ChemSpider: An Online Chemical Information Resource

by Harry E. Pence*

Department of Chemistry and Biochemistry, SUNY College at Oneonta, Oneonta, New York 13820
*pencehe@oneonta.edu

by Antony Williams

ChemSpider, Royal Society of Chemistry, U.S. Office, Wake Forest, North Carolina 27587

The World Wide Web is having a profound effect on the availability of chemical information. A modern chemist may wish to know a variety of information about a given compound, including physical and chemical properties, molecular structure, spectral data, synthetic methods, known reactions, safety information, and systematic nomenclature. In the past, having access to such a range of information required a small library of different reference works because no single resource contained all this data. This was problematic in terms of both cost and physical space for storage.

Now, however, a single Web site provides all this information for millions of compounds and is also free.

What Is ChemSpider?

ChemSpider is a free, online chemical database offering access to the type of information described in the previous paragraph for almost 25 million unique chemical compounds sourced and linked out to almost 400 separate data sources on the Web *(1)*. ChemSpider is not just a search engine layered on terabytes of chemistry data; it is also a crowdsourcing community for chemists who contribute their data, skills, and knowledge to the enhancement and curation of the database. ChemSpider, therefore, resembles Wikipedia by encouraging participation and contributions from the community.

A Brief History

ChemSpider was created as a hobby project by one of the authors (A.W.) and a few associates. Its success in permeating the domain of online chemistry and its contributions to community-based chemistry attracted the attention of the Royal Society of Chemistry (RSC), which acquired the system in May 2009. This purchase made good sense because ChemSpider gained major infrastructure support and access to a wealth of materials provided by the RSC: journals, books, supplementary materials, and databases. RSC also uses the ChemSpider database to expand the features in Project Prospect, a relatively new initiative in journal publication. Project Prospect, which is focused on semantic markup, won the 2007 ALPSP/Charlesworth Award for Publishing Innovation *(2)*.

ChemSpider as an Aggregator of Chemical Information

In conversations, the authors have heard ChemSpider described as the Google for Chemistry and a Wikipedia for chemists. In reality, and to add the confusing hyperbole, it is neither and both. By aggregating data from nearly 400 different data sources and connecting them by means of chemical structure as the primary record in the database, ChemSpider has been able to link Wikipedia *(3)*, PubChem *(4)*, Chemical Entities of Biological Interest (ChEBI) *(5)* and The Kyoto Encyclopedia of Genes and Genomes (KEGG) *(6)*, chemical vendors, a patent database, and both open- and closed-access chemistry journals. Where possible, each chemical record retains the links to the original source of the material, thereby associating a microattribution. These links let a ChemSpider user source information of particular interest, including where to purchase a chemical, chemical toxicity, metabolism data, and so on. Aggregating that level of connected information via a classical search engine such as Google would be very time-consuming.

Curating the Information on ChemSpider

Curation, which involves ensuring the accuracy of the data in a digital database *(7)*, is an essential problem for any reference source. ChemSpider allows registered users to enter information and annotate and curate the records. The requirement to register and login is to prevent anonymous acts of vandalism. The chemical community has been forthcoming in adding information, including new chemical structures, associations between structures and publications, addition of analytical data such as spectra, and the curation of chemical identifiers and property data.

Currently, the standard chemical resource is the Chemical Abstracts Service (CAS), which has been in the business of aggregating chemistry-related data for 102 years in order to create the CAS registry. CAS recorded its 50 millionth chemical structure just last year *(8)*. But in just over three years, ChemSpider has aggregated nearly 25 million unique chemical entities. New additions to the database are made daily, especially because it is now integrated with the RSC publishing process whereby new compounds identified in prospected RSC articles are deposited and released to the community as the article is published. Many of the compounds in the ChemSpider database have already been curated, and the process is ongoing. More than a million of the name−identifier relationships have been robotically or manually curated. This has produced highly qualified chemical dictionaries that can be used as the basis of entity extraction engines as explained recently in an article by Hettne et al. *(9)*.

Comparing ChemSpider to Other Free Online Sources

What other free online sources might be compared to ChemSpider? There are certainly a number of other chemistry

10.1021/ed100697w Published on Web 08/30/2010

databases online. For example, PubChem *(4)* from the National Institutes of Health contains well over 25 million structures linked to various data and sources. However, PubChem's data are not curated and are contaminated with errors from the various depositors, realities described in multiple presentations and in various online forums *(10)*. NIH lacks the resources to curate the data within the PubChem repository, so it depends on the depositors to curate their own data. This is significantly different from ChemSpider, which depends on the crowdsourcing activities of the community.

ChemSpider has a number of advantages over a simple Google search. The variety of information about a compound provided at ChemSpider is hard to match on any other free Web site. The data continue to be validated and updated by practicing chemists, and in many cases, they have been reviewed for accuracy. ChemSpider also provides links to many other online sources for further information. This plethora of links now includes Google Books, Scholar, and Patents; Microsoft Academic Search; the RSC Databases, Books, and Publishing Web site; and an ever-increasing number of government, commercial, and academic databases.

Searching the Web using one of the new types of search engines such as Wolfram Alpha or Google Squared is less useful than ChemSpider. Those services provide neither structure-based searching of the Internet nor systematically organized data curation. The closest comparison in terms of validated and crowdsourced contributions to the domain of chemistry are the chemical pages in Wikipedia; however, Wikipedia has information on far fewer compounds and supports only text searching, not structure searching.

ChemSpider's Usefulness for Teaching, Learning, and Research

ChemSpider is already becoming an important resource for teaching, learning, and research. Specifically, the spectroscopic data (more than 2000 spectra in total) are the basis for the Spectral Game, which has already been used by more than 8000 students in nearly 100 countries *(11)*. This game lets students learn how to interpret NMR spectra by validating either ^{1}H or ^{13}C spectra against two or more structures. The game increases in complexity as it progresses; ultimately, students must choose a spectrum match from among five structures. Students can also participate in data curation. Such exercises are already being set by academics to train their students to research and validate chemical data online *(12, 13)*.

The recently added ChemSpider SyntheticPages, which are created by the community, for the community, provide an online database of chemical synthesis procedures. Chemists can now populate an online database with one of their chemical reactions and outline how to perform a reaction. Each reaction has a digital object identifier (DOI) issued so that a student can add this online "publication" to their résumé.

Future ChemSpider Initiatives

Work will soon begin on a project called ChemSpider Education. This feature will provide a subset of data and new interface elements focused on chemistry students from secondary schools through the undergraduate years. ChemSpider is already a major resource for teaching and learning that should be covered in the undergraduate chemistry curriculum. As more data are added and curated, and as new capabilities are developed, ChemSpider can become the major chemistry Internet portal for chemical educators and chemists everywhere.

Literature Cited

1. ChemSpider Home Page. http://www.chemspider.com/ (accessed Aug 2010).
2. Project Prospect Home Page. http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp (accessed Aug 2010).
3. Wkipedia Home Page. http://www.wikipedia.org/ (accessed Aug 2010).
4. PubChem Home Page. http://pubchem.ncbi.nlm.nih.gov/ (accessed Aug 2010).
5. ChEBI Home Page. http://www.ebi.ac.uk/chebi/ (accessed Aug 2010).
6. KEGG Home Page. http://www.genome.jp/kegg/ (accessed Aug 2010).
7. Definition of Digital Curation on Wikipedia. http://en.wikipedia.org/wiki/Digital_curation#cite_note-dccdefn-0 (accessed Aug 2010).
8. Schulz, W. G. *Chem. Eng. News* **2009**, *87* (37), 3.
9. Hettne, K. M.; Williams, A. J.; van Mulligen, E. M.; Kleinjans, J.; Tkachenko, V.; Kors, J. A. *J. Cheminform.* **2010**, *2* (1), 3; DOI: 10.1186/1758-2946-2-3. http://www.jcheminf.com/content/2/1/3 (accessed Aug 2010).
10. The ChemSpider Blog. http://www.chemspider.com/blog/hacking-pubchem-technology-easy-quality-difficult.html (accessed Aug 2010).
11. Bradley, J.-C.; Lang, A.; Williams, A. Spectral Game Home Page. http://www.spectralgame.com/ (accessed Aug 2010).
12. Bradley, J.-C. Open Notebook Science. http://usefulchem.wikispaces.com/ (accessed Aug 2010).
13. Moy, C. L.; Locke, J. R.; Coppola, B. P.; McNeil, A. J. *J. Chem. Educ.* **2010**, *87*; DOI: 10.1021/ed100367v.

