FAIR datasets for acid dissociation constants

by Jonathan Zheng, Ye Li, and Leah McEwen

Acid dissociation constants (pKas), the physicochemical properties that govern how acidic a chemical is in a solution, are an important type of reference information for research in chemistry and related fields as they govern the behavior of pharmaceutical drugs in the human body, environmental impact of molecules, and manufacturability of chemical products. To enable the use of reliable pKa data in broad practical and theoretical applications, IUPAC compiles values published in the literature, critically evaluates these based on established methods of assessment and publishes these in printed reference works [1-6].

A digital pKa dataset based on high quality compilations from IUPAC that is FAIR (Findable, Accessible, Interoperable, and Reusable) can be a trustworthy reference for chemists to build machines learning models to predict pKas of any conceivable organic molecule. Other openly accessible online pKa data sources lack critical context and metadata regarding the measured properties, such as the temperature at which the pKa was determined, the experimental method or assumptions used, the estimated reliability of such data points, and so on.

With permission from IUPAC, an effort to digitize several printed IUPAC pKa collections was initiated as part of a PhD research project by Jonathan Zheng, a PhD candidate in the laboratory of Professor William H. Green at the Massachusetts Institute of Technology (MIT). Members of the Green Research Group, MIT Libraries, the IUPAC Committee on Publications and Cheminformatics Data Standards (CPCDS), IUPAC Division V (Analytical Chemistry) and the team at PubChem were consulted at various points throughout the project.

More than 20,000 pKa values were extracted from scanned text and compiled into a structured dataset. Entries were checked both manually and programmatically to review for accurate transcription and ensure all metadata and provenance were included. Machine readable chemical notations including InChI strings, InChIKeys, and SMILES strings were also incorporated to enhance reusability and interoperability. Figure 1 illustrates the workflow for the digitization and curation process.

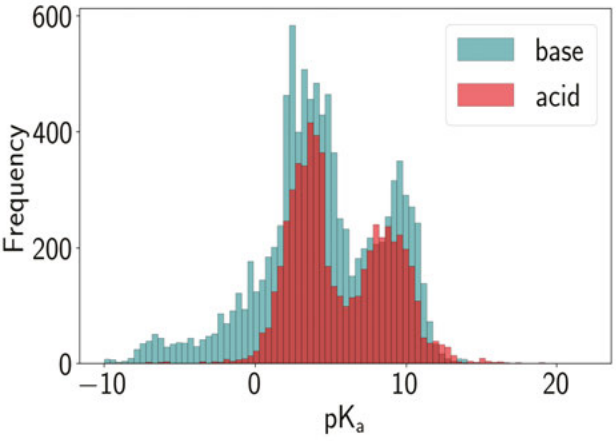

Figure 2. Distribution of pKa values in the curated dataset.

The resulting curated dataset includes **21,147 entries** covering **8,843 unique molecules** from three different volumes. Figure 2 illustrates the distribution of pKa values across the dataset. The data is available to the community under CC BY-NC 4.0 license from the IUPAC public GitHub repository and also from Zenodo [7], and includes further details on the digitization process and accompanying tables of the original literature references and evaluation methods. IUPAC groups will continue to work on curating the data scanned from the additional printed volumes, including disambiguation of less familiar chemical names and solution composition information. Additional data will become available as well from ongoing evaluation of acid pKa values in polar aprotic solvents (IUPAC project 2015-020-2-500).

This collection of IUPAC evaluated dissociation constants has now become more Findable from citations to the DOI, more Accessible through open repositories, more Interoperable across data analysis platforms in the open CSV format, and more Reusable through the open license and descriptive documentation. The international scientific community can apply this FAIR dataset to many areas of research. The Green Research group

continued on page 29

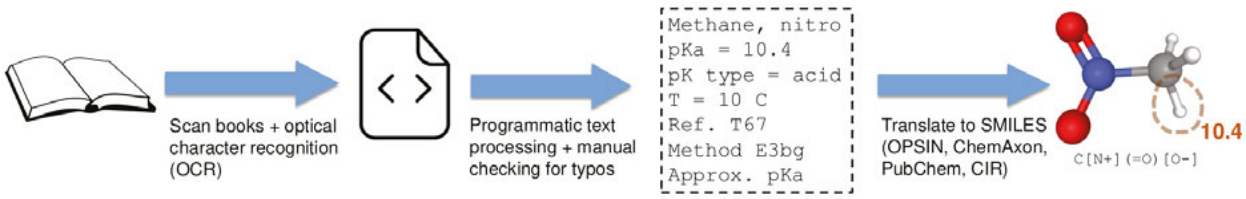

Figure 1. Workflow of the digitization and curation process.

is water purification, pharmaceutical products in the environment, photochemical degradation of anthropogenic chemicals, plastic pollution, thorium chemistry …. whatever …, chemists should join forces and work out presentations topic by topic and add them to the new Resource section on the IUPAC website. When high-quality material is already available on the web, there is no reason to duplicate it; include it via link(s) (with comments if required) so that the presentations on iupac.org become excellent gates to high-quality, exhaustive coverage of relevant and urgent topics for people with a wide range of backgrounds and competences in chemistry. These presentations should for instance be perfect to use for educational purposes by teachers who will be able to make their own presentations adjusted to the students' level of competence.

4) Make the NAO Forum a bottom-up meeting arena

This new contact point represents a step in the right direction, but the agenda for the meetings should not consist of items the President wants to discuss. A genuine problem in the relation between the IUPAC leadership and the NAOs has always been the top-down nature of the communication between the parties. Unlike the larger countries in the Union, the NAOs in most of the smaller countries are operated by volunteers that rotate so often that IUPAC remains overwhelming even after years of engagement. A good indicator for the closeness of contact and the ease of communication is the lack of interventions from the members at the biannual Council Meeting; very few asks for the floor, and when important issues are given a time slot of only 5-10 min, even seasoned council-meeting attendees don't feel invited to intervene.

In my opinion, implementation of these four action items will make IUPAC a more interesting organization to join, serve, and be associated with. Another likely consequence is that IUPAC will become better known outside the chemical circles and therefore better positioned to make an impact through the application of the chemical sciences to the betterment of humankind. That will indeed be needed in the years to come.

Leiv K. Sydnes is professor emeritus at Department of Chemistry, University of Bergen, Norway. He was president of IUPAC 2004-2005 and chaired the CHEMRAWN committee from 2008-2015.

Project Place (cont. from p. 26)

at MIT are already utilizing this dataset alongside other openly available thermodynamic data in calculations to determine the energies of ions in water. The resulting set of calculated hydration energies for over 300 ionic solutes will enable further modeling of ions in solution across a range of settings.

Ensuring that scientific data is FAIR for the broadest possible use and benefit of the global society is a community endeavor that involves many stakeholders in the research data ecosystem—chemists measuring and publishing original research data; other chemists working on machine learning projects for compelling use cases; IUPAC expert volunteers evaluating and curating property data; chemical information professionals who can facilitate testing, documenting, depositing and sharing data. IUPAC has a critical role in coordinating with the broader community around FAIR chemical data, providing standards and expertise in robust chemical data reporting and enabling worldwide access and use across disciplines [8-9].

References

1. Ionisation Constants of Organic Acids in Aqueous Solution; E P Serjeant and Boyd Dempsey; Oxford/Pergamon (1979) (Oxford IUPAC chemical data series)
2. Dissociation Constants of Organic Bases in Aqueous Solution; DD Perrin; Butterworths (1965)
3. Dissociation Constants of Organic Bases in Aqueous Solution, Supplement 1972; DD Perrin; Butterworths (1972)
4. Dissociation Constants of Organic Acids in Aqueous Solution; G Kortum, W Vogel and K Andrussow; Butterworths (1961)
5. Dissociation Constants of Inorganic Acids and Bases in Aqueous Solution; D D Perrin; Butterworths (1969)
6. Acid-Base Dissociation Constants in Dipolar Aprotic Solvents; Izutsu, K; Blackwell (1990)
7. Jonathan Zheng. (2022). IUPAC/Dissociation-Constants: v1.0 (v1-0_initial-release) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7236453
8. Bruno, I.; Coles, S.; Koch, W.; McEwen, L.; Meyers, F.; Stall, S. (2021) FAIR and Open Data in Science: The Opportunity for IUPAC. *Chem Int*, 43(3), 12-16. https://doi.org/10.1515/ci-2021-0304
9. McEwen, L. and Mustafa, F. (2023). WorldFAIR Chemistry: Making IUPAC Assets FAIR. *Chem Int*, 45(1), 14-17. https://doi.org/10.1515/ci-2023-0104

Jonathan Zheng is a PhD student in the Green Group at MIT; Ye Li is the Librarian for Chemistry, Chemical Engineering, Materials Science and Engineering at MIT and Chair of the Division of Chemical Information of the American Chemical Society; Leah McEwen is the Chemistry Librarian at Cornell University and Chair of the IUPAC Committee on Publications and Cheminformatics Data Standards.