Considerations for Management of Laboratory Data

2003 Scientific Computing & Instrumentation LIMS Guide, November 2003

Michael H Elliott

Drowning in a sea of data? Nervous about 21 *CFR* Part 11? Worried about longer and longer product development times? Having a difficult time just finding the information you need to make decisions? Worried about the protection of your intellectual property?

These are some of the many questions managers and scientists are facing today that motivate them to look at new ways to manage data and information in the laboratory. Companies want to move away from a paper-based operation to a more modern electronic workflow to improve efficiencies and reduce time to market for new compounds.

Traditionally, laboratories have looked to Laboratory Information Management System (LIMS) to assist them in managing the ever-increasing information workload. LIMS have performed admirably to manage *structured* data – those data records that have a fixed format, such as database records on samples, their related tests, and fixed formatted results. However, new analytical technologies, reporting requirements and regulations have forced a dramatic increase in the amount of *unstructured* electronic records. These records, such as instrument data files, spreadsheets, reports and image files have no common structured format between them. Laboratories have faced the challenge of not only collecting these records, but also managing them in a way that insures long-term preservation and knowledge retention.

There are many types of systems that attempt to address the growing need to manage these unstructured records which are also known as "knowledge assets." Some of these are: knowledge engineering/management systems, document management systems, content management systems, scientific data management systems, data archival systems and hierarchical storage management systems. Each type of system varies in its focus and capabilities. The growing problem of managing electronic records across a wide range of industries will likely increase the number of technologies and vendors over time. With such a wide range of alternatives available, it is important to understand your current and future business drivers, any regulatory requirements, your existing infrastructure and the level of investment you can afford before selecting a system. Once you have your requirements clearly defined, you can begin narrowing your selection of the type of system that best suites your needs. A few of the important considerations for this selection process are outlined below.

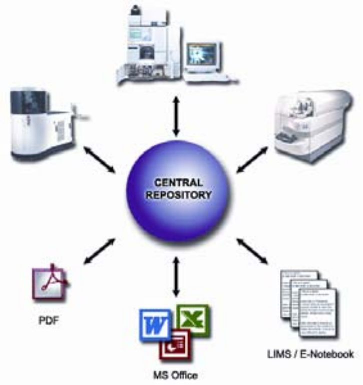

Figure 1 – A data management system centralizes data capture and utilization

Determine the type of electronic records you wish to manage

The primary consideration for selecting the type of system to use is to determine the *types* of electronic records you wish to store and manage. As stated earlier, the typical laboratory has a wide range of electronic records from instrument raw data files to processed data that is presented in a report format. Document management systems, for example, are designed specifically to manage reports, such as Microsoft Word documents or the industry standard Portable Document Format (PDF) files. Several of these systems do an excellent job of managing document workflow and document consolidation. However, these systems do not have an understanding how to manage instrumentation data. Some can store these files passively, but cannot extract content for database searching. An example would be the peak, area and concentration information of a chromatography data file. You wish to store that information to find the file at a later date, or to report on it, so it has to be extracted and placed in the database. In addition, some instrument vendors have very complex data structures with multiple files per single injection, which requires special interfaces. This management challenge is different from that of pure document management. Some products, such as a select number of knowledge engineering/management systems, are designed to manage both the “human readable” (i.e. documents) and “machine readable” (i.e. binary instrumentation records) in one integrated solution.

Determine the policies for your electronic records

The second consideration is to determine the *policies* for your electronic record management. The policy will determine the life cycle, workflow or the “birth to grave” of your records. This life cycle is comprised of 5 phases: Creation, Collection, Warehousing/Organization, Archiving, and Destruction. Your policy, which can be dependent on the class or types of records, determines the movement of records through these various phases. For example, records for regulatory purposes may need to be kept for 30 years, while records for a drug discovery project may only need to be kept for the duration of that project. Some of the systems available allow for complete automation of the life cycle by configuring your individual policies into the system.

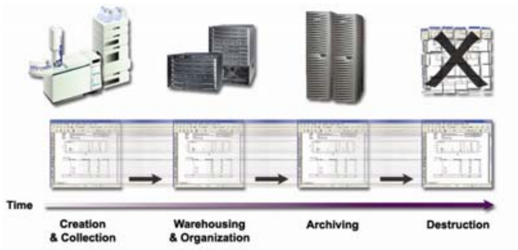

Figure 2 – The life cycle of an electronic record

The descriptions of each phase are:

Creation – Electronic records can be created in many different ways. For example: automatically, through an automated process such as an instrument data system; manually, such as saving a document from an MS Office application; or records can be created through a print driver such as Adobe Distiller which automatically creates a PDF document from what could have been printed to a paper printer. As stated previously, some instrument vendors create complex data structures and multiple data files per injection. How the files are created impacts the *Collection* phase.

Collection – Some Knowledge Engineering Systems (KES), Scientific Data Management

Systems (SDMS) and Hierarchical Storage Management Systems (HSM) have automated methods to collect or "harvest" electronic records from client workstations or servers. Your policy will determine the timing of the collection, such as every day at 4:00pm or immediately as a file is created, and whether you wish to have a "push" or "pull" collection. "Push" record collection is typically a small "agent" program that is running on the client, such as an instrument PC. This agent can "wake up" at a certain period of time and "push" data files to the repository. A "pull" agent sits on a server and monitors or "sweeps" multiple clients simultaneously. The advantage of the "push" method is that there is very little network traffic versus a "pull" method which is constantly engaging the network. The advantage of the "pull" method is that there are fewer programs to install and manage. Your policy will determine both the frequency and the type of collection method.

Warehousing and Organization – Records that are harvested during the *collection* phase are typically stored in repository, organized into a logical structure and "warehoused." In this phase, files become immediately accessible to users of the system, by direct access or by searching a database. This database is built from the metadata of the electronic records which is the descriptive information about the file (such as data time created, source computer, etc.) and the contents of the file (component information, sample id, results, full text index of documents, etc.). These files are "filtered" during the upload process to copy the appropriate contents from the file and to update the database. There should be no limit as to what you can extract from the files, nor should you be restricted from updating extracted information. You never know what you need to search and report on in the future. The individual electronic records are usually stored in a secure file storage location with the database pointing to the files. Due to cost, administration overhead, scalability and validation complexity, the collected electronic records are rarely stored directly in the database itself. Disk arrays, Network Attached Storage (NAS) and Storage Area Networks (SAN) are the most common storage systems used. During the period the files are most often accessed, they live in the warehouse and can then be automatically *archived* to another media, such as magnetic tape or optical disk.

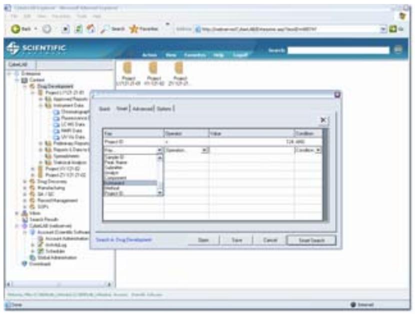

Figure 3 – Searching for records from a repository

Archiving – Archiving is the process of moving files from the warehouse storage location to another media for long-term record retention. This involves moving the designated records to a media such as CD, DVD, magnetic tape, Magneto Optical (MO) or Content Addressed Storage (CAS) and updating the database to catalog the files' new storage location. Your policy should determine the timing of such a move. For example, all files older than 3 years are archived to an optical disk. This process can be as simple as manually burning and moving data to a single DVD or it can be as complicated as automatically moving files to a robotic magnetic tape management system. The advantages of a single DVD are cost. It is very inexpensive to produce a DVD. However, the data is not immediately accessible when that DVD is not mounted and is not practical for a large laboratory. You can implement DVD jukeboxes, but with changing format standards, you must decide your level of risk you wish to accept for both failure rates and data migration in the future. Newer technologies, such as the self-healing and self-replicating CAS technologies from companies such as EMC, appear to be the future of long-term retention. The costs of these newer technologies have not yet dropped to be competitive with low-end tape and DVD burners, though this is improving yearly.

Destruction – There comes a time when all good records must die. Your retention policy should determine the life span of your records and be based on the class or types of records you are managing. Some technologies on the market allow you to automate the destruction of the records after they have reached the end of their retention period. Properly designed systems will not just freely delete the records and their corresponding database entries, but will allow for a review and sign off by a select group of reviewers to insure proper checks and balances. You should also be able to assign files to a new class that may have a different retention period. For example, a project that is completed in drug discovery and now its resulting data needs to be moved into a drug development project which has a different retention period.

Determine your regulatory requirements

Another important consideration is to determine any regulatory requirements associated with the management of your electronic records. For example, if you are a biotechnology, medical device or pharmaceutical company that sells into the United States, your areas of product development and manufacturing are covered under US Food and Drug Administration (FDA) regulations. Applicable regulations are: Good Laboratory Practices (GLP), Good Clinical Practices (GCP) and Good Manufacturing Practices (GMP). Collectively, these are known as cGxP or the “predicate” rules. If you operate in a cGxP environment *and* you are storing electronic records in lieu of paper records *and* these records are required for compliance with the predicate rules, these records fall under the FDA’s Part 11 requirements for electronic records and electronic signatures.

US 21 *CFR* Part 11 outlines a comprehensive set of requirements for the security, control, auditing and accessibility of these records. Since the trend in the industry is to move away from paper to electronic record management to improve operational effectiveness, Part 11 is a good set of guidelines to have in place. Currently, the FDA is reviewing Part 11, but the concepts of proper security and controls over your electronic records are important whether you are in a regulated FDA industry or not.

Despite the claims of some vendors, data management systems will *not* make you compliant with *any* of the federal regulations. If the data management system contains a features set that meets the technical requirements of Part 11, it can only *assist* you in becoming compliant. Software systems are *not* compliant – organizations are or are not compliant. It is the responsibility of your company to have the proper training, documentation, policies, validation and internal controls to achieve compliance. There is no “magic bullet” in achieving compliance!

After you assess your regulatory requirements, it is important to screen out those technologies that will or will not assist in your compliance efforts. Many general content management or archival systems do not have the required controls. In addition, many of those systems do not have the concept of electronic signatures – or if they do, few meet the requirements of Part 11. Key considerations when reviewing systems are: security and access to records, measures to insure validity of the data, electronic signatures and consistency of approach, revision control and the ease of system validation.

Data organization and collaboration

A major driver of many data management projects is the sharing and collaboration of scientific information between researchers. Many times this involves sharing data between research sites that previously did not share project information. It is an important consideration to determine the types of data that needs to be shared, how you want to organize the data and the accessibility of the information. This will help to determine if a solution can be tailored to meet the collaborative business objectives.

For example, a typical data archival or storage management solution also performs the collection and archiving of records. However, for those departments that need immediate access to information, these systems fall short. In many cases, you must know what you are looking for and must contact an archive administrator who will enable the restore functions of the system to locate your records and retrieve them. This could take minutes, hours or days.

In addition, traditional archival systems do not extract sufficient content for performing searches, reports or mining of the large data collections found in the scientific environment. If you wish to trend the results of the last three years of mass spectrometry analysis for a particular component, the information must be available and accessible for reporting.

Knowledge engineering, document management and content management solutions are designed for collaboration and data sharing and are better suited to the task at hand. They typically allow for more robust data organization and for different levels of searching and data mining.

Determine your infrastructure requirements

Your current and future IT infrastructure has an influence on the type of technology used to manage your electronic records. It is difficult to automatically collect data from laboratory instruments if there is no network in place!

Your network infrastructure will help you to determine the strategies you employ for system implementation. With some TOF MS instruments creating 300MB data files, you can slow down a 10Mbit per second corporate network very quickly if you are transmitting hundreds of these files. You need to determine the available bandwidth and latency of your network. This will influence the policies you set in place for data collection and whether you are going to “push” or “pull” data into the repository. This will also determine the responsiveness of the system to the end user when they are accessing data. For a high volume, data intensive laboratory, a 100Mbit per second is recommended as a minimum.

Determining the number of data files, their average size and the desired length of time to keep the records in the warehouse (before moving to a long term repository) is important. It will determine the size of the database and the amount of storage hardware required. Though many users don’t know how much data they generate, it is worth the time to investigate this upfront to avoid having a vendor sell you a solution you will have to upgrade three months after installation.

An understanding of the integration needs of the system into the corporate enterprise is helpful before implementation. A frequent request of users is to have the data management system "integrated" with their LIMS, existing document management system, SAP or other enterprise-wide system. Ask users what the business requirements are for integration, many times the response is "I'm not sure, but I want it integrated." For any integration project to be successful there has to be a clear understanding and agreement between the integrator and users on the requirements and objectives of the integration project. It is worth the investment of time to analyze the required data flows and to specify the desired interface before any work begins or before a vendor tries to sell you a "standard" interface.

Your organization's database standard also has an influence on the selection of a solution. Oracle and Microsoft's SQL Server are two of the most common. Each one has it own unique advantages. However, if your organization is standardized on Oracle, don't spend the energy on other database solutions! Some of more modern commercially available record management systems offer a choice of database back ends.

Summary

As science-based organizations evolve from paper-based operations to improve their productivity and effectiveness, there is a need to manage the resulting electronic records. These records of varying types and sizes can be managed by a number of commercially available technologies. However, it is important to understand your business requirements, policies and infrastructure while embarking toward the goal of the "paperless laboratory." As Henry Ford once said, "Obstacles are those frightful things you see when you take your eyes off your goal."

Michael Elliott is President of Atrium Research & Consulting, a market research and analysis company focused on the scientific informatics market. He can be reached at info@atriumresearch.com