INFORMATICS

Chemistry's web of data expands

Patent information to be made publicly accessible amid worries about data quality.

BY RICHARD VAN NOORDEN

Other areas of science feast on free online data, but chemistry has been late to the party. Now it is catching up.

In the latest effort to provide free access to chemical information, the London-based company SureChem (owned by Digital Science, a sister company to Nature Publishing Group) said this week that it would release data on 10 million molecules patented by the pharmaceutical industry since 1976. Harvested automatically from some 20 million patents, the data could lower barriers to drug discovery by academic researchers.

The announcement, made on 26 March at the spring meeting of the American Chemical Society (ACS) in San Diego, California, follows a similar move by computing giant IBM last December. IBM deposited computer-harvested data on about 2.4 million small molecules into PubChem, the world's largest free chemistry repository, which is run by the US National Library of Medicine in Bethesda, Maryland.

Both data releases serve in part to promote the companies' subscription services for patent and structure analysis. But Michael Walters, a chemist working in academic drug discovery at the University of Minnesota in Minneapolis, thinks that the initiatives could mark "a sea change in the way in which patent data are accessed and analysed". The data should make it easier for chemists to see which bioactive molecules have drawn the attention of the drug industry — and to explore new drug targets by designing compounds that are not named in patents.

CHEMISTRY ON THE INTERNET

Academic drug discovery will get another boost in September, when a consortium of eight pharmaceutical firms, three biotechnology companies and a number of leading informaticians releases its own free, online drug-discovery platform, the Open Pharmacological Concepts Triple Store (OpenPHACTS). Supported in part by a €10-million (US$13-million) grant from the European Union's Innovative Medicines Initiative, the website will link data on small molecules and their biological effects, to provide a library of compounds that anyone can download and explore.

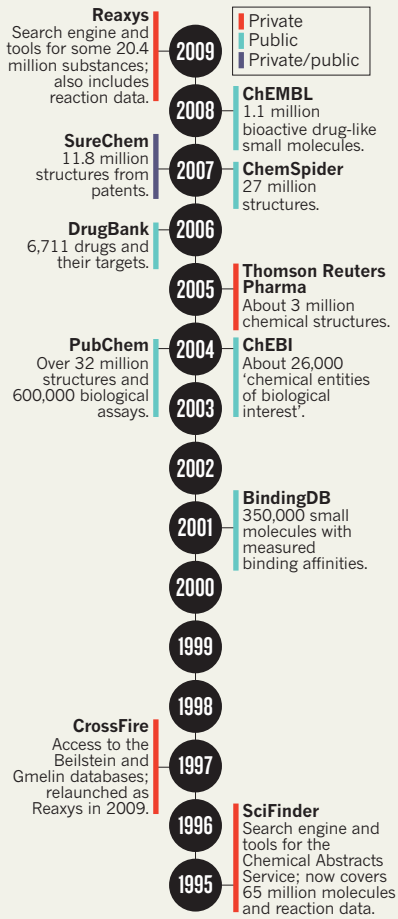
Unlike biologists, who are swamped by free databases on genes and proteins, chemists have always expected to pay for their data. Until a few years ago, the market in chemical information was monopolized by the ACS Chemical Abstracts Service, a manually curated registry that now holds more than 65 million structures, charges individual users thousands of dollars a year for access and does not allow large downloads or repurposing of its information. Its SciFinder service offers tools to make sense of the data. Similar analytical services are sold by firms such as IBM, Thomson Reuters and Elsevier in Amsterdam, which offers the Reaxys tool (see 'Chemistry breaks free').

But in 2004, the US National Institutes of Health (NIH) created PubChem, into which anyone can deposit data on structures and their biological activity. In 2005, the ACS sought to restrict PubChem's reach to molecules characterized by NIH-funded researchers, but was unsuccessful. The database has now grown to more than 32 million structures and, according to PubChem, has roughly 100,000 unique users per day. In 2007, another free repository, ChemSpider, was created by chemist Antony Williams; in 2009, it was purchased by the UK Royal Society of Chemistry in London and it now holds 27 million structures.

These two databases are now the Internet's main chemistry hubs, linking out to other sources of free online information, such as ChEMBL, a database of about 1 million bioactive drug-like small molecules hosted by the European Bioinformatics Institute in Hinxton, UK. The result is a web of interconnected free data, contrasting with high-quality but closed-off subscription databases.

CHEMISTRY BREAKS FREE

A web of free, online chemical databases has sprung up over the past decade, alongside established subscription offerings.



QUALITY CONTROL

But as biologists already know, free online data can be poorly curated — and chemical data is no exception. In a project presented at the ACS meeting in San Diego, Williams and his colleagues showed how five large online databases disagreed on the structures of 150 top-selling drugs: the best got 99% of structures correct, whereas the worst managed only 76%. In fact, notes Williams, Wikipedia proved the most reliable source of structural information in that experiment — mostly because of an effort to clean up the site's 13,000 pages about chemicals.

Williams says that more chemists need to concentrate on data standards and start actively correcting information online. Christopher Southan, a chemical-information consultant in Gothenburg, Sweden, who previously worked for drug giant AstraZeneca, agrees: "The danger is that now people are connecting all this online chemical and biological information together, and there's so much noise and imprecision that they're building a house of cards."

For now, cheminformatics pioneers are excited by the potential of free online information, and are keen to raise awareness of the possibilities. "The average medicinal chemist was weaned on SciFinder," says Southan. "I can't see them rushing into online data — but slowly but surely, anyone working in academic drug discovery will start to use it." ■

CORRECTION

The graphic 'Frequent fliers' in the News story 'Activists ground primate flights' (*Nature* **483,** 381–382; 2012) should have listed the American Anti-Vivisection Society instead of PETA in its source list.