update of the ThermoML schema necessary.

Accordingly, TRC is developing an updated schema, version 5.0, and associated forward and backward conversion software in the context of IUPAC. Previous recommendations related to ThermoML:


Kenneth Kroenlein <kenneth.kroenlein@nist.gov> is Group Leader, Thermodynamics Research Center, National Institute of Standards and Technology, in Broomfield, Colorado, USA.
Scientific publishing is changing; Open Access allows for published information to be freely shared, and Open Data repositories [1] are improving our ability to share and (re-)use data. These and initiatives such as RDA, FORCE11, and FAIR, [2-4] are changing the way we perceive the notion of publishing science. In the laboratory, equipment is becoming interconnected and new technologies are producing vast amounts of data; transformative change is underway. New publication practices are required that address the needs of the data producers and the data users. Many community-based grassroots projects and initiatives have already successfully set up infrastructures for storing and sharing data.

With this background of change and disruption, we must not lose sight of the fact that science can only thrive if the research produces high quality, reproducible results and data. ‘Publish or perish’ is the result of false incentivization. More and faster often translates into superficial and incremental. Thus, while new technologies should be embraced, they need to be implemented in ways that ensure good data reporting and allow validation, verification, and sharing.

The Beilstein-Institut publishes two platinum Open Access journals: the *Beilstein Journal of Organic Chemistry* and the *Beilstein Journal of Nanotechnology*. These two journals are unique in their fields, in that neither have APCs nor any other fees for authors or readers and both have been awarded the DOAJ Seal, [5] confirming an exceptionally high level of publishing standards and best practices. The Beilstein-Institut was one of the first publishers to check all incoming manuscripts for text similarity. In addition to the standard peer-review process, submitted manuscripts are checked by PhD scientists for consistency and plausibility. Through our internal editorial team, we see the many advantages that electronic publishing brings. However, there are disadvantages, such as the contribution of easy copy and paste to the prevalence of text plagiarism, and many potential advantages have yet to be realized, for example the routine reporting of validated experimental data.

*Figure 1. The workflow for STRENDA-DB. The linking of validated data input with peer-reviewed article publishing ensures that the data are not only complete but have also been reviewed by experts.*
With its longstanding history in high-quality chemistry data, the Beilstein-Institut has initiated and runs two data standards projects: STRENDA and MIRAGE, [6,7] guidelines for data reporting in the areas of enzyme chemistry and glycomics. Both projects aim at proposing reporting guidelines to allow the readers to analyse, validate, corroborate, and interpret the findings in the publications. The guidelines are developed by a panel of leading international scientists in each of the corresponding fields in a consensus-driven process and in close consultation with the wider community. The STRENDA reporting guidelines are recommended by the major journals in biochemistry and the MIRAGE guidelines are starting to be adopted by the major journals in glycobiology. To improve their practical use by both authors and journals, the STRENDA guidelines have recently been implemented as a web-based front-end for STRENDA-DB. [8] This supports authors by providing a data submission form that automatically checks the manuscript data for compliance with the STRENDA guidelines prior to or during the publication process. The successful formal assessment is documented in a fact sheet that can be submitted with the manuscript to the journal. In addition, each dataset is assigned a DOI to allow the easy tracking and referencing of data. The data become publicly available in the open access database only after the corresponding article has been peer-reviewed and published in a journal. The workflow for manuscript and data processing is shown in Figure 1.

STRENDA-DB demonstrates how a community-driven initiative can not only produce data reporting guidelines, but, when combined with innovative software, can turn those guidelines into a practicable data validation and reporting system for research scientists, putting the data at the center of the publication process.

References
1. www.biosharing.org
2. RDA: Research Data Sharing without barriers. https://rd-alliance.org
4. www.nature.com/articles/sdata201618
8. www.beilstein-strenda-db.org

Martin G. Hicks <mhicks@beilstein-institut.de> is at the Beilstein-Institut, Frankfurt, Germany. Carsten Kettner is at the Beilstein-Institut, Frankfurt, Germany.

Celebrating 17 Years of InChI

Status and Future of the IUPAC InChI: context and use cases

Building on the past and ongoing work of InChI working groups, a three-day meeting will be held next year on 16-18 August 2017 (Wed-Friday) on the main National Institutes of Health (NIH) campus in Bethesda, MD (short Metro/Subway ride from downtown Washington DC). The meeting will bring together the current InChI community and working groups that define the current state of the InChI project, together with other interested stakeholders. The aim is to discuss what is needed for the chemical, biomedical, materials, and related academic and industry communities for proper and useful structure standard representation of both small and large molecules, and the future direction and activities of InChI development will be a major goal of the meeting.

There is no registration fee to attend the meeting!
If you wish to attend, please sign up at www.inchi-trust.org

We welcome suggestions for the meeting agenda and also offers of assistance in planning the meeting—please contact Steve Heller, the project director, at steve@inchi-trust.org
For almost 100 years, IUPAC has been well known around the world for its efforts in standardizing nomenclature in chemistry. At the start of the present century, it became clear to all involved in chemical structure representation work that, with the extensive use of computers and electronic information in all aspects of chemistry and related sciences, an IUPAC standard was necessary. From this critical need, the IUPAC International Chemical Identifier—InChI—project was launched in cooperation with the US standards agency NIST. The result of this effort has been the development, maintenance, and expansion of capabilities of the open source nonproprietary International Chemical Identifier (InChI), first by NIST and now by the InChI Trust, a not-for-profit UK charity. Over 100 chemical information specialists and computational chemists volunteer to test the software before a public release; this optimal quality control by a world-wide user community has led to improvements to and releases of the software with very few problems. The reliance on input from many volunteers enables the project staff to be restricted to two part-time contractors, a project director and a programmer, thus minimizing the running costs of the Trust.

This brief discussion of InChI will highlight ongoing efforts to strengthen and extend this standard for chemical structures and its hashed form, the InChIKey. Information standards are critical to enable effective and efficient communication of scientific content. Validation and reproducibility of research results are critical to advances in science. Without a chemical structure standard, it was becoming impossible to find and share all the reported results needed for a particular purpose. The costs of experiments are ever increasing, hence the need for increased efficiency in labs around the world. Open Access, Open Data, and Open Standards are areas that are expanding rapidly and are facilitating faster and more effective research discovery. However, before you can share data about a chemical, you need to find where the information has been made available on the Internet. Collaborative, interoperable, and global dissemination standards are essential in a more networked world.

The InChI is an open-source, widely adopted standard found in most chemical information-containing databases, including those from Chemical Abstracts, Reaxys, ChemSpider, ChEMBL, Open PHACTS, PubChem, DrugBank, PDB, SigmaAldrich, and many others, including internal Chemical and Pharma corporate databases. The InChI distills diverse chemical representations into a single form, and is a string that enables easier linking and integration of scientific content, especially with printed and electronic data sources. For example, one can easily look up chemical structures in internet search engines such as Google, Bing, and Yahoo using an InChIKey. InChI is continually and actively being extended to increase its applicability and usability. The initial version, released in 2009, was able to handle almost 99% of the chemicals which scientists are concerned with every day. Additional work is underway to improve the treatment of inorganics and organometallics and to handle biopolymers and their positional isomers and chemical mixtures. The latest release, in 2017 (version 1.05), added polymer support and multithreading capabilities. InChI has also been incorporated in a chemical reaction identifier (RinChI) and its use in labeling via QR codes is being explored (see Chem Int Nov 2016, p. 22; https://doi.org/10.1515/ci-2016-0616). Funding to maintain and enhance InChI comes from most major chemistry publishers (CAS, Elsevier, Wiley, RSC, Springer-Nature, Taylor & Francis) and database and chemical suppliers and providers (Sigma-Aldrich, ChemAxon, BioRad, ACD/Labs, OpenEye, RELX group), as well as from US governmental agencies (NIH, FDA, and NIST). This funding helps to ensure that the future development of InChI meets the needs of the scientific community and it also helps to support other potential avenues for its use.

InChI is a valuable addition to other compound identifiers (e.g., systematic and trivial names, registry numbers, and various versions of SMILES) in a database; it is not intended to be a replacement. With the implementation of the ISO identification of medicinal products (IDMP) and the related ISO 11238 standards, adding and having an InChI will allow for an easier, effective, and more complete search for information on a particular chemical, be it a drug, a pollutant, or a chemical for other commercial and/or noncommercial use.

Details on the project can be found at https://www.iupac.org/inchi and http://www.inchitrust.org


Ray Boucher and Alan McNaught are Directors of the InChI Trust, Cambridge, UK, and Steve Heller <steve@inchi-trust.org> is its Project Director. Ray Boucher <ray@inchi-trust.org> is also member of IUPAC Polymer Division Subcommittee on Polymer Terminology. ORCID.org/0000-0002-4786-4223

by Ray Boucher, Stephen Heller, and Alan McNaught

The Status of the IUPAC InChI Chemical Structure Standard
Election of IUPAC Officers and Bureau Members

At the 49th IUPAC General Assembly in São Paulo, Brazil, on Wednesday 12 and Thursday 13 July 2017, the Council will be asked to elect a Vice President and members of the Bureau to fulfill the vacancies created by retiring members. IUPAC National Adhering Organizations have been invited to submit nominations. (https://iupac.org/election-iupac-officers-bureau-members-call-nominations/)

On 1 January 2018, Qi-Feng Zhou (China), Vice President and President-Elect of IUPAC, will become President. Natalia Tarasova (Russia), current President, will become Past President and remain an officer and a member of the Bureau for a period of two years. Meanwhile, Marc Cesa (USA), current Past President, will retire. Secretary General Richard Hartshorn (New Zealand) and Treasurer Colin Humphris (UK) were both elected by the Council in August 2015 for a four-year term and will continue their service for two more years.

Vice-President
The Vice-President to be elected at the 49th Council will be President-Elect and will become President on 1 January 2020.

The nominations received for Vice President are as follows:

- Christopher M.A. Brett (Portugal)
- Javier García-Martínez (Spain)

Elected Members of Bureau
Elected Members serve four-year terms, and are eligible for re-election to a second four-year term. No National Adhering Organization shall have more than one Elected Member on the Bureau, and the principle of fair geographical representation of Members shall be taken into account, as stipulated in the IUPAC Statutes.

The following are Members whose terms continue to the end of 2019:

- Mei-Hung Chiu (China/Taipei) (2016-2019)
- Ehud Keinan (Israel) (2016-2019)
- Kew-Ho Lee (Korea) (2016-2019)
- Pietro Tundo (Italy) (2016-2019)

The nominations received for Elected Members of the Bureau are as follows:

- Russell J. Boyd (Canada) (2014-2017); reappoint
- Tavarekere K. Chandrashekar (India) (2014-2017); reappoint
- Javier García-Martínez (Spain)
- Mary Garson (Australia)
- Christopher K. Ober (USA) (2014-2017); reappoint
- Ken Sakai (Japan)

At the conclusion of the 48th Council at Busan in August 2015, there were ten Elected Members on the Bureau. At the 49th Council in São Paulo, the Bureau will make recommendations to the Council as to the number of Elected Members (no less than ten) who

What Does the Bureau Do?

The Bureau is established by the Council to act for the Union during intervals between meetings of the Council; it therefore fulfills important functions by ensuring continuity. The principal duties of the Bureau—as quoted in the statutes—are as follows:

- to ensure the strict observance of statutes and bylaws
- to prepare the agenda for meetings of the Council and in particular to make provision for elections
- to make recommendations thereon to the Council
- to attend the meetings of the Council
- to implement the decisions of the Council and execute the program of the Union as directed by the Council
- to take steps to ensure that international congresses of pure and applied chemistry are held
- to take decisions about the holding of scientific meetings as proposed by the division committees
- to take all other steps necessary for the good conduct of the affairs of the Union

See Statutes S7 and Bylaws for more details.
should be on the Bureau for the succeeding two years. At least four Elected Members will be elected at the Council in São Paulo, i.e., the minimum number of ten Elected Members, less the six Elected Members who continue in office until 2019.

In addition to the five officers and the ten Elected Members, the Bureau also includes the eight Division Presidents (each elected by their individual Division), and the chairs of the following Standing Committees: the Committee on Chemistry Education (CCE); the Committee on Chemistry and Industry (COCI); CHEMRAWN (the Committee on CHEMical Research Applied to World Needs); the Interdivisional Committee on Terminology, Nomenclature and Symbols (ICTNS); and the Committee on Publications and Cheminformatics Data Standards (CPCDS).

https://iupac.org/2017-election-of-iupac-officers

IUPAC Announces the Winners of the 2017 IUPAC-Solvay International Award for Young Chemists

The International Union of Pure and Applied Chemistry and Solvay announce the winners of the 2017 IUPAC-Solvay International Award for Young Chemists, presented for the best Ph.D. theses in the chemical sciences, as described in 1000-word essays.

The five winners are:

- **Christopher Michael Lemon, Ph. D.**, Harvard University, USA
- **Fergus Eoin Poynton, Ph. D.**, Trinity College Dublin, Ireland
- **Leonardo Scarabelli, Ph. D.**, Universidade de Vigo, Spain
- **Chenjie Zeng, Ph. D.**, Carnegie Mellon University, USA
- **Nan Zhang, Ph. D.**, Fuzhou University, China

The winners, who are from five different countries, will each receive a cash prize of USD 1000 and travel expenses to the 46th IUPAC World Chemistry Congress, 9-14 July 2017, in São Paulo, Brazil. Each winner will also be invited to present a poster at the IUPAC Congress describing his/her award-winning work and to submit a short critical review on aspects of his/her research topic, to be published in Pure and Applied Chemistry. The awards will be presented to the winners of the 2016 (see CI Sep 2016, p. 19) and 2017 competitions during the Opening Ceremony of the Congress.

The titles of the winners’ theses are:

- Dr. Poynton: “Spectroscopic Investigations into the Excited-State Processes and Reactivity of Ruthenium(II) Polypyridyl Complexes”
- Dr. Scarabelli: “Synthesis and Self-Assembly of Anisotropic Plasmonic Nanoparticles”
- Dr. Zeng: “Precision at the Nanoscale: on the Structure and Property Evolution of Gold Nanoclusters”
- Dr. Zhang: “Design and Synthesis of Composite Photocatalysts and Their Structure-Activity Relationship Study”

There were 34 applications from individuals receiving their Ph.D. degrees from institutions in 19 countries. The award selection committee, chaired by Dr. Mark C. Cesa, IUPAC Past President, comprised members of the IUPAC Bureau and a senior science advisor from Solvay, all of whom have a wide range of experience in chemistry.

In view of the many high-quality applications, the Committee also decided to award an Honorable Mention to:

- **Jian He**, The Scripps Research Institute, USA
- **Liana Hie**, University of California, Los Angeles, USA
- **Junzhi Liu**, Max Planck Institute for Polymer Research, Germany
- **Pit Losch**, University of Strasbourg, France
- **Charalambos Pappas**, University of Strathclyde, United Kingdom
- **Derrick Andrew Roberts**, The University of Cambridge, United Kingdom
- **Stafford Wheeler Sheehan**, Yale University, United States
- **Feng Yang**, Peking University, China

Mark Your Calendar

Upcoming IUPAC-endorsed events
See also www.iupac.org/events for links to specific event websites

2017 (after 1 August)

13-17 August 2017 • 200 Years of Selenium Research • Stockholm, Sweden
The 11th International Symposium on Selenium in Biology and Medicine and The 5th International Conference on Selenium in the Environment and Human Health (Se2017)
Prof. Elias Arnér, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, SE-171 77 Stockholm, Sweden, E-mail: Elias.Arner@ki.se, www.se2017.se

16-18 August 2017 • Chemical Identifier • Bethesda, MD, USA
The IUPAC International Chemical Identifier, InChI, 10th anniversary workshop
Steve Heller, workshop coordinator, Division VIII InChI Subcommittee Chair and the InChI Trust project director, E-mail: steve@inchi-trust.org, www.inchi-trust.org

28-31 August 2017 • Macromolecular Complexes • Tokyo, Japan
17th IUPAC International Symposium on Macromolecular Complexes (MMC-17)
Hiroyuki Nishide, Chair of Program Committee, Department of Applied Chemistry, Waseda University, Tokyo 169-8555, Japan, E-mail: nishide@waseda.jp; Kenichi Oyaizu, Chair of Local Organizing Committee, Waseda University, E-mail: oyaizu@waseda.jp, www.waseda.jp/assoc-mmc17

10-14 September 2017 • Organic Materials for Electronics and Photonics • Prague, Czech Republic
81st Prague Meeting on Macromolecules (PMM): Polymers and Organic Materials for Electronics and Photonics: Science for Applications
Prof. Jiří Pfeifer, Pogram chair, E-mail: pfleger@imc.cas.cz ; and Daniella Ilinerová, Institute of Macromolecular Chemistry, Czech Academy of Sciences, E-mail: sympo@imc.cas.cz; www.imc.cas.cz/sympo/81pmm

17-20 September 2017 • BloodSurf • Clemson, SC, USA
Blood-biomaterial interface: where medicine and biology meet physical sciences and engineering
Ilya Reviakine (U Washington/Seattle, WA), E-mail: reviakin@uw.edu and Robert Latour (Clemson University), E-mail: latourr@clemson.edu, co-organizers; www.ireviakine.net/Bloodsurf

17-22 September 2017 • Ionic Polymerization • Durham, United Kingdom
International Symposium on Ionic Polymerization – IP 2017
Professor Lian Hutchings, Chair of Local Organizing Committee, E-mail: l.r.hutchings@durham.ac.uk; Dr Mike Shaver, Chair of Program Committee, E-mail: michael.shaver@ed.ac.uk; www.dur.ac.uk/soft.matter/ip2017/

27-29 September 2017 • Bioorganic Chemistry • Konstanz, Germany
11th International Symposium on Bioorganic Chemistry (ISBOC-11)
Program Committee Chair: Andreas Marx, University of Konstanz, Dept. of Chemistry, Universitäetsstr. 10, Postf. 726, D-78457 Konstanz, Germany, T:+49 7531 88-5290, andreas.marx@uni.kn, www.uni-konstanz.de/isboc-11/

2-5 October 2017 • Green Chemistry • Moscow, Russian Federation
7th IUPAC International Conference on Green Chemistry
Prof. Natalia P. Tarasova, Conference Chair, D. Mendeleev University of Chemical Technology, Moscow. Dr. Anna S. Makarova, Chair of Local Organizing Committee, E-mail: annmakarova@mail.ru; http://greeniupac2017.muctr.ru

8-11 October 2017 • Chemistry Education • Sétif, Algeria
ACRICE 2017, 3rd African Conference on Research in Chemistry Education
Prof Djafar Benachour, Department of Process Engineering, Ferhat Abbas University SETIF 1, Sétif 19 000, Alegria; E-mail: bendjafer@univ-setif.dz
www.univ-setif.dz/OC5/FT/ACRICE

9-13 October 2017 • Advanced Materials • Kuala Lumpur, Malaysia
Ong Eng Long, Organizing Chair, E-mail: ongelong@gmail.com; ikmhq@ikm.org.my. 25th POLYCHAR 2017 Secretariat, Institut Kimia Malaysia, 127B, Jalan Aminuddin Baki, Taman Tun Dr Ismail, 60000 Kuala Lumpur, Malaysia, Tel.: +603 77283272, E-mail: secretariat@25polychar.org.my, www.25polychar.org.my
11-13 October 2017 • Smart Materials • Jeju Island, Korea
IUPAC-FAPS 2017 Polymer Congress on Smart Materials for Emerging Technology
Jungahn Kim, Chair of the Organizing Committee, Department of Chemistry, Kyung Hee University, Seoul, Korea, E-mail: jakim05@khu.ac.kr; www.faps2017.org

5-9 November 2017 • HPLC 2017 • Jeju Island, Korea
46th International Symposium on High Performance Liquid Phase Separations and Related Techniques
HPLC 2017 Secretariat contact: Haengdo Lee, Department of Chemistry, Seoul National University, Seoul 151-747, Korea, E-mail: hplc2017@gmail.com; www.hplc2017-jeju.org

2018

21-23 February 2018 • Chemistry Conference for Young Scientists • Blankenberge, Belgium
Chemistry Conference for Young Scientists (ChemCYS 2018)
Koninklijke Vlaamse Chemische Vereniging vzw; E-mail: support@chemcys.be
www.chemcys.be

4-7 June 2018 • Polymers and Organic Chemistry • Montpellier, France
Polymers and Organic Chemistry 2018 (POC 2018)
Dr Ghislain David (Chair), Institute Charles Gerhardt, School of Chemistry of Montpellier, 8 rue de l’Ecole Normale, F-34296 Montpellier Cedex 5, France, E-mail: ghislain.david@enscm.fr

4-7 June 2018 • Isotopes and Isotopically Labelled Compound • Prague, Czech Republic
13th International Symposium on the Synthesis and Applications of Isotopes and Isotopically Labelled Compounds
Prof. Tomáš Elbert (Chair of the Local Organising Committee), E-mail: elbert@uochb.cas.cz
www.iis-prague2018.cz

1-5 July 2018 • MACRO2018 • Cairns, Australia
World Polymer Congress
Prof. Sébastien Perrier and Prof. Martina Stenzel (conference co-chairs); Conference Coordinator: Taylor Mills, Leishman Associates, E-mail: taylor@leishman-associates.com.au
www.macrol18.org

8-13 July 2018 • Photochemistry • Dublin, Ireland
27th IUPAC International Symposium on Photochemistry
Dr. Miguel A. Garcia-Garibay (Conference co-chair), Department of Chemistry and Biochemistry, University of California, Los Angeles, E-mail: mgg@chem.ucla.edu, and Dr. Susan Quinn, School of Chemistry, University College Dublin, Belfield, Dublin 4, Ireland, E-mail: susan.quinn@ucd.ie

10-14 July 2018 • Chemistry Education • Sydney, Australia
International Conference on Chemistry Education (ICCE) 2018
Chair of the Program Committee: Prof Siegbert Schmid, University of Sydney, School of Chemistry, Siegbert.schmid@sydney.edu.au

16-21 September 2018 • Organic Synthesis • Florence, Italy
22nd International Conference on Organic Synthesis (22-ICOS)
Professor Alberto Brandi (Conference Chair) and Professor Maurizio Taddei (Vice-Chair), E-mail: secretariat@22-icos-florence.it
www.22-icos-florence.it
Big Chemistry

What is Big Data? It is symptomatic of the question that the first place where I looked for an answer was Wikipedia, which defines it as “data sets that are so large or complex that traditional data processing software is inadequate to deal with them.” [1] Handling massive amounts of information poses a number of technological challenges, including the search, analysis, transfer, and visualization of data sets whose size is simply hard to comprehend. For example, the Chemical Abstracts Service (CAS) SciFinder database, which contains detailed records for more than 127 million organic and inorganic substances, is only a few terabytes in size (tera = $10^{12}$). Compare that with the millions of transactions processed each day by retailers such as Amazon or Walmart, which are imported into databases containing several petabytes of data (peta = $10^{15}$).

It is perhaps not surprising that the appraisal of such data sets is increasingly being used as a predictive tool to monitor consumer trends, optimize production and distribution costs, protect computer systems (“cybersecurity”), and improve many other facets of business and manufacturing.

Notes

Written by Daniel Rabinovich <drabinov@uncc.edu>.
The International Union of Pure and Applied Chemistry is the global organization that provides objective scientific expertise and develops the essential tools for the application and communication of chemical knowledge for the benefit of humankind and the world. IUPAC accomplishes its mission by fostering sustainable development, providing a common language for chemistry, and advocating the free exchange of scientific information. In fulfilling this mission, IUPAC effectively contributes to the worldwide understanding and application of the chemical sciences, to the betterment of humankind.

President: Natalia Tarasova (Russia)
Vice President: Qi-Feng Zhou (China)
Past President: Mark C. Cesa (USA)
Secretary General: Richard M. Hartshorn (New Zealand)
Treasurer: Colin Humphris (UK)

NATIONAL ADHERING ORGANIZATIONS

Asociación Química Argentina (Argentina)
Australian Academy of Science (Australia)
Österreichische Akademie der Wissenschaften (Austria)
Bangladesh Chemical Society (Bangladesh)
The Royal Academies for the Sciences and Arts of Belgium (Belgium)
Brazilian Chemical Society (Brazil)
Bulgarian Academy of Sciences (Bulgaria)
National Research Council of Canada (Canada)
Sociedad Chilena de Química (Chile)
Chinese Chemical Society (China)
Chemical Society located in Taipei (China)
 Croatian Chemical Society (Croatia)
Sociedad Cubana de Química (Cuba)
Czech National Committee for Chemistry (Czech Republic)
Det Kongelige Danske Videnskabernes Selskab (Denmark)
National Committee for IUPAC (Egypt)
Finnish Chemical Society (Finland)
Comité National Français de la Chimie (France)
Deutscher Zentralausschuss für Chemie (Germany)
Association of Greek Chemists (Greece)
Hungarian Academy of Sciences (Hungary)
Indian National Science Academy (India)
Royal Irish Academy (Ireland)
Israel Academy of Sciences and Humanities (Israel)
Consiglio Nazionale delle Ricerche (Italy)
Caribbean Academy of Sciences—Jamaica Chapter (Jamaica)

Science Council of Japan (Japan)
Jordanian Chemical Society (Jordan)
Kazakh National Academy of Science (Kazakhstan)
Korean Chemical Society (Korea)
Kuwait Chemical Society (Kuwait)
Institut Kimia Malaysia (Malaysia)
Academy of Sciences of Mozambique (Mozambique)
Nepal Polymer Institute (Nepal)
Koninklijke Nederlandse Chemische Vereniging (Netherlands)
Royal Society of New Zealand (New Zealand)
Chemical Society of Nigeria (Nigeria)
Norsk Kjemisk Selskap (Norway)
Chemical Society of Pakistan (Pakistan)
Polska Akademia Nauk (Poland)
Sociedade Portuguesa de Química (Portugal)
Colegio de Químicos de Puerto Rico (Puerto Rico)
Russian Academy of Sciences (Russia)
Comité Sénégalais pour la Chimie (Sénégal)
Serbian Chemical Society (Serbia)
Slovak Chemical Society (Slovakia)
Slovenian Chemical Society (Slovenia)
National Research Foundation (South Africa)
Spanish IUPAC Committee (Spain)
Institute of Chemistry, Ceylon (Sri Lanka)
Svenska Nationalkommittén för Kemi (Sweden)
Swiss Academy of Sciences (Switzerland)
Ministry of Science and Technology (Thailand)
Türkiye Kimya Derneği (Turkey)
Royal Society of Chemistry (United Kingdom)
National Academy of Sciences (USA)
Programa de Desarrollo de Ciencias Básicas (Uruguay)
50th General Assembly
&
47th IUPAC World Chemistry Congress
« Frontiers in Chemistry:
Let’s create our Future!
100 years with IUPAC »

JULY 5-12 2019

IUPAC will celebrate its Centenary holding its General Assembly and World Congress in Paris, France, along with dedicated sessions and events.