IUPAC Facilitating Chemistry Data Exchange in the Digital Era

by Leah Rae McEwen

The scientific community faces an unprecedented communication challenge as larger volumes of research data are published and data storage moves into digital space. Most communication tasks in chemistry involve representing molecules. Traditionally, these have formalized around the need to register, search, view, and publish information about chemicals for human readers. Data collection and analysis are further described by formalized domain vocabularies. In the digital environment, machine-readability, as well as human interpretation of data, is a significant factor in the level of accuracy and completeness of data exchange. More than ever, there is a need for standard and robust protocols that support reliable interoperability: the transfer of depictions and descriptions of chemicals between systems without loss or distortion of information.

IUPAC has long recognized that consistent use of chemical representation and terminology is critical for documenting and reporting chemical data. The former Committee on Printed and Electronic Publications (CPEP) recently re-envisioned part of its remit, becoming the Committee on Publications and Cheminformatics Data Standards (CPCDS). [1] With this shift, the committee aims to leverage IUPAC's deep institutional expertise and authority in chemical nomenclature and representation to support expanding digital applications of chemical data. CPCDS is liaising with the Research Data Alliance (RDA), [2] as well as CODATA (International Council for Science: Committee on Data for Science and Technology), [3] to complement chemical information expertise with international forums for digital data exchange strategies.

In July 2016, IUPAC CPCDS co-sponsored a workshop with the RDA Chemistry Research Data Interest Group (CRDIG), [4] engaging researchers, cheminformatics specialists, educators, publishers, and librarians to identify key opportunities for developing unified approaches to communicating chemical information in the digital environment. Hosted by the United States Environmental Protection Agency (EPA) National Center for Computational Toxicology (NCCT) [5] in Research Triangle Park, NC, the workshop explored the potential of IUPAC scientific definitions to function digitally as machine-readable standards for robust automated processes. Focusing on IUPAC's strengths in developing standards for describing molecules, measurements, and properties, this joint effort aims to support cheminformatics by expanding the reach and impact of IUPAC standards globally.

Interoperability among many sophisticated data software programs in chemistry is key to the accurate publication, re-use, and exchange of data throughout the research cycle. Exchanging information about digital data involves communicating, not just with other chemists, but with computer systems, websites (via APIs), and databases. Computers have different requirements than professional chemists for interpreting meaning; everything must be explicitly captured in the form of rule sets managed by software algorithms. Once codified, these rules can govern the management of astounding numbers of items that meet the criteria, such as tens of millions chemical compounds based on specified parameters of atoms and bonds. The IUPAC InChI (International Chemical Identifier) [6] algorithm is an example of an interoperable and machine-readable rule-set for identifying molecular structure.

For a long time, well-developed practices for the careful documentation of experiments and rigorous systematic nomenclature have been central features of chemistry. [7] Despite these long-established mechanisms for assuring clear communication among (human) chemists, managing the reliable translation of concepts from humans to machines and among computer systems has proved as elusive as consistent depiction of organometallics. Compiling data from multiple published sources into large collections has uncovered hundreds of variations in the representation of enantiomers, ions, salt forms, *etc.* Generating consistent representations of molecular structures for the same compound regardless of the software used is one of the major challenges in chemical information. With emerging machine-readable standards for structure notation and nomenclature, accurate management and manipulation of large inventories of chemical data is greatly improving. However, there is much inconsistency in defining the input to, and the applications of, these computable standards.

The goal of the joint IUPAC-RDA effort is to facilitate interoperability and the exchange of information at the machine level, as well as among researchers and consumers of data. The workshop participants at EPA identified several key opportunities associated with machine-readable chemical structure representation. These include: 1) augmenting IUPAC graphical representation guidelines to support reliable interpretation by machines, 2) reviewing criteria for file formats conveying information about chemical structures, and 3) standardizing protocols for translation and normalization of variations in chemical depiction. Underpinning these opportunities is a need for engagement with stakeholders in the chemistry community and in the data publishing and reporting industry. The workshop also considered digital applications of standard terminologies, such as those published in the IUPAC Color Books. [8]

Graphical Representation Recommendations were devised and published in *Pure and Applied Chemistry* (*PAC*) in 2006 [9] and 2008 [10] for "the display of two-dimensional chemical structure diagrams." These guidelines are intended to establish conventions to prevent ambiguity in communicating chemical structure information in publications. There are a few alerts to drawing software issues for end users. However, best practices for software encoding are not addressed. Expanding the guidelines to consider machine interpretation of chemical depictions can prevent the corruption of chemists' intentions when converting to computer-defined chemical structures. Harmonizing the guidelines with other structure standardization and nomenclature considerations would improve consistency and reduce translation error in drawing software and databases.

While there are recognized systems for registering characterized chemical substances, there are no officially sanctioned file formats for transferring molecular structure information. A series of file formats originally published by Molecular Design Limited in 1992 has become a *de facto* community standard for representing single molecules (MOLfile), multiple molecules and data (SDfile), and reactions (RXNfile and RDfile). [11] The documentation for these files types is available upon request, and a recent format upgrade greatly expands the potential to encode more complex chemical scenarios. [12] However, these are propriety formats and there are no machine-readable templates or community agreements in place for interpreting these options in practice. Files are re-used, re-parsed, and re-constituted using different conventions for encoding and decoding myriad chemical features of interest beyond basic connectivity. File formats and software for interpreting them are full of idiosyncratic pre-formulated short-cuts devised for different applications that not only vary in their operations but, unknown to downstream users, also obfuscate basic atom-level interpretation.

Machine-readable, parsable structure notations are plagued by the same challenges. SMILES (Simplified Molecular-Input Line-Entry System) and SMARTS (*i.e.*, substructure pattern extensions) are popular chemical notations particularly useful for depicting functional patterns. SMILES rules were published in 1988 in basic form [13] and the copyrighted theory manual is still available for personal use. [14] There has been some community effort to coordinate the development of an open specification. [15] However, numerous diverse extensions and translating schemes persist and there is no formal process or review. The InChI serves a different role as a more formalized structure-based identifier that effectively supports automated structure validation and linking. The standard form of InChI remains limited to small covalent organic molecules, with recent extensions for polymers and reactions and several current projects for organometallics, large molecules, and mixtures. [16] All of these rely on being able to unambiguously represent the chemical structures and systems for which an InChI or other machine-readable notation is desired. The interoperability of chemical data could be improved by standardizing a small number of open chemical file formats.

Normalizing chemical structure is a necessary part of the process to transfer data between systems. Currently, many approaches to normalization exist, usually employed by and for the local needs of the importing system. With an increasing complexity of structures and variability in representation across tens of millions of substances, critical information, such as stereochemical configuration or electron delocalization, can get lost in translation. "Reconstituting" structures in human-readable form is subject to contextual human interpretations and preferences (think of the variety of ways in which chemists draw benzene). When such structures are exported once more, these local preferences can yield a different digital representation than the one that was originally imported. Similar issues may arise when rendering structures most appropriate

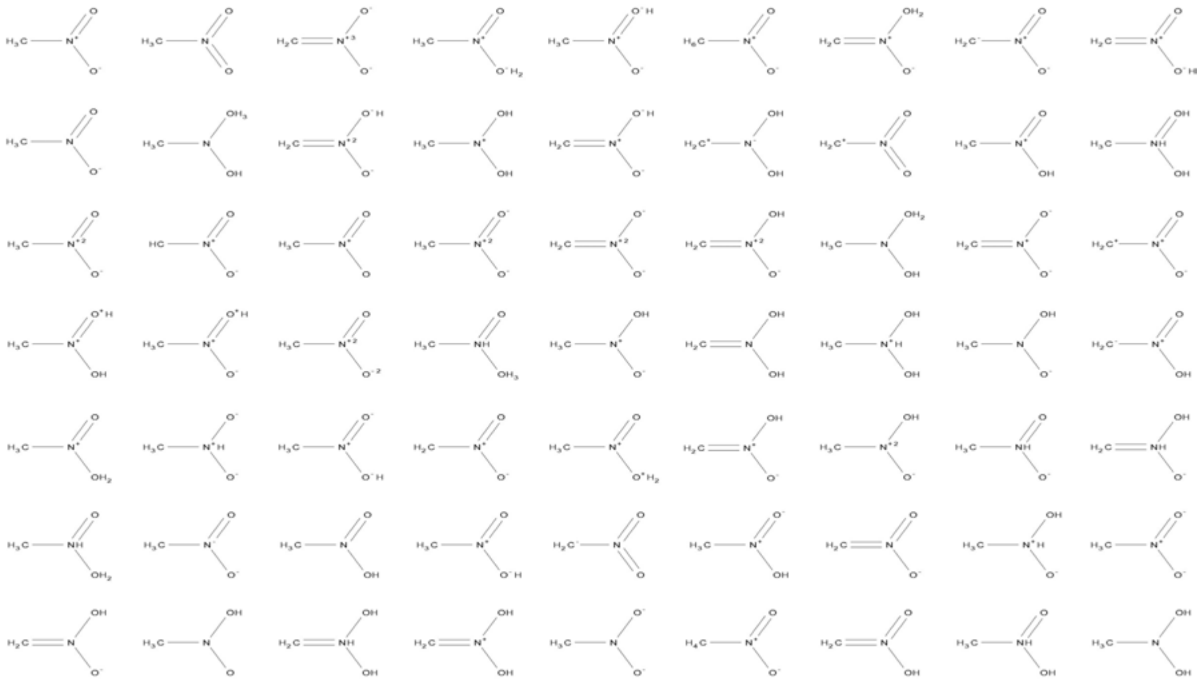

Figure 1. Variations on nitromethane (courtesy of PubChem) [17]

for local conditions, resulting in chemically related but not identical structures, such as tautomers. Iterative interpretations can introduce inaccuracies with implications downstream for the re-use of data by both chemists and computer applications (see Figure 1). Standard protocols for translation, coupled with guidelines for interpreting input and open file formats, could greatly facilitate the accurate exchange of chemical data.

Engaging researchers in the documentation of their research outputs is critical for ensuring integrity in increasingly automated information processes. Outreach efforts can help the chemical information community better understand why chemists draw molecules the ways that they do and where the crucial points exist in communicating chemistry among humans and machines. Targeted standardization of practice can benefit authors, readers, publishers, reviewers, and educators, as well as system and software developers. The research and publishing communities are embracing standard document identifiers such as DOI (Digital Object Identifier), [18] including schema for citing data, [19] and implementing author identifiers, such as ORCID (Open Researcher and Contributor ID). [20] Use of these identifiers is improving the accuracy of citations and cross-linking. [21] Workflows that support further coupling of chemical identifiers, author determined molecular structure, and characterization data files will enhance the chemical record and support research funding mandates to share supporting data. [22] Ideally, standards should operate invisibly, but training the next generation about the positive outcomes from incorporating updated standards in their workflows will result in an improvement in data quality over time.

Engaging researchers in the documentation of their research outputs is critical for ensuring integrity in increasingly automated information processes.

The IUPAC Color Books describe a diversity of measurement methods and properties, as well as molecular nomenclature. [23] Many of these terms, along with others from PAC, are collected in the Gold Book Compendium of Chemical Terminology to facilitate discovery across the corpus of concepts defined by IUPAC. Several IUPAC projects have focused on supporting the Gold Book in an online format, including planning in conjunction with the new IUPAC web presence. Each term in the most recent iteration has a permanent DOI for easier citation and dynamic linking back to IUPAC. [24] However the data are not systematically structured and it is difficult to automatically retrieve linked terms. In order to make the Gold Book reliable and sustainable in the longer term, chemical terms and the scientific relationships that connect them to each other should be drawn dynamically from IUPAC source publications and systematically structured to ensure consistent resolution to the authoritative IUPAC source. Discussions on this point at the EPA workshop fed into a special digital vocabularies session about the Gold

Book at the 8th RDA Plenary in Denver, CO in September 2016. [25]

Formulating the Gold Book for digital use involves the conversion of terms from the current web-display format into a more automated machine-interpretable form. Individual terms can be associated with a URL (or URI – Universal Resource Identifier) pointing to a human-readable form with the term definition and citation. This approach meets requirements for both human and machine access to the information, without compromising functionality. It facilitates both internal and external linking to terms, and additional functionality can be readily incorporated, such as links to provenance. Meaningful relationships among terms could be indicated (*e.g.* synonyms), and feedback and review commentary incorporated into the workflow. Further analysis of the definitions could identify potential overlap among terms, and comparison with the text of the Color Books and PAC could help identify gaps in Gold Book coverage. One use of such a machine-readable compendium of IUPAC terminology could be automated referral from other documentation that incorporates these chemical concepts, such as chemical patents, textbooks, experimental methods, and computational models.

There has been much talk of the potential of Big Data and the Internet of Things, linking measurement instruments to lab notebooks to publication templates. Much of chemical research lies on the long tail of small scale experiments. Although this research may enter the scientific record in countless individual publications, many experimental data files can be classified into a few types that are amenable to substance characterization and property determination. Aggregated across tens of millions of compounds and substances, this is a huge amount of data that should, in principle, be accessible for re-use. If its accessibility and reliability could be assured, this data would become a core resource for chemistry research as a global endeavor, contributing to crucial trans-national impacts in the areas of biomedicine and pharmacology, climate change, pollution, and public health. Opportunities to leverage data and digital technologies in facilitating chemical communication is of utmost interest to the community, and central to IUPAC's mission.

Clearly, there is compelling need for authoritative IUPAC chemical descriptions to be machine-accessible for scalable data processing. As data exchange via the Cloud reaches a global scale, IUPAC is well positioned as an international scientific union to bridge scientific meaning and automated processes with functional elements for data exchange, including open file formats and chemical descriptors. Coordination and collaboration with other international scientific data initiatives, such as CODATA and the RDA, provides exciting opportunities for expanding the impact of IUPAC in the global scientific community. Look for further discussion on this topic in a special issue of *Chemistry International* on Big Data in July 2017 and at a special symposium at the IUPAC World Congress in São Paulo, Brazil. [26]

References

1. www.iupac.org/body/024
2. www.rd-alliance.org
3. www.codata.org
4. www.rd-alliance.org/groups/chemistry-research-data-interest-group.html
5. www.epa.gov/aboutepa/about-national-center-computational-toxicology-ncct
6. https://iupac.org/who-we-are/divisions/division-details/inchi/
7. https://iupac.org/who-we-are/our-history
8. https://iupac.org/digital-data-challenges-in-chemistry
9. http://doi.org/10.1351/pac200678101897
10. http://doi.org/10.1351/pac200880020277
11. http://doi.org/10.1021/ci00007a012
12. http://accelrys.com/products/collaborative-science/biovia-draw/ctfile-no-fee.html
13. http://doi.org/10.1021/ci00057a005
14. www.daylight.com/dayhtml/doc/theory
15. http://opensmiles.org/opensmiles.html
16. www.inchi-trust.org
17. https://pubchem.ncbi.nlm.nih.gov/compoundnitromethane
18. www.doi.org
19. www.datacite.org
20. http://orcid.org
21. http://crossref.org
22. http://doi.org/10.1515/ci-2016-3-408
23. https://iupac.org/what-we-do/books/color-books/
24. http://goldbook.iupac.org
25. www.rd-alliance.org/ig-chemistry-research-data-working-rda-8th-plenary-meeting
26. www.iupac2017.org/special-symposia.php

Leah Rae McEwen <lrm1@cornell.edu> is chemistry librarian at Cornell University, USA. ORCID.org/0000-0003-2968-1674. 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.

iupac.org/body/036