

# DSSTox Website Launch: Improving Public Access to Databases for Building Structure-Toxicity Prediction Models

**Ann M. Richard**

US Environmental Protection Agency, Research Triangle Park, NC, USA

***Distributed:*** Decentralized set of standardized, field-delimited databases, each separately authored and maintained, that are able to accommodate diverse toxicity data content;

***Structure-Searchable:*** Standard format (SDF) structure-data files that can be readily imported into available chemical relational databases and structure-searched;

***Tox:*** Toxicity data as it exists in widely disparate forms in current public databases, spanning diverse toxicity endpoints, test systems, levels of biological content, degrees of summarization, and information content.

## INTRODUCTION

The economic and social pressures to reduce the need for animal testing and to better anticipate the potential for human and eco-toxicity of environmental, industrial, or pharmaceutical chemicals are as pressing today as at any time prior. However, the goal of predicting chemical toxicity in its many manifestations, the 'T' in 'ADMET' (adsorption, distribution, metabolism, elimination, toxicity), remains one of the most difficult and largely unmet challenges in a chemical screening paradigm [1]. It is widely acknowledged that the single greatest hurdle to improving structure-activity relationship (SAR) toxicity prediction capabilities, in both the pharmaceutical and environmental regulation arenas, is the lack of sufficient high quality data, for sufficient diversity of chemical structures, for the many and varied toxicity endpoints of potential concern. That toxicity endpoints can range from gross phenomenological disease measures (e.g., cancer, developmental malformations, hepatotoxicity), to hormone disruptions caused by highly specific receptor interactions (e.g., androgen or estrogen receptor binding), and can be metabolism-dependent and species/sex/tissue-specific, add further layers of complexity and challenge to this problem [2]. With the added recognition that toxicity data, particularly from whole animal studies, are a diminishing resource that will not likely be expanded significantly in the future, it is paramount to be able to fully mine the chemical toxicity data that currently exists [3].

There are two general informatics trends in biology that have facilitated explosive advances in genomics and structural biology, as well as many other areas of study: 1) data standardization and 2) on-line, open-access to well-documented data. Data standardization enables the collation of large amounts of data from disparate sources into a usable form for searching across standardized metrics. On-line, open access to data brings broad and varied intellectual capabilities to bear on data analysis, providing the fuel to feed the engine of scientific advancement. Two examples that support this assertion are the PDB (Protein Data Bank, <http://www.rcsb.org/pdb/>), a widely used public repository of crystallographic structure data [4], and the NCBI (National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/>), a treasure trove of on-line databases and bioinformatics data mining tools for exploring public genomics information. From an informatics standpoint, however, historical toxicity data present some difficult and unique challenges that have confounded efforts to create a centralized public data repository. Not only do these data exist in many formats and locations in the public domain, but they also span many levels of biological organization, detail, degrees of summarization and annotation, and disciplines of toxicological study [5,6]. The only common thread and shared information metric that truly has the potential to span and unify these disparate data is the molecular structure of the test chemical, and the underlying chemistry that it represents.

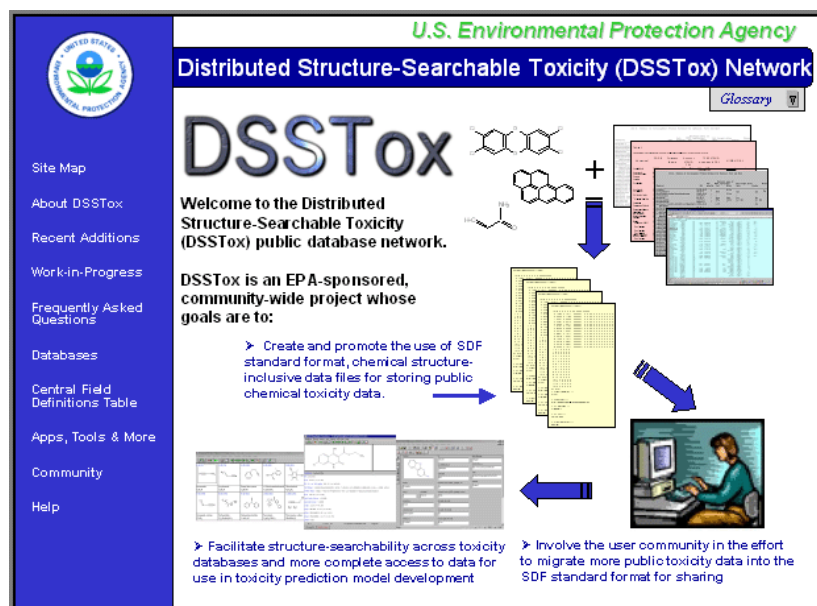


Figure 1. DSSTox website home page.

## DSSTOX DATABASE NETWORK

A primary objective of the DSSTox website [<http://www.epa.gov/nheerl/dsstox>] is to serve as a central community forum for publishing standard-format, structure-annotated chemical toxicity data files for open-access, public use (see Figure 1). In this initial launch phase, data files cannot be structure-searched on the DSSTox website itself, but the data files can be downloaded in their entirety and freely used by individuals, corporations, government agencies, commercial vendors, or other public websites to enhance in-house or public SAR and structure analogue search capabilities. The website is scheduled to launch in early March, 2004 with the publication of four distinct DSSTox databases, each representing a separate toxicity data source collaboration:

- CPDB – Carcinogenic Potency Project Summary Tables for Rats & Mice (CPDBRM, 1254 chemicals), Hamsters (CPDBHA, 80 chemicals), Dogs (CPDBDO, 5 chemicals), and Non-human Primates (CPDBPR, 26 chemicals); Source, L.S. Gold, Carcinogenic Potency Project, UC Berkeley, <http://potency.berkeley.edu> [7-9].
- DBPCAN – EPA Disinfection By-products Carcinogenicity Estimates Database, 207 chemicals; Source, Y.T. Woo, US EPA, Office of Pollution Prevention & Toxics [10,11].
- EPAFHM – EPA Fathead Minnow Acute Toxicity Database, 617 chemicals; Source, C. Russom, US EPA, Mid-Continental Ecology Division-Duluth [12,13].
- NCTRER – FDA NCTR Estrogen Receptor Binding Database, 232 chemicals; Sources, W. Tong and H. Fang, US FDA, National Center for Toxicological Research, <http://edkb.fda.gov/index.html> [14,15].

Additional collaborations are in progress or have been initiated with researchers in varied toxicology disciplines to enlarge the list of DSSTox database offerings. To encourage and support community participation in further expanding this list, the DSSTox website provides a wide assortment of tools, templates, tutorials, outside links, contacts, and reference materials that can be used by any person or group wishing to construct and publish a DSSTox database, either by itself or as a supplement to an external website or journal publication. A goal of this project is to encourage potential database authors (i.e., DSSTox Sources) to adopt the standard data file format, standard chemical fields, and minimum documentation requirements, while retaining primary authorship, serving as the Main Contact, and presenting the database in the way that conveys the most essential toxicity information to chemists, modelers and others interested in using these files for improving SAR and analogue search capabilities. This construct ideally will

foster both broader access to more useful data, as well as greater communication and linkages among experts in diverse toxicology domains, chemists, modelers, and others interested in using these data [16].

## DSSTOX FILE, DATA, & DOCUMENTATION STANDARDS

Each DSSTox database is published as a separate and distinct module that adheres to standard conventions in SDF data file format, file names, chemical structure fields, and minimum documentation requirements. “SDF” (Structure Data Format) is a text-based file format originally developed by Molecular Design Limited [17] that currently serves as a *de facto* public standard for the import and export of chemical structure data in virtually all chemical relational database applications and molecular modeling programs. SDF files adhere to strict formatting rules pertaining to chemical structure representation and field names and values. These files can store 2D or 3D molecular structures and an unlimited number of chemical records and corresponding data or text fields. Hence, SDF files are very versatile: they can accommodate many types of data, are easily edited and manipulated by programming scripts, and could be easily ported to other types of standard formats, such as the mark-up languages, XML and CML.

Field Name	Allowable Values	Brief Definition
Structure	MDL SDF format	2D graphical structure
Formula	text	Empirical formula of displayed <b>Structure</b>
MolWeight	#	Molecular weight of displayed <b>Structure</b>
StructureShown	tested form simplified to parent predicted form general form active ingredient	Description of displayed <b>Structure</b>
CAS	##### ## #	CAS Registry Number of displayed <b>Structure</b>
SMILES	text	SMILES molecular text code for displayed <b>Structure</b>
DSSTox_ID	#	Sequential ID number of record in database that uniquely identifies database record when reported with DSSTox_FileName.
DSSTox_FileName	text	Full DSSTox file name without file extension
ChemName	text	Common or trade name of chemical listed in original Source database
SubstanceType	defined organic inorganic organometallic mixture or unknown	Nature of chemical listed in original Source database
TestedForm	parent salt complex unknown or multiple forms	Tested form of chemical listed in original Source database
CAS_TestForm	##### ## #	CAS Registry Number of <b>ChemName</b> chemical listed in original Source database, provided for reference purposes only in specialized Defined Organic Parent (DOP) file.
SMILES_TestForm	text	SMILES code of <b>ChemName</b> chemical listed in original Source database, provided for reference purposes only in specialized DOP file.
AddToParent	text	For a <b>SubstanceType</b> =defined organic and <b>TestedForm</b> =salt or complex, field entry lists chemical moieties that are removed to create simplified parent structure; information provided to document changes in <b>Structure</b> field from main SDF file to DOP file.
ChemNote	text	Text note field for providing miscellaneous additional chemical information to the database record
ChemCount	1 # of #	Counter field for locating and counting replicate structures or CAS numbers within the database

Table 1. Abbreviated Definitions of Standard Chemical Fields used by DSSTox.










The range of possible toxicity data fields is as diverse as the field of toxicology is broad. However, the chemical structure annotation of toxicity databases can be made to adhere to rather narrow standards in reporting that can serve to more accurately convey the chemical content of a particular toxicity database and greatly facilitate data parsing, analogue structure-searching, and SAR modeling efforts. DSSTox defines a set of standard chemical data fields (Table 1) for possible use that attempt to capture minimum desired chemical annotation for these data, with consideration for the ways in which toxicity databases are currently annotated and used by both toxicologists and modelers. For example, the **SubstanceType** field classifies a chemical substance according to the broad chemical categories: *defined organic*, *inorganic*, *organometallic*, or *mixture or unknown*. This allows the DSSTox SDF format to publish the full chemical content of toxicity databases, while at the same time allowing for easy segregation, in a single search step, of the portion of the data that is more amenable to SAR modeling. Similarly, the parent structure form of the tested chemical (i.e., the neutral, non-salt, non-complex form), or active ingredient of a formulation might be represented graphically in the structure field to facilitate structure-searching across diverse datasets. In this case, the retained knowledge in the database of the original tested form of the chemical might be very pertinent to proper interpretation and use of these data in SAR analysis. These standard chemical fields, or some subset of these fields applicable to the content of a particular database, will span all published DSSTox databases and accompany the diverse toxicity data fields (see Figure 2).

CPDBRM	DBPCAN	EPAFHM	NCTRER ....
DSSTox Standard Chemical Fields			
SAL CPDB	ChemClass DBP	ChemClass FHM	NCTRlogRBA
TD50 Rat	Concern Level	MOA	ER RBA
TD50Mouse	Rationale	MOACONF	ChemClass ERB
Target Sites Rat Male	Rational Source	CLOGP	Activity Group ERB
Target Sites Rat Female	Analog ChemName	LC50	Rationale ChemClass ERB
...	AnalogCAS	LC50NOTE	MeanChem Class ERB RBA
Other Species	Analog SMILES	LC50RATIO	LogP
		MIXMOA	F1, F2, ...F6
		TOXINDEX	
		FATS	
		BEHAVIOR	

**Figure 2. DSSTox standard chemical fields spanning diverse toxicity databases.**

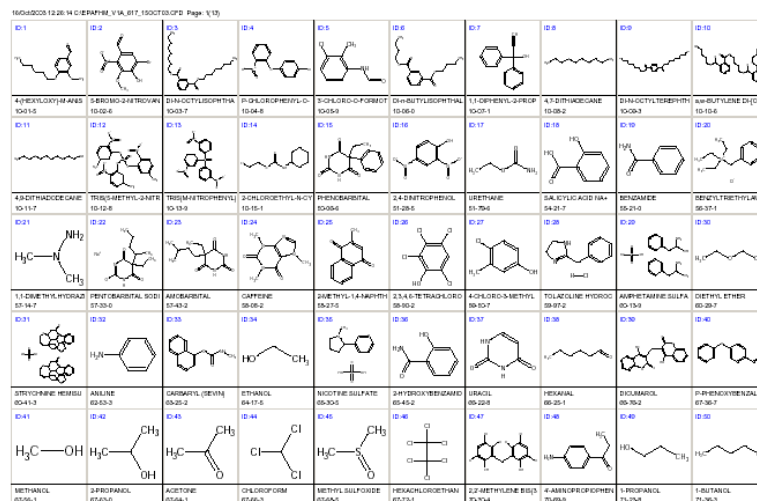
DSSTox standard chemical fields listed in Table 1 (or some subset) are included in all DSSTox database files, whereas toxicity data fields are particular to each toxicity database and typically nonstandard and nonoverlapping. Sample toxicity data fields are listed for four DSSTox databases.

In addition to the standards incorporated into the DSSTox SDF data file(s), each published DSSTox database will adhere to standard file naming conventions and documentation requirements. A sample data/documentation file download table, appearing on the main “Source SDF Download Page” for the DSSTox EPAFHM database, is shown in Figure 3.

File Type	Description	File Size	Format
Document Files			
Log File	EPAFHM_LogFile_19Oct03.pdf	63KB	
Field Definition File	EPAFHM_FieldDefFile_19Oct03.pdf	137KB	
	EPAFHM_FieldDefFile_19Oct03.doc	109KB	
Data Files: <i>EPAFHM - Main File</i>			
SDF Structure/Data File	EPAFHM_v1a_617_15Oct03.sdf	*.zip 516KB	
• Data Table (no structures)	EPAFHM_v1a_617_15Oct03_nostructures.xls		
• Structures Table	EPAFHM_v1a_617_15Oct03_structures.pdf		
Data Files: <i>EPAFHM – Defined Organic Parent Structures Only</i> (i.e., excluding inorganics, organometallics, mixtures, and representing salts and complexes in simplified parent form):			
SDF Structure/Data File	EPAFHM_DOP_v1a_610_15Oct03.sdf	*.zip 506KB	
• Data Table (no structures)	EPAFHM_DOP_v1a_610_15Oct03_nostructures.xls		
• Structures Table	EPAFHM_DOP_v1a_610_15Oct03_structures.pdf		
<a href="#">File Error Report</a>			

**Figure 3. DSSTox data/document file download table for EPAFHM.** Download table that appears on the DSSTox EPAFHM Source SDF Download Page listing all available documentation and data files for the EPA Fathead Minnow Acute Toxicity Database. Main files include all chemicals, whereas the DOP files include only defined organics, with salts and complexes represented in parent structure form

The DSSTox SDF file naming convention includes a 6-letter NAMEID (e.g., EPAFHM), the version number and revision letter of the file (e.g., v1a), the total number of chemical records (e.g., 617), and the date of file creation (e.g., 15Oct03). This file name is used for primary reference in outside reporting, and any future updates or modifications to the file will be documented in the Log File document according to file name. The Field Definition File is intended to be a primary reference document for the SDF Structure/Data File. It consists of a summary description followed by a table defining each of the fields contained in the database, and including units of measure and allowable field entries. As an adjunct to the SDF Structure/Data File, for users unable to view the SDF, we include a Microsoft Excel spreadsheet Data Table that contains all data fields except the graphical **Structure** field. To supplement the Excel Data Table and provide users with a quick visual overview of the structural content of the databases, the Structures Table pdf file provides a tiled graphics view of all structures contained in the SDF Structure/Data File (see Figure 4).



**Figure 4. A portion of the DSSTox Structures Table pdf for EPAFHM.**

## CONCLUSIONS

A larger, more broadly encompassing goal of “chemo-bioinformatics” and computational toxicology is to facilitate data integration and exploration across both chemical and biological data domains. The DSSTox project is a first step in this direction to the extent that it encourages greater data standardization, increased access to diverse public toxicity data, and structure-searchability through these data. Future goals of the DSSTox project are to expand the list of published toxicity databases, particularly in traditionally underrepresented areas of toxicology (e.g., such as immunotox and neurotox), partner with outside public efforts to provide on-line structure-analog searching capabilities through DSSTox databases, and coordinate with other public data standardization efforts in the fields of toxicology and genomics. Annotating genomic databases and historical toxicity databases with the same set of DSSTox standard chemical fields would provide a common search metric for exploring these two large data domains from a chemical structure perspective. The ability to gather diverse biological data relative to common structural analogues has the potential to greatly expand and deepen our ability to generate useful, biologically-based SAR hypotheses. The DSSTox project is building the data foundation that will enhance opportunities to explore common chemistry and structural correlations spanning traditionally disconnected areas of toxicology and pharmaceutical research.

## ACKNOWLEDGMENTS

A large number of persons have contributed to the development of the DSSTox concept, website, and databases. Many others have assisted in the quality review of website materials and provided many helpful and constructive suggestions. A full listing of all database collaborators and contributors of published DSSTox databases are provided in refs. [9,11,13,15]. Additional acknowledgements are listed on the DSSTox website, at <http://www.epa.gov/nheerl/dsstox/Acknowledgments.html>. *This manuscript has been reviewed by the US Environmental Protection Agency 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

1. **Johnson D.E. and G.H.I. Wolfgang.** 2000. Predicting human safety: screening and computational approaches. *Drug Discovery Today* 5:445-454.
2. **Richard, A.M.** 1999. Application of Artificial Intelligence and Computational Methods to Predicting Toxicity. *Knowl. Eng. Rev.* 14:307-317.
3. **Richard, A.M., R. Benigni.** 2002. AI and SAR approaches for predicting chemical carcinogenicity: Survey and status report. *SAR and QSAR in Environ. Res.* 13:1-19.
4. **Schmidt, T.S.** 2002. Banking on Structures: An explosion of structural information is on the horizon and the Protein Data Bank – the single international repository for data on the three-dimensional structures of biomolecules is ready. *Bio-IT World*, Oct. 9, <http://www.bio-itworld.com/archive/100902/banking.html>
5. **Richard A.M. and C.R. Williams.** 2003. Public sources of mutagenicity and carcinogenicity data: Use in structure-activity relationship models. In *QSARs of Mutagens and Carcinogens*, Ed. R. Benigni, CRC Press, NY, pp. 51-179.
6. **Richard A.M. and C.R. Williams.** 2002. Improving structure-linked access to publicly available chemical toxicity information. *Curr. Opin. Drug Disc. Devel.* 5:136-143.
7. **Gold, L.S., T.H. Slone, B.N. Ames, N.B. Manley, G.B. Garfinkel, and L. Rohrbach.** 1997. Chapter 1: Carcinogenic Potency Database. In: *Handbook of Carcinogenic Potency and Genotoxicity Databases*,



Eds. Gold, L.S., and E. Zeiger, Boca Raton, FL: CRC Press, pp. 1-605.  
<http://potency.berkeley.edu/text/methods.html>

8. **Gold, L.S., N.B. Manley, T.H. Slone, and L. Rohrbach.** 1999. 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.* 107(Suppl. 4):527-600. <http://ehpnet1.niehs.nih.gov/docs/1999/suppl-4/toc.html>
9. **Gold L.S., T.H. Slone, C.R. Williams, J.M. Burch, T.W. Stewart, A.E. Swank, J. Beidler, and A.M. Richard.** 2003. DSSTox Carcinogenic Potency Database Summary Tables for Rats and Mice, Hamsters, Dogs, and Non-human Primates (**CPDBRM**, **CPDBHA**, **CPDBDG**, **CPDBPR**): SDF Files and Documentation. [www.epa.gov/nheerl/dsstox/](http://www.epa.gov/nheerl/dsstox/)
10. **Woo, Y.T., D. Lai, J.L. McLain, M.K. Manibusan, and V. Dellarco.** 2002. Use of mechanism-based structure-activity relationships analysis in carcinogenic potential ranking for drinking water disinfection by-products. *Environ. Health Perspect.* 110 Suppl 1:75-87.
11. **Woo Y.T., C.R. Williams, N.Fields, and A.M. Richard.** 2003. DSSTox EPA Water Disinfection By-Products With Carcinogenicity Estimates (**DBPCAN**): SDF Files and Documentation. [www.epa.gov/nheerl/dsstox/](http://www.epa.gov/nheerl/dsstox/)
12. **Russom, C.L., S.P. Bradbury, S.J. Broderius, D.E. Hammermeister, and R.A. Drummond.** 1997. Predicting modes of action from chemical structure: Acute toxicity in the fathead minnow (*Pimephales promelas*). *Environ. Tox. Chem.* 16:948-967.
13. **Russom C.L., C.R. Williams, T.W. Stewart, A.E. Swank, and A.M. Richard.** 2003. DSSTox EPA Fathead Minnow Acute Toxicity Database (**EPAFHM**): SDF Files and Documentation. [www.epa.gov/nheerl/dsstox/](http://www.epa.gov/nheerl/dsstox/)
14. **Fang, H., W. Tong, L.M. Shi, R. Blair, R. Perkins, W. Branham, B.S. Hass, Q. Xie, S.L. Dial, C.L. Moland, and D.M. Sheehan.** 2001. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Tox.* 14:280-294.
15. **Tong W., H. Fang, C.R. Williams, J.M. Burch, and A.M. Richard.** 2003. DSSTox National Center for Toxicological Research Estrogen Receptor Binding Database (**NCTRER**): SDF Files and Documentation. [www.epa.gov/nheerl/dsstox/](http://www.epa.gov/nheerl/dsstox/)
16. **Richard A.M. and C.R. Williams.** 2002. Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network: A Proposal. *Mutat. Res.* 499:27-52.
17. **Dalby A., J.G. Nourse, W.D. Hounshell, A. Gushurst, D.L. Grier, B.A. Leland, and J. Laufer.** 1992. Description of several chemical-structure file formats used by computer-programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* 32:244-255.

#### **Address Correspondence to:**

Ann M. Richard  
Mail Drop B143-06  
National Health & Environmental Effects Research Lab  
U.S. Environmental Protection Agency  
Research Triangle Park, NC USA 27711  
e-mail: [Richard.Ann@epa.gov](mailto:Richard.Ann@epa.gov)