



BEYOND LIMS: THE INTEGRATED DATA PIPELINE

Andrew Bowen
TESSELLA SUPPORT SERVICES PLC

Issue V1.R0.M0
January 2005



Introduction

Over the last 25 years the Laboratory Information Management System (LIMS) has been developed to become the centrepiece of laboratory IT. A LIMS of some kind, whether it be a commercial system or a bespoke (custom-built) system, is now essential for all but the smallest of analytical laboratories.

However, whilst the LIMS is firmly established at the heart of the laboratory, it is still only part of the solution that most organizations now require. A standard LIMS package may be enough in a basic contract laboratory (samples in, analysis, results out), but if the laboratory is part of a wider business activity, for example research and development, then it is only one part of a more complex whole. Hence the LIMS cannot be regarded as an island.

The demands placed upon the analytical laboratory grow year by year. In the discovery process leads are getting harder to find, more samples must be processed and more data is generated. In the production plant quality standards and regulations are frequently changing, becoming more and more demanding. In short, there is much more data to deal with and more stringent requirements for quality and traceability. Furthermore, in order to meet business objectives, there is real pressure for work to be completed more quickly and to be undertaken in innovative ways.

An integrated data pipeline joins up the LIMS and other systems to enable the efficient flow of information around the organization. In doing this it provides an important part of the solution to the ever-increasing demands placed on the laboratory and the business.

The data pipeline

The pipeline analogy for data is borrowed from the pharmaceutical industry; the drug discovery pipeline that moves from disease target, to lead identification, through characterization and on to development (there are equivalent pipeline analogies in other industries). However, the analogy is an imperfect one; a pipeline contains a flowing, unchanging material and has no memory of what has passed through before. A data pipeline has to be much more than just a conduit. This pipeline continuously adds value to what flows through it. As the data flows it is transformed; raw data is analyzed to become information and collected information gives rise to knowledge. The data pipeline must also have a memory. Rather than just being a mechanism for transporting data it has to be a repository, so that everything that underpins the derived knowledge is stored accessibly. The LIMS is just one element in a complete data pipeline.

Exploring the boundaries

The typical LIMS is increasing in functionality year on year, reaching further out from its original roots in sample management and reporting. A LIMS will now often include interfaces to analytical instruments and some data analysis capabilities. But a LIMS is not meant to be a whole data pipeline. A LIMS has boundaries; points at which the supplied functionality ends and within which data is contained. To integrate the LIMS into a data pipeline one must either extend it or enable the free flow of information across the boundaries to other systems.

To see the context in which the LIMS exists we can look immediately upstream, in the laboratory, downstream and then at the wider environment. In an integrated pipeline data must flow easily between the LIMS and all of these other functions. The following are examples of what may need to be integrated with a LIMS in this way:

- Upstream:
 - ❖ Portfolio planning/resource allocation
 - ❖ Experimental design

- In the laboratory:
 - ❖ Instrument interfaces (sequence files and output files)
 - ❖ Automation of laboratory equipment
 - ❖ Electronic laboratory notebooks

- Downstream:
 - ❖ Data analysis
 - ❖ Reporting
 - ❖ Quality assurance
 - ❖ Long-term data archiving – ePreservation

- Over the whole data pipeline:
 - ❖ Decision support systems
 - ❖ Portfolio management
 - ❖ Integration with other databases
 - ❖ Document management
 - ❖ Workflow management
 - ❖ Regulatory submission

To identify the key boundaries for a LIMS, it is necessary to explore where data moves in and out. For example:

- Where is data being manually keyed into the LIMS system?
- Where are spreadsheets used to import data into, or export data out of, the LIMS?
- What output from the LIMS is sent downstream on paper reports?

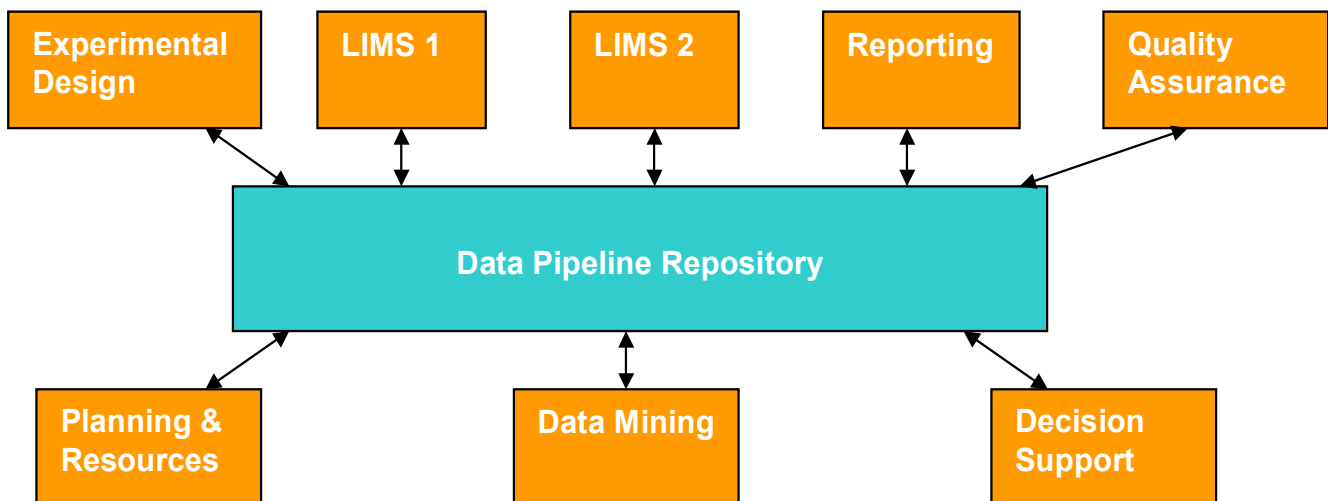
Manual keying, manual data import/export and paper reporting will all tend to reduce efficiency and data quality. In a closely regulated environment the data quality problems that can result from manual data entry and handling can be particularly costly to deal with. Integration will help to overcome these inefficiencies and problems.

Building an integrated data pipeline

Integration of systems and processes into a data pipeline should be driven by business requirements rather than information technology. There is no single IT solution to the problem; integration can be achieved in many different ways. However, there are two common modular patterns for providing this integration. Which of the two is best to use can depend on many things. If you are starting with a clean sheet of paper then you will have the luxury of choice, if you are faced with integrating existing systems then the choice may effectively have already been made for you. A pragmatic solution may need to use both patterns in different parts of the pipeline.

The shared data repository

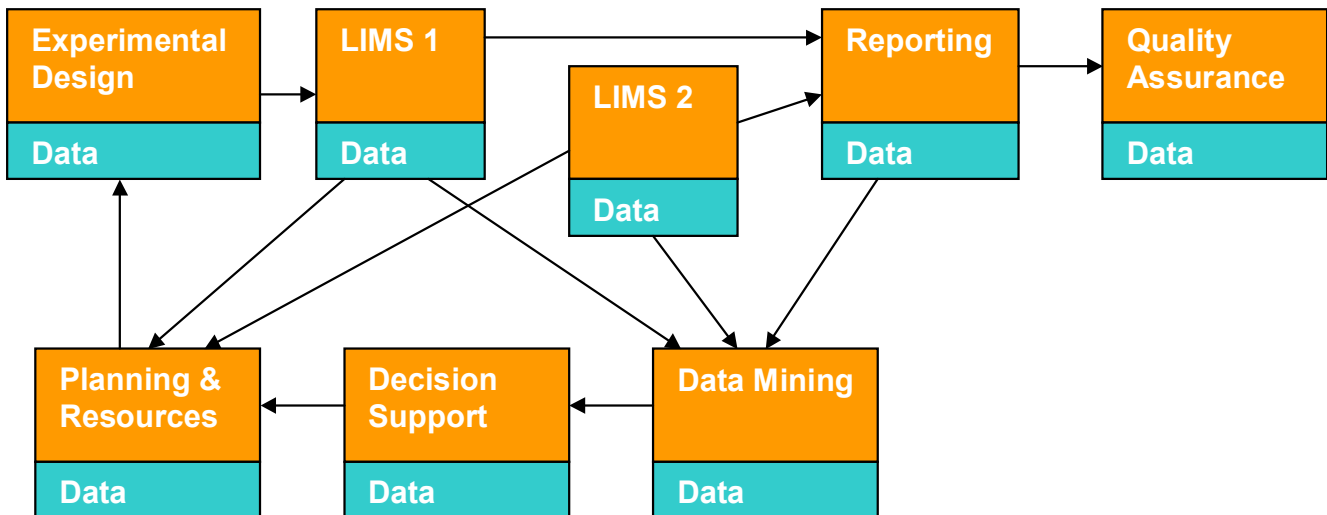
In this pattern a single central repository holds all of the pipeline data and information. Independent modules access the repository through their own particular 'view' of the data. Most modules are only concerned with a subset of the shared data but those data subsets will intersect.



This approach presents some challenges; it is best if the repository is mature and stable. Even small changes to the data storage schema could potentially impact a number of different modules. However, careful design of the module 'views' to provide an abstraction from the underlying data can provide quite a high-level insulation against repository changes and enhancements. The benefit of the shared data approach is that new modules can be created and existing modules modified without impacting other modules in the system. It also allows centralized security and backup of data. Any system using a shared data repository is best designed that way from the outset. It may not be suitable for integrating systems that have been hitherto independent. The typical technical implementation of this approach would be using a shared enterprise-wide database, such as Oracle or SQL Server. The modules may then be implemented in any number of different ways; they don't all have to use the same technology.

Interfaces and services - a federated pattern

In this pattern the pipeline data is distributed between the different modules. Data moves between modules via formal interfaces and services. A module may 'push' data to downstream modules or 'pull' data from upstream modules. Query-only services provided by modules allow higher-level functions, like decision support, to gather data from around the pipeline.



This approach can enable the connection of widely differing systems and could be used to integrate existing legacy systems; this is a distinct advantage if you are not starting with a clean sheet of paper. However, distributed data is more difficult to manage and back up. This arrangement will also probably involve replication of data in different places. Keeping several copies of data in synchronization when changes are made can be quite challenging. This approach is however more modular than a shared data repository approach. Provided the interfaces remain unchanged then whole modules can be replaced, data storage included, without impacting the rest of the system. Interfaces for this type of arrangement are increasingly implemented using web services, XML and emerging grid technologies.

Barriers to LIMS extension and integration

When integrating an existing LIMS into a data pipeline one can encounter any number of obstacles. If you have a commercial LIMS then you will be dependent on how open and extensible that product is. Some LIMS vendors actively encourage customization, providing open databases and built-in development languages, whilst other vendors may lock down their products. If you have your own bespoke LIMS then you might suppose that it will be easy to extend and integrate. However, even bespoke (custom built) systems can be very difficult to extend depending on how they have been built in the first place. These are just some of the problems that you may encounter:

- **Closed Systems:** Some vendors will lock down their systems and not provide any external interfaces. A product may support data import and export, perhaps via a text file or spreadsheet, but this is hardly a useful interface if it requires manual intervention to use it. To get real integration you need to be able to move data into and out of a system automatically using programming interfaces or standard services and some vendors may not support this
- **Closed Data Formats:** Commercial systems of all kinds may employ closed proprietary data formats. This could apply to the LIMS itself or the other systems that you want to integrate with. Analytical instruments often use proprietary binary file formats that may change with each release of the instrument software
- **Pipeline Bottlenecks:** You may achieve a high level of integration only to find that bottlenecks limit the overall performance of the pipeline. Almost by definition there is always a bottleneck somewhere, but they do vary in severity. If the bottleneck is within the system then buying more hardware may fix it, but this assumes the technology is scalable. The worst kind of bottlenecks may require whole processes or systems to be re-engineered, especially if the rate-limiting step for a process is a manual one
- **Design Failures:** The lack of a good overall design to the system can mean that making and testing changes is a very lengthy process. In the worst case it may not be practical to extend or integrate at all. This is particularly relevant if you have your own home-grown bespoke systems. Many bespoke systems evolve over a period of time to meet changing requirements, sometimes without a clear design intention

- **Testing Environment:** Before you make changes to any production system you need to be able to thoroughly test out those changes in a realistic, parallel environment. If you don't have a separate testing environment then rolling out changes on a production system is an extremely risky activity

Data management in the pipeline

An integrated data pipeline could be implemented simply for the benefit of efficiency gains and to eliminate process bottlenecks, but other benefits can be realized with the right data management capabilities within the pipeline. Most important of all is the ability to record the relationships between data throughout the whole pipeline. So, for example, a user would be able to look at information in a final report and trace back the origin of that information and all of the experiments, data and analyses that give rise to it.

If you work in a highly-regulated industry, like pharmaceuticals, then a high level of traceability will be a feature of your business already. But there are potentially large benefits in having this implemented within an integrated system rather than, perhaps, having a paper trail to follow. A well-integrated system ensures data quality and reduces the costs of auditing projects and studies.

But, in addition to efficiency, data quality and traceability benefits, integrated data management opens up new possibilities. Integration lets you look at your information resources as a whole with data mining, portfolio management and decision support systems. Being able to start from a big picture and then examine the underlying detail and data allows you to extract more value from a precious business resource.

New ways of working

Integration can also enable new ways of working at the laboratory level. One possible advance is the concept of dynamic or adaptive experimentation. A conventional experimental programme may rely on a sequence of separate experiments to reach a desired outcome. The first experiment is designed, performed and then analyzed in a sequential fashion. The findings from the first experiment are then used to design the second experiment and the cycle repeated. An approach like this might typically be taken in exploring an experimental parameter space (e.g. formulation design) with each successive experiment narrowing on an optimal outcome.

Dynamic experimentation involves speeding up the feedback process so that early results become available before the whole experiment is complete. This may allow you to dynamically adjust the design of your experiment even whilst it is still running to better target the experimental resources on a specific region of the experimental space. With sufficient levels of integration you may even be able to automate this process. Dynamic or adaptive experimentation has already been used to good effect in clinical trials and will surely also find an important place within the laboratory.

Protecting your investment – digital preservation

An archive is an important resource for many businesses. Long-term storage of experimental data, study reports and the like may be vital to comply with regulation and essential to defend patents and protect your intellectual property. The image of a traditional archive is of locked rooms with endless shelves of paper files and folders. But more and more data is now ‘born digital’ and the emphasis is shifting to electronic archives. As the final resting place for your important data it is logical to integrate long-term archiving with an integrated data pipeline.

Digital preservation is not just about document management, neither is it just a technology. Long-term preservation of digital data and documents requires a strategy, active management and the right technology to support it. If you archive a paper report then you can be fairly certain that someone will still be able to read it in 100 years time. In the digital world you can find that a file format from as little as five years ago is impossible to read with your most recent software.

Thus, whatever system is used for storing data, it is important to ensure that the data can still be understood when the system is upgraded, replaced or retired. In some cases, the way in which the data is visualized is as important as the data itself. For example: a huge data set stored as a file of comma-separated values is almost useless: data mining and visualization tools are necessary to understand it, and the ability to use such techniques needs to be retained. In general, if the lifetime of the data exceeds that of the software used to help create or analyze it, which is typically just a few years, then there is a potential problem.

The solutions to such problems are part of an emerging discipline known as 'digital preservation'. Over the last few years, techniques have been developed to help reduce many of these problems, including centralized storage of data (e.g. a hierarchical storage system), data migration (to make data accessible to new technologies) and data standardization (using long-term formats such as XML to reduce the rate of technological obsolescence).

Summary

Integration of the LIMS with other systems to build a data pipeline can provide business benefits far beyond simple efficiency and quality gains. Integration is an enabler for a whole range of technologies and business processes from the laboratory to the enterprise level.

Tessella has many years' of experience of joining up disconnected systems and building integrated solutions in and around the laboratory. We have established ourselves as world leaders in providing practical, flexible and future-proof solutions in the field of digital archiving.

Tessella - providing innovative software solutions to scientific, technical and engineering problems

Tessella specializes in the application of innovative software solutions to scientific, technical and engineering problems. Our services cover software design and development, IT consultancy, infrastructure support and project management.

Other Technical Supplements published by Tessella include:

- | | |
|---|--|
| <input type="checkbox"/> Active Server Pages | <input type="checkbox"/> Integrated Lab Systems |
| <input type="checkbox"/> Archiving of Electronic Info | <input type="checkbox"/> J2EE |
| <input type="checkbox"/> Automated GUI Testing | <input type="checkbox"/> Java |
| <input type="checkbox"/> Bayesian Statistics | <input type="checkbox"/> Linux |
| <input type="checkbox"/> Beowulf Clusters | <input type="checkbox"/> Microsoft .NET |
| <input type="checkbox"/> Beyond LIMS | <input type="checkbox"/> Object Oriented Programming |
| <input type="checkbox"/> C++ | <input type="checkbox"/> Open Source and Free Software |
| <input type="checkbox"/> Choosing and Using a LIMS | <input type="checkbox"/> Pocket PC |
| <input type="checkbox"/> COM | <input type="checkbox"/> Portable GUI Development |
| <input type="checkbox"/> Computational Fluid Dynamics | <input type="checkbox"/> Real Time Systems |
| <input type="checkbox"/> Computer Image Processing | <input type="checkbox"/> Regression Testing |
| <input type="checkbox"/> Decision Support Systems | <input type="checkbox"/> Security and the Internet |
| <input type="checkbox"/> Development for the Mobile Platform | <input type="checkbox"/> Simulation |
| <input type="checkbox"/> Digital Preservation Practical Experiences | <input type="checkbox"/> Soft Computing |
| <input type="checkbox"/> e-GIF | <input type="checkbox"/> Software Development Cycle |
| <input type="checkbox"/> Electronic Data Capture | <input type="checkbox"/> Software Documentation |
| <input type="checkbox"/> Electronic Lab Notebooks | <input type="checkbox"/> Software Portability |
| <input type="checkbox"/> Evolutionary Computing | <input type="checkbox"/> Software Re-engineering |
| <input type="checkbox"/> Excel | <input type="checkbox"/> Software Specification |
| <input type="checkbox"/> Extending the Life of Software | <input type="checkbox"/> SQL |
| <input type="checkbox"/> FDA 21 CFR Part 11 | <input type="checkbox"/> UNIX Inter-Process Comms |
| <input type="checkbox"/> Formulation | <input type="checkbox"/> UNIX Systems Performance |
| <input type="checkbox"/> FORTRAN 90 | <input type="checkbox"/> Web Services |
| <input type="checkbox"/> Grid Computing | <input type="checkbox"/> Windows 2000 Services |
| <input type="checkbox"/> High Throughput Experimentation | <input type="checkbox"/> XML |
| <input type="checkbox"/> High Throughput Screening | <input type="checkbox"/> X Windows |
| <input type="checkbox"/> Instrumentation | |



INVESTOR IN PEOPLE

Tessella Support Services plc

3 Vineyard Chambers, Abingdon, Oxon, OX14 3PX, England

Tel: (+44) (0) 1235 555511 Fax: (+44) (0) 1235 553301

info@tessella.com www.tessella.com