The background of the page is a collage of laboratory-related images. On the left, several test tubes are arranged diagonally, containing liquids of various colors. In the bottom right, a multi-well microplate is visible, with a pipette tip positioned over one of the wells. The overall color palette is a mix of soft pinks, purples, and blues, creating a scientific and clinical atmosphere.

Having information about a
chemical compound
summarized in one place
saves time.

A New Era in Chemical Information:

PubChem, DiscoveryGate, and Chemistry Central

By Svetla Baykoucheva

Selecting relevant information and suppressing details is the sort of pragmatic fudging everyone does every day. It's a way of coping with too much information. For almost everything you see, hear, taste, smell, or touch, you have the choice between examining details by scrutinizing very closely and looking at the "big picture" with its other priorities.

—Lisa Randall (2006), *Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions*, p. 29

HOW we go about finding the information we need, whether it's chemical or not, depends on how detailed we want this information to be and how we want it organized. Researchers and information specialists are accustomed to searching bibliographic databases to find references to articles published in scientific journals. In some cases, though, having the information about a chemical compound summarized in one place rather than trying to extract specific data from different papers saves a lot of time. To use an analogy from Lisa Randall's book, *Warped Passages: Unraveling the Mysteries of the Universe's Hidden Dimensions* (Harper Perennial, 2006), we could live with one-dimensional information (titles and abstracts of papers) or we could add extra dimensions by using resources that provide summarized reports of the physical, chemical, and biological properties of chemical compounds.

This article will discuss how three important events that have happened in the past 2–3 years—the emergence of PubChem, the introduction of DiscoveryGate, and the creation of Chemistry Central—are changing the field of chemical information.

PUBCHEM

PubChem (<http://pubchem.ncbi.nlm.nih.gov>) is a free service that was launched by the U.S. National Institutes of Health (NIH) in 2004. It consists of three databases (PubChem Compound, PubChem Substance, and PubChem Bio-Assay) linked together and incorporated in the Entrez information retrieval system of the National Center for Biotechnology Information (NCBI). PubChem is also an important part of the NIH's Molecular Libraries Roadmap Initiative (<http://nihroadmap.nih.gov>).

PubChem Compound contains more than 10 million unique structures and

provides biological property information for each compound through links to other Entrez databases.

PubChem Substance contains more than 17 million records of substances that were deposited by other organizations—publishers, suppliers of chemicals, and other vendors. It provides descriptions of chemicals and links to PubMed, protein 3-D structures, and biological screening results. If the composition of a chemical sample is known, the description includes links to PubChem Compound.

While the PubChem Compound database contains records of chemicals that have already been identified, the PubChem Substance database contains records of all chemical substances that have been submitted to PubChem. If a component of a sample can be identified as an individual compound, it becomes a candidate for inclusion in the PubChem Compound database.

PubChem BioAssay can be searched to find information about bioassays using specific terms pertinent to the bioassay. It is also possible to browse or download PubChem BioAssay results. The bioassays' descriptions are also searchable.

You can search PubChem by names of chemicals, elements, or groups of elements. Other capabilities include searching by property range, such as temperatures between one number and another, or drawing a structure and using the drawing as a query. It is also possible to identify compounds that are similar to those studied.

PubChem has a limited scope of coverage—its goal is to provide information on small organic molecules. With its 17 million substances and 10 million unique structures, it is smaller than the CAS Registry File, which has 31 million organic and inorganic substances and more than 58 million sequences. You can access the latter through STN, SciFinder, and SciFinder Scholar, which are subscription-based products from the Chemical Abstracts Service (www.cas.org), a branch of the American Chemical Society. The rate, though, with which PubChem has been growing is really impressive. Many vendors are now feeding information about chemical compounds into it. In 2005, Nature Publishing Group was the first big commercial publisher to start doing so, followed by other vendors.

PubChem is a free service. Regardless of whether you are a librarian, a researcher, a physician, a teacher, a student, or someone who is simply concerned about a chemical in the new shampoo that you just bought, PubChem can provide you with information about the physical, chemical, and biological properties of many chemicals, including some drugs.

DISCOVERYGATE

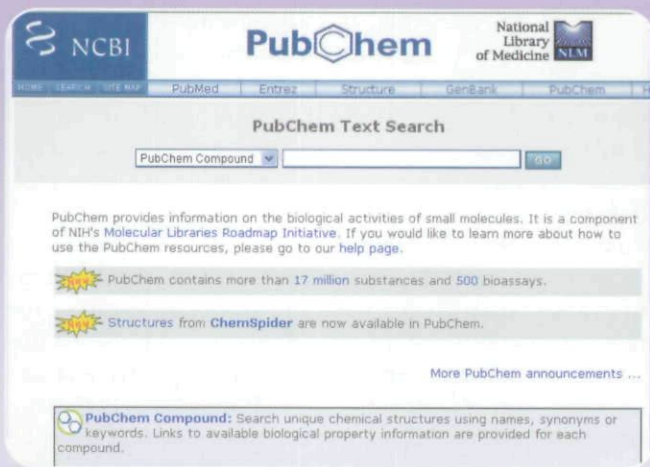
DiscoveryGate (DG; www.discoverygate.com) is an online service from Elsevier MDL that provides access to more than 20 databases through a unified platform called Compound Index. The databases in DG contain extensive experimental

data derived from journals, patents, chemical catalogs, major reference works, and Food and Drug Administration (FDA) documents. They cover more than 27 million structures, 17 million reactions, and more than 500 million experimentally proven values, which makes DG the largest collection of chemical properties.

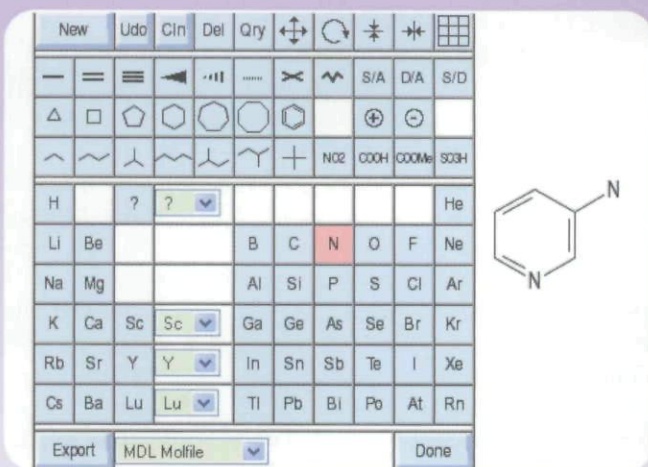
In DG it is possible to search for information on chemical, physical, and biological properties of chemicals (including new drugs at different stages of development); to find suppliers and compare prices of chemicals; to discover methods to synthesize new and existing compounds; to associate structures with experimentally determined property data retrieved from articles and patents; and to view, in one document, data retrieved from different sources. You search DG via a series of forms, with each form having its own set of command buttons. DG also allows exporting the results to reports that can combine data from multiple searches in one report that can be printed, saved, or exported to an HTML file. Some of the most important databases in DG are discussed in what follows. A comprehensive list of the DG databases is available at the MDL Web site (www.mdli.com/solutions/solutions_for/academics/dg_academics.jsp).

CrossFire Beilstein provides access to the huge Beilstein database, which is the crown jewel of DG. It is the world's most comprehensive database for compounds, reactions, properties, and citations. It covers organic chemistry dating from 1771 and contains more than 9.9 million chemical structures (with 6.5 million of them being unique). It is the world's largest reaction database, with more than 10 million reactions. Most of the 320 million chemical properties in it have been experimentally verified. The database also provides access to 900,000 abstracts dating from 1980.

CrossFire Beilstein allows conducting structure searching, reaction searching, and data searching. Until recently, the Beilstein database was not included in the Compound Index and could be searched only separately. Because MDL



The main search screen of PubChem



The structure drawing screen in PubChem

acquired the Beilstein database in March 2007 from the non-profit Beilstein-Institut (although Elsevier had been involved with the database's production and marketing since 1998), it is now integrated into the Compound Index of DG.

CrossFire Gmelin is another important database in DG, with information about inorganic chemicals. It has not yet been integrated with the Compound Index and must be searched separately.

MDL Comprehensive Medicinal Chemistry Database is derived from the Drug Compendium in Pergamon's Comprehensive Medicinal Chemistry (CMC). The MDL Comprehensive Medicinal Chemistry database provides 3-D models and important biochemical properties for more than 8,400 pharmaceutical compounds (1900 to the present). It allows scientists to view more than 7,500 bioactive compounds used as medicinal agents.

MDL Toxicity Database is a structure-searchable bioactivity database of toxic chemical substances. It contains data from studies on the toxicity, mutagenicity, skin and eye irritation, tumorigenicity and carcinogenicity, and effects on reproduction. References to the original publications reporting the toxicity data and to relevant review articles when applicable are also included.

MDL Metabolite Database consists of a database, a registrar, and a browser. All data in this system is organized into schemes that depict the transformations and fate of parent compounds. These could be drugs, agricultural chemicals, industrial chemicals, or environmental contaminants. While compiling xenobiotic transformations and metabolic schemes from reliable sources, the vendor also allows companies to register and store results of in-house metabolism studies and to share these results with researchers in the same enterprise. It is possible for a researcher to create structure-searchable metabolic schemes on a desktop and to upload them to the corporate database. This allows organizations to get better insights into metabolic activity and the fate

of candidates for drugs and to coordinate their discovery programs. Using the browser, scientists can search by structure to retrieve molecules of interest, as well as structural fragments or specific transformations.

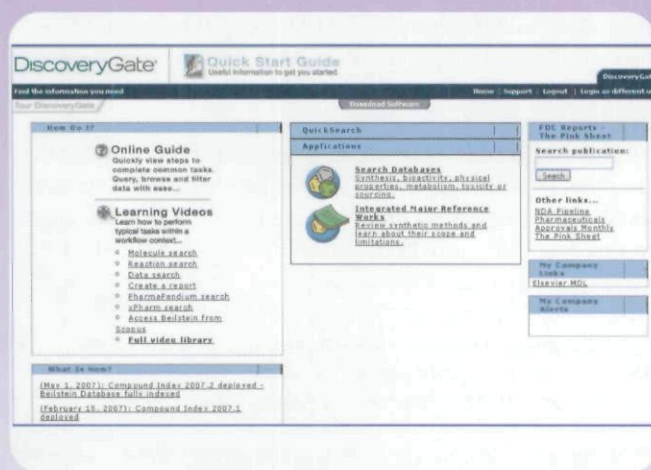
xPharm is a unique resource providing pharmacological data to researchers involved in the preclinical or discovery phase of drug development. A detailed article (Marie K. Saimbert, "xPharm: A Preclinical Information Portal for Discovery Researchers in Institutions of Higher Learning," *Journal of Electronic Resources in Medical Libraries*, 3(1), 2006, p. 63; www.mdl.com/products/pdfs/xpharm.pdf) provides information about this resource.

The main users of DG are the pharmaceutical, chemical, and biotechnology companies, as well as some academic institutions that can afford it. With its rich content, DG could be very useful to researchers not only in the chemistry field, but also to those working in the life sciences.

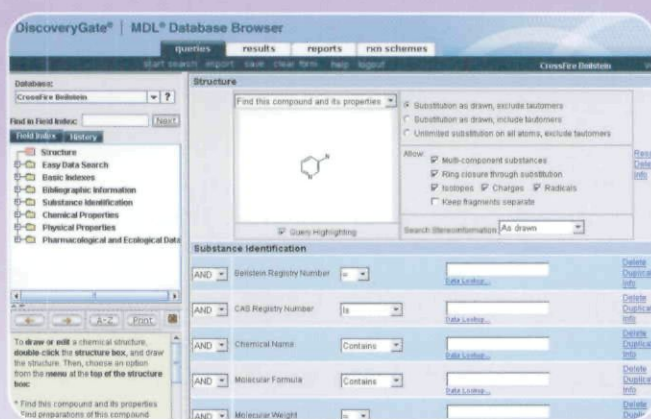
CHEMISTRY CENTRAL

Chemistry Central (www.chemistrycentral.com) is a new open access service for chemistry and related disciplines that was launched by BioMed Central in 2006. It allows free access to peer-reviewed articles published in the *Chemistry Central Journal* and the BioMed journals, as well as to articles published in independent journals that are using the BioMed publishing services.

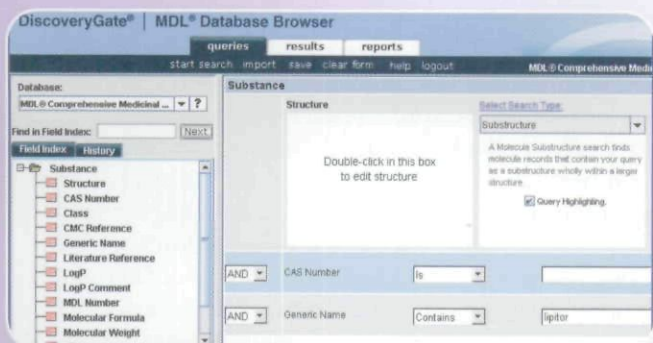
Chemistry Central currently provides access to the *Beilstein Journal of Organic Chemistry*, *BMC Biochemistry*, *BMC Chemical Biology*, *BMC Structural Biology*, *Carbon Balance and Management*, *Chemistry Central Journal*, and *Geochemical Transactions*. As with many other open access ventures, Chemistry Central tests new business models. Brian Vickery, editorial director of Chemistry Central, discussed these more fully in a recent interview (Svetla Baykoucheva, "Paving the Road to More Open Access for Chemistry: An Interview with Brian Vickery, Editorial Director of Chemistry,"



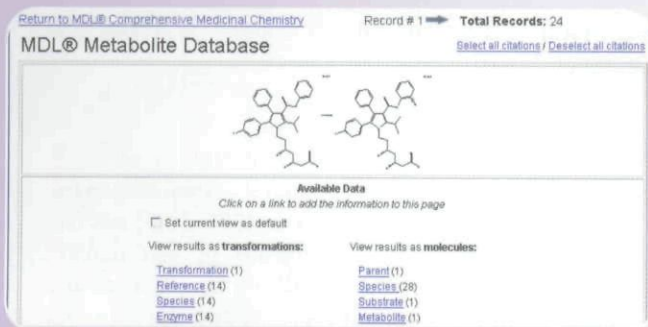
DiscoveryGate's main interface



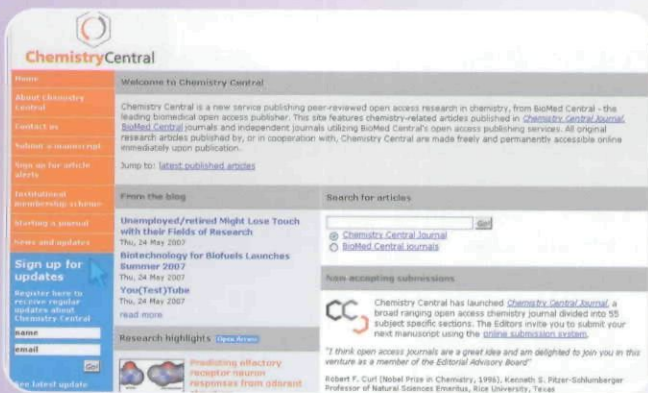
Search screen of CrossFire Beilstein in DG. The structure of beta-amino pyridine was drawn and used to search for information on this compound rather than searching by its name.



Search screen of the MDL Medicinal Chemistry database in DG. The search is for Lipitor, also known as Atorvastatin, a widely prescribed cholesterol-lowering drug.



The properties page of the MDL Metabolite Database, showing information about the metabolism of Lipitor



Chemistry Central's home page

Chemical Information Bulletin, V. 59, No. 1, pp. 13–15; <http://accinf.org/docs/publications/Interviews/Vickery/2007>.

NOT JUST FOR CHEMISTS

Although I have often used the phrase “chemical information” throughout this article, I would like to emphasize that chemistry and the life sciences have become so intertwined that the resources I discuss here can greatly benefit researchers in the life sciences too. Very often, though, they are not familiar with these resources and/or are not aware that their organizations subscribe to them. To help, a

comprehensive list of chemistry databases has been posted (www.chembiogrid.org/related/resources/about.html).

Researchers and practitioners in the biomedical field can find useful information about drugs through PubChem, which is now cross-indexed with the Compound Index of DG. Users of one of these two information systems can immediately see if related information exists in the other system. Only subscribers to DG, though, could get full access to both systems. PubChem users get free searching of both systems and full access to PubChem results.

ACCEPTANCE FACTORS

Several factors will have an impact on how DG is accepted by librarians and end users. It is an expensive service that many organizations cannot afford, especially with ever-shrinking library budgets. At the University of Maryland, which is not one of the wealthiest academic institutions, we have a campuswide license for DG that allows everyone affiliated with the university to use this service, even from home (through a virtual private network, or VPN). We were able to switch, for the same price, from the Commander platform for CrossFire Beilstein and CrossFire Gmelin to DiscoveryGate, which is a much richer resource. I include DG in all my classes for the chemistry courses I teach, and the students are fascinated by what they can do with it. I have also done several seminars for librarians, showing them how to find relevant and unbiased information about drugs.

The other issue with DG—learning how to use it—can be a long and frustrating experience, even for chemists. Elsevier MDL provides WebEx and on-site training, as well as excellent instructional materials, but DG is a sophisticated product that requires a lot of time and dedication on the part of anyone wanting to use it. The rewards for researchers, though, are enormous. There are some librarians and faculty members who are still reluctant to embrace DG; they continue to access CrossFire Beilstein and CrossFire Gmelin through the old Commander interface rather than through DG.

The collaboration between a free (PubChem) and a proprietary (DG) service is a new model for providing scientific information. It is of great benefit to researchers, as the content of different information systems can be used together, thus eliminating unnecessary, repetitious searches performed in separate databases.

Not that long ago, you would often hear people saying that the chemistry field had been lagging behind other disciplines in enjoying free access to information. The events discussed in this article are proving that this cliché is outdated.

Svetla Baykoucheva (sbaykouc@umd.edu) is head of the White Memorial Chemistry Library of the University of Maryland–College Park and the editor of the *Chemical Information Bulletin*, published by the Chemical Information Division of the American Chemical Society (ACS). For 8 years, from 1997 until 2005, she was manager of the ACS Library and Information Center in Washington, D.C.

Comments? Email letters to the editor to marydee@xmission.com.

Copyright of Online is the property of Information Today Inc. and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.