WHITEPAPER

Environmental Data Exchange and Standardization

Environmental Measurement Symposium /
National Environmental Monitoring Conference
August 9, 2012
Washington, D.C.

Robert J. Whitehead
Vice President
ChemWare, Inc.

CHEMWARE
LABORATORY INFORMATION MANAGEMENT

“The nice thing about
standards is that we have
so many to choose from.”

-**Andrew Tanenbaum**
Computer Scientist

CONTENTS

This whitepaper is a synopsis of Environmental Data Exchange and Standardization, presented at the Environmental Measurement Symposium/National Environmental Monitoring Conference on August 9, 2012 in Washington, D.C. The paper provides an historical perspective on environmental industry efforts toward data standardization, with recommendations on an approach to establishing a common data exchange standard that preserves existing databases and automated data review systems. See ChemWare's website (http://www.chemware.com/NEMC2012.aspx) for related conference materials from private and public sector panelists.

INTRODUCTION

In response to the passage of the 1980 federal law known as Superfund, the United States Environmental Protection Agency's Analytical Services Branch established the Contract Laboratory Program contracting vehicle to leverage the commercial laboratory industry's growing capacity in trace level analytical services and data management.

The CLP Statement of Work standardized the electronic data file format to be produced by contract laboratories in order to create the ability to programmatically validate the enormous amount of data being consumed by the agency. The adoption rate of the resulting "Format A" and "Format B" standards was 100% — in order to be awarded a CLP contract and get paid for testing services, a laboratory was required to submit (and pre-validate) analytical data in one of these two formats for every project.

Over the next ten years, the Department of Energy, Department of Defense, and the EPA worked on the Department of Energy Environmental Management Electronic Data Deliverable Master Specification (DEEMS), a new "universal" EDD standard. Published in 1995, DEEMS "face[d] much controversy and resistance to implementation specifically because of the size of its deliverable, its structure being hierarchical verses relational, and the breadth of information…not readily available electronically through current Laboratory Information Management Systems (LIMS)."

Ten years after that, the EPA and U.S. Army Corps of Engineers expanded upon the DEEMS standard with the Staged Electronic Data Deliverable (SEDD), part of an initiative to establish uniform processes for delivery, review, storage, and retrieval of chemical and radiological data. Conceptually, SEDD was intended to address the superset of all data that might be required by environmental data consumers across all agencies, "staged" to allow users to choose the formats (stages) necessary to meet their individual data reporting and validation (usability) needs.

INTRODUCTION

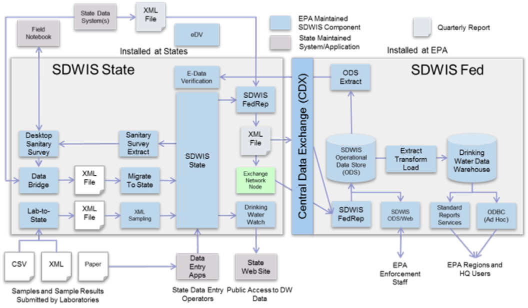


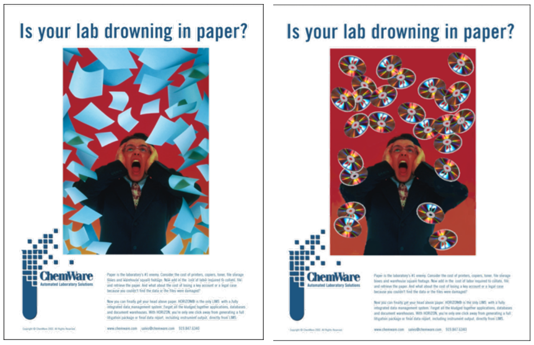
Figure 1: Excerpt from a presentation by Greg Fabian, PMP, at the May 30, 2012 Exchange Network National Meeting, Philadelphia, PA (http://www.exchangenetwork.net/en2012-agenda/)

While HORIZON® LIMS users have always been able to capture and generate the data necessary to comply with DEEMS and SEDD standards, environmental data consumers have made only marginal progress over the past twenty years toward convergence on a single EDD standard. Commercial environmental laboratories still commonly generate over one hundred different EDD formats for their public and private customers; the Safe Drinking Water Information System (SDWIS) components, databases, and data flows still vary between state and federal SDWIS agencies (Figure 1), and several EPA regional Superfund offices still require conversion of the CLP single file format into their own region-specific multiple file formats. The vision of a “super EDD” appears ephemeral at best.

THE PUSH TO PAPERLESS

The primary data management problems faced by most environmental laboratories are (1) the ability to convert paper-based records to secure, manageable electronic records; (2) the ability to quickly identify out-of-control results or other data processing excursions, in time to take corrective action and still meet customer turnaround times; and (3) the ability to mine and transform the data into the electronic format required by the customer. In 2002, ChemWare launched the first LIMS with a fully integrated Scientific Data Management System (SDMS), devised to solve the paper problem (Figure 2). Despite widespread success in capturing data electronically and extracting data from unstructured documents, regulatory-driven data consumers still demanded hardcopy reports as well as electronic data files. The laboratories formerly drowning in paper were now drowning in EDDs (Figure 3).

The primary data management problems faced by the regulatory agencies and other environmental data users are (1) the ability to programmatically consume the analytical data into a database in order to avoid manual data entry; (2) the ability to automatically verify the data conforms with contractual and/or regulatory requirements for electronic data submissions; and (3) the ability to validate the data against data quality objectives (sometimes referred to as "measurement quality objectives"). In order to solve the first problem, many agencies provide laboratories with spreadsheet templates or web-based data entry forms (Figure 4), which just shifts the data entry burden (and costs) to the laboratory. These templates and forms provide no mechanism for integration with LIMS or for parsing a standardized data file into the spreadsheet.


Is your lab drowning in paper?
ChemWare
Is your lab drowning in paper?
ChemWare

Figures 2 and 3: In 2002, "drowning in paper" was a common theme with laboratories. In 2012, the proliferation of EDDs has nullified the gains of going paperless.

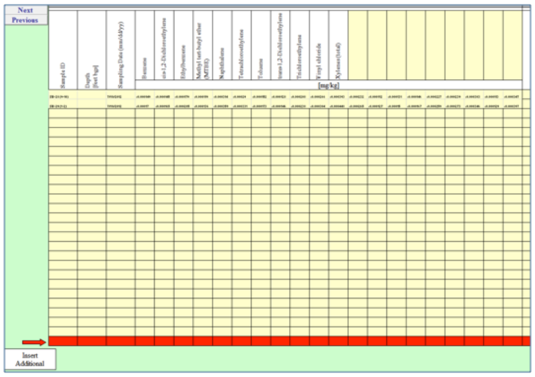


Figure 4: The North Carolina Department of Environment and Natural Resources (NCDENR) provides an Excel-based data entry template for laboratories to use when reporting data in support of the North Carolina Dry-Cleaning Solvent Cleanup Act Program (DSCA).

HEALTHCARE DATA PRECEDENT

For over twenty years, the healthcare industry has been exchanging clinical demographic and test results data electronically using the Health Level Seven (HL7) standard. While the standard is not perfect, the adoption rate has been nearly universal – it is used throughout the world by hospitals, clinical laboratories, public health agencies, and even manufacturers of diagnostic instrumentation (which commonly use HL7 protocols to exchange data with laboratory information systems). Medicare/Medicaid and other billing and payment systems use HL7 as well, which was a critical driver in accelerating technology adoption.

Today, most clinical data generators and consumers use integration engine software (also known as "information brokers") to transform and map variations in vocabularies and message segment formats in order to facilitate the automated exchange. These variations exist between different versions of HL7, which continue to evolve over time, and also because some users have created their own "variants" of the standard. Deviations, however, are easily handled through an integration engine and are relatively insignificant when compared with the variety of electronic formats and content across the environmental data universe.

THE PROBLEM WITH DOWNSTREAM MAPPING AND VALIDATION

If the data is transformed as it is being transmitted from the laboratory, and then electronically validated as it is being consumed by the regulator, the laboratory loses control over the quality of the data. According to Lean principles , non-value-added steps are to be excised from the process. To get it right the first time in the laboratory, data validation must occur as close as possible to the data generation step (Figure 5). This means that the measurement quality objectives (MQOs) must be known in advance and utilized by the LIMS (or laboratory personnel, if manual data processing) prior to generating the electronic results data file. This was the original SEDD model incorporated into HORIZON LIMS – a model that was eventually abandoned due to lack of standardization across the industry.

Regardless of the model, the same data source, business rules, and data reduction rules must be applied to both hardcopy and EDD generation in order to avoid discrepancies. However, as pointed out by other presenters in this session, the rules provided by regulators often differ from one program to the next and may preclude the electronic and hardcopy data from matching. This can occur even if all the data are mined from the same LIMS database using identical queries.

Taking a page from the healthcare industry, and recognizing that the environmental industry is no closer to standardization today than it was twenty years ago, the integration engine seems the more prudent approach. While MQOs and general quality control data validation are still handled within the LIMS, we recognize that data consumers will want to preserve their automated data review processes and existing and historical databases. As evidenced by the data checking and exchange processes occurring between the state and federal SDWIS programs (Figure 1), too much has been invested by too many disparate data consumers to expect them to converge on a single standard. And unlike the healthcare industry, not since the Contract Laboratory Program has there been any financial incentive for environmental consumers to move toward standardization. Within even the smallest laboratories and most basic LIMS – if the system can export to a common file format – the data mapping, validation, and routing can be handled by an integration engine. In HORIZON LIMS, this functionality and configuration is being handled through the API and built-in EDD framework , evolving toward a

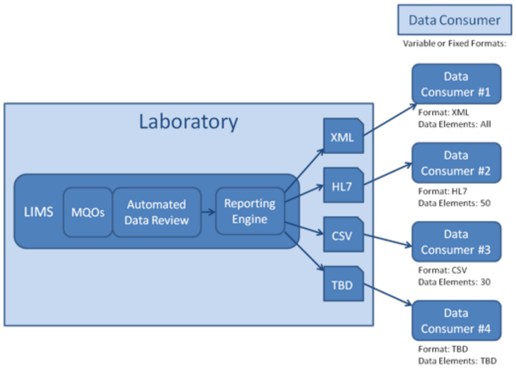


Figure 5: If the MQOs, automated data review, and data transformation/mapping processes are managed by the LIMS and handled by the laboratory prior to data transmission, consumers would theoretically not need information brokers or require re-validation of the EDD received from the laboratory.

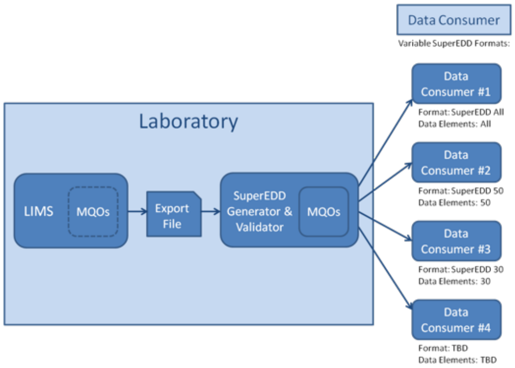


Figure 6: LIMS modules and third-party EDD generation/validation tools would be commercially viable if data generators and consumers standardized on the superset of all data elements needed by all environmental programs. The laboratory need only capture and export the data required for this standard "SuperEDD," and an integration engine would handle all the mapping, validation, formatting, and routing to each of the laboratory's customers.

THE PROBLEM WITH DOWNSTREAM MAPPING AND VALIDATION

technology-neutral (LIMS-neutral) service-oriented “superEDD generator/validator” (Figure 6). The pace of technology would accelerate dramatically once data generators and consumers standardized on the superset of all data elements needed by all environmental programs.

10

AN ECONOMIC INCENTIVE TO STANDARDIZE?

In the healthcare industry, in order to get paid quickly and accurately, the laboratory must comply with standardized diagnostic procedure and observation code nomenclatures (known as LOINC and SNOMED codes). Clinical laboratories are paid (or reimbursed through insurance companies and Medicare/Medicaid) through the same electronic messaging standards (HL7) as those used in receiving inbound electronic test requests and submitting outbound electronic test results. The laboratories eagerly comply with these standards because they can (1) automate the sample registration (accessioning) process from the electronic test request; (2) avoid producing (and mailing) hardcopy data reports; and (3) get paid quickly and efficiently without printing or mailing an invoice.

Cooperation is the only thing that prevents the environmental industry from achieving the same level of progress. Commercial tools for translating and validating environmental data files such as SEDD, ADaPT, and EQuIS, have been available for years. While the pressure and responsibility (for manual data entry and electronic data transformation) continue to shift toward the data generator, the data consumer has much more to gain in this bargain. The USEPA, Army Corps, and other environmental regulators and data consumers manage decade's worth of "data silos" – with minimal inter-agency (or even inter-program) visibility and interoperability. Vast programming resources are tied up in developing and maintaining these custom, disparate systems. Within the laboratory, scientists spend more time transcribing data into spreadsheets and data entry forms and less time ensuring the validity of the analytical data itself. Instead of improving quality systems and implementing lean practices, valuable QA/QC resources focus efforts on verifying that hardcopy and electronic data actually match. The primary objective of computerization – automated (and presumably error-free) data processing – seems to have been lost along the way. The lack of standardization, and the misguided belief that more data is always better, have conspired to prevent environmental laboratories from achieving the same efficiencies as their paperless clinical laboratory counterparts. The data consumer will eventually pay for these inefficiencies one way or another.

11

CONCLUSION AND RECOMMENDATIONS

Most efforts at standardizing environmental laboratory data formats have originated from the data consumer's perspective – that is, with the assumption that related field and laboratory data would be aggregated and validated at the consumer's end. There was a general belief in the 1980s and 1990s that LIMS were not sufficiently robust to handle the data management requirements. As a result, data consumers built a myriad of custom databases and automated data review tools, with little or no standardization across dozens of environmental programs.

If the laboratory captures all the necessary data in LIMS, the electronic data deliverable (EDD) can be programmatically generated and transmitted to the data consumer. Either custom software code can be written for each end-user's preferred EDD format, or a middleware application (e.g., an integration engine or information broker) can be used to map and transform data through a graphical interface. The middleware approach is preferred because it minimizes customization and uses a single data source and data mining operation, thereby reducing potential sources of error. Because of a lack of standardization across environmental data consumers, commercialization of data management solutions has been stifled. Instead, laboratories create hundreds of custom programs and/or manually enter data into spreadsheets in order to accommodate ever-changing requirements.

To use the middleware approach, the LIMS (or similar data source) should be capable of capturing the superset of all the data used by all of the laboratory's data consumers. Ideally, a single standard would be developed to describe the content and format of this data superset. In that way, laboratories and laboratory data management companies could build solutions capable of efficiently exchanging data between all components of the environmental data lifecycle – field equipment, sample submitters, LIMS, analytical instrumentation, mobile applications, integration engines, regulators, public health officials and partners, public information portals (e.g., water supply customers), and other data consumers.

Additional copies of “Environmental Data Exchange and Standardization” and other industry white papers are available at www.chemware.com.

last edited: October 5, 2012

ChemWare
Laboratory Information Management