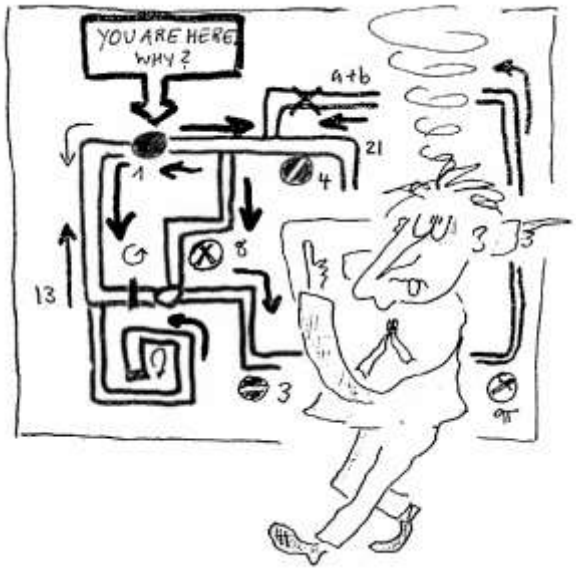
A Century of Nomenclature for Chemists and Machines

Evan Hepler-Smith
Boston College History Department
Leah McEwen
Cornell University

The chemist's tongue is sticking out and his eyes are upturned in concentration. Spirals circle above his head; the chemist is confused. His body twisted into a knot, he slowly backpedals, one finger pointing straight up, the other pointing backwards. To his side, a chart displays a maze, or perhaps a diagram of a complex logic circuit, illustrating an arcane web of decisions and procedures. A dot sits at rest within the diagram, marking the stymied chemist's progress through this maze. It is annotated with a comment: "You are here. Why?"



M. Volkan Kısakürek, "Chemistry Journals and Nomenclature, 1892-1930," in *Organic Chemistry: Its Language and Its State of the Art*, ed. M. Volkan Kısakürek (Basel: VHCA, 1993), 55-75, on 71. Courtesy Wiley / Verlag Helvetica Chimica Acta. [**PERMISSIONS PENDING**]

A good-humored manuscript author sent this sketch to M. Volkan Kisakürek, editor of *Helvetica Chimica Acta* and member of the IUPAC Commission on Nomenclature of Organic Chemistry from 1987–2001, as an illustration of the challenges of systematic chemical nomenclature. Following nomenclature rules, the cartoon suggests, was a task more fit for a machine than for the poor confused chemist.

From the founding of IUPAC a century ago, codifying rules of chemical nomenclature has been one of the Union's central functions. Since 2001, IUPAC has concurrently pursued the parallel project of establishing a method for generating unique, machine-readable chemical

identifiers, the notation now known as the IUPAC International Chemical Identifier (InChI). The impressive InChI project combines IUPAC’s deep institutional experience developing chemical nomenclature standards with the distinctive challenges of a digital, globally networked information landscape [1-3].

Though the InChI project is relatively new, the history of IUPAC efforts to standardize chemical representation in the digital world stretches back to the 1940s. Whereas the history of IUPAC nomenclature testifies to the role of the Union in maintaining the continuity and accessibility of chemistry’s massive treasury of past results, the history of IUPAC efforts in machine documentation shows how chemists have made the Union an instrument for grappling with technological change. In contrast to the broad uptake of InChI, IUPAC’s past machine documentation efforts had relatively little visible impact on the practice of chemistry. Yet they nevertheless provide valuable clues as to the enabling role of IUPAC-based work in the development of global chemistry over the past century. Studying the history of how IUPAC addressed chemical structure representation as a job fit for machines can shed light on the stewardship opportunities and challenges for IUPAC’s next 100 years.

* * *

We tend to think about the impact of standards-setting organizations in terms of the enduring standards that they manage to set. This is a particularly sensible starting point for summing up the work of the IUPAC nomenclature commissions, which publish cumulative collections of guidelines in their respective “color books”: the Blue Book (organic nomenclature), the Red Book (inorganic nomenclature), and so on [4]. For example, starting from the 1500 pages of guidelines contained in the latest edition of the Blue Book, one can trace their genealogy back to the resolutions of the 1892 International Congress of Geneva for the Reform of Chemical Nomenclature. (It is worth noting, however, that decisions *not* to sanction proposals or practices are just as important a part of the work of standards organizations. The intellectual memoirs of the Dutch chemist P.E. Verkade (1891–1979, chair of the IUPAC Commission on Nomenclature of Organic Chemistry 1934–1971), are packed with naming schemes that his commission considered but dismissed [1].)

The “Geneva Nomenclature” launched a century and a quarter of negotiations in which the pendulum of standards-setting swung back and forth between pristine logic and pragmatism. Though guided by a systematic ideal of rule-bound correspondence between chemical names and structural formulas, the Commission on Nomenclature of Organic Chemistry most often proceeded in a spirit of “systematic flexibility” [5]. By allowing for alternative naming procedures, the commission accommodated the differing conventions of publications like *Beilsteins Handbuch* and *Chemical Abstracts* and provided workarounds for the shortcomings of specific nomenclature rules. E. J. Crane (1889–1966), who worked extensively with IUPAC nomenclature commissions during his long term as editor-in-chief

of *Chemical Abstracts* (1915–1958), summed up this attitude nicely in a 1937 letter: “Rules are sometimes to be ruled instead of always to apply” [6].

Recovering the history of IUPAC efforts to standardize chemical identification on machines requires starting in the past rather than working backwards from the present. As IUPAC officers and committee members returned to Union activities after the end of World War II, one of the first subjects that attracted their attention was how to facilitate the continued production of chemists’ precious reference compendia. Between 1947 and 1952, an Advisory Council on *Beilstein* and *Gmelin* chaired by the British chemist Alexander Todd (1907–1997) set to work getting the war-torn German publications *Beilsteins Handbuch der organischen Chemie* and *Gmelins Handbuch der anorganischen Chemie* (*Beilstein* and *Gmelin*) back into production.

But the specter of information overload threatened even publications that were not directly affected by the war, such as *Chemical Abstracts*. The war-catalyzed growth in chemical research and the postwar declassification and publication of large quantities of state-sponsored research brought additional urgency to this challenge. Wartime advances in mechanical and electronic computing, coupled with increasing use of business machines such as punched card sorters and tabulators for managing myriad recordkeeping problems, suggested a solution: notations designed for handling molecular structure on machines [7].

G. Malcolm Dyson (1902–1978), technical director at the British pharmaceutical firm Genatosan, firmly believed that the day of machine-based chemical documentation had come. Dyson reasoned that a machine-readable code or “cipher” offered two distinct advantages over systematic names. First, by starting from the ground up rather than from names already in use, a system of ciphers could aim for the logical consistency that eluded IUPAC nomenclature, which had to accommodate existing usage. Second, machine-readable ciphers would enable reference libraries and publications like *Chemical Abstracts* and *Beilstein* to use machines to overcome a bottleneck in the work of chemical documentation: the expert labor involved in writing and interpreting systematic chemical names.

Dyson’s publication of his cipher system in 1947 earned him widespread attention among chemists working with nomenclature and documentation. He was named to the IUPAC Commission on Nomenclature of Organic Chemistry in 1947; he remained a member through 1971, cultivating friendships with Verkade and the American organic nomenclature guru Austin Patterson (1876–1956). Dyson was also appointed chair of a new IUPAC Commission on Codification, Ciphering, and Punched-Card Techniques (1947–1961). Other members included Howard Nutting (1901–1986) of Dow Chemical, the British nomenclature expert Alec Duncan Mitchell (1888–1963), and James W. Perry (1907–1971) of MIT, a detergent chemist turned punched-card expert. Verkade was the lone non-Anglo-American among the commission’s nine members. The commission also worked with collaborators who were not formal commission members, including Madeleine (Berry) Henderson (1922-2011) of MIT [8].

Numerous other chemists saw the same opportunity as Dyson in the blank slate afforded by machine-oriented ciphers. Chemists proposing alternative schemes were unhappy to see one of their competitors in charge of an IUPAC commission to determine which scheme would gain international sanction. Sure enough, in 1951, the IUPAC commission provisionally adopted Dyson’s notation, a decision which became permanent in 1961. By that time, however, the decision had little direct impact. Chemical firms and other organizations that dealt with chemical information had adopted one or the other of various notations according to local needs and preferences. Still, Dyson parlayed his reputation into a position as research director at Chemical Abstracts Service (CAS). In this role, he oversaw the initial stages of the computerization project that became the CAS Chemical Substances Registry and gave birth to the soon-to-be-ubiquitous CAS Registry Number.

* * *

In 1969, in light of the increasingly widespread application of computers to the management of chemical information, IUPAC’s Executive Committee commissioned a report on machine documentation in chemistry. Working in consultation with *Chemical Abstracts* staff, the authors of this report recommended that IUPAC appoint a committee to pursue standardization in machine handling of chemical information. The primary task of this commission was to be “the machine handling of chemical structures and the computer generation of nomenclature,” including “a unique definition of chemical structure which is understandable on the printed page and yet logical, unambiguous to a computer program” [9].

The Union accordingly formed an International Committee on Machine Documentation in the Chemical Field, with the kind of international membership that was typical of IUPAC, in contrast to Dyson’s predominantly Anglo-American group. The abbreviated description of the duties of this committee left substantial room for alternative interpretations of its purpose. The French physical chemist and cheminformatics pioneer Jacques-Émile Dubois (1920–2005), nominated as chair of this commission, focused on the mandate to study *machine handling of chemical structures*. To Dubois, this meant establishing requirements for computer-based exchange of information about molecular structure among a diverse international constituency searching for chemical information and analyzing chemical data. To committee member and CAS associate director Fred Tate (1920–1980), the committee’s job was *computer generation of nomenclature*, that is, identifying and promoting opportunities to revise IUPAC nomenclature to accommodate it to processing and handling on machines. French nomenclature commission member Noël Lozac’h (1915–2003) saw the machine documentation committee as an opportunity to swing human-oriented chemical names back toward the pristine logic of the Geneva Congress. Nobody gave much thought to Dyson’s IUPAC-sanctioned notation. The committee continued to meet, correspond, and liaise with IUPAC nomenclature commissions through the mid-1970s without the members ever quite agreeing what their job was supposed to be [10].

IUPAC continued to sponsor work in this domain, including an International Symposium on Techniques for the Retrieval of Chemical Information, held in London in 1976 [11]. Union commissions pursued various projects in other areas involving computerization and automation. However, IUPAC did not return in earnest to work on machine-readable chemical identifiers until the InChI project.

* * *

Neither Dyson's nor Dubois's commissions seems to have had much of a direct impact on chemical information management or on chemistry more broadly. Nevertheless, juxtaposing these histories with those of IUPAC nomenclature commissions and the InChI project can teach us something about IUPAC and its relationship to the development of global chemistry over the past 100 years.

First, IUPAC was no technological laggard. The Union took up the question of machine-readable notation in 1947, around the same time that national chemical organizations did so, just a few years after punched cards migrated out of accounting departments into chemists' broader awareness. The machine documentation committee was convened within a few years of the launch of large-scale computer-based information systems like the CAS Registry.

Second, the Union's quick response to emerging information technologies proved less effective (or at least less enduring) than standardization efforts addressing well-established information technologies: reference works like *Chemical Abstracts* and *Beilstein*, in the case of the Commission on Nomenclature of Organic Chemistry, and the host of chemical databases and software widely relied upon by the year 2000, in the case of InChI. This would seem to be counterintuitive. Shouldn't a standardization effort be more effective when able to start from a clean slate in a new medium, as Dyson and other members of the ciphering committee hoped would be true for an IUPAC punched card notation?

It turns out that the snarled, frustrating limitations of nomenclature—trivial names sanctioned by longstanding use, the inconsistent schemes of *Beilstein* and *Chemical Abstracts*, the conflicts of 1920s international politics—were opportunities as well as obstacles. The Commission on Nomenclature of Organic Chemistry managed to codify a set of rules in 1930, a landmark achievement for the young Union, precisely *because* the respective conventions of *Chemical Abstracts* and *Beilstein* were so entrenched. The broadly-acknowledged importance of fostering incremental harmonization of the nomenclatures used by these major reference works allowed IUPAC's nomenclature efforts to seem well worth pursuing. The added stakes of nomenclature's entanglement with questions of interwar politics just encouraged more interest and engagement in the subject.

The ubiquity of electronic databases, chemical drawing programs, and other computer-based resources around the turn of the 21st century meant that the InChI effort began

under similar conditions. While InChI, in marked contrast to nomenclature, did in fact start fresh, it was operating in a well-established field of machine-based analysis of chemical structures. There was widespread appreciation for the challenges imposed by the divergent notations and file formats for representing chemical structure.

This brings us to the third lesson of this story. Within a consensus-based international organization like IUPAC, standard-setting efforts seem best positioned to take hold when conflicting practices have already taken root. This has been the case for chemical structure representation, at least; it is likely true more generally. It is easier to recognize what it will take to mitigate a community-wide problem, and to secure broad-based support for doing so, after things have already gotten into a muddle. In the terminology of information scholars, IUPAC does not typically traffic in *systems* (rationally planned, centrally controlled, constraint-heavy but smoothly operating environments). IUPAC more often addresses *networks* and *gateways*, that is, protocols for coordinating different systems so that users can hop, perhaps a bit awkwardly, from one to another and back. This is not to say that Dyson's and Dubois' efforts were ill-founded. IUPAC has been a valuable site for studying emerging problems in information management, too. But for getting rules in place that can be made to stick, history suggests that you sometimes have to wait until heads are already spinning.

February 23, 2019

References

1. Pieter Eduard Verkade, *A History of the Nomenclature of Organic Chemistry*, trans. S. G. Davies (Boston: Reidel, 1985).

2. Leah McEwen, "InChI'ng Forward: Community Engagement in IUPAC's Digital Chemical Identifier," *Chemistry International* 40, no. 1 (2018): 27–31, https://doi.org/10.1515/ci-2018-0109.

3. Ray Boucher, Stephen Heller, and Alan McNaught, "The Status of the IUPAC InChI Chemical Structure Standard," *Chemistry International* 39, no. 3 (2017): 47, https://doi.org/10.1515/ci-2017-0316.

4. "Color Books," International Union of Pure and Applied Chemistry, https://iupac.org/what-we-do/books/color-books/, accessed 14 Feb 2019.

5. Evan Hepler-Smith, "Systematic Flexibility and the History of the IUPAC Nomenclature of Organic Chemistry," *Chemistry International* 37, no. 2 (2015): 10–14, https://doi.org/10.1515/ci-2015-0232.

6. Crane to Leech, 26 Oct 1937, William A. Noyes Papers, 15/5/21, University of Illinois Archives, Urbana, IL, Box 14, Folder "General Correspondence, 1937–38."

7. Bonnie Lawlor, “The Chemical Structure Association Trust,” *Chemistry International* 38, no. 2 (2016): 12–15, https://doi.org/10.1515/ci-2016-0206.

8. Robert V. Williams, “Madeline M. Henderson: From Chemical Information Science Pioneer to Architect of the New Information Science,” *Libraries and the Cultural Record* 45, no. 2 (2010): 167–84.

9. J. W. Barrett, H. K. Livingston, and Byron Riegel, “Machine Documentation in the Chemical Field: Report to the Bureau of IUPAC,” 4 July 1969, Addenda to the Records of the Union of Pure and Applied Chemistry, Science History Institute, Philadelphia, PA, Box 91.

10. Papers of International Committee on Machine Documentation in the Chemical Field, 1968–1978, Addenda to the Records of the Union of Pure and Applied Chemistry, Science History Institute, Philadelphia, PA, Box 91.

11. “International Symposium on Techniques for the Retrieval of Chemical Information,” *Pure and Applied Chemistry* 49, no. 12 (1977): 1779–1900.

Evan Hepler-Smith <heplers@bc.edu> is Visiting Assistant Professor of History at Boston College. His book in progress, *Compound Words: Chemists, Information, and the Synthetic World*, explores how nomenclature systems and information management have shaped the history of modern chemistry.

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), co-chair of the CPCDS Subcommittee on Cheminformatics Data Standards, and secretary of the InChI Subcommittee. ORCID.org/0000-0003-2968-1674