A Brief Introduction to ToxML

Reference:	TSO/2012/11/1
Authors:	P N Judson, M A Ali
Version:	1.1
Created:	19th November 2012
Last modified:	28th November 2012
Status:	Final

c/o Lhasa Limited
22-23 Blenheim Terrace
Woodhouse Lane
Leeds LS2 9HD
England

A Brief Introduction to ToxML

Background

Scientific data generated from the biomedical disciplines that could be used for data mining, analytics and modelling is continually growing. The challenge with exploiting this resource is that data are held in a multitude of different formats. Restructuring this information to visualise, share and combine is laborious, error prone and time consuming.

To support the wide exchange of data in the absence of an exchange standard, each software developer would need to negotiate with every other to develop specific data exchange modules. This is illustrated in figure 1a, where each of four database systems needs to incorporate three different data exchange interface modules. Figure 1b illustrates how having a common exchange standard means that only one data exchange interface module is needed for each application. A developer wishing to support exchange for an additional application needs only to know the specification for the standard. A general purpose editor for entering data into standard files might or might not be needed because, as Figure 1b illustrates, users of different applications can edit data locally using the editor they are familiar with.

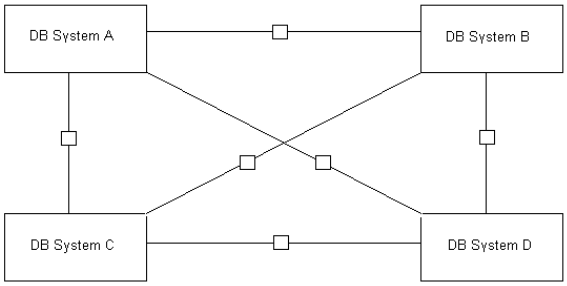

Figure 1a: no exchange standard

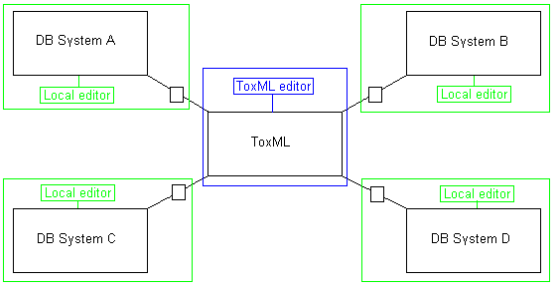

Figure 1b: common exchange standard

What ToxML is and what it is not

ToxML is used by a number of organisations, including the US FDA, Leadscope Inc., and Lhasa Limited. It is a data exchange standard based on the Extensible Mark-up Language (XML), for toxicological and related data (including chemical structures). It defines the structure of a data file for communication between software applications and contains a controlled vocabulary.

ToxML does not impose standardisation of toxicological terms between different users. Terms are standardised within ToxML for data transfer but can be converted during export/import to terms appropriate to a particular site or application.

The ToxML specification editor is freely available on the ToxML website (www.toxml.org) allowing users to view, and to submit changes to, the structure of the standard. It is not a data entry tool and it is not part of the remit of the ToxML Standard Organisation (TSO) to provide one. The expectation is that software vendors will provide ToxML export/import functions for their database management applications and a stand-alone ToxML data entry tool covering some areas of toxicology is already available from Leadscope Inc.

The entire ToxML definition is large and will become much larger. However, a ToxML file will normally include only the parts of the standard required for the data it contains. The import function of an application should tolerate, and normally ignore, fields in a ToxML file that are not relevant to it and treat the absence of fields as simply an absence of data.

Existing standards

The SEND format has been developed for submission of registration data to the US FDA and it covers data relevant to registration. As a basis for development of a wider-ranging, long term standard it has the disadvantages of not supporting hierarchical data structures and of depending on transmission of sets of files rather than a single file.

Structure data format (sdf) files are widely used for transmission of chemical structures with limited amounts of associated data. The format is used, for example, for DSSTox files from the US EPA National Center for Computational Toxicology. The sdf format was designed primarily for chemical structure data. While it supports the inclusion of, for example, biological data it is not ideal for large amounts of such data and it does not support hierarchical data.

The IUCLID database schema is used in Europe for regulatory submissions. IUCLID was the first full implementation of the OECD harmonised templates (OHTs). Some changes have been made since but work is in hand to make the standards the same. IUCLID and the OECD harmonised templates, like SEND, are primarily intended for regulatory submission data.

ToxML therefore is expected to have broader coverage than SEND and IUCLID, but the objective is for ToxML to cover all the fields included in the above formats and for there to be good consistency with them for those fields.

ToxML and the choice of XML

Although it has only become more widely known recently, the project to develop ToxML started alongside early work on the OHTs. The two have been developed side by side and have much in common. Work has been done on both by the same individuals. However, the primary intended use of ToxML is for the exchange of data relevant to (Q)SAR modelling, which requires a broader coverage of data than that required for regulatory submissions. ToxML is also intended to cover a wider range of end points than the templates and IUCLID. The aim is for those parts that the standards have in common to be fully compatible. Discussions are in progress to provide easy interconversion also between ToxML and SEND.

Like the OHTs, ToxML uses the XML format. XML was chosen because it supports a hierarchical data structure. It is well-suited to modern data communication, allowing all data to be contained in a single file format and it is not bound to a specific software application or programming language. Program development platforms provide code and functions to allow software developers to write to and read from XML files easily, and there are free XML browsing applications to allow a user to look at the contents of an XML file (and hence a ToxML file).

Lhasa Limited have developed a modified sdf format which supports hierarchical data and this was considered as the possible basis for the standard. However XML was preferred because it is so widely used, whereas the extended sdf is peculiar to Lhasa Limited and offers no advantages over XML.

Incremental, community development of ToxML

> ñThe secret of getting ahead is getting started. The secret of getting started is breaking your complex, overwhelming tasks into small, manageable tasks and then starting on the first one.ò
>
> Attributed to Mark Twain

To develop an exchange standard for toxicity data covering all possible end points is a huge task. Creating such a standard as a complete package, even a preliminary one, by committee would be very costly and slow. Much time might be spent on developing parts of the standard that, in practice, were never used. Therefore a wiki site model approach has been chosen. This allows two or more parties who wish to exchange data that are not already covered by ToxML to make provisional additions to a draft of the standard. Other ToxML users can comment on the proposals and suggest modifications. A curator reviews them, checking for and resolving conflicts or duplications, and adds the proposals to the live standard. Thus the standard will grow according to need but in a controlled way.

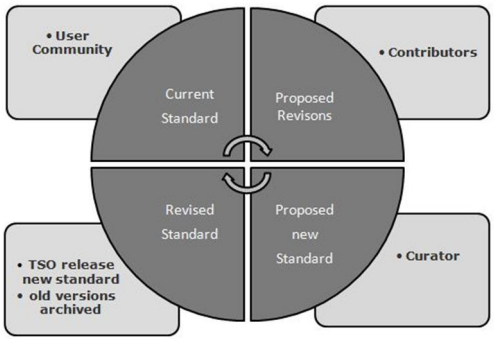

Figure 2: controlled development of standard via wiki website

The TSO has been set up to develop and maintain the standard. It is governed by an Advisory Board elected by ToxML users. Curation of the standard and operation of the wiki site are the responsibilities of a management organisation appointed by the Advisory Board.

Further information and becoming involved

The current standard, general information, the TSO constitution, a discussion forum, and contact details can be found on the ToxML website: **www.toxml.org.** You can get access to the editor, and submit changes and additions to the standard, by registering on the site.

TSO is keen to encourage use of ToxML and its expansion through voluntary use of the editor on the website. If ToxML does not currently meet your needs please log onto the site and make additions to the draft standard. The success of the standard depends on development in this way by people like you. Proposals will be reviewed promptly by a curator and will be incorporated into the new release version. If changes to your proposals

are needed, for consistency with existing components of the standard or because of differing proposals from different contributors, the curator will discuss them with you.

Philip N Judson
philip.judson@blubberhouses.net
Mohammed A Ali
ash.ali@lhasalimited.org