ChemSpider webinar: Data standards in chemistry Q&A report
Watch the recording

Speakers: Kamil Dzuibek (University Assistant, University of Vienna, Austria), Robert Hanson (Professor of Chemistry, Emeritus, St. Olaf College, USA), Wendy Patterson (Scientific Director, Beilstein-Institut, Germany)

Q: **Great to see use of AI in here, spectral analysis which AI tools or LLM do you find useful?**

Robert Hanson: I wish they existed. But so far, to my knowledge, they do not.

Q: **Dear Sir Thank you so much for very inspiring session. I am concerned about which molecular software which can represent a conjugated venom peptide molecules such as Mastoparan gold nanoparticle aptamer?**

Robert Hanson: This is a very good question. As you probably know, there is no molecular way of describing a gold nanoparticle. Here is where flexibility in representations is important. Perhaps the best you can do it is to draw the structure of mastoparan and replace an H with Au. And also include an image of the drawing. The point is to have a representation that is as close as possible to what you use yourself. See our recommendations at https://www.degruyterbrill.com/document/doi/10.1515/pac-2025-0409/html

Q: **Dear Professor Hanson, is it possible to apply your hydrogen atom orbital generator to develop a tutorial in chemistry for year-1 students? what reference should be included?**

Robert Hanson: Sure. S. P. Tully, T. M. Stitt, R. D. Caldwell, B. J. Hardock, R. M. Hanson, and P. Maslak, "Interactive Web-Based Pointillist Visualization of Hydrogenic Orbitals Using Jmol" J. Chem. Educ. 2013, 90, pp 129–131. http://dx.doi.org/10.1021/ed300393s

Q: Till now how much and also upto what year data has been disseminated for use of machine learning.

Kamil Dzuibek: I can only comment on the crystallographic data. All AlphaFold versions are trained on PDB data; a recent ML tool for structure identification from X-ray pair distribution functions (https://doi.org/10.1039/D4DD00001C) has been trained on PDFs simulated from crystallographic structures obtained from the Crystallography Open Database (COD); another program for finding relevant papers based on uploaded

powder diffraction patterns (https://doi.org/10.1107/S2053273322007483) was trained on the IUCr archive for powder CIFs uploaded by authors with papers published in IUCr journals. In conclusion, I believe that open access to data is more crucial than the restrictive year.

Q: **for Beilstein tool - Do industry use the tools to help FIARify the data? Or is it avoided as it becomes open. Is there any application for industry?**

Wendy Patterson: the tools and software we develop are released with either the MIT license or CC0 (in the case of look-up tables), thus industry is certainly welcome to use-reuse-further develop these tools according to their needs. All domains in chemistry industry work with and use CDX files so I certainly believe these tools will be applicable for them as well.

Q: **Thank you for your great work on these data standards and FAIR practices. Please continue to work with data curators so we can help spread the word on our campuses and incorporate these standards in our work. The Data Curation Network in the United States creates primers as well that can be another avenue for dissemination.**

Kamil Dzuibek: Thank you for this excellent comment and making me aware of the DCN! I checked the website https://datacuration.network and discovered that the Mass Spectrometry Primer is the only resource in the collection directly related to the chemical sciences.

Wendy Patterson: Thank you so much for bringing this network and the primers to our attention. Absolutely agree that data curation is key to ensuring that we are working with not only FAIR data but quality data. I did find the Mass Spectrometry primer where several colleagues were involved in the creation.

Q: **Thank you for the insightful presentation on FAIR data principles in chemistry. As an MSc research student, I'd like to ask how can early-stage researchers ensure our data is both standardized and reusable, especially when we don't have access to advanced institutional repositories?**

Robert Hanson: No repository is necessary. You can do this yourself. See our open-access paper: https://www.degruyterbrill.com/document/doi/10.1515/pac-2025-0409/html

Q: **Would databases like FAIR data reduce spurious data in publications?**

Robert Hanson: We certainly hope so. Because if the raw data is made available, then it is less likely (but not impossible) to forge.

Q: **Dear Dr. Patterson, is there a tool that selects among various published crystallographic structures, for the same material, which is the most accurate?**

Kamil Dzuibek: This is an interesting question, which falls squarely into the category of data quality. Indeed, while the FAIR principles have advanced data openness and reusability, they do not by themselves guarantee the accuracy, reliability, or representativeness of scientific data. *check*CIF is a good validation tool (remember to compare the structural data together with structure factors and analyse carefully the *check*CIF report!), but it is not an exhaustive one. Simple comparison of R-factors is not always helpful, significance tests, like e.g. the Hamilton's R-ratio test are often recommended instead (see https://onlinelibrary.wiley.com/iucr/itc/Cb/ch8o4v0001/sec8o4o2). PLATON, which I mentioned in my presentation, is a multipurpose crystallographic tool program that can provide you various validation routines (https://www.platonsoft.nl/platon). Besides, I am aware that Julian Henn is working on new quality indicators for diffraction data and runs his own company, which you can find at https://www.dataqintelligence.com Please note that these solutions are available for a fee. However, you can read his articles to see if you find them helpful.