# Chemical structure indexing of toxicity data on the Internet: Moving toward a flat world

Ann M Richard[1]*, Lois Swirsky Gold[2] & Marc C Nicklaus[3]

**Addresses**
[1]National Center for Computational Toxicology
US Environmental Protection Agency
MD D343-03
Research Triangle Park
NC 27711
USA
Email: richard.ann@epa.gov

[2]University of California, Berkeley
Department of Molecular and Cell Biology
EO Lawrence Berkeley National Laboratory
Life Sciences Division
MS 946
Berkeley
CA 94720
USA

[3]Center for Cancer Research
National Cancer Institute
National Institutes of Health, DHHS
Laboratory of Medicinal Chemistry
NCI-Frederick
376 Boyles Street
Frederick
MD 21702
USA

*To whom correspondence should be addressed

**Current Opinion in Drug Discovery & Development** 2006 **9**(3):314-325

*Standardized chemical structure annotation of public toxicity databases and information resources is playing an increasingly important role in the 'flattening' and integration of diverse sets of biological activity data on the Internet. This review discusses public initiatives that are accelerating the pace of this transformation, with particular reference to toxicology-related chemical information. Chemical content annotators, structure locator services, large structure/data aggregator web sites, structure browsers, International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChI) codes, toxicity data models and public chemical/biological activity profiling initiatives are all playing a role in overcoming barriers to the integration of toxicity data, and are bringing researchers closer to the reality of a mineable chemical Semantic Web. An example of this integration of data is provided by the collaboration among researchers involved with the Distributed Structure-Searchable Toxicity (DSSTox) project, the Carcinogenic Potency Project, projects at the National Cancer Institute and the PubChem database.*

***Keywords*** Carcinogenic Potency Database, chemical structure, chemoinformatic, data mining, structure annotation, structure databases, toxicity data models, toxicity prediction

## Introduction

In his recently published book, *The World is Flat*, Thomas Friedman describes a climate of rapidly transforming globalization, extending from all areas of commerce to education and politics [1]. Central drivers to this transformation include the following: (i) the break down of barriers to the access, exchange and integration of ideas and information; (ii) the adoption of standards and technologies to fuel such changes; and (iii) 'glocalization', a term coined by Friedman that alludes to the integration of localized expertise into the global economy. These drivers and trends apply to the spectrum of scientific inquiry in terms of the needs and advances of chemoinformatics, bioinformatics, drug discovery, systems biology, etc [2,3,4•,5,6]. In contemplating the status of publicly available chemical toxicity databases, in their varied complexity, composition and formats, and the corollary objective of improving the toxicity predictions that these databases fuel, the concepts outlined by Friedman could not be more apt: the world is 'flattening', and chemical structure annotation is playing an increasingly important role as a key 'flattener' and top-level integrator of diverse biological activity data.

Advances are currently being made on several fronts that are beginning to break down the traditional barriers that exist between disparate information domains in the study of the toxicity of chemicals. These advances are expected to lead to enriched public resources and capabilities, and are driven partly by cross-cutting technologies (eg, genomics and bioassay profiling), partly by the increasing participation of toxicity domain experts who contribute to formal data model construction, and partly by the increasing levels of aggregation, annotation and integration taking place at the level of chemical identification. Toward the greater goal of screening chemicals for a wide range of toxicity endpoints of potential interest, publicly available resources encompassing a wide spectrum of biological and chemical data space must be effectively harnessed using existing and evolving Internet technologies (ie, data must be systematized, integrated and mined), if long-term screening and prediction objectives are to be achieved. This article builds on recent reviews by other authors and focuses particularly on the increasing role played by standardized chemical structure annotation as a top-level indexing and mining metric for public sources of toxicology data.

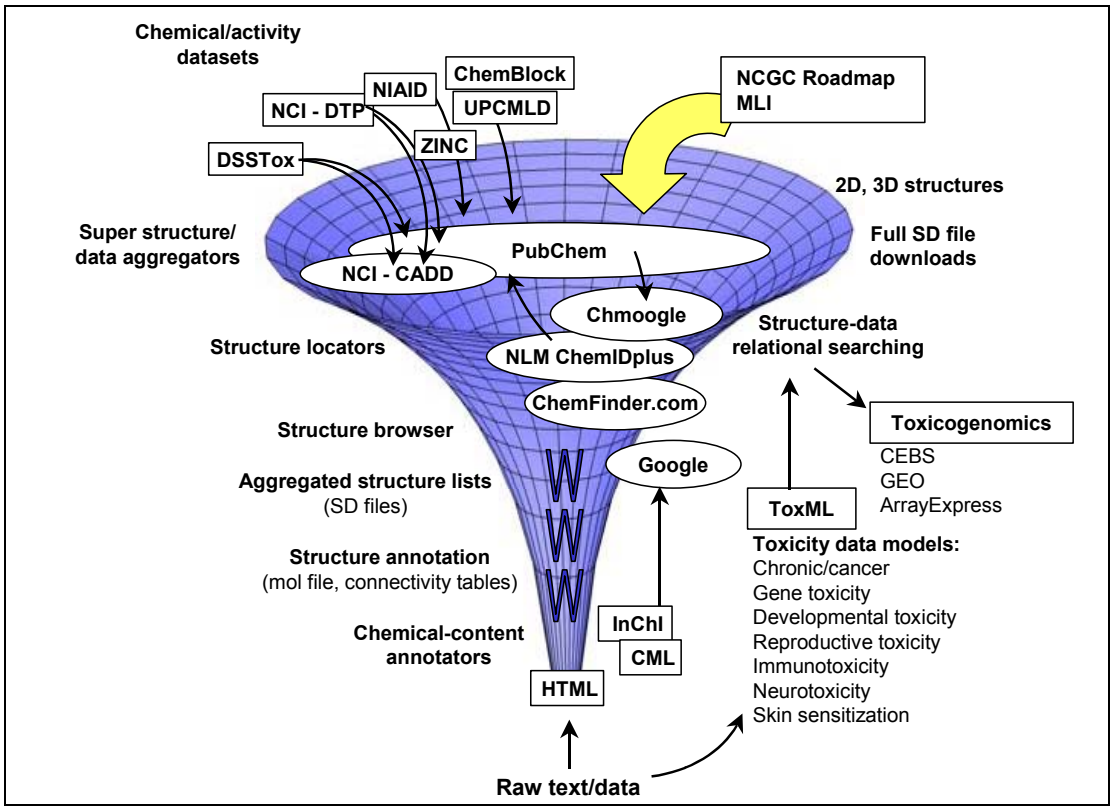
## Chemical annotation and aggregation

Figure 1 portrays a series of steps involved in the transformation of chemical information on the World Wide Web from raw text to curated to structure-searchable formats, at various levels of annotation and aggregation. Outlined below are a number of advances and public initiatives that are accelerating the pace of this transformation, both in general and with particular reference to toxicology-related chemical information.

### *Chemical content annotation*

The base of Figure 1 represents the vast majority of the World Wide Web chemical-related toxicology content as it currently exists, that is, in HyperText Markup Language (HTML) raw text format with no formal, standardized chemical annotation. Many prominent web researchers are advocating that the next generation of the Internet,

**Figure 1. Schematic of structure-annotated chemical/activity data on the World Wide Web.**




The figure provides an illustration of the steps involved in the progression of chemical/structure annotation and data aggregation that lead to broad-based chemical structure searching and data mining of chemical/biological information on the Internet.
**2D** Two-dimensional, **3D** three-dimensional, **CADD** computer-aided drug discovery, **CEBS** Chemical Effects in Biological Systems, **CML** Chemical Markup Language, **DSSTox** Distributed Structure-Searchable Toxicity database, **DTP** Developmental Therapeutics Program, **GEO** Gene Expression Omnibus, **HTML** HyperText Markup Language, **InChI** International Union of Pure and Applied Chemistry International Chemical Identifier, **MLI** Molecular Libraries Initiative, **NCGC** National Institutes of Health Chemical Genomics Center, **NCI** National Cancer Institute, **NIAID** National Institute of Allergy and Infectious Diseases, **NLM** National Library of Medicine, **SD** structure-data, **ToxML** an XML database standard based on toxicity controlled vocabulary, **UPCMLD** University of Pittsburgh Center for Chemical Methodologies and Library Development, **ZINC** database of commercially available compounds for virtual screening.

termed the 'Semantic Web' [7], should move away from a heavy reliance on pure textual content and text searching toward greater systematic content annotation, to enable relational search querying and data mining. The long-term objectives of greater relational capability with respect to chemical content on the World Wide Web can be viewed as being part of this much wider goal for the Internet of the future. The first technological hurdle to be overcome is in the processing and annotation of raw textual data containing chemical content.

The technology currently exists to automatically process raw text information, and to recognize and annotate anything resembling a chemical name with a chemical structure or a text-based representation of the chemical structure. Researchers at IBM Corp have demonstrated this capability using the IBM WebFountain software [8], which is based in part on the LexiChem and Picto technologies from OpenEye Scientific Software [9], by applying it to the chemical structure-annotation of a large public web site, that is, the 8 million pages of the US Patent Library [10]. In consideration of the common and anticipated problems in processing raw text (ie, irregularities and inaccuracies in chemical name reporting), it is perhaps surprising that existing technologies can yield as high as 90% accuracy in chemical name recognition and accurate structure annotation for many chemistry-related documents [OpenEye Scientific Software, personal communications]. Although typically applied to individual documents, these technologies and similar technologies being developed for public use [2] could, in principle, be applied to the chemical-content annotation of much larger segments of the public Internet, or to the large body of unstructured (ie, unformalized) textual data pertaining to chemical toxicity that are internal to government regulatory agencies such as the US Food and Drug Administration and the US Environmental Protection Agency (EPA).

There are two elements to this processing of raw text information: (i) chemical name recognition; and (ii) name-to-structure conversion. Both the Picto software from OpenEye [9] and the ACD/Name software from Advanced Chemistry Development Inc [11] provide name-to-structure conversion capabilities, primarily based on chemical name recognition using International Union of Pure and Applied Chemistry

(IUPAC) systematic naming conventions. The current industry standard formats of chemical structure representation and exchange are the 'mol file' connectivity table format from MDL (in ASCII) [12], and text-string-based representations, such as SMILES strings [13,14] or IUPAC International Chemical Identifier (InChI) strings [15••]. Full structure annotation using mol file content, or the equivalent, is the ultimate goal for enabling fully functional structure and substructure searching capabilities; however, text-based representations of chemical structure play an essential intermediary role in the primarily text-centric information world of the present.

InChI strings [15••] are the most recent entry into the field of chemical representation and have a number of noteworthy advantages compared with other types of commonly used identifiers: (i) an InChI string is generated directly from a mol file or another standardized representation of chemical structure and provides a unique textual representation of that structure; (ii) InChI strings are hierarchical text strings that are writable either in compact single-line or in eXtensive Markup Language (XML) notation, which can support many levels of increasingly detailed chemical description (eg, tautomeric forms, chiral centers, stereochemistry, charge state); and (iii) InChI-generation software is open-source and freely available [15••]. Murray-Rust *et al* recently proposed a comprehensive and feasible plan for the adoption and incorporation of systematic chemical annotation of published chemical literature through the use of Chemical Markup Language (CML) [16••], which represents a move toward a truly chemical Semantic Web [17]. InChIs are proposed to play a central role in providing a functionally useful, unique public domain chemical structure 'tag' on an HTML page [18,19]. Since its public launch in April 2005, InChI code display and generation has been incorporated into a number of software applications (eg, ACD/ChemSketch from Advanced Chemical Development [20••,21••,22••], PipeLine Pilot from SciTegic [23], the CACTVS Chemoinformatics Toolkit from Xemistry GmbH [24] and Marvin from ChemAxon Ltd [25]), has spawned third party InChI-to-structure open-source conversion utilities (eg, BKChem [26•]), and has been used to annotate major public Internet sites hosting chemical/biological databases (eg, PubChem [27••], the web site of the National Cancer Institute (NCI)'s computer-aided drug discovery (CADD) group [28••], and also the InChI sites listed in reference [15••]).

### *Structure locators and structure browsing*

Another approach to structure annotation involves the use of externally maintained listings of chemical inventories, referred to here as 'structure locator services' or 'structure index files'. These resources are intended to provide functional structure and substructure searching for locating chemical-related information. Two prominent examples of public resources acting primarily as structure locators of chemical toxicity information are ChemFinder.com from CambridgeSoft Corp [29] and the ChemIDplus system from the National Library of Medicine (NLM) Specialized Information Services [30]. Structure locator services maintain a central aggregated listing of chemical structures and provide full structure searching capability. These services use an internal lookup table that points to either external information pages associated with a particular web site or database containing the desired chemical (web site home indexed), or information web pages that pertain specifically to the chemical of concern (chemically indexed). Such resources have provided an important bridge between the primarily textual, historical chemical toxicity information posted on the Internet, and structure and structure analog searching capabilities for locating information on a specific chemical or related chemicals. One major limitation to these resources from the vantage point of a user or chemoinformaticist, however, is that the full structural content of a database or web site is unavailable, that is, the structure index files are inaccessible to the user except through the host search interface. For example, to access the chemical structures viewed in ChemFinder.com as individual mol files requires the capabilities of the commercial CambridgeSoft ChemDraw [31] application, whereas the ChemIDplus [30] service provides only a pictoral (gif) representation of the chemical structure to the user. In addition, the host structure locater service wholly determines which external web sites are to be included in their master structure index file.

Chmoogle [32••], a free, open-access chemistry search engine, launched in November 2005, is the most recent Internet entry into what might be termed 'global structure locator services'. The Chmoogle mission is to 'discover, curate and index all of the public chemical information in the world' [33], providing fast and simple structure searching across the Internet. Large structure-annotated collections (such as PubChem [27••]) have been initially targeted for inclusion, with over 6.7 million structures currently indexed. Beyond its global intentions, Chmoogle is distinguished from the large structure locator services previously mentioned, in that it provides open-access tools, such as Free Chmoogle [34••], which can be used to either bring local structure searching capability to a web site containing chemical content that has been structure-indexed within Chmoogle [4•], or allow a web site or application to freely access Chmoogle structure-searching capabilities (see, eg, ACD/ChemSketch [20••]).

Whereas a particular chemical structure with associated two- and three-dimensional information can be accurately represented as an industry-standard mol file with a graphical chemical structure component, a collection or aggregation of such structures with or without text/data fields is most commonly represented and exchanged in open structure-data (SD) file format, which is described by its originator MDL as simply being 'many mol files combined with data for each' [12]. The functional interface between an SD file and the Internet (most often after conversion of these files to a binary structure database format that is more efficient for searches, in particular (sub)structure searches) is termed a 'structure browser', of which there are many examples on publicly accessible web sites (see, eg, ChemIDplus [30], PubChem [27••], the NCI/CADD web site [28••] and Chmoogle [32••]) and available for use with public web sites (see, eg, the CACTVS Chemoinformatics

Toolkit [24], Marvin [25] and Free Chmoogle [34••]). The structure browser typically provides a user with a graphical structure drawing interface and structure search functionality across the information maintained or indexed by the web site. For chemical information resources on the Internet to be optimally useful for modeling and data mining applications, however, the chemical structure information should be freely available and downloadable in a variety of formats, including SD file, for either part or the entirety of a database or structure-index file. In addition, freeware SD viewer applications for desktop PC use, such as PowerMV from the National Institute of Statistical Sciences [35••] and ChemFileBrowser from Hyleos.net [36•], are available to complement these Internet structure browsing capabilities and to encourage broader use of such files.

### *Super aggregators, open data access and chemical profiling*

Application of the term 'super structure/data aggregator' is restricted in this review to those few extremely large-scale public chemical database projects that include functional structure searching as well as open access to full structure-indexed database content.

The web site of the CADD group at the NCI's Laboratory of Medicinal Chemistry [28••] is an example of a public resource that provides sophisticated chemical canonicalization (ie, creating uniform representations of chemical structures, including new hashcode-based identifiers) and property annotation (eg, PASS descriptors [37]), as well as relational structure, text and data searching. The databases offered on this web site include the large NCI Developmental Therapeutics Program (DTP) chemical screening database [38•] and additional public databases primarily derived from government sources; the combined data from these databases totals several million chemical records. Similarly to ChemFinder.com [29] and ChemIDplus [30], this resource from the CADD research group maintains central files of structure-annotated records, but unlike ChemFinder.com and ChemIDplus, the results of all searches and the whole of the data contained in almost all of the databases are fully downloadable in SD file format. The databases available on this web site include structure-only content, as well as structure-assay content.

The NLM National Center for Biotechnology Information (NCBI) PubChem project [27••] is a relatively new entry onto the public chemical/bioactivity database scene, but because of its large size, scope and pace of growth (to more than 5 million unique chemical structures in less than two years), it is likely to have a huge impact on the way in which chemical and biological activity summary data are represented, made publicly available and relationally accessed in the future. PubChem is first and foremost an online chemical data management model. This database differs significantly from the previously mentioned resources in that it is a 'user depositor' system that invites chemical-structure-annotated data submissions [39], preferably with summary bioassay data (although it also includes large structure-only datasets, eg, ZINC, which is a free database of commercially available compounds for virtual screening [40]). PubChem represents a significant extension of National Institutes of Health (NIH) and NLM data and data management resources [41], and incorporates the full ChemIDplus [30] structure library as well as the contents of the NCI DTP database [38•]. Chemical data deposited into PubChem include the chemical structural content as submitted by the depositor, which is assigned a PubChem depositor substance identification (SID) code. PubChem, however, enforces some basic standards of uniform canonicalized chemical data representation. After assigning an SID to any submission that can be chemically parsed (without attempting to correct chemically meaningless connectivities), PubChem attempts to extract one compound (or several compounds in the case of, eg, mixtures or salts) from the submission. A PubChem compound identification (CID) code is then assigned to each compound extracted, either by finding a match with an existing PubChem reference chemical, or, if no match is found, by assigning a CID code *de novo*, possibly after some limited correction is conducted on the basis of heuristic tests of chemical 'reasonableness' to detect obvious errors. In this way, the providers of PubChem can enforce some quality control on the internal referencing of submitted chemical structures and their representation within the database, without endorsing or otherwise reviewing the accuracy of user-submitted chemical information beyond simple 'reasonableness' standards.

PubChem [27••] is the database component of the NIH Molecular Libraries Initiative (MLI). One of the key motivations in the development of PubChem was that this resource should serve as a large central public repository of chemical bioactivity data that would be generated from the Molecular Libraries Screening Center Network [42••,43,44••]. The MLI project, launched in mid 2003 and now well underway, comprises the screening of an extremely large library of small molecules (> 100,000) [45•] in hundreds of high-throughput bioassays, and depositing the resulting data into the PubChem database. At one of the screening centers of the MLI, the NIH Chemical Genomics Center (NCGC), these compounds are screened at multiple dose dilutions and concentration-response curves are generated for every compound. This technique represents a significant departure from the typical pharmaceutical high-throughput screening paradigm, supporting an increased reliability of 'hits' and the ability to generate $IC_{50}$ values from the primary screen. Researchers involved in the project are screening the MLI small-molecule library to enable the development of 'chemical probes' of gene, pathway and cellular functions, with the broad objective of advancing understanding of the relationship between chemical structure and biological function. To the extent that this designed chemical library will sample sufficient chemical space with associated reference toxicity data, there is also potential for the results to be relevant to toxicity inferences. This project represents a significant investment for the NIH, relying on the use of top-level chemical indexing and chemical probes, and supporting the development of a general chemical/biological profiling capability in the public domain (for recent examples of biological profiling applications to toxicity screening, see references

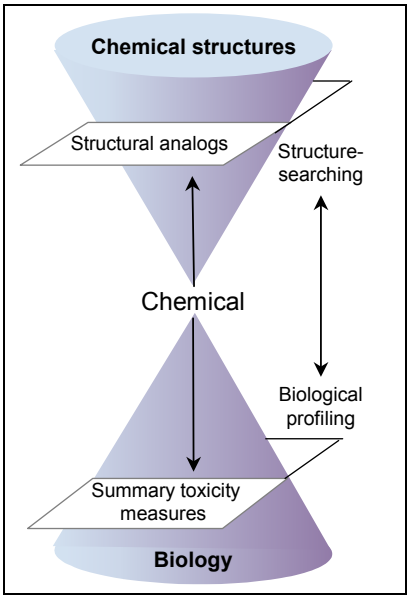
[46•,47•,48•]). The sheer size and scope of the MLI (all of the data from which are being deposited in PubChem) provides a huge impetus for other types of public biological activity and toxicity database resources to move toward chemical structure indexing as a way of aligning with the central PubChem repository. Evidence of this trend was provided during a recently held workshop of the National Toxicology Program on High-Throughput Screening Assays [49], the goal of which was to solicit input and recommendations from a broad range of experts from government and industry on proposed interactions between the National Toxicology Program (NTP) and the NIH MLI project. To be included in the summary report from the workshop [49] are recommendations to add chemical structure indexing to the NTP online database and also to better codify summary toxicity measures within the historical NTP database [50•] to interface with PubChem and the new data to be generated within the NIH MLI project.

## Public toxicity data models

The concept and utility of top-level chemical structure indexing is broadly applicable to all types of chemical toxicological data or experiments, from toxicogenomics and microarray experiments to traditional toxicology experiments categorized by assay-type, test area (ie, ecological versus health) and/or discipline (eg, cancer, subchronic toxicology, genetic toxicology, neurotoxicology, immunotoxicology, developmental toxicology or reproductive toxicology). However, significant challenges remain in delving deeper into the data morass, for example, in migrating and systematizing large amounts of past toxicological data that are in principle 'public', but which are not readily accessible to the public (eg, from literature, government agencies or public archives). A recent review by Yang *et al* surveyed the current status of public toxicity databases in terms of their diverse content, database structure, download accessibility, relational and read-across capabilities, and suitability for data mining [51••]. This review articulated the need to actively engage toxicology experts in the process of creating systematic vocabularies and data standards (see also reference [5]). A second recent review by Yang *et al* goes one step further [52••] in defining the formal needs of a fully relational and mineable toxicity data model, and describes the public ToxML data schema project [53•,54••], which is specifically designed to meet those needs.

The fundamental importance of these efforts to systematize data warrants further elaboration. Figure 2 provides a simple schematic illustration of a data model, patterned after the ToxML schema [53•,54••], in which two main types of data aggregation occur with respect to chemical indexing and biological/toxicological activity. The first type of aggregation enables a user to survey and restrict the overall chemical space to structure analog or similarity space, which requires the availability of sufficient standardized chemical structure and property representations across toxicology experiments. The second type of aggregation involves the ability to derive intermediate aggregations and summarizations of the toxicity data. This is accomplished by using a built-in hierarchical data structure, informed by toxicity domain experts, that captures essential experimental details (eg, dose-response data) at the lowest level. These details are then organized within categories or aggregation layers, for example, species, tissue or organ effects, or biochemical or bioassay endpoints, which can be either quantitative (potencies) or qualitative (categorical yes/no calls). These intermediate layers of data summarization [55] feed into modeling efforts and create the potential to 'read-across' a database or a collection of databases. Data models such as these will provide an essential link to the toxicity domain experts, encouraging their active involvement and participation, and serving as a sanctioned conduit for transfering unstructured data from internal databases, government archives and literature studies into a data format that can be integrated, mined and potentially modeled. An example of such a current effort is provided by the International Life Sciences Institute (ILSI) Developmental Toxicity Database Workgroup [56•,57], which comprises developmental toxicologists, modelers and experts from both industry and government. This workgroup is attempting to develop a functional public data model for use in capturing data from studies reported in the developmental toxicology literature to ultimately support the development of improved structure-activity models. The top-level chemical indexing and intermediate layers of summarization associated with this public data model will encourage ready integration of data into existing chemical inventories, such as the PubChem repository [27••], for expanded relational and data mining uses.

**Figure 2. Chemical toxicity data models.**




The figure illustrates a model construction that would enable 'read-across' relational data mining and biological profiling at intermediate levels of data summarization. Based on a figure in reference [52••].

### *Pharmaceutical online chemical resources*

There are a number of public data models and resources available (not exclusively toxicity related) that are specifically focused on the chemical and bioactivity domains of interest to the pharmaceutical industry and drug

discovery researchers, and focused toward relational searching and data mining. An early entry into the field of chemical genomics was the ChemBank database [3,58•], which was proposed as a public model that would bridge biology and chemistry knowledge space at multiple levels, spanning data content ranging from raw, unprocessed data to formatted bioactivity and chemistry data, to public pathway and genomics databases. DrugBank [4•,59•] is an example of a newer public database that provides reasonable coverage of the known pharmaceutical chemical space. It places equal emphasis on characterizing the chemical and the biological/target properties of known pharmaceuticals, as well as provides functional and chemical integration of these two information domains. Both the ChemBank and DrugBank databases place emphasis on chemical indexing of the web site content. DrugBank offers chemical structure (similarity) and chemical name searching of its web site, with the ability to download individual hits in mol file or SD file format, but not multiple hits or the entire database (except as a flat file with SMILES included). ChemBank offers SMILES-based structure searching and full SD file download of chemically indexed information. The public resources of the NCI CADD group mentioned above, are currently in the process of being greatly expanded and will offer web services and chemically indexed databases that span both the toxicology and drug discovery fields [28••,60•].

### *Toxicogenomics and chemical indexing*

In the area of toxicogenomics, the Chemical Effects in Biological Systems (CEBS) project [61•] is intended to provide a broad-based, fully relational public database resource that will incorporate top-level chemical indexing (in collaboration with the EPA Distributed Structure-Searchable Toxicity (DSSTox) database project [62,63•]), standardized data dictionaries [64] and ToxML data schema [53•,54••]. The CEBS project will also offer exploration tools that will link toxicogenomics data with external Internet resources and historical toxicology data across toxicology domains. Prominent public microarray databases, such as the NCBI's Gene Expression Omnibus (GEO) [65], the European Bioinformatics Institute (EBI)'s ArrayExpress [66] and the Environment, Drugs and Gene Expression (EDGE) database [67], demonstrate significant read-across and relational capabilities with respect to genomics experiments and integration with external public Internet resources, but provide little or no chemical indexing of database content. The Comparative Toxicogenomics Database (CTD) [68] is an exception to this, in that it provides chemical name indexing of the site content, pictoral structure representation and links to the ChemIDplus [30] chemical data page, but with no structure searching capability of the site, and no availability of mol file or SD file chemical-content downloads. The general lack of chemical indexing in these bioinformatics databases and information resources effectively isolates them from other types of chemically indexed toxicology data, and inhibits the assessment of database chemical coverage and chemical searching and aggregation efforts. The situation is well-captured by Wishart *et al* in the following words: 'This state of affairs largely reflects the 'two solitudes' of cheminformatics and bioinformatics. Neither discipline has really tried to integrate with the other' [4•].

## Chemical information data quality

A major issue with respect to the above-mentioned chemical indexing technologies and advances, which has received relatively little formal recognition or attention to date, is the quality of chemical information associated with toxicological data in the public domain. The standards applied for the reporting of chemical information vary widely; errors and incomplete reporting of chemical information on tested compounds are widespread within public toxicology literature and databases, and there is no central public entity charged with ensuring the quality or completeness of the chemical information associated with toxicological information. Chemical Abstracts Services (CAS) and CAS Registry Numbers (RNs) [69] have played an important historical role in enforcing quality standards for the reporting of chemical information in chemical and biological literature; however, a commercial proprietary registration system of chemical information, such as CAS, that does not provide ready linkage to actual chemical structure content in the public domain is becoming an increasingly obsolete model [16••,70,71]. Neither CAS RNs nor chemical names can directly enable generalized structure (ie, similarity) or substructure searching (ie, exact matching of substructural features) across public domain resources. In addition, the assignment of CAS RNs to results from chemical toxicity experiments reported in the literature or in compiled public databases is most often derived from secondary public sources of aggregated chemical information (eg, ChemFinder.com [29], ChemIDplus [30] or PubChem [27••]) rather than from direct reference to the costly, subscription-only CAS information resources [69]. Hence, errors in the association of CAS RNs with chemical names or structures in the literature or in compiled public toxicity databases (either through incorrect assignment of the CAS RN or through incorrect representation of structure) are perhaps more common than is generally realized.

The large-scale resources available for aggregating chemical structure information in the public domain (eg, ChemIDplus [30]) are often perceived by the biological and toxicological community as being definitive sources of accurate and reliable chemical information; however, these sources are themselves secondary aggregators of public information and, therefore, inevitably incorporate inaccuracies from public information resources. These large public aggregators of chemical structures serve as primary resources of chemical information for the majority of public chemical database annotation efforts, and, hence, errors and imprecise structure representations (eg, without stereochemistry) tend to propagate across public toxicity databases (see also reference [72]). During the course of the EPA DSSTox Database project [63•], chemical structure, chemical name and CAS RN information for > 8000 unique chemical records from a variety of toxicity databases were reviewed for accuracy and consistency through organic chemistry expertise, and verified using multiple public information sources [73] (this dataset included several databases and structure index files that are soon to be published on the DSSTox web site). Secondary toxicity databases (which are most often compiled by non-chemists) were found to have

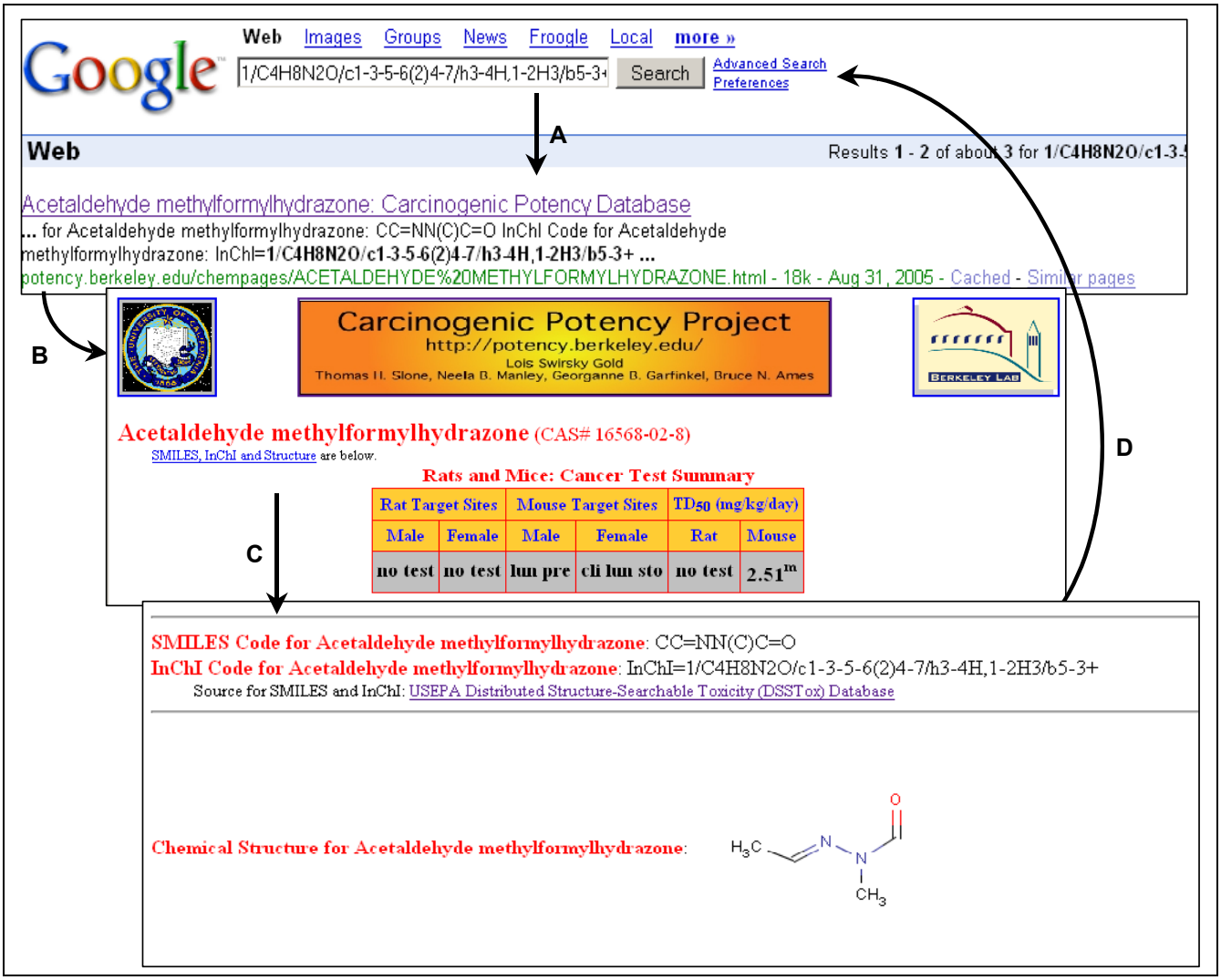
the highest rates of chemical identification errors, whereas chemistry aggregator sites such as ChemFinder.com [29] and ChemIDplus [30] had 1 to 2% errors (some of these errors were minor, involving incorrect or missing stereochemistry, and others were more major, involving incorrect atomic configuration). The present status of the NCBI PubChem project [27••], with respect to the quality of the chemical information presented with bioassay data, is best characterized as 'user beware', as the chemical structure content is posted as submitted by the depositor without additional review; however, users can make approximate informed judgments on the quality of data based on knowledge of the PubChem data submitter [39] and published quality review practices (see, eg, the published DSSTox quality assurance procedures [73•]). The only way to minimize these errors is to increase the degree of vigilance and review on the part of chemical aggregators, to submit error reports to central chemical aggregation sites, to provide prominent posting of quality review procedures (or lack thereof), and to keep the users of public chemical information better informed. Less stringent quality control of chemical structure information might be required of global structure locator services, such as Chmoogle [32••], because these services act only as pointers and are more remotely associated with the corresponding biological activity data. Frequent re-use and migration of data from one database to another calls for vigilance, even in the case of strong 'majority opinions', with respect to chemical structures, names, identifiers, etc, across multiple databases. Over time, however, an increasingly interlinked, flat chemical world should facilitate the detection of inconsistencies and errors in chemical databases, resulting from more frequent cross-checks among data/structure collections.

## Case study: DSSTox and the Carcinogenic Potency Database

Chemically related toxicity information, offered in different public venues on the Internet and associated with different types and levels of information content, can be effectively linked and integrated through the use of chemical annotation, structure locator files and structure browsing tools. Herein is provided an example of a long-term collaboration between the EPA DSSTox Database project [62,63•] and the Carcinogenic Potency DataBase (CPDB) project (based at the University of California, Berkeley) [74-76,77•], which has more recently been broadened to include collaboration with the NCI/CADD group's public web services [28••] and the PubChem project [27••]. The goal of the original collaboration between the DSSTox and the Carcinogenic Potency projects was to provide a documented, standardized and fully structure-annotated SD file of the 'CPDB Summary Table – All Species' [78], which would be available for public download from the DSSTox web site [79]. This SD file was created to serve the needs of structure-activity modelers and to augment the detailed toxicological content of the CPDB web site [77•] with DSSTox standard chemical fields [80]. The DSSTox SD file and the CPDB web site are maintained as separate entities. In the early published versions (CPDBAS_v1a and CPDBAS_v2a), the DSSTox SD files for CPDB summary tables did not effectively link to the full content of the CPDB web site; the CPDB web site hosts a large number of detailed data tables and plots pertaining to chemical carcinogenicity studies and, although the CPDB previously included chemical names and CAS RNs, these data were not effectively indexed from a chemical structure searching perspective. External structure locator services, such as ChemFinder.com [29], maintained an internal structure index file for the CPDB web site, but were (and continue to be) indexed only to the main CPDB home page and not to the toxicological results for particular chemicals on the CPDB web site.

Motivated by this collaboration with researchers working on the DSSTox project, and the increasing availability of chemical information (such as InChIs, SMILES, etc), additional indexing by chemical structure has been added through the Carcinogenic Potency project to augment the detailed data content of the CPDB web site [81]. Separate chemical data pages with a distinct web site URL address for each of the > 1450 chemical substances contained in the CPDB, containing summary data, analyses of individual experiments and links to information throughout the CPDB web site, are now provided [81]. On each individual CPDB chemical substance data page (except in cases of undefined substances or mixtures), several chemical identifiers are provided (see Figure 3), including a pictorial representation of the chemical structure, the CAS RN, the SMILES string and the InChI code, which correspond to the DSSTox CPDB SD file record contents for that chemical substance. With the addition of InChIs and chemically indexed web site content, the CPDB web site has, in effect, been internally 'chemically activated', that is, the CPDB web site host has added sufficient chemical structure identifier content for its pages to be 'structure located' by a general Internet text search, without the need for outside intervention or posting of separate structure locator files on third-party web sites. Figure 3 illustrates this capability using an actual Google text search of the InChI code for acetaldehyde methylformylhydrazone, which successfully locates the chemically indexed web page on the CPDB web site [18,81]. Since a web site URL is now assigned to each chemically indexed page on the CPDB web site [81], these URLs have been added as a distinct field to the latest version of the DSSTox CPDB SD file (CPDBAS_v3a; currently posted on PubChem [27••], and soon to be posted on the updated DSSTox web site [63•]). Hence, any third-party user, structure locator service or structure aggregator can acquire both the CPDB summary table with full chemical structure annotation and also the URLs that confer a structure locator capability to the CPDB web site, from either the DSSTox web site or PubChem [27••]. The URLs that confer a structure locator capability to the CPDB web site can directly link a user to the detailed data content of the CPDB web site, which includes data from analyses of thousands of individual chemical carcinogenesis experiments.

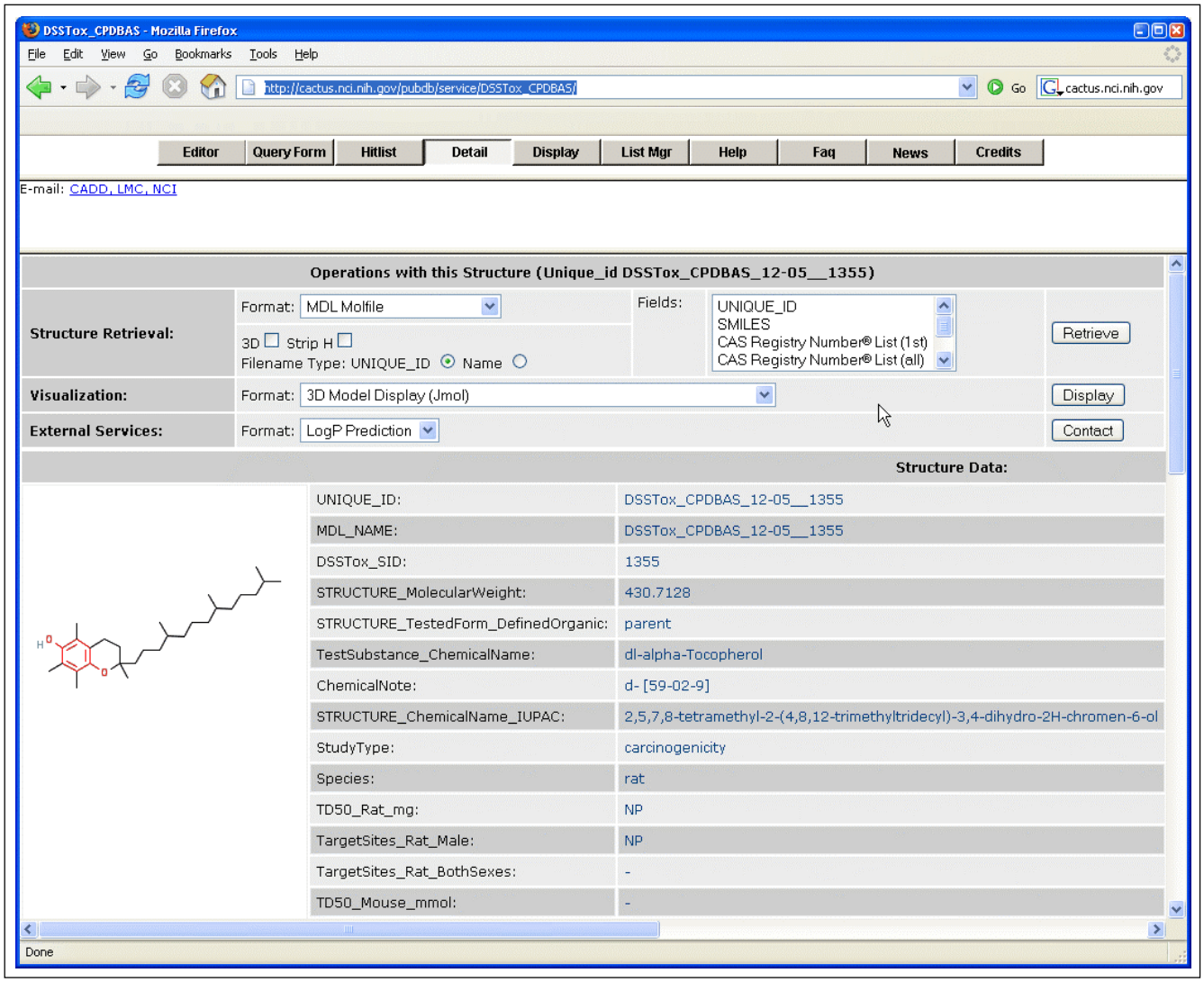
**Figure 3. Locating an InChI structure-indexed web page starting from a Google search.**




The figure illustrates how a chemical information page on the Carcinogenic Potency Database (CPDB) web site can be located, using a general Internet text search of the InChI code. The example used in this case is a search for the InChI code for acetaldehyde methylformylhydrazone (only a portion of the CPDB web page is shown). Inserting the InChI code for acetaldehyde methylformylhydrazone in Google and initiating a search, results in a hit (**A**), which links to the CPDB web site (**B**). From this web site a link to the SMILES, InChI and structure is provided (**C**), and it can be seen that the InChI code is displayed exactly as originally searched (**D**).

Ongoing collaborations between researchers working on the DSSTox project [63•] and those involved with the PubChem project [27••] and NCI/CADD group's public web services [28••], have been directed toward providing structure searching capability, property enhancements and improved data context to the DSSTox published toxicity data files. The PubChem and NCI/CADD web sites separately process and host the DSSTox published data files in a structure-searchable format, and these are fully downloadable as SD files [60•]. The PubChem/DSSTox collaboration additionally motivated revision of the DSSTox standard chemical fields for the latest DSSTox SD file versions [79,80], providing an increased distinction between structure-related and test-substance-related fields, as well as the addition of a unique DSSTox substance identification field (DSSTox_SID) for cross-referencing with the PubChem SID codes. In the case of the DSSTox CPDBAS database, which is posted in a structure-searchable form on both the NCI/CADD (Figure 4) and PubChem web sites, this means that the chemical locator URLs that link each DSSTox chemical record to the more extensive, chemically indexed CPDB web site content are additionally provided to the user. Hence, these URLs act as pointers to the 'glocalized' [1] content of the CPDB web site, which means that the localized expertise and chemical carcinogenicity content of the CPDB are more effectively chemically indexed, and are accessible through global Internet structure locator services. Finally, given the open availability of the SD structure index file for the CPDB web site, a future enhancement could include web-site-hosted substructure- and structure-searchability functions on the CPDB web site, through the use of a freeware tool such as Free Chmoogle [34••].

**Figure 4. Screen shot of the NCI/CADD interface for the DSSTox CPDB data file.**




The figure shows a detailed display page for one example entry of the CPDB resulting from a substructure search. The substructure that was searched was 1,4-dihydroxyphenyl. (Only a portion of the web page is shown; this screenshot is from a β-test version of the web service [60•].)

## Conclusions

Chemical content annotators, structure locator services, structure/data aggregator web sites, structure browsers and InChI codes are all playing important roles in overcoming barriers to the integration of toxicity data and are bringing users closer to the reality of a mineable chemical Semantic Web [16••,17]. Additionally, the increasing levels of implementation of data models, data standardization and curation efforts, and the increasing availability of large resources of biological profiling data are enriching the biological activity content that can be meaningfully associated with chemicals. The ability for users to broadly integrate toxicological information across biological information domains using chemical structure as the primary annotation and search metric will expand the possibilities, and change the paradigms for toxicity data mining and screening. Looking out over the flattened world of toxicology from this chemically indexed vantage point, the future looks bright.

## Acknowledgments

The authors wish to thank Chihae Yang (Leadscope Inc), Maritja Wolf (Lockheed Martin Corp, contractor to the EPA), ClarLynda Williams (EPA, North Carolina State University student training cooperative), Markus Sitzmann (NIH/NCI post doctoral fellow), Stephen Bryant and Jane Tseng (NCBI PubChem project), Wolf-Dietrich Ihlenfeldt (CACTVS, NCI/CADD and NCBI PubChem projects), Thomas Slone (Carcinogenic Potency Project, University of California, Berkeley), Klaus Gubernator (Chmoogle), Christopher Austin (NIH Chemical Genomics Center), and members of the IUPAC/National Institute of Standards and Technology (NIST) InChI development team (Stephen Heller, Stephen Stein, Dmitrii Tchekhovskoi). In addition, Lois Swirsky Gold acknowledges support from the US National Institute of Environmental Health Sciences through the EO Lawrence Berkeley National Laboratory, interagency agreement (YIES101901) with the US Department of Energy (DE-AC-03-76SFO0098), and from a grant at the University of California,

Berkeley for research in disease prevention through the Dean's Office of the College of Letters and Science.

## Disclaimer

This manuscript has been reviewed by US EPA's National Center for Computational Toxicology and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

## References

•• of outstanding interest
• of special interest


1. Friedman T: *The World is Flat*. Farrar, Straus and Giroux, New York, NY, USA (2005).

2. Cavalla D: **Web alert: The management of chemical and biological information.** *Curr Opin Drug Discovery Dev* (2003) **6**(3):294-295.

3. Strausberg RL, Schreiber SL: **From knowing to controlling: A path from genomics to drugs using small molecule probes.** *Science* (2003) **300**(5617):294-295.

4. Wishart DS, Knox C, Guo AC, Shrivastava S, Stothard P, Chang Z, Woolsey J: **DrugBank: A comprehensive resource for *in silico* drug discovery and exploration.** *Nucleic Acids Res* (2006) **34**(Database issue):D668-D672.

• *Describes a new public database resource that is particularly focused on drugs, and which offers both chemical and target information in a relational format*.

5. Salter AH: **Large-scale databases in toxicogenomics.** *Pharmacogenomics* (2005) **6**(7):749-754.

6. Cassman M: **Barriers to progress in systems biology.** *Nature* (2005) **438**(7071):1079.

7. Salamone S: **Masters of the Semantic Web.** *Bio-IT World* (2005) October:29-30. *http://www.bio-itworld.com/issues/2005/oct/cover-story-semantic-web/*

8. **IBM Almaden research Center: WebFountain:** IBM Corp, New York, NY, USA. *http://www.almaden.ibm.com/webfountain/*

9. **OpenEye Scientific Software: Products:** OpenEye Scientific Software, Santa Fe, NM, USA. *http://www.eyesopen.com/products/*

10. Boyer SK: **Document/patent – analysis based on molecular diversity.** *4th Meeting on US Government Chemical Databases*, National Institutes of Health, Frederick, MD, USA (2005).

11. **Chemical naming: Expert desktop software:** Advanced Chemistry Development, Toronto, ON, Canada (2006). *http://www.acdlabs.com/products/name_lab/*

12. **MDL: MDL® CTFile formats no-fee:** Elsevier MDL, San Leandro, CA, USA (2006): *http://www.mdli.com/downloads/public/ctfile/ctfile.jsp*

13. **SMILES home page: A collection of SMILES-related hyperlinks and information:** Daylight Chemical Information Systems, Aliso Viejo, CA, USA (2006). *http://www.daylight.com/smiles/f_smiles.html*

14. **Input of structures by means of SMILES strings:** National Cancer Institute, Frederick/National Institutes of Health, Bethesda, MD, USA. *http://cactus.nci.nih.gov/services/smiles.html*

15. **The IUPAC International Chemical Identifier (InChI™):** International Union of Pure and Applied Chemistry, Research Triangle Park, NC, USA (2006). *http://www.iupac.org/inchi/*

•• *Provides links to open-source InChI generation software, third-party information sites, frequently asked questions, InChI interpreters, etc*.

16. Murray-Rust P, Mitchell JB, Rzepa HS: **Communication and re-use of chemical information in bioscience.** *BMC Bioinform* (2005) **6**:180-196.

•• *Describes the current status of chemical information on the Internet, including problems and challenges, and presents a cogent vision and proposal for moving toward a truly public chemical Semantic Web*.

17. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y: **Enhancement of the chemical Semantic Web through the use of InChI identifiers.** *Org Biomol Chem* (2005) **3**(10):1832-1834.

18. Murray-Rust P, Rzepa HS, Zhang Y: **Googling for InChIs: A remarkable method of chemical searching.** W3C: World Wide Web Consortium, Massachusetts Institute of Technology, Cambridge, MA, USA (2004). *http://lists.w3.org/Archives/Public/public-swls-ws/2004Oct/att-0019/*

19. Prasanna MD, Vondrasek J, Wlodawer A, Bhat TN: **Application of InChI to curate, index and query 3-D structures.** *Proteins* (2005) **60**(1):1-4.

20. **ACD/ChemSketch 8.0 freeware:** Advanced Chemistry Development, Toronto, ON, Canada (2005). *http://www.acdlabs.com/download/chemsk.html*

•• *This reference, along with references [21••] and [22••], provides downloadable structure-drawing freeware, including InChI code conversion to structure or generation from drawn structure capabilities, and Chmoogle structure-searching access.*

21. **What is InChI?** Advanced Chemistry Development, Toronto, ON, Canada (2005). *http://www.acdlabs.com/download/inchi_more.html*

•• *This reference, along with references [20••] and [22••], provides downloadable structure-drawing freeware, including InChI code conversion to structure or generation from drawn structure capabilities, and Chmoogle structure-searching access.*

22. **ACD/Labs integrates ChemSketch to Chmoogle search engine:** Advanced Chemistry Development, Toronto, ON, Canada (2005). *http://www.acdlabs.com/clients/pr_chmoogle1105.html*

•• *This reference, along with references [20••] and [21••], provides downloadable structure-drawing freeware, including InChI code conversion to structure or generation from drawn structure capabilities, and Chmoogle structure-searching access.*

23. **PipeLine Pilot:** SciTegic, San Diego, CA, USA (2006). *http://www.scitegic.com/products_services/pipeline_pilot.htm*

24. Ihlenfeldt WD, Takahashi Y, Abe H, Sasaki S: **Computation and management of chemical-properties in CACTVS: An extensible networked approach toward modularity and compatibility.** *J Chem Inf Comput Sci* (1994) **34**(1):109-116.

25. **Marvin:** ChemAxon, Budapest, Hungary. *http://www.chemaxon.com/marvin/*

26. **BKChem.org: InChI support in BKchem:** BKChem.org (2006). *http://bkchem.zirael.org/inchi_en.html*

• *Provides a downloadable, freeware, open-source chemical drawing program with InChI-to-structure conversion capabilities*.

27. **NCBI: PubChem:** National Center for Biotechnology Information, Bethesda, MD, USA (2006). *http://pubchem.ncbi.nlm.nih.gov*

•• *Provides a large user-depositor, structure-activity data aggregator with relational searching capabilities, full structure searching and open data access with SD file download*.

28. Ihlenfeldt WD, Voigt JH, Bienfait B, Oellien F, Nicklaus MC: **Enhanced CACTVS browser of the open NCI database.** *J Chem Inf Comput Sci* (2002) **42**(1):46-57.

•• *Describes a large structure-activity data aggregator with relational searching capabilities, full structure searching capabilities, chemical and predicted chemical property annotations, and open data access with SD file download*.

29. **ChemFinder.com: Database and internet searching:** CambridgeSoft Corp, Cambridge, MA, USA (2006). *http://chemfinder.cambridgesoft.com/*

30. **National Library of Medicine Specialized Information Services: ChemIDplus Advanced:** National Library of Medicine, Bethesda, MD, USA (2006). *http://chem.sis.nlm.nih.gov/chemidplus/chemidheavy.jsp*

31. **ChemDraw:** CambridgeSoft Corp, Cambridge, MA, USA (2005). *http://www.cambridgesoft.com/products/family.cfm?FID=2*

32. **Chmoogle: Searching the world's chemistry:** eMolecules, Del Mar, CA, USA (2006). *http://www.chmoogle.com/*

•• *This reference, along with reference [34••], provides a new public resource for Internet chemical searching and free tools to make individual web sites with chemical content structure searchable*.

33. Bradley D: **Oogling for chemists.** *Reactive Reports: Chemistry web magazine* (2005). *http://www.reactivereports.com/50/50_1.html*

34. **Bring Chmoogle's power home: Free Chmoogle:** eMolecules, Del Mar, CA, USA (2006). *http://www.chmoogle.com/doc/products.htm*

•• *This reference, along with reference [32••], provides a new public resource for Internet chemical searching and free tools to make individual web sites with chemical content structure searchable.*

35. Liu, J, Feng, J, Young, S: **PowerMV: A software environment for statistical analysis, molecular viewing, descriptor generation and similarity search:** National Institute of Statistical Sciences, Research Triangle Park, NC, USA (2005). *http://www.niss.org/PowerMV/*

•• *Provides a publicly available SD file viewer that can be freely downloaded for unlimited PC desktop use. Major limitations are that there is no explicit substructure searching and no user help manual, but the system is easy to use and contains many useful viewing and simple search features*.

36. **ChemFileBrowser:** Hyleos.net (2006). *http://www.hyleos.net/?s=applications&p=ChemFileBrowser*

• *Provides a publicly available SD file viewer that can be freely downloaded for unlimited PC desktop use. SD content can be printed and viewed, but there are no text or structure search functions.*

37. Poroikov VV, Filimonov DA, Ihlenfeldt WD, Gloriozova TA, Lagunin AA, Borodina YV, Stepanchikova AV, Nicklaus MC: **PASS biological activity spectrum predictions in the enhanced open NCI database browser.** *J Chem Inf Comput Sci* (2003) **43**(1):228-236.

38. **NCI/NIH Developmental Therapeutics Program:** National Cancer Institute, Frederick/National Institutes of Health, Bethesda, MD, USA. (2005). *http://dtp.nci.nih.gov/index.html*

• *Provides antitumor screening assay data on > 200,000 chemicals.*

39. **NCBI: PubChem substance data source information:** National Center for Biotechnology Information, Bethesda, MD, USA (2006). *http://pubchem.ncbi.nlm.nih.gov/sources/*

40. **A free database for virtual screening: ZINC:** University of California San Francisco, San Francisco, CA, USA (2005). *http://blaster.docking.org/zinc/*

41. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Geer LY *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* (2006) **34**(Database issue):D173-D180 doi:10.1093/nar/gkj158.

42. Austin CP, Brady LS, Insel TR, Collins FS: **NIH Molecular Libraries Initiative.** *Science* (2004) **306**(5699):1138-1139.

•• *Describes a major new public initiative in the area of chemical/biological activity profiling.*

43. Pincock S: **Chemical genomics collaborations heat up.** *Scientist* (2005) **19**(18):22-26.

44. **NIH Roadmap: Molecular libraries and imaging:** National Institutes of Health, Bethesda, MD, USA (2005). *http://nihroadmap.nih.gov/molecularlibraries/*

•• *Provides a description of the Molecular Libraries Screening Center Network, the chemoinformatics capabilities within PubChem and technology development in areas of chemical and assay diversity, instrumentation and predictive absorption, distribution, metabolism, excretion and toxicology*.

45. **NIH molecular libraries small-molecule repository:** Discovery Partners International, San Diego, CA, USA. *http://mlsmr.discoverypartners.com/*

• *Discovery Partners International are involved in chemical acquisition, management and distribution for the NIH Molecular Libraries Screening Center Network. Entering the depositor code 'DPISMR' (Discovery Partners International Small Molecule Repository) in a PubChem search provides a listing of the current library of 66,000 chemicals*.

46. Fliri AF, Loging WT, Thadelo PF, Volkmann RA: **Biological spectra analysis: Linking biological activity profiles to molecular structure.** *Proc Natl Acad Sci USA* (2005) **102**(2):261-266.

• *Describes a proof-of-principle study illustrating the concept of chemical/biological profiling as a tool for screening.*

47. Whitebread S, Hamon J, Bojanic D, Urban L: **Keynote review: *In vitro* safety pharmacology profiling: An essential tool for successful drug development.** *Drug Disc Today* (2005) **10**(21):1421-1433.

• *Provides a good overview of the state-of-the-art in activity profiling of drugs for safety assessment and toxicity.*

48. Roter AH: **Large-scale integrated databases supporting drug discovery.** *Curr Opin Drug Discovery Dev* (2005) **8**(3):309-315.

• *Provides a survey of existing genomics and informatics databases spanning a wide range of capabilities.*

49. **Summary report.** *National Toxicology Program: High-Throughput Screening Assays Workshop*, Arlington, VA, USA (2005): manuscript in preparation.

50. **National Toxicology Program database search application:** National Institutes of Health, Bethesda, MD, USA (2005). *http://ntp-apps.niehs.nih.gov/ntp_tox/*

• *Provides a wealth of toxicological information indexed by chemical name and includes details of the study level, but with no structure indexing capability and little relational read-across capability at present.*

51. Yang C, Benz RD, Cheeseman MA: **Landscape of current toxicity databases and database standards.** *Curr Opin Drug Discovery Dev* (2006) **9**(1):124-133.

•• *Provides a survey of a wide range of Internet toxicity data resources, and discusses their compatibility with data mining objectives*.

52. Yang C, Richard AM, Cross KP: **The art of data mining the minefields of toxicity databases to link chemistry to biology.** *Curr Comput Aided Drug Des* (2006): in press.

•• *Presents a formal description of data mining objectives and the requirements of toxicity databases, with illustrative examples*.

53. Yang C, Arvidson K, Benz RD, Matthews E, Cheeseman MA, Kruhlak N, Mayer J, Nelson C, Twaroski ML, Hollingshaus G, Aveston S *et al*: **ToxML, 'a domain intelligent database standard for linking toxicology to chemistry'.** (2006): manuscript in preparation.

• *Describes an important public initiative addressing the fundamental data structuring requirements for fueling improved and expanded toxicity data mining in support of screening and prediction*.

54. **About ToxML:** Leadscope Inc, Columbus, OH, USA (2006). *http://www.leadscope.com/toxml.php*

•• *Describes an important public initiative addressing the fundamental need for the availability of improved data models. It is anticipated that toxicity schema for genetic toxicology and chronic toxicity will be made publicly available on this site*.

55. Richard AM, Williams CR: **Public sources of mutagenicity and carcinogenicity data: Use in structure-activity relationship models.** In: *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*. Benigni R (Ed), CRC Press, New York, NY, USA (2003):145-173.

56. Julien E, Willhite CC, Richard AM, DeSesso JM: **Challenges in constructing statistically based structure-activity relationship models for developmental toxicity.** *Birth Defects Res Part A-Clin Mol Teratol* (2004) **70**(12):902-911.

• *Provides an evaluation of the current state of predictive toxicology models in developmental toxicity, with recommendations to address basic data model needs as a prescription for the future improvement of such models*.

57. **Improving the Use of Toxicity Data in Statistically Based Structure-Activity Relationship Models for Developmental Toxicity:** International Life Sciences Institute, Washington, DC, USA (2006). *http://rsi.ilsi.org/Projects/DevToxSAR*

58. **ChemBank: Initiative for chemical genetics:** Broad Institute, Chemical Biology Program, Cambridge, MA, USA (2005). *http://chembank.broad.harvard.edu/*

• *Provides a database with SMILES structure searchability and full SD file download availability*.

59. **DrugBank:** University of Alberta, Calgary, AB, Canada (2006). *http://redpoll.pharmacy.ualberta.ca/drugbank/*

• *Provides a new public resource having relational and open-data content, chemical structure search capability, and a wide range of information available for approved, experimental, biotechnological and neutraceutical drugs*.

60. **Frederick/Bethesda data and online services: Early beta test version.** National Cancer Institute, Frederick/National Institutes of Health, Bethesda, MD, USA. (2005). *http://cactus.nci.nih.gov/PubDBs/*

• *At the time of writing, this site is only available as the β-test version, but will soon be released publicly. The site contains a comprehensive set of links to publicly available databases.*

61. Waters M, Boorman G, Bushel P, Cunningham M, Irwin R, Merrick A, Olden K, Paules R, Selkirk J, Stasiewicz S, Weis B *et al*: **Systems toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base.** *Environ Health Perspect* (2003) **111**(6):811-824.

• *Describes a fully integrated public database repository that links toxicogenomics with historical toxicity data. The database is available in the β-version for public access.*

62. Richard AM: **DSSTox web site launch: Improving public access to databases for building structure-toxicity prediction models.** *Preclinica* (2004) **2**(2):103-108.

63. **Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network:** US Environmental Protection Agency, Washington, DC, USA (2006). *http://www.epa.gov/nheerl/dsstox/*

• *Provides access to fully downloadable SD files with data for a variety of toxicity and chemical domains, standardized chemical annotation and documentation. In the future, downloadable structure index files for a variety of public online toxicity data resources, for example, the NTP, the EPA Integrated Risk Information System and the EPA High Production Volume program, will also be posted on this site*.

64. Fostel J, Choi D, Zwickl C, Morrison N, Rashid A, Hasan A, Bao W, Richard A, Tong W, Bushel PR, Brown R *et al*: **Chemical effects in biological systems – data dictionary (CEBS-DD): A compendium of terms for the capture and integration of biological study design description, conventional phenotypes and 'omics data.** *Toxicol Sci* (2005) **88**(2):585-601.

65. **NCBI: GEO – Gene Expression Omnibus:** National Center for Biotechnology Information, Bethesda, MD, USA (2006). *http://www.ncbi.nlm.nih.gov/geo/*

66. **EMBL-EBI: ArrayExpress at the EBI:** European Bioinformatics Institute, Cambridge, UK (2006). *http://www.ebi.ac.uk/arrayexpress/*

67. **Environment, Drugs and Gene Expression Database (EDGE):** McArdle Laboratory for Cancer Research, Madison, WI, USA (2006). *http://edge.oncology.wisc.edu/*

68. Mattingly CJ, Colby GT, Forrest JN, Boyer JL: **The Comparative Toxicogenomics Database (CTD).** *Environ Health Perspect* (2003) **111**(6):793-795.

69. **CAS:** Chemical Abstract Services, Columbus, OH, USA (2006). *http://www.cas.org/*

70. Marris E: **News feature: Chemical reaction.** *Nature* (2005) **437**(7060):807-809.

71. Heller SR, Stein SE, Tchekhovskoi DV: **Open source/open access/open data and the IUPAC International Chemical Identifier - InChI.** *American Chemical Society National Meeting*, Washington, DC, USA (2005):CINF-60.

72. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI: **WOMBAT: World of molecular bioactivity.** In: *Cheminformatics in Drug Discovery*. Oprea TI, Mannhold R, Kubinyi H, Folkers G (Eds), Wiley-VCH, Wienheim, Germany (2004):223-239.

73. **DSSTox Quality Chemical Information Review Procedures:** US Environmental Protection Agency, Washington, DC, USA (2006). *http://www.epa.gov/nheerl/dsstox/ChemicalInfQAProcedures.html*

• *This web site is expected to be available from May 2006.*

74. Gold LS, Slone TH, Ames BN, Manley NB, Garfinkel GB, Rohrbach L: **Carcinogenic Potency Database.** In: *Handbook of Carcinogenic Potency and Genotoxicity Databases*. Gold LS, Zeiger E (Eds), CRC Press, Boca Raton, FL, USA (1997):1-606. *http://potency.berkeley.edu/CRCbook.html*

75. Gold LS, Manley NB, Slone TH, Rohrbach L: **Supplement to the Carcinogenic Potency Database (CPDB): Results of animal bioassays published in the general literature in 1993 to 1994 and by the National Toxicology Program in 1995 to 1996.** *Environ Health Perspect* (1999) **107**(Suppl 4):527-600.

76. Gold LS, Manley NB, Slone TH, Rohrbach L, Garfinkel GB: **Supplement to the Carcinogenic Potency Database (CPDB): Results of animal bioassays published in the general literature through 1997 and by the National Toxicology Program in 1997 and 1998.** *Toxicol Sci* (2005) **85**(2):747-808.

77. **The Carcinogenic Potency Project:** Carcinogenic Potency Database Project, Berkeley, CA, USA (2005). *http://potency.berkeley.edu/*

• *Provides a wealth of data from thousands of published experiments on chemical carcinogenicity for > 1500 chemical substances tested in several species, and also includes chemically indexed summary data pages*.

78. **Summary table by chemical of the Carcinogenic Potency Database:** Carcinogenic Potency Database Project, Berkeley, CA, USA (2005). *http://potency.berkeley.edu/chemicalsummary.html*

79. **Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network: SDF download page: CPDBAS:** US Environmental Protection Agency, Washington, DC, USA (2006). *http://www.epa.gov/nheerl/dsstox/sdf_cpdbas.html*

80. **Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network: More on DSSTox standard chemical fields:** US Environmental Protection Agency, Washington, DC, USA (2006). *http://www.epa.gov/nheerl/dsstox/MoreonStandardCF.html*

81. **Carcinogenic Potency Project: All results for each chemical:** Carcinogenic Potency Database Project, Berkeley, CA, USA (2005). *http://potency.berkeley.edu/chemnameindex.html*