

STARLIMS®



WHITE PAPER

Content Acquisition and Content Management: The STARLIMS Approach to Implementing a Unified and Fully Integrated Scientific Data Management System

March, 2008

Simon Wood, PhD
Executive Director, Marketing and Education

NOTE:

This White Paper is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for STARLIMS products remains at the sole discretion of STARLIMS.

TABLE OF CONTENTS

THE DATA MANAGEMENT PROBLEM FACING TODAY'S LABORATORIES	3
RESOLVING THE PROBLEM	5
DEVELOPING AN INTEGRATED APPROACH	7
STARLIMS SDMS	8
APPLICATION OF STARLIMS SDMS WITHIN THE LABORATORY	12

THE DATA MANAGEMENT PROBLEM FACING TODAY'S LABORATORIES

Data-Rich, Information-Poor

One of the major problems facing today's laboratories is this: An abundance of data but insufficient means to effectively use it. Today's laboratories are often characterized by being data rich but information poor. The data is there but there are insufficient means to effectively manage, deliver and utilize it. The data exists, but trying to find and use it in a meaningful way can be a frustrating task.

The issue is made worse by the volume of data that is constantly being created. Work carried out by the University of California Berkeley School of Information Management and Systems indicates that in 2000 new data was being created at the rate of 250 Mbytes per person, per year. By 2003, this had grown to 800 Mbytes. Within the laboratory and scientific arena, it is likely that the volume of data is growing even more rapidly. For example, it is estimated that the amount of genomic sequencing data is doubling each year, according to *Towards 2020 Science*, a report published by Microsoft Corp. in 2006.

Changes in laboratory techniques and equipment account for much of this growth. The adoption of automated high-throughput instruments enables far larger sample throughput with consequently larger volumes of data. In addition, these instruments normally have their own data systems that can create complexities of their own. In today's laboratory it is not unusual to find several, or all, of the following informatics systems: Laboratory Information Management Systems (LIMS), Scientific Data Management Systems (SDMS), Electronic Laboratory Notebooks (ELN), Chromatography Data Systems (CDS) Mass Spec Data systems and many other instrument data systems. Furthermore, given the key role that the laboratory plays within many organizations, the laboratory may receive information from, or provide information to, many other enterprise systems: Enterprise Resource Planning (ERP), Manufacturing Resource Planning (MRP), Human Resource (HR) and accounting or financial systems.

Silos of Information

Typically, there is no single platform unifying these data management systems. Silos of information are created and users may need to search each system separately to find the required data and information—if they even know that the data exists in the first place. Today's laboratories need a simple and effective way to access all the data available to them.

Another issue facing today's laboratories is this: just what is laboratory data? Obviously this includes data generated by laboratory equipment that relates to the testing of the samples that the laboratory handles. Such data is usually highly structured in terms of the samples, tests and results hierarchy; the traceability on these entities; and the additional information associated with them (the so-called metadata or data about data). However, look at the data and information that a laboratory handles on a regular basis and it becomes obvious that laboratory data includes much more. It may include anything from Certificates of Analysis, Material Safety Data Sheets or Product Specification details, through customer and client-testing requests, to complex written

reports, detailing research findings, and even email threads between researchers. All of this information needs to be managed in some way. However, unlike sample, test and result information, much of this data may be unstructured, with few or no logical links or relationships.

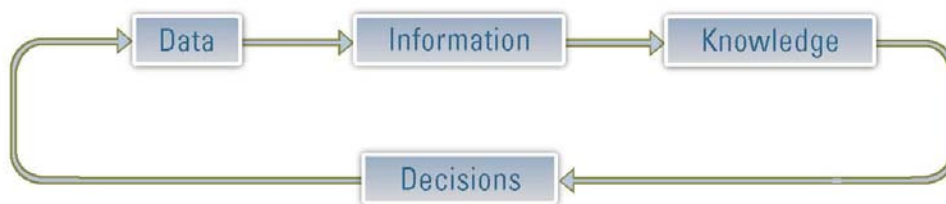
Dissemination

This type of data may also require an effective delivery and distribution method, either to other systems or to individuals. For example, training record information in a Human Resources system may be needed in a LIMS; sampling point information relating to a water distribution network may be needed in a LIMS or ERP; and invoicing information can be important to the laboratory if payments are in arrears.

Increasing globalization makes data availability an even more urgent issue. With multiple sites around the world manufacturing products for the global market, it is crucial to ensure that they are all working according to the correct specifications. If those specifications change, the changes must be made available and implemented by all the manufacturing plants; and the QA/QC laboratories must ensure that they test the product against the valid specifications. This is a classic data distribution and delivery problem, which can be solved with an integrated laboratory informatics solution.

The Information Value Chain

There is a further aspect to information management, and it is this. Probably the major asset of the laboratory is the data and information that it generates. However, the data and information is only valuable if it is used to create knowledge, and that knowledge itself is only valuable if it is used to drive the decision-making process. However, if the data and information is not understood and properly managed, then this data, information, knowledge decision-making cycle cannot be effectively implemented.



Finally, laboratories need to face the issue of data obsolescence. This does not mean that the data itself becomes obsolete, but rather that the data *storage format* becomes obsolete, making the data difficult or impossible to use. This may occur, for example, as instrument suppliers change and upgrade their instrument systems, or if the lab replaces its instrument suppliers.

RESOLVING THE PROBLEM

To work effectively, laboratories need to spend less time storing, managing and retrieving information, and more time using the data and information that they have. To achieve this, all relevant data within the laboratory must be captured, processed and stored in some way that ensures it is easily retrievable over the long term. This should apply to data from any source—whether the data is structured or unstructured. However, attention must also be paid to regulatory issues and any such storage and retrieval mechanism must support the users' regulatory requirements (adherence to FDA 21CFR Part 11, support for GxP and ISO17025, etc.).

Achieving all this means that relevant information from a multitude of different systems will be extracted, made available, and managed from a single source. In other words, there will no longer be a need to conduct multiple searches across different systems to find the required structured and unstructured data. Of course, to do this, the strategy must be supported by the correct data and information management systems.

Multiple Discrete Lab Informatics Systems

Within the laboratory informatics area, various systems can be used for the management of laboratory data: Laboratory Information Management Systems (LIMS), instrument data systems such as Chromatography Data Systems, Electronic Laboratory Notebooks (ELN) and Scientific Data Management Systems. Commercial LIMS have been on the market for about 20 years, roughly since the advent of sophisticated instrument data systems. LIMS are well suited for managing structured data (particularly sample management), and were initially developed in response to the needs of QA/QC and product-testing laboratories. ELNs and SDMS are newer products. ELNs have been developed in response to the needs of research type organizations to replace traditional paper notebooks with electronic versions that allow for more efficient management of research work and other unstructured data.

While LIMS and ELN have been considered separate systems meeting the different needs of managing data within routine labs and research labs respectively, this distinction is becoming blurred due to changes in the nature of some research work. In bio-banking, for example, sample management is essential to store the large volumes of samples, such as DNA—suggesting that bio-banking applications would rely on LIMS. However, the matter is complicated because the research findings themselves may be more effectively managed using an ELN. Developments such as this make it even more imperative that data and information can be consolidated to truly leverage the information available.

Integrated Laboratory Data Management

SDMS, on the other hand, have been developed to integrate laboratory data and information from whatever source—by enabling extraction of information from relevant systems and documents and making it available to the users. The STARLIMS SDMS module provides unparalleled levels of SDMS functionality within a system that is seamlessly integrated with a LIMS. It provides a single unified laboratory informatics solution that can resolve the information management problems described above.

Eliminating Special Interfaces

This seamless integration means that many of the issues associated with utilizing both SDMS and LIMS are eliminated. These are real issues facing users who wish to create an integrated laboratory informatics solution because current LIMS and SDMS solutions have been developed as stand-alone products. Each solution has its own architecture; interface and administration functions and some sort of interfacing mechanism must be developed and maintained between the systems. All of these add to the overall training and support overheads. In addition, to work effectively as stand-alone products, it is quite possible that the chosen SDMS and LIMS may share common functionality, leaving users with the decision of what parts of which system to use. With a fully integrated system, data and information will be stored as efficiently as possible and any possible data duplication can be reduced or eliminated.

DEVELOPING AN INTEGRATED APPROACH

Any fully integrated system of this sort needs to be technically sophisticated. The system needs to read unstructured data and to extract relevant information from it. This information may be used for the creation of structured data or to help with knowledge management. For example, a CDS will take the raw data from a chromatographic run, process it and output a report or file. However, for a fully integrated data, information and knowledge-management system to work effectively, it must:

- Identify and process the file itself to extract relevant data and metadata, such as parameter and set-up information;
- Extract and store the data in a neutral or common format (i.e. unified XML). Make it possible to use a single search mechanism to access all lab data, resolving the “silos of information” problem;
- Ensure data longevity (resolve data and format obsolescence issues);
- Use the data and metadata to automatically update results in LIMS or any other system;
- Transfer the file to long-term electronic storage, possibly a Document Management System;
- Make the information available for future searching, extraction and processing.

It must be emphasized that in this context, the terms “files” and “documents” mean much more than just instrument data files. An integrated system of this sort will be able to manage, process and fully utilize the value of the information in a multitude of formats. This means being able to extract data from written reports so that the information is readily available—rather being locked away on a hard drive or, even worse, in a filing cabinet as part of a paper-based lab notebook.

Parsing of Data, Documents and Unstructured Data

A sophisticated and intuitive parsing mechanism can be applied to any file, document or other unstructured data—allowing the data and information to be effectively extracted and managed and moved into the knowledge management cycle. Such a system can also be used to control the distribution of data and information within the organization. For example, information in centrally managed Material Safety Data sheets could be automatically distributed as required. In the same way, centrally managed product specifications could be automatically uploaded to the systems that require the information; and personnel data such as training records could be automatically maintained.

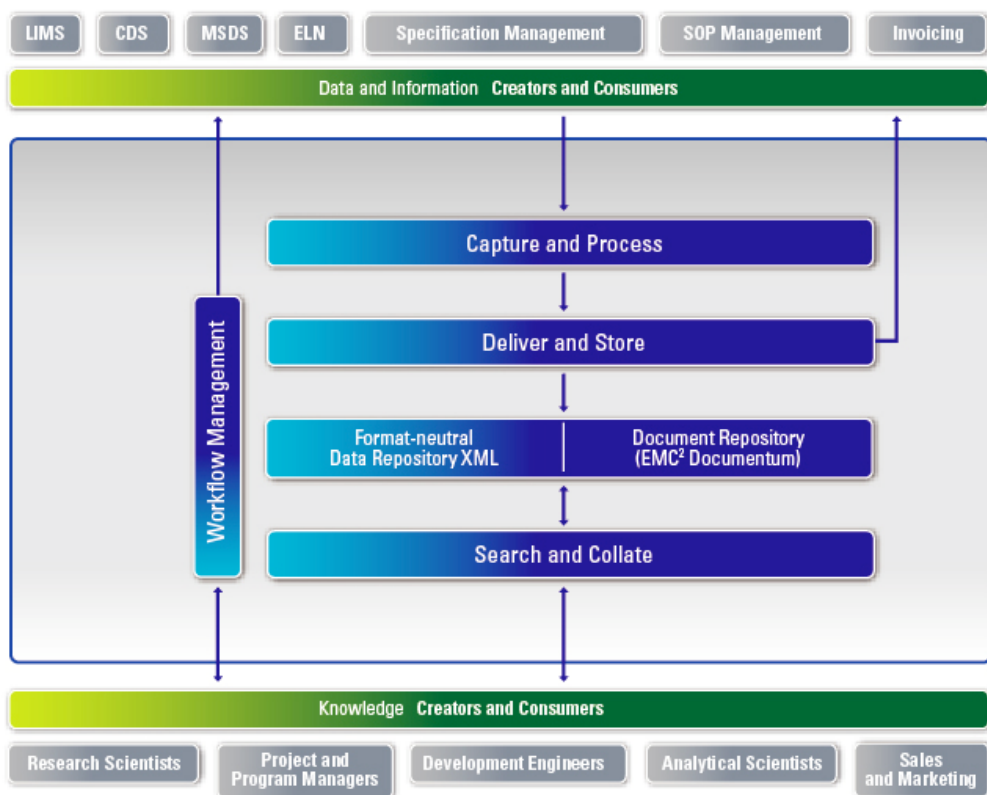
Workflow Management

Once this file handling mechanism has been created, the next logical step is to build workflow management for the files. This would enable documents to undergo a controlled process of creation, review and approval. Upon approval, the data and information enters the content management lifecycle, becoming available within the data, information and knowledge-management system.

Such a system would create a central data and information management system for the laboratory—a single unified repository of the information needed to make effective decisions.

The STARLIMS SDMS module is a highly sophisticated web-based system developed using the latest .NET technology and fully incorporating the use of Web services. It includes advanced pattern recognition and data mining technologies, as well as sophisticated indexing and searching facilities to ensure that information can be easily retrieved once it has been processed. It is completely integrated with the STARLIMS V10 web-based LIMS, and fully supports workflow management. STARLIMS SDMS is integrated with the EMC Documentum document management system for storing documents and files. The product has been designed to meet the regulatory requirements and demands placed on many types of laboratories, including 21 CFR Part 11.

SDMS Processes



File Capture

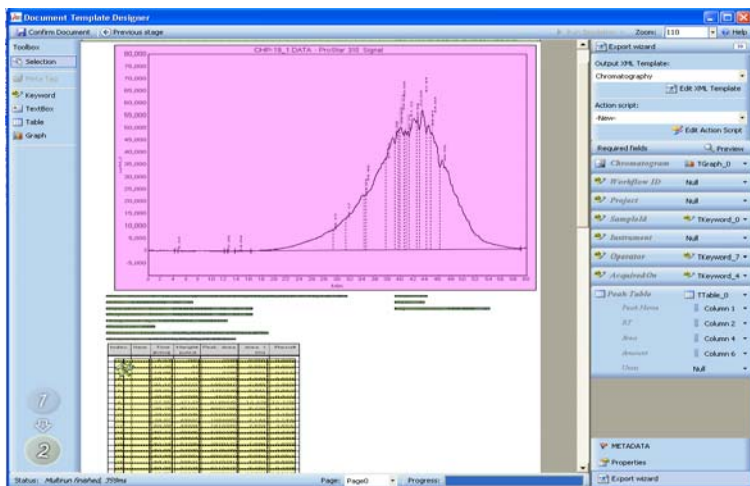
The first step in the process is the capture of files that need to be processed and stored. These files may come from any source: instrument output, Certificates of Analysis provided by suppliers, product specifications used within the organization, standard operating procedures, etc.

A File Grabber application that runs in real time is used to automatically capture the required files. The File Grabber monitors the file system and is configured to look for the required files or documents and to pass them onto the File Recognition and Processing Application. The File Grabber application creates a 21 CFR Part 11-compliant “layer” around the application, which eliminates any possible risk that a file could be altered between its creation and the time it is securely placed in the SDMS repository.

Intelligent Parsing

The File Processing Application is responsible for extracting information from the files captured by the File Grabber. The File Processor parses data and information from the files and can handle most common file types including PDF, Word, CSV and other text files.

Intelligent Parsing routines and file templates are defined within the system for known file types to extract the required data and metadata. Even within PDF files, data and information can be extracted from tables. Graphical images including graphs and spectra can also be identified and extracted. Multiple recurring data and information within a file (i.e. multiple results for different samples that are part of an instrument run) are automatically processed. All known types of documents are processed automatically. Should an unknown type of document be detected, the system will attempt to define the data extraction routine required based on intelligent pattern recognition algorithms and previously defined document types. The unknown document will also be placed in a queue where the required parsing routine must be confirmed by a user; in this way the identification of an unknown file type does not interrupt the processing of other files.



Intelligent Parsing automatically identifies charts, keys and tables.

The templates for known files are built using the Template Designer tool that allows the definition of structures, objects and data that need to be extracted. Users can define keywords that identify specific data and metadata within the file.

Data and Metadata Storage

File processing may require data storage, and this can be accomplished by storing data as unified XML output. The advantage of storing the data in this way is that it can be read and used by any application that understands the XML schema. This allows a system such as LIMS to use the data, for example to enter results from an instrument into the LIMS. However, it is much more than just an effective instrument integration tool. It allows the original document to be recreated based on the XML output, and facilitates building reports across data and information. In other words, information can be collected, collated and presented from multiple sources and from multiple original formats through a single unified source—effectively eliminating the problem of silos of information and data obsolescence. STARLIMS SDMS includes a unified XML designer that allows for the fast and simple creation of XML templates, eliminating the need for manual creation of these inputs.

File Storage

Once file processing is complete the original document may be stored within a file storage mechanism if required. STARLIMS SDMS is integrated with the EMC Documentum document management application, providing laboratory users with all the functionality of a fully featured document management system.

Information Retrieval and Usage

Far more than a glorified storage system, STARLIMS SDMS enables laboratories to maximize the value of the information at their disposal, by allowing information to be efficiently searched and retrieved. It provides full integration with Microsoft Desktop® products, so that data can be instantly imported into applications such as Word® and Excel®. In addition, when integrated with a product such as EMC Documentum®, all the original files are available together with the built-in search and retrieval capabilities of this fully functioning document management system.

Workflow Manager

STARLIMS SDMS also allows for the definition and management of document workflows. A workflow is associated with the processing template for the type of document. Documents that are associated with workflows are only processed by the SDMS once their specific workflows have been completed.

Users can be defined with various roles that correspond to the workflow steps. These roles are Creator, Approver, Reviewer and Consumer.

Creators can create the original document, but may also recreate the document if required;

Approvers can approve or reject a document; approval may trigger another approval and or review by a different user;

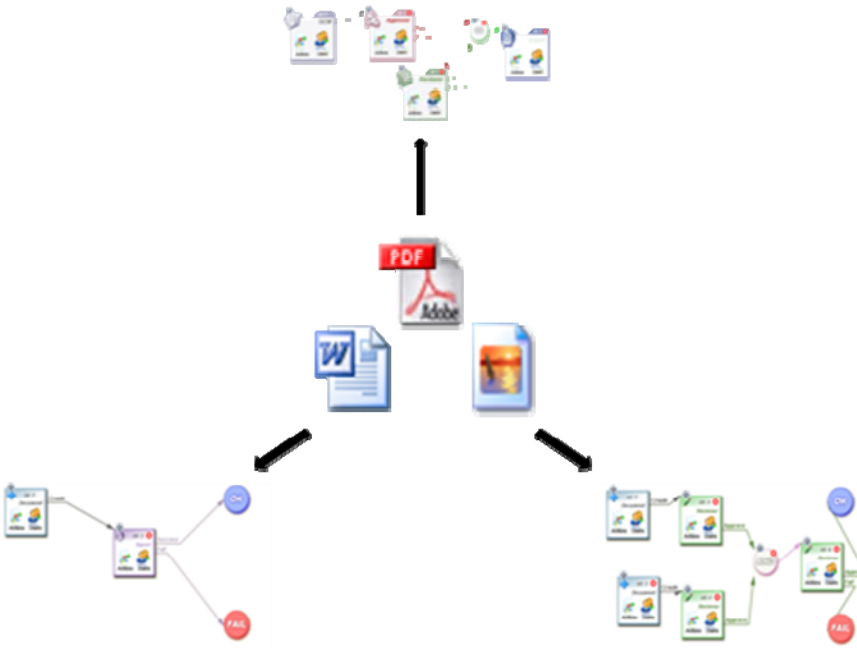
Reviewers can modify the document as well as approve, or reject it; approval may trigger another approval and/ or review by a different user;

Consumers may receive the document and may choose to acknowledge receipt or ignore it (the last consumer identified in a workflow does not have the option to ignore the document).



The Workflow Editor allows custom workflows to be defined for the various document types and to specify the routing rules. Each creation, recreation and edit of the document creates a new version, and integration with the Documentum® document management system means that each version can be recorded and maintained. Each document is also digitally signed—ensuring that a complete audit trail of changes will be available. Users are made aware of workflow steps outstanding for them through the STARLIMS SDMS user interface; however it is also possible to include a real-time reminder within a user’s STARLIMS V10 dashboard.

Implementation of the Workflow Manager allows documents to be created, reviewed, modified and approved within a controlled environment that will meet regulatory requirements. It also helps ensure that any required data and information can be extracted and utilized as required by the laboratory’s informatics strategy.



Automatic routing to the appropriate workflow is based on document type.

APPLICATION OF STARLIMS SDMS WITHIN THE LABORATORY

STARLIMS SDMS is applicable to any laboratory that handles and manages files and documents. The integration with the web-based STARLIMS V10 Laboratory Information Management System provides a unique way of seamlessly integrating structured and unstructured data. Implementation of the Workflow Manager adds an additional layer of functionality and control. The STARLIMS SDMS has many possible applications within the laboratory environment, and a few examples are outlined below.

Instrument Integration

This is probably the most obvious application of STARLIMS SDMS. Data can be captured from any instrument that outputs a file of known structure, whether it is a single result, multiple results for a single sample, or multiple samples. Information can be extracted to unified XML, (which also allows the storing of graphical images) and imported to other systems. In addition, the original file can be stored using the content management functionality and the document can be recreated using XML. In other words, STARLIMS SDMS provides not only instrument integration, but also a mechanism of storing the information in a format-neutral way that ensures long-term data availability.

Certificates of Analysis and Reports

Materials supplied to the laboratory or organization may come with Certificates of Analysis from the manufacturer. If these are supplied in electronic format, they can be automatically processed and stored. This may include extraction of relevant test data and results, and automatic import of that information into any system: LIMS, an ERP, manufacturing systems, etc.

If the laboratory itself creates Certificates of Analysis based on its own testing, the STARLIMS workflow manager becomes very useful. Once review and approval is completed the Certificate of Analysis can be distributed as required and processed in the same way as any other document or file.

The same type of process may also be applicable to other reports that the laboratory generates.

Product Specifications

Maintaining, managing and updating manufacturing product specifications is always difficult, particularly in global multi-site environments. The specifications need to be created, updated, approved and distributed. Obviously the workflow manager can be used to manage the review and approval mechanism. In addition, the specification details can be converted to XML format—making them available for automatic upload to any system within the organization that requires the information.



Standard Operating Procedures and Methods (SOPs)

The use of STARLIMS SDMS for managing SOPs and methods follows the same process described above for Specifications. Workflow management can control the creation, review and approval; and on approval data can be extracted, processed and made available as required.

Research Reports and Output from Electronic Laboratory Notebooks

One of the major sources of data, information and knowledge that can be difficult to unlock are research reports. As more and more research is managed using Electronic Laboratory Notebooks, these can also be a rich source of data and information. STARLIMS SDMS opens the real possibility of integrating this research information seamlessly with other information, and ensuring that its long-term value can be maintained. This will be a major step forward, compared to situations where research work may be isolated in existing paper notebooks or locked up in research reports that cannot be easily accessed.

Knowledge Management

Of course, once file and document management has been implemented using STARLIMS SDMS, data and information is available within a single unified system. This makes Knowledge Management a real possibility, allowing simple access to all the data and information created and used by the laboratory. Implementing Knowledge Management requires a unified laboratory informatics strategy and STARLIMS SDMS can form the underlying basis for the implementation and execution of this strategy.