

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.

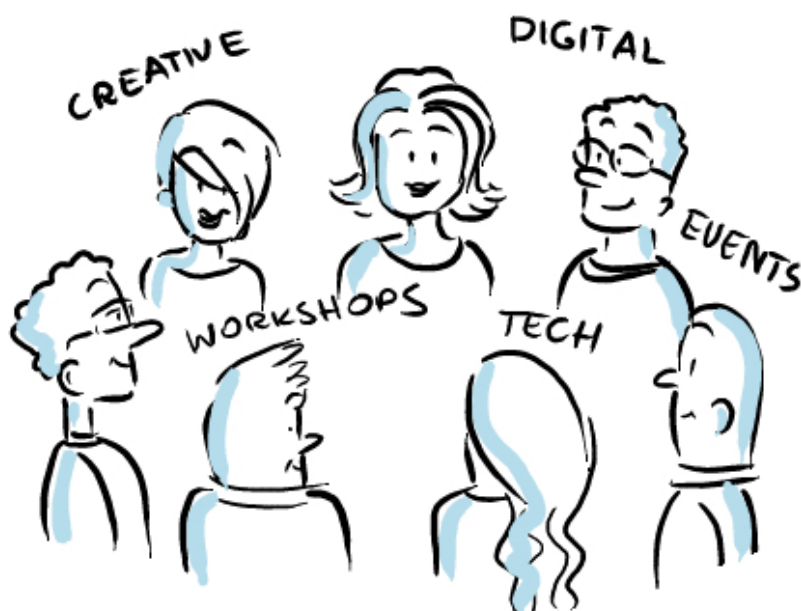
Global Data Standards

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

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.

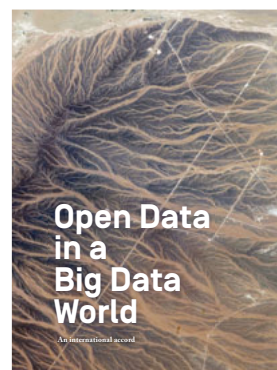
Discussion on chemistry terminologies were continued at a CRDIG session at the 8th RDA Plenary meeting held during International Data Week 2016 in Denver [22] and at a VoCamp event in Washington DC. [23] Both events provided an opportunity to seek input from terminology experts in the wider data science 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


Open Data is a Big Data World, but without suitable standards and arrangements to ensure interoperability, that data will not be useful.
www.science-international.org/#accord



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! 

References

1. H.W.J. Rittel, M.M. Webber, Dilemmas in a general theory of planning, *Policy Sci.* **4**(2):155-169, 1973.
2. European Commission, *Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data. A submission to the European Commission.* www.fosteropenscience.eu/sites/default/files/pdf/831.pdf (Accessed 3 Mar 2017).
3. Science International, *Open Data in a Big Data World.* www.icsu.org/science-international/accord/open-data-in-a-big-data-world-long (Accessed 3 Mar 2017).
4. The Royal Society, *Science as an open enterprise.* <https://royalsociety.org/-/media/policy/projects/sape/2012-06-20-saoe.pdf> (Accessed 3 Mar 2017).
5. National Institute of Standards and Technology. www.nist.gov (Accessed 3 Mar 2017).
6. CODATA: ICSU Committee on Data for Science and Technology. www.codata.org (Accessed 3 Mar 2017).
7. International Council for Science—ICSU. www.icsu.org (Accessed 3 Mar 2017).
8. World Data System: Trusted Data Services for Global Science. www.icsu-wds.org (Accessed 3 Mar 2017).
9. S.R. Hall, B. McMahon, The implementation and evolution of STAR/CIF ontologies: Interoperability and preservation of structured data, *Data Sci. J.*, **15**:1-15, 2016. <http://doi.org/10.5334/dsj-2016-003>
10. IUPAC Committee on Publications and Cheminformatics Data Standards. <https://iupac.org/body/024> (Accessed 5 Mar 2017).
11. IUPAC Subcommittee on Cheminformatics Data Standards. <https://iupac.org/body/036> (Accessed 5 Mar 2017).
12. The Research Data Alliance. www.rd-alliance.org (Accessed 17 Feb 2017).
13. The Research Data Alliance. *RDA in a Nutshell*, February 2017. www.rd-alliance.org/sites/default/files/attachment/RDA_in_a_nutshell_February_2017_updated.pptx (Accessed 17 Feb 2017).
14. The Research Data Alliance, *RDA Recommendations and Outputs.* www.rd-alliance.org/recommendations-and-outputs/all-recommendations-and-outputs (Accessed 17 Feb 2017).
15. C.C. Austin, T. Bloom, S. Dallmeier-Tiessen, V.K. Khodiyar, F. Murphy, A. Nurnberger, L. Raymond, M. Stockhause, J. Tedds, M. Vardigan, A. Whyte, Key components of data publishing: using current best practices to develop a reference model for data publishing, *Int. J. Digit. Libr.* 1-16, 2016. <https://dx.doi.org/10.1007/s00799-016-0178-2>
16. SCHOLIX. www.scholix.org (Accessed 21 Feb 2017).
17. Data Literature Interlinking Service. <https://dliservice.research-infrastructures.eu> (Accessed 21 Feb 2017).
18. RDA/WDS Scholarly Link Exchange (Scholix) Working Group. www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg (Accessed 21 Feb 2017).
19. D.Y. Esther, et al., *Wheat Data Interoperability Guidelines.* www.rd-alliance.org/system/files/Wheat%20Data%20Interoperability%20Guidelines_0.pdf
20. RDA Chemistry Research Data IG. www.rd-alliance.org/groups/chemistry-research-data-interest-group.html (Accessed 17 Feb 2017).
21. The Research Data Alliance. *We DIG Chemistry!* www.rd-alliance.org/we-dig-chemistry (Accessed 17 Feb 2017).
22. The IUPAC Color Books: Translation of Chemistry Terminologies to Digital Vocabularies—Working Meeting. www.rd-alliance.org/ig-chemistry-research-data-working-rda-8th-plenary-meeting (Accessed 17 Feb 2017).
23. GeoVoCampDC2016. <http://vocamp.org/wiki/GeoVoCampDC2016> (Accessed 17 Feb 2017).
24. Backup, Maintenance, and Redevelopment of the IUPAC Gold Book Website. <https://iupac.org/project/2016-046-1-024> (Accessed 17 Feb 2017).
25. S. Heller, A. McNaught, S. Stein, D. Tchekhovskoi, I. Pletnev, InChI - the worldwide chemical structure identifier standard, *J. Cheminform.* **5**:7 (2013).
26. Status and Future of the IUPAC InChI: Context and Use Cases. www.inchi-trust.org/status-future-iupac-inchi-context-use-cases/ (Accessed 8 May 2017). See details on page 46.

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