

# **Commercial Laboratory Perspectives on Automated Analytical Data Validation**

## **Author(s) Names**

Brad Mosakowski and Anne Wilhoit  
Air Toxics Ltd.

## **ABSTRACT**

In the past decade, many portable web based data solutions were introduced which improved efficiency, reduce costs or increase the quality and speed of complex data handling. However, a general lack of standardization in the environmental industry has prevented stakeholders from realizing any real benefit from these significant software developments. Stuck in the technology of the 1990's, most environmental data is currently shared via hardcopy report and pdf. Those attempting electronic file sharing are faced with a staggering number of possible electronic deliverable formats. The complexity of diverse formats, require continued maintenance, needlessly slow down project timelines and significantly increase project costs. The introduction of Staged Electronic Data Deliverables (SEDD) attempts to correct this problem by offering a universal complete and consistent data transfer format. The SEDD format can fulfill a number of objectives: improved transparency (supplying a complete set of fundamental raw data), web based transportability and immediate data usability when combined with tools enabling independent software validation. Agency oversight and public scrutiny in many environmental programs have created a need for more efficient and cost effective ways to provide data of known and documented quality. Presented is a commercial laboratory perspective supporting the SEDD as a universal format for direct linkage into such business standard software applications as fault tree analysis enabling point of use data validation with an unprecedented degree of confidence.

## **INTRODUCTION**

The rapid growth of the environmental industry in conjunction with the technology advances of recent decades, created a multitude of potential data formats for the transfer of analytical results. Although the technology revolution has continued to evolve along industry standards acceptability (ie Microsoft, SQL, WORD) the environmental industry has continued to lag behind and resist the acceptance of standards. It is not uncommon for a commercial laboratory to receive requests for analytical data in formats created more than a decade ago; i.e. ERPIMS. The Staged Electronic Data Deliverable (SEDD) has offered a credible alternative to become the standard format for the delivery and review of analytical laboratory data. SEDD was first introduced in 2003 but has not been widely accepted. The EPA's website ([www.epa.gov/superfund](http://www.epa.gov/superfund)) currently lists only 30 labs able to produce the format and 4 commercial entities accepting the format.

## **HISTORY**

With the advent of personal computers and internet service providers, many types of data can be generated, transferred, reviewed and used for decision making purposes. At the

analytical level, many differing electronic deliverable formats (EDDs) were designed to transfer the instrument results to the end-user. Unfortunately, these efforts simulated the existing limitations of the data systems and instrumentation. These EDDs varied not just for each client, but at times for each project. Air Toxics Limited provided over 300 distinct formats in the 1990s. Although the computing industry has matured since that time, the environmental EDD formats used today are still based on the technology available at the time of their conception. The current trend in computing is to offer unique services in conjunction with open platforms which are based on market consensus. This approach allows professionals to worry less about legacy protocols and data formats, extending their efforts to creating new ways to interpret the data.

## **THE FUTURE**

The EPA has developed and released a new EDD layout: SEDD (Staged Electronic Data Deliverable). This format is based in the XML (eXtensible Mark-up Language) originally developed to enhance the exchange of data. Its primary purpose is to facilitate the sharing of data across different information systems, particularly via the Internet[1]. The developers of the major computer operating systems recognized the importance of creating universal data standards. The SEDD format was created to universalize electronic data deliverables and data portability. The SEDD captures the required data elements needed for complete analytical results evaluation. There are many providers of software products within the environmental industry – databases, data visualization, modeling, data validation, LIMS and many others. All of these products could be set up to import and export data based on the SEDD specification.

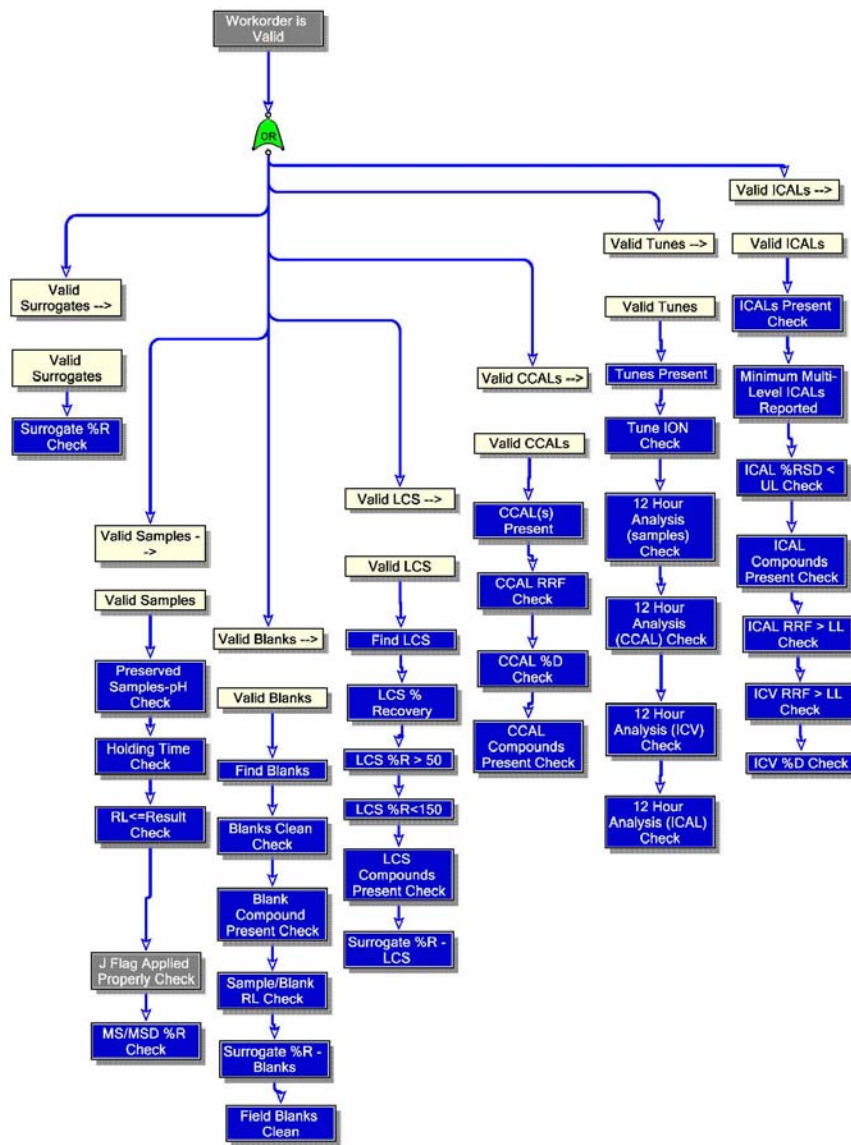
As SEDD contains every possible data element associated with a laboratory sample, it actually houses more information than any other commonly used layouts. Thus, all databases can be populated using SEDD and the extraneous information ignored. This universal file can be shared between the laboratory and the sampling company, between the sampling company and the governing agency, and between agencies. SEDD is also the only available format that is capable of containing enough information to validate data to EPA's Level IV designation, thus providing the basis for electronic data validation routines.

## **AUTOMATED DATA VALIDATION**

As the possibilities are examined, why does the industry continue to pursue manual data verification and validation? It is not uncommon for a laboratory today to receive requests for antiquated hard-copy reproductions of the complete analytical process. The State of New Jersey just recently released a deliverable guidance document in April of 2007 requiring all laboratories to abort their paperless LIMS review systems and regress to two chemist signatures per page of computer print out to be copied, bound and shipped back to New Jersey data validator(s)[4]. This begins the tedious and time consuming process of human data validation. Paper does not offer labor saving bookmarks, searching routines, internet or CD portability, data mining capabilities, or any level of efficiency. Paper simply offers the ability to make annotations that reside in a file cabinet for years and consume endless hours of manual human inspection.

The value of computers is that they can quickly and more efficiently complete repetitive tasks especially searches, comparisons and calculations. They are most useful in areas where they can be programmed to evaluate data based on specific criteria. The process of Data Verification/Validation is essentially a process of asking well defined questions. Data is then qualified according to a prescribed set of rules called “Fundamental Guidelines for Data Validation”. If software was developed that incorporated all of Functional Guidelines and could instantaneously deliver a summarized set of discrepancies against these rules, then more human time could be spent on project set up, quality objectives, QAPP review, data interpretation, defining usability, and client education. Effective use of the computer resource enables the human resource to provide truly value added services that require professional skills and judgment.

Figure 1: Fault Tree Data Validation Logic



There are commonly accepted Business Process Management software applications which fulfill a similar role. These applications allow businesses to relegate the details of managing process integrity and exception localization to a computer. Software automatically evaluates business rules and assigns process tasks to the correct system, group, or individual at the appropriate time. Data Verification/Validation by definition is a business process. In conjunction with an open, portable data format such as SEDD, there is an opportunity to fully realize the labor saving benefits of computerization.

The use of software to perform common, well-defined tasks also provides a level of consistency and transparency offering benefits to both the consumer and the provider. Projects with the greatest potential hazard require the highest level of scrutiny, thus requiring the most time and effort to prepare a data usability summary. With the possible health risk decisions to be made, maximizing the efficiency of this process would have a dramatic impact on how quickly conclusions and corrective actions could be made. There is an opportunity to provide not just data management, but also support for distributed process execution. With the right software, a computer can automatically validate an SDG in several seconds from a file created at the laboratory and sent immediately through the internet. The software can identify data flaws in a highly consistent manner, recalculate 100% of the values, add correct qualifying flags, and alert the project chemist of any areas that require professional judgment. Electronic data validation is nothing more than an independent software routine (software independent of that used to derive the result) checking the results produced by the primary software systems.

Figure 2: Example Electronic Validation Report

<p><b>VII. Laboratory Control Spikes (LCS)/ Control Spike Duplicates (LCSD)</b></p> <p>Data for LCS standards are generated to determine to determine the instrument/analyst precision and accuracy. These standards are derived from a source other than that used for the initial calibration to determine accuracy. The frequency of LCS per analytical batch is project specific. Evaluation criteria include verification that:</p> <ol style="list-style-type: none"> <li>1. the LCS and LCSD were analyzed at the required frequency</li> <li>2. re-calculation of all LCS and LCSD results match those reported</li> <li>3. compare calculated results against defined acceptance limits</li> </ol> <p>All LCS/LCSD compounds were within QC limits with the following exceptions:</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;">Compound</th> <th style="text-align: left;">Instrument</th> <th style="text-align: left;">Date</th> <th style="text-align: left;">% Rec</th> <th style="text-align: left;">Bias</th> <th style="text-align: left;">Affected Samples</th> <th style="text-align: left;">Flags</th> </tr> </thead> <tbody> <tr> <td>Chloroethane</td> <td>msd4.i</td> <td>4/8/2005</td> <td>62%</td> <td>n/a</td> <td>IA-018-I-03, IA-018-I-04, IA-018-I-05, IA-018-I-05 Duplicate, IA-018-I-06, IA-018-I-07, IA-018-I-07 Duplicate, IA-018-I-08, OA-018-G-02</td> <td>Q3</td> </tr> </tbody> </table>							Compound	Instrument	Date	% Rec	Bias	Affected Samples	Flags	Chloroethane	msd4.i	4/8/2005	62%	n/a	IA-018-I-03, IA-018-I-04, IA-018-I-05, IA-018-I-05 Duplicate, IA-018-I-06, IA-018-I-07, IA-018-I-07 Duplicate, IA-018-I-08, OA-018-G-02	Q3
Compound	Instrument	Date	% Rec	Bias	Affected Samples	Flags														
Chloroethane	msd4.i	4/8/2005	62%	n/a	IA-018-I-03, IA-018-I-04, IA-018-I-05, IA-018-I-05 Duplicate, IA-018-I-06, IA-018-I-07, IA-018-I-07 Duplicate, IA-018-I-08, OA-018-G-02	Q3														
<p><b>VIII. Internal Standards</b></p> <p>Internal standard performance criteria ensures the GC/MS sensitivity and response are stable during each sample, blank and standard analysis. Internal standards are non-target volatile organic compounds selected to mimic the target compounds and spiked at known concentrations prior to analysis. The internal standards are used to monitor instrument performance during actual sample analysis; and to quantify tentatively identified compounds (unknowns). Evaluation criteria include verification that:</p> <ol style="list-style-type: none"> <li>1. IS retention time and areas are within criteria for all samples and blanks</li> <li>2. IS areas meet defined acceptance limits</li> </ol> <p>All internal standard compounds were within QC limits with the following exceptions:</p> <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="padding: 5px;">There were no exceptions. All data was within validation criteria.</td> </tr> </table>							There were no exceptions. All data was within validation criteria.													
There were no exceptions. All data was within validation criteria.																				

The following examples show how SEDD and electronic data validation can be used in real project situations:

**Example 1:** Air Toxics Ltd., has utilized Quality Software Tools Data Assessor™ application to perform 100% data validation in our laboratory since 2004. When data is

released to the client, 100% of the values have already been computer validated following National Functional Guidelines. This is a much stricter validation than a project chemist could manually perform along with eliminating >95% of chemist/QA review time. The software was utilized on an EPA oversight residential indoor air vapor intrusion project in Hartford, IL. Approximately 6000 TO-15 analyses were analyzed over a three year period of time. All of the samples underwent traditional manual independent validation. Although we have received the occasional call usually two or three months after data submission regarding minor compilation issues in the data packages (ie. missing pages), the secondary validation effort did not report any significant data quality findings.

**Example 2:** The major driver behind current independent validation programs is laboratory data fraud. The fear that data from the lab cannot be trusted fuels the perceived need for an independent set of eyes. Utilizing the SEDD deliverable and software validation, an automated search for common inappropriate practices are quickly and easily performed. These include BFB scan selection, CCV manipulations, wrong ICAL or no ICAL, and manual integrations. Even the most recent data fraud settlement involving aborted sample run times [2] could have been discovered much earlier if automated software validation had been employed. The lab had been performing the action for several years yet went undiscovered by human data validation for some time. The laboratory involved is on the list of SEDD participants. A simple software rule created in ten minutes or less to compare sample run time to CCV final target compound run time would have saved thousands of dollars in resampling costs and court fees.

**Example 3:** It is also possible to bypass the SEDD deliverable and embed the decision tree validation rules into any laboratory LIMS. Validation at the point of data generation identifies common lab mistakes in data analysis or reporting and prompts user correction prior to release. These include such routine errors as incorrect compound sublist, incorrect initial calibration, expired hold time to analysis, incorrect dilution factor and wrong sample identifier. The Air Toxics Limited LIMS was embedded with Quality Software Tools Data Assessor validation routines in 2004. The three top pre-release errors identified by the embedded software included:

1. dilution factors miscalculated
2. CCV/LCS flagging applied incorrectly
3. incorrect ICAL referenced

## **CONCLUSION**

The acceptance and utilization of SEDD is a critical component in the continuing evolution of the environmental industry. Timely EPA release of SEDD Stage III is essential along with industry wide use of SEDD by all environmental stake holders. Testing laboratories must accept the challenge and help move this initiative forward by becoming proponents of the complete SEDD deliverable. Continuing business and scientific practices adverse to the internet age inhibits our industry's natural progression towards efficiency and sustainability. Businesses and agencies able to transport a

complete and transparent data set will establish new levels of service and data acceptability.

Note: The author will be available after the presentation to demonstrate electronic validation using a personal computer. Anyone with a compliant SEDD Stage III 8260B file is encouraged to bring a CD or memory key version for a demonstration validation and continued discussion.

## **REFERENCES**

1. Bray, Tim; Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau (September 2006). [Extensible Markup Language \(XML\) 1.0 \(Fourth Edition\) - Origin and Goals](#). World Wide Web Consortium. Retrieved on October 29, 2006
2. EPA Office of Inspector General (May 2007). EPA-350-R-07-002 Semiannual Report To Congress.
3. Guidance on Environmental Data Validation and Verification, QA/G-8, Final, November 2002, EPA//240/R-02/004, Office of Environmental Information, Quality Staff, U.S. Environmental Protection Agency, Washington, DC
4. [NJDEP-SRWM Low Level USEPA Method TO-15 \(NJDEP- LLTO-15\)](#)  
<http://www.state.nj.us/dep/srp/guidance/vaporintrusion/>
5. Quality Software Tools; <http://www.qualitysoftwaretools.com>