



A perspective of publicly accessible/open-access chemistry databases

Antony J. Williams

ChemZoo Inc., 904 Tamaras Circle, Wake Forest, NC 27587, United States

The Internet has spawned access to unprecedented levels of information. For chemists the increasing number of resources they can use to access chemistry-related information provides them a valuable path to discovery of information, one which was previously limited to commercial and therefore constrained resources. The diversity of information continues to expand at a dramatic rate and, coupled with an increasing awareness for quality, curation and improved tools for focused searches, chemists are now able to find valuable information within a few seconds using a few keystrokes. This shift to publicly available resources offers great promise to the benefits of science and society yet brings with it increasing concern from commercial entities. This article will discuss the benefits and disruptions associated with an increase in publicly available scientific resources.

Ask a chemist how often they use the Internet to search for science-related information online and it would be fair to expect the answer would be daily. Certainly the web now dominates as the primary portal to query for general information and data. Yet, despite the tremendous growth in scientific Internet resources, the potential for providing useful chemistry information and data have only recently started to be tapped. Bioinformatics has led the charge in providing online access to data far ahead of the efforts in Chemistry. Open-access databases, like GenBank and the Protein Data Bank, have been assisting biologists to translate gene and protein sequences into biological relevance for well over two decades. Some of the responsibility for the differences in efforts has commonly been put onto the shoulders of publishers in chemistry, whether for scientific articles or commercial chemistry databases. Nevertheless, societal thrust, evangelists and group efforts are forcing both free and open access (*vide infra*) to chemistry-related information.

There are many indexes of chemistry databases online and this article is not intended to be yet another. Rather, this article will review both the availability and capabilities of online chemistry databases, particularly those offering free or open access. The progress in the availability of freely accessible information is highly enabling and advantageous for the advancement of science and our future well-being but probably seen as a disruptive force for commercial bodies but, this shift is needed, not

only to facilitate improved access to information in academia and government laboratories but also in commercial organizations feeling the pressure of poor performance in the discovery and development processes. Pharmaceutical companies, in particular, should welcome improved access to chemistry-related information as their business dominance is taken to task with drugs coming off patent and no replacement blockbuster drugs in the pipeline.

In keeping with the web-based nature of this article, the majority of references will actually be to Internet resources. At the time of submission all references were active but, as is the nature of the Internet, these resources will age and may disappear.

What is open versus free access?

There is much confusion around the differences between open access (OA) versus free access (FA). This is also accompanied by corporate protectionism and political battles which have found their way to Capitol Hill. Both OA and FA offer a great opportunity to the advancement of science by sharing data, information and knowledge as it is created. With these in place, publishing houses and institutional repositories of information, specifically structure databases with related information, are threatened by the potential impact on their business model and associated revenues.

The first major international statement on open access was the Budapest Open Access Initiative (BOAI), in February 2002 (FAQs

online). The definition of open access from the BOAI frequently asked questions website is as follows: “By ‘open access’ to this literature, we mean its free availability on the public Internet, permitting any user to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited.” With this definition in mind, one can understand the potential impact to publishing business models. There are other concerns in the OA model but these are reviewed elsewhere.

Free access is not equivalent to OA, but, by definition, OA assumes free access as part of its model. This author could not find an equivalent definition for free access online, but suggests the following ‘free access is access that removes price barriers but not necessarily any permission barriers.’ Many organizations provide free access to their publications, content and data but the rules under which the information is made available can differ between the hosts. Two specific examples are cited:

- (1) The Royal Society of Chemistry offers access to thousands of articles *via* its free access policies. Copyright remains with the society.
- (2) The SureChem patent portal (*vide infra*) provides chemical structure-searchable access to patents online, free of charge. They do, however, sell their content to organizations, provide secure portal access and data export for organizations for a fee and have a business model enabling both free access and for profit activities.

To most people free access is not inferior to open access – just different. Most scientists are overjoyed to have free access to information and data previously not available and will willingly use such resources.

Political decision makers are now directly engaged in reviewing the benefits of OA (and even FA) to science and society and limiting open access. In November of 2007 President Bush vetoed a bill that aimed to make all National Institutes of Health (NIH)-funded research publications freely available on the web and the furor that arose from the appearance of the PubChem database as competition to other repositories (specifically the Chemical Abstracts Service) has caused a significant rift between members of the ACS and the management team. While these challenges are yet to be fully navigated, it is clear that publishers will need to modify their business models to address the drive toward more FA and OA resources. Meanwhile, groups such as the Public Library of Science and Chemistry Central, discussed recently, are leading the charge for more open access.

For the purpose of this article, both free access and open access will be defined as no barriers to using the system in order to derive value: no forced registration or logins in order to use the system. This does not necessarily mean that there might not also be fee-based services associated with the resource – a site can have free and fee-based services on offer simultaneously.

Free- and open-access online chemistry databases

Depending on the definition of a database, there are many freely available on the web. It is not uncommon to hear chemists comment about a downloadable database of structures. This commonly refers to a file containing a number of structures, commonly tens to tens of thousands in the Structure Data Format (SDF). In general, these files are then imported to a database for viewing. There are many hundreds of SDF files available online and, based on experience, a week of downloading, importing and de-duplication could easily provide at least 15 million unique structures. As this article goes to press the PubChem dataset now numbers 18 million unique structures and can be downloaded as a single data source. These SDF files are commonly provided by chemical vendors for the purpose of facilitating commercial sales. Such files can contain structure identifiers (names and numbers), experimental or physical properties, file-specific identifiers and, commonly, pricing information. Since they are assembled in a heterogeneous manner, such data are plagued with inconsistencies and data quality issues.

There are, however, a number of online database resources offering access to valuable data and knowledge. Some of these could be thought of as ‘linkbases’, a term which for the sake of this article can be considered as a repository of molecular connection tables linking out to multiple sources of data and associated information. Although this review cannot be exhaustive we will examine a number of these free resources and the value they are starting to deliver to chemists.

PubChem

The highest profile, online database is certainly PubChem. NIH launched the database in 2004 as part of a suite of databases supporting the New Pathways to Discovery component of their roadmap initiative. PubChem archives and organizes information about the biological activities of chemical compounds into a comprehensive biomedical database to support the Molecular Libraries initiative component of the roadmap. PubChem is the informatics backbone for the initiative and is intended to empower the scientific community to use small molecule chemical compounds in their research.

PubChem consists of three databases (PubChem Compound, PubChem Substance and PubChem Bio-Assay) connected together and incorporated into the Entrez information system of the National Center for Biotechnology Information (NCBI). PubChem Compound contains 18 million unique structures and provides biological property information for each compound through links to other Entrez databases (see Fig. 1 for an example). PubChem Substance contains records of substances from depositors into the system. These are publishers, chemical vendors, commercial databases and other sources. It provides descriptions of chemicals and links to PubMed, protein 3D structures and screening results. The PubChem Compound database contains records of individual compounds. PubChem BioAssay contains information about bioassays using specific terms pertinent to the bioassay.

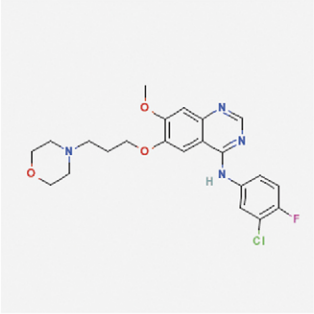
PubChem can be searched by alphanumeric text variables, such as names of chemicals, property ranges or by structure, substructure or structural similarity. As of December 2007, its content is approaching 38.7 million substances and 18.4 million unique structures.

NCBI PubChem Information on biological activities of small molecules

HOME SEARCH SITE MAP PubMed Entrez Structure GenBank PubChem Help

Search PubChem Compound [] GO

Compound Summary:



CID: 123631 [] []

BioActivity: Summary []
All: 3 Links
Active: 3 Links

Protein Structures: 3 Links []

NLM Toxicology: Link []

Substances: []
All: 55 Links
Same: 19 Links
Mixture: 36 Links


Related Compounds: []
Same, Connectivity: 3 Links


Similar Compounds: 526 Links []


Structure Search []

MeSH Synonyms Properties Descriptors Category Exports

Medical Subject Annotations: (Total: 1) []

 **gefitinib**
Pharmacological Action:
Antineoplastic Agents
Protein Kinase Inhibitors

 PubMed via MeSH

 PubMed MeSH Keyword Summary []

Depositor-Supplied Synonyms: (Total: 27) []

FIGURE 1

The compound summary page for Gefitinib in PubChem. Page 1 only is shown (<http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=123631>).

The value of PubChem has been lauded by a number of advocates, especially those evangelizing open access and open data initiatives. Certainly, such a source of data opens up new possibilities in regards to data mining and extraction. Zhou *et al.* [1] concluded that the system has an important role as a central repository for chemical vendors and content providers. This enables evaluation of commercial compound libraries and saves biomedical researchers from the work associated with gathering and searching commercial databases. They concluded that the hit-to-lead decision-making process in drug discovery programs can certainly benefit from the ongoing annotation service provided by PubChem. They identified that over 35% of the 5 million structures from chemical vendors or screening centers currently found in the PubChem database are not present in the CAS database and suggested that parties, such as CAS, could benefit from inclusion of the open-access chemical and biological data found in PubChem into their own repositories.

PubChem continues to grow in stature, content and capability. The number of both substances and unique chemical structures continues to grow on an annual basis. The bioassay data resulting from the NIH Roadmap initiative are likely to continue to grow and PubChem will assume a prominent role in distributing the data in a standard format. Despite the obvious value of PubChem,

the platform has not been without its detractors. There have been numerous heated debates regarding the position of CAS relative to the resource and the reader is referred elsewhere for opinions and commentaries. Others have commented on their concerns regarding the quality of the data content within PubChem. Shoi-chet [2] has commented that the screening data are less rigorous than those in peer-reviewed articles and contain many false positives. Deposited data are not curated and so mistakes in structures, units and other characteristics can, and do, occur. Shoi-chet worries that chemists who use PubChem may be sent on a wild goose chase. Williams and others have pointed to the accuracy of some of the identifiers associated with the PubChem compounds. The problems arise from the quality of submissions from the various data sources. There are thousands of errors in the structure-identifier associations because of this contamination and this can lead to the retrieval of incorrect chemical structures. It is also common to have multiple representations of a single structure because of incomplete or total lack of stereochemistry for a molecule. PubChem is not resourced to resolve these issues and these unfortunate errors have, in general, arisen due to chemical vendors seeing PubChem as a potential path to commercial gain rather than its primary focus of supporting the Molecular Libraries initiative.

eMolecules

eMolecules, originally released as Chmoogle before legal wranglings with a famous search engine provider, is a commercial entity that offers a free online database of over 7 million unique chemical structures derived from more than 150 suppliers and offers scientists and procurement staff a path to identifying a vendor for a particular chemical compound. By providing access to compounds for purchase they are providing a free access online service, similar to those of commercial databases, such as Symyx Available Chemical Directory, CAS's ChemCats and Cambridge-Soft's ChemACX, to name just three. They also offer a service to chemical suppliers to customize a search page for the vendor and request online quotes.

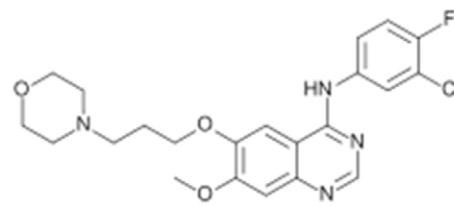
The system offers access to more than 4 million commercially available screening compounds and hundreds of thousands of building blocks and intermediates and access *via* ChemGate to NMR, MS and IR spectra from Wiley-VCH for over 500,000 compounds, a fee-based service. eMolecules also provides links to many sources of data for spectra, physical properties and biological data, including the NIST WebBook, the National Cancer Institute, DrugBank and PubChem.

eMolecules is presently fairly limited in its scope to offering a path to the purchase of chemicals and links to the more popular government databases. Nevertheless, the site is popular with chemists who are searching for chemicals and the interface is intuitive and easy to use, a key element in attracting users.

Wikipedia

For this author, Wikipedia certainly represents an important shift in the future access of information associated with small molecules. A wiki is a type of computer software that allows users to easily create, edit and link web pages. A wiki enables documents to be written collaboratively, in a simple markup language using a web browser, and is essentially a database for creating, browsing and searching information. There are now thousands of pages describing small organic molecules, inorganics, organometallics, polymers and even large biomolecules. Focusing on small molecules in general, each one has a Drug Box or a Chemical infobox. The drug box provides identifier information, in this case each identifier linking out to a related resource, and chemical data, pharmacokinetic data and therapeutic considerations. At present there are approximately 8000 articles with a chembox or drugbox [3], with between 500 and 1000 articles added since May. The detailed information offered on Wikipedia regarding a particular chemical or drug can be excellent, see Fig. 2, or weak.

There have been comments regarding the quality of Wikipedia in recent months. Nevertheless, there are many dedicated supporters and contributors to the quality of the online resource. The chemistry articles are not without issue and the drug and chemboxes, specifically, have been shown to contain errors. However, the advantage of a wiki is that changes can be made within a few keystrokes and the quality is immediately enhanced. This community curation process makes Wikipedia a very important online chemistry resource, whose impact will only expand with time. Wikis, and their usage for the purpose of communication in chemistry, will be discussed in more detail in a separate article in this publication.



Gefitinib

Systematic (IUPAC) name

N-(3-chloro-4-fluoro-phenyl)-7-methoxy-6-(3-morpholin-4-ylpropoxy)quinazolin-4-amine

Identifiers

CAS number 184475-35-2

ATC code L01XE02

PubChem 123631

DrugBank APRD00997

Chemical data

Formula $C_{22}H_{24}ClFN_4O_3$

Mol. mass 446.902 g/mol

Pharmacokinetic data

Bioavailability 59% (oral)

Protein binding 90%

Metabolism Hepatic (mainly CYP3A4)

Half life 6–49 hours

Excretion Faecal

Therapeutic considerations

Pregnancy cat. C (Au), D (U.S.)

Legal status S4 (Au), POM (UK), R-only (U.S.)

Routes Oral

FIGURE 2

The DrugBox for Gefitinib from Wikipedia (<http://en.wikipedia.org/wiki/Gefitinib>).

DrugBank

DrugBank [4] is a very valuable resource blending both bioinformatics and cheminformatics data and combines detailed drug (i.e. chemical) data with comprehensive drug target (i.e. protein) information. The Drugbank is hosted at the University of Alberta, Canada and supported by Genome Alberta & Genome Canada, a private, non-profit corporation, whose mandate is to develop and implement a national strategy in genomics and proteomics research. At the time of release in 2006, the database

contained >4100 drug entries including >800 FDA-approved small molecule and biotech drugs as well as >3200 experimental drugs. Additionally, >14,000 protein, or drug target sequences, are linked to these drug entries. Each DrugCard entry contains >80 data fields, with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data. Many data fields are linked to other databases (KEGG, PubChem, ChEBI, PDB and others). The database is fully searchable, supporting extensive text, sequence, chemical structure and relational query searches.

DrugBank has been used to facilitate *in silico* drug target discovery, drug design, drug docking or screening, drug metabolism prediction, drug interaction prediction and general pharmaceutical education. The version 2.0 release of DrugBank [5] was scheduled for January 2008 and will add approximately 800 new drug entries and each DrugCard entry will contain over 100 data fields, with half of the information being devoted to drug/chemical data and the other half devoted to pharmacological, pharmacogenomic and molecular biological data.

DrugBank has committed to a semi-annual updating schedule to allow information on newly approved and newly withdrawn drugs to be kept current. The intention is to acquire experimental spectral data (NMR and MS specifically) and to expand the coverage of nutraceuticals and herbal medicines.

Human Metabolome Database

The Human Metabolome Database (HMDB) [6] contains detailed information about small molecule metabolites found in the human body and is used by scientists working in the areas of metabolomics, clinical chemistry and biomarker discovery. The database is developed by the same group responsible for DrugBank and, in many ways, parallels the capabilities. The database links chemical data, clinical data, and biochemistry data. The database currently contains nearly 3000 metabolite entries and each MetaboCard entry contains more than 90 data fields devoted to chemical, clinical, enzymatic and biochemical data. Many data fields are hyperlinked to other databases, as with DrugBank, and the database supports extensive text, sequence, chemical structure and relational query searches.

HMDB certainly offers a valuable resource to scientists investigating the metabolome. In general, the data are of high quality, but unfortunately have inherited some of the quality issues from PubChem regarding identifiers associated with chemical structures (see page 17 of this online presentation).

Zinc

Zinc is a free database of commercially available compounds for virtual screening. The library contains over 4.6 million molecules, each with a 3D structure and gathered from the catalogs of compounds from vendors. All molecules in the databases are assigned biologically relevant protonation states and annotated with molecular properties. The database is available for free download in several common file formats and a web-based search page including a molecular drawing interface allows the database to be searched. This database facilitates the delivery of virtual screening libraries to the community and continues to grow on an annual basis as new chemicals are supplied by vendors.

SureChem

SureChem provides chemically intelligent searching of a patent database containing over 8 million US, European and World Patents. Using extraction heuristics to identify chemical and trade names and conversion of the extracted entities to chemical structures using a series of name to structure conversion tools SureChem has delivered a database integrated to nearly 9 million individual chemical structures (see Fig. 3 for an example search result on Gefitinib). The free access online portal allows scientists to search the system on the basis of structure, substructure or similarity of structure as well as text-based searching expected for patent inquiries. SureChem provides a 24-hour turnaround on the deposition of new patent data as they are issued thus providing essentially immediate access to structure-based searching of patent space.

The SureChem database contains a large number of fragment structures, as well as normal structures, and this is to be expected because of the nature of patents. Search options allow exclusion of these fragments before searching. Although name to structure conversion is not perfect and there will be some errors as a result of the process the, SureChem team have already indicated their intention to enable real-time curation of the data by users to facilitate annotation of the data and an ongoing community improvement process as the system is used. The SureChem structure database has also been published to ChemSpider (*vide infra*) facilitating searches from multiple systems.

ChemSpider and ChemRefer

ChemSpider was released, by this author and associated team, to the public in March 2007 with the lofty goal of 'building a

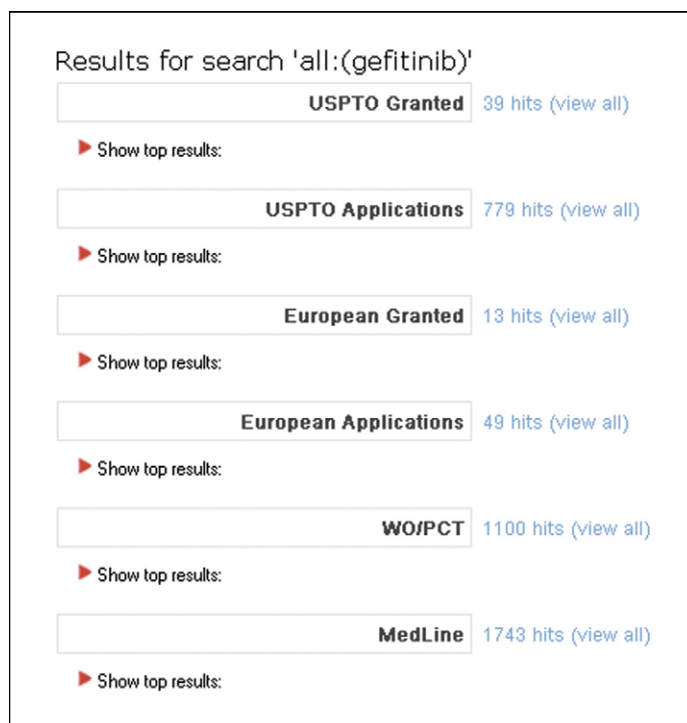


FIGURE 3

An example patent results page for SureChem (<http://www.surechem.org/>), a chemically intelligent, free access online patent search. A search on Gefitinib provided an abundance of Patent information to examine.

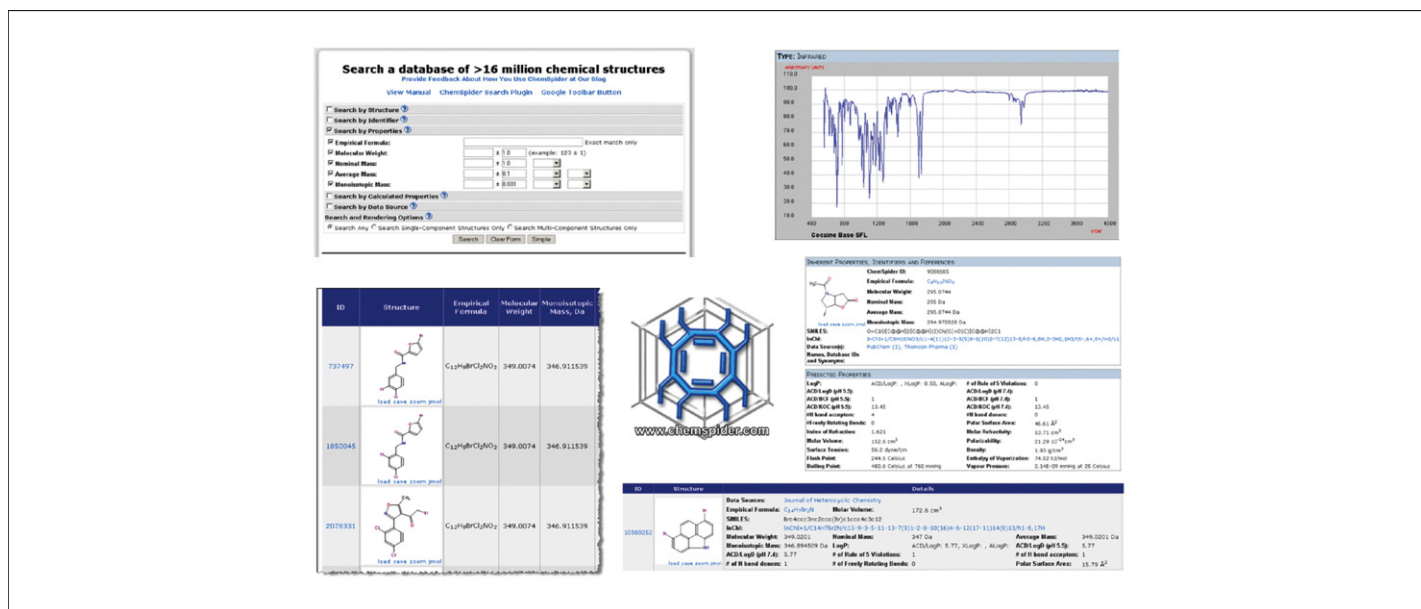
ChemSpider has enabled capabilities not available to users of any of the other systems. These include real time curation of the data, association of analytical data with chemical structures, real-time deposition of single or batch chemical structures (including with activity data) and transaction-based predictions of physico-chemical data. The ChemSpider developers have made available a series of web services to allow integration to the system for the purpose of searching the system, as well as generation of InChI identifiers and conversion routines.

The system integrates text-based searching of open access articles. This capability, known as ChemRefer, presently searches over 150,000 OA Chemistry articles. The index is estimated to grow to a quarter of a million articles by early 2008. Efforts are presently underway to extract chemical names from the OA articles and convert these names to chemical structures using name to structure conversion algorithms. These chemical structures will be deposited back to the ChemSpider database, thereby facilitating structure and substructure searching in concert with text-based searching.

ChemSpider has an intention different from all other resources reviewed in this article – a focus on, and commitment to, community curation. The social community aspects of the system,

The list of databases and resources reviewed above is representative of the type of information available online. Other highly regarded databases frequented by this author include The Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network hosted by the US Environmental Protection Agency, KEGG and CheBI. There are also many other resources available and the reader is referred to one of the many indexes of such databases available on the Internet to identify potential resources of interest, for example, the [here](#) and [here](#).

An increasing number of public databases will continue to become available but the challenge, even now, is how to integrate and access the data. In a recent article, Willighagen *et al.* [7]



Graphical elements associated with ChemSpider (<http://www.chemspider.com/>), a structure centric community for chemists. The system presently contains over 20 million unique chemical compounds.

introduce userscripts to enrich biology and chemistry-related web resources by incorporating or linking to other computational or data sources on the web. They showed how information from web pages can be used to link to, search, and process information in other resources, thereby allowing scientists to select and incorporate the appropriate web resources to enhance their productivity. Such tools connecting open chemistry databases and user web pages is the ideal path to more highly integrated information sharing.

Will access to free resources challenge the commercial sector?

A number of organizations generate sizeable revenues from the creation of chemistry databases for the life sciences industry. The annual fees for accessing this information probably exceeds half a billion dollars. The Chemical Abstracts Service alone generates revenue in excess of \$250 million dollars. The primary advantage of commercial databases is that they have been manually examined by skilled curators, addressing the tedious task of quality data-checking. Certainly, the aggregation of data from multiple sources, both historical and modern, from multiple countries and languages and from sources not available electronically, is a significant enhancement over what is available via Internet searching. The question remains how long will this remain an issue?

Modern scientists are primarily concerned with what has happened in the more immediate history, specifically in new areas of science. Interest in antique historical records rarely exists for modern forms of science. In many ways what is not available online is probably not of interest to more and more researchers, whether this be appropriate or not. Increasingly librarians at universities and large pharma are giving away their print collections in favor of electronic repositories of chemical journals. Despite some views that search engine models may be parasitic in their nature Internet search engines are likely to increasingly be the first port of call for the majority of scientists for two simple reasons – they are fast and they are free. As has already been discussed, chemically searchable patents are now available online, at no charge. The present database, while containing errors, still provides value to its users and offers access to information not available in other curated patent systems, specifically access to prophetic compounds and 48-hour turnaround on the updating of patent data after they are issued. Such immediate access to new data, the potential of the web to provide alerts on the basis of particular structure class and the interest of the modern scientist in more recent information, offers a significant business advantage to such offerings. In terms of data quality

issues the Internet generation has already demonstrated a willingness to curate, modify and enhance the quality of content as modeled by Wikipedia. With the appropriate enhancements in place online curation and markup of the data in real time can quickly address errors in the data as has already been demonstrated by the ChemSpider system.

With the improvements promised by the semantic web then if there are data of interest to be found the search engines will facilitate them. As the President of CAS, Robert Massie commented 'Chemical Abstracts has to be better than Google, better than our competitors, so that we can charge a premium price.' It is this author's judgment that this will become increasingly difficult as science and technology progress and redesigning commercial database business models will become more necessary.

Conclusion

Increasing access to free- and open-access databases of both chemistry and biological data is certainly impacting the manner by which scientists access information. These databases are additional tendrils in the web of Internet resources that continue to expand in their proliferation of freely accessible data and information such as patents, open and free access peer-reviewed publications and software tools for the manipulation of chemistry-related data. As data-mining tools expand in their capabilities and performance the chemistry databases available online now, and in the future, are likely to offer even greater opportunities to benefit the process of discovery. As these databases grow in both their content and quality there will be challenging times ahead for those parties presently in tension regarding the commercial business models of publishers versus the drive toward more freely available data. This author remains hopeful that the journey ahead can bring benefit to all parties.

It should be noted that following the initial preparation of this article President Bush signed a bill requiring the National Institutes of Health to mandate open access for NIH-funded research. This act is likely to lead to a catalytic growth in open access exposure.

Acknowledgements

The author would like to acknowledge the active participation of many Open Data, Open Source, Open Standards (ODOSOS) contributors for participating in multi-forum discussions regarding the changing environment of open and free access. I also extend my thanks to the many passionate individuals and teams who have contributed to the databases and systems reviewed herein. Your efforts are facilitating better science and deserve acknowledgment.

References

- 1 Zhou, Y. *et al.* (2007) Large-scale annotation of small-molecule libraries using public databases. *J. Chem. Inf. Model.* 47, 1386–1394
- 2 Baker, M. (2006) Open-access chemistry databases evolving slowly but not surely. *Nat. Rev. Drug Discov.* 5, 707–708
- 3 Martin Walker, member of Wikipedia: Chemistry (private communication)
- 4 Wishart, D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672
- 5 Wishart, D.S. *et al.* (2007) DrugBank: a knowledgebase for drugs, drug actions and drug targets, *Nucleic Acids Res.* Nucleic Acids Research Advance Access published online on November 29, 2007 (<http://nar.oxfordjournals.org/cgi/content/abstract/gkm958v1>)
- 6 Wishart, D.S. *et al.* (2007) HMDB: The Human Metabolome Database. *Nucleic Acids Res.* 35, D521–D526
- 7 Willighagen, E.L. *et al.* (2007) Userscripts for the Life Sciences, BMC Bioinformatics, 8, 487. Published online on December 2007 (<http://www.biomedcentral.com/1471-2105/8/487>)