Towards a Digital IUPAC

by Leah R. McEwen

Communicating chemical knowledge is at the core of the IUPAC mission and underlies the success of the chemistry enterprise. In the global economy of the 21st century, this involves exchange among computer systems along with the expert scientists who use them. To enable the appli-

cation of IUPAC outputs in the digital environment, IUPAC must augment its efforts to enable accessibility and interpretation by machines as well as humans. The Union must adapt to a digital work culture to engage in its mission of sustainable development, common language and free exchange of scientific information. [Frey 2014, <https://doi.org/10.1515/ci.2014.36.1.14>]

The Committee on Publications and Cheminformatics Data Standards (CPCDS) (<https://iupac.org/body/024>) is charged to develop standards that enable and “promote interoperable and consistent transmission, storage, and management of digital [chemical information] content.” Since 2016, the CPCDS Subcommittee on Cheminformatics Data Standards has been tasked to explore the needs of the chemical community with the objective of coordinating the collective expertise of relevant IUPAC Divisions and Committees and external global organizations. A special issue of *Chemistry International* on “Research Data, Big Data and Chemistry” was edited by the Subcommittee for the 49th General Assembly in São Paulo (<https://iupac.org/etoc-alert-chemistry-international-jul-sep-2017/>)

As demonstrated in related communities of practice such as crystallography, machine readable scientific definitions and standard data formats facilitate accurate reporting, further scientific analysis and processing of measurements. Collective sharing of data within a domain enables the generation of new insights that are applicable more broadly. The adoption of standard file formats and standard identifiers across the community and stakeholders greatly aids in workflows to accurately publish and share data in digital venues. [Bruno 2020, <https://charlestonlibraryconference.com/here-come-the-data/>]

Developing and disseminating digital representations of IUPAC intellectual assets is not simply a software problem. Criteria for machine readability needs to be robust, function consistently across many different computer systems, and be based on accepted Internet

protocols. The FAIR Data Principles describe high level criteria for enabling data and associated information to be Findable, Accessible, Interoperable and Re-usable for both humans and machines in a distributed digital environment. These principles provide a good starting point for understanding what is required to enable data to be effectively shared and allows IUPAC to tap into many motifs for digital exchange emerging in the data sciences and informatics expert communities. [Wilkinson *et al.*, <https://doi.org/10.1038/sdata.2016.18>]

The goal of the committee in the coming biennium will be to formulate machine-processable technical descriptions that build on the authoritative scientific definitions developed by the scientific Divisions of IUPAC. From an information perspective, IUPAC outputs may be classed into three pillars that support communication of chemical principles and knowledge: definitions of terms, names and symbols; critically evaluated standard data values; and specifications for chemical structures and other data representation. Describing the chemical world is too complex to accurately communicate through a single motif and there are different aspects of what to frame explicitly in machine depictions. The challenge of this work will be to break down this problem into discrete interoperable functionalities that are essential for accurate exchange of critical information and can enable broader utility collectively.

Towards a Digital IUPAC: CPCDS Chair's Statement

CPCDS has been launching projects in conjunction with a number of Divisions as well as active user groups with a goal to showcase the application of IUPAC assets to global problems in digital science.

Digital Terminology

One of the most significant undertakings for CPCDS in collaboration with the Divisions is development and stewardship of the digital form of the IUPAC Compendium of Chemical Terminology. Colloquially known as the Gold Book after its first editor, Victor Gold, the electronic edition (<https://goldbook.iupac.org>) is a visible face of the significant investment that members of the IUPAC Divisions have made over the years to formally define many important chemical terms. A recent project has stabilized the content and provided the groundwork for more active curation and use of the terms (<https://iupac.org/project/2016-046-1-024>). Term definitions may now be downloaded, accessed through an Application Programming Interface (API), and cited with automatic links through Digital Object Identifiers (DOIs).

The improved availability of Gold Book terms for computer applications has generated interest in extending machine representation of chemical concepts to enable new capabilities. Through a newly formed project (<https://iupac.org/project/2019-032-1-024>), IUPAC seeks to support the development of terminology, nomenclature, and symbols for chemistry commensurate with the digital environment. This necessitates a more efficient mechanism for managing terms that supports rigorous articulation and approval processes and ensures this rigor and provenance in the digital space. This project will provide a secure system and engage all the Divisions to develop a sustainable process for promulgating and reviewing terms.

Machine readable critically evaluated data

The Periodic Table is one of the most well known chemical information constructs, and a joint project of CPCDS with the IUPAC Commission on Isotopic Abundances and Atomic Weights (CIAAW) to develop a machine readable specification of this resource exemplifies the complementary role of the standing committee in expanding the utility of authoritative IUPAC output (<https://iupac.org/project/2019-020-2-024>).

The CIAAW has been streamlining the process for managing the speed and accuracy of the data evaluation and communication of updates to official IUPAC approved standard values through their website (<https://ciaaw.org/>). To ensure these data are accessed accurately by machine systems, the values, associated

uncertainties and other descriptive information must be consistently expressed in formats that can be parsed without human interpretation or intervention. CPCDS is working with CIAAW to augment curation practices for digital dissemination in adherence to the FAIR Data Principles that will facilitate more accurate computation, maintain links to provenance, and expose this content more broadly across disciplines.

FAIR description of measurement data

IUPAC has stewarded for many years a standard “Data Exchange” format for spectroscopic information originally developed by the Joint Committee on Atomic and Molecular Physical Data, known as JCAMP-DX [Grasselli 1991, <https://doi.org/10.1351/pac199163121781>]. A project building on IUPAC’s extensive expertise is being formalized to apply the FAIR Data Principles through description of digital data objects that will facilitate the processing of raw and derived spectroscopic data from instruments through publication and review to further study and analysis. In addition to specification of a standard format for the metadata, the project will seek to formulate validation criteria to enable systems to check files for machine readable and interoperable representation based on the standard.

Communicating information about chemical structures

The IUPAC International Chemical Identifier (InChI) is a chemical descriptor that notates structure information in a layered format and canonically identifies discrete compounds. It has become an essential standard for communicating chemistry in the Internet era. InChI facilitates the accurate matchup of chemical records for discrete compounds when linking and exchanging across computer systems [McEwen 2018, <https://doi.org/10.1515/ci-2018-0109>]. The InChI algorithm is jointly stewarded by Division VIII and the InChI

Towards a Digital IUPAC: CPCDS Chair's Statement

Trust, an independent nonprofit charity established to develop and promote use of the standard (<https://www.inchi-trust.org/>). Several CPCDS projects are incorporating InChI as a core feature of the metadata to facilitate interoperability.

The SMILES (Simplified molecular-input line-entry system) family of chemical representation notation is a common digital motif for automated retrieval of structural information that supports substructure searching, molecular patterns and reaction transforms. The continued ubiquitous use of dated SMILES documentation is limiting the accurate global exchange of chemical information and a project is underway to develop open reference documentation that articulates a standard interpretation of SMILES (<https://iupac.org/project/2019-002-2-024>). SMILES plays a complementary role to the standard InChI identifier in cheminformatics and formalizing the specification will enhance the accuracy of input used to generate canonical InChIs.

These projects represent some of the many opportunities in which CPCDS will engage over the course of the new biennium. CPCDS members are also participating in the Interdivisional Subcommittee on Critical Evaluation of Data (<https://iupac.org/body/505>) to harmonize formats for the archiving of both compiled and evaluated data. CPCDS is launching a task force to develop white papers that will focus on emerging technologies, new areas of science, and current issues in global chemistry—including current and proposed applications of Artificial Intelligence, Machine Learning, and Blockchain Technologies in the chemical sciences.

IUPAC is strategically placed to connect these collective efforts into open and FAIR data initiatives globally and across disciplines through participation on the International Science Council (ISC) Committee on Data (CODATA) (<http://www.codata.org/>). The Digital Revolution is one of the four domains of the ISC Action Plan to Advance Science as a Global Public Good (<https://council.science/actionplan/>). The Secretary General of IUPAC is a member of the CODATA Executive Committee and has highlighted on a number of occasions the strategic importance of Cheminformatics and of IUPAC work in developing the tools and standards that will be needed by chemists and all those who use chemical data in the world of big data. International IUPAC related collaborations were recently highlighted in a special issue of Data Intelligence on the FAIR Data Principles. 🏛️

References

1. Simon J. Coles, Jeremy G. Frey, Egon L. Willighagen and Stuart J. Chalk; Taking FAIR on the CHIN: The Chemistry Implementation Network; https://doi.org/10.1162/dint_a_00035
2. Shelley Stall, Leah McEwen, Lesley Wyborn, Nancy Hoebelheinrich and Ian Bruno; Growing the FAIR Community at the Intersection of the Geosciences and Pure and Applied Chemistry; https://doi.org/10.1162/dint_a_00036

Leah R. McEwen, of Cornell University, is the chair of the IUPAC Committee on Publications and Cheminformatics Data Standards, <http://orcid.org/0000-0003-2968-1674>

Save the Date

International FAIR Convergence Symposium & CODATA General Assembly
convened by CODATA and GO FAIR
Paris, 22-24 October 2020