Enabling collaborative science via the integration of analytical data into organisational systems

The storage of data drawn from the analytical process has for years been the stumbling block in the overall 'knowledge management' of the discovery process. This article argues that current scientific data management packages offer promise in addressing the needs of an analytical laboratory and its users, while having the ability of progression in offering an ultimate Analytical Laboratory Notebook.

By Mark Bayliss

What is the value of analytical data? What value can be ascribed to the interpretations of the data made by experts? What value can be attached to the ability to rapidly recall the knowledge that was originally derived from the analytical data buried within our organisations, which can be used to solve organisational challenges such as the problem of a new impurity in a manufacturing batch of drug material? What value does knowledge play in the search for the next market-leading product? How can we better track the supporting knowledge for the inception date of novel structures that then go on to be the next blockbuster drug? Is there value to enable scientists to access their historical knowledge to make better informed decisions?

Much of the corporate domain excellence that has been built up over the years is founded in major part on analytical science. It is for the most part a regulatory requirement that all drugs and most marketed chemistries are supported by a truly bewildering wealth of analytical data. And yet, we often perceive this data to be of questionable value, and do little to store the data in a rapidly accessible medium. Scientists are typically not satisfied to release the true structure of a compound until they have subjected it to a battery of analytical techniques that, as a whole, provide the final unambiguous result. How much of the supporting data in this process can really be ascribed to knowledge and actually be useful? How much of what is done in the analytical process has to be duplicated because the previously learned knowledge cannot be found quickly or the person who used to know about it has left the company or changed departments?

'Knowledge Management' has very much become an industry buzzword and often abused. However, to understand what this phrase means for certain parts of our organisations, we need to consider the extents of the knowledge creation, capture and management processes. We should also consider whether this data has value, and how facile access can be made to this knowledge

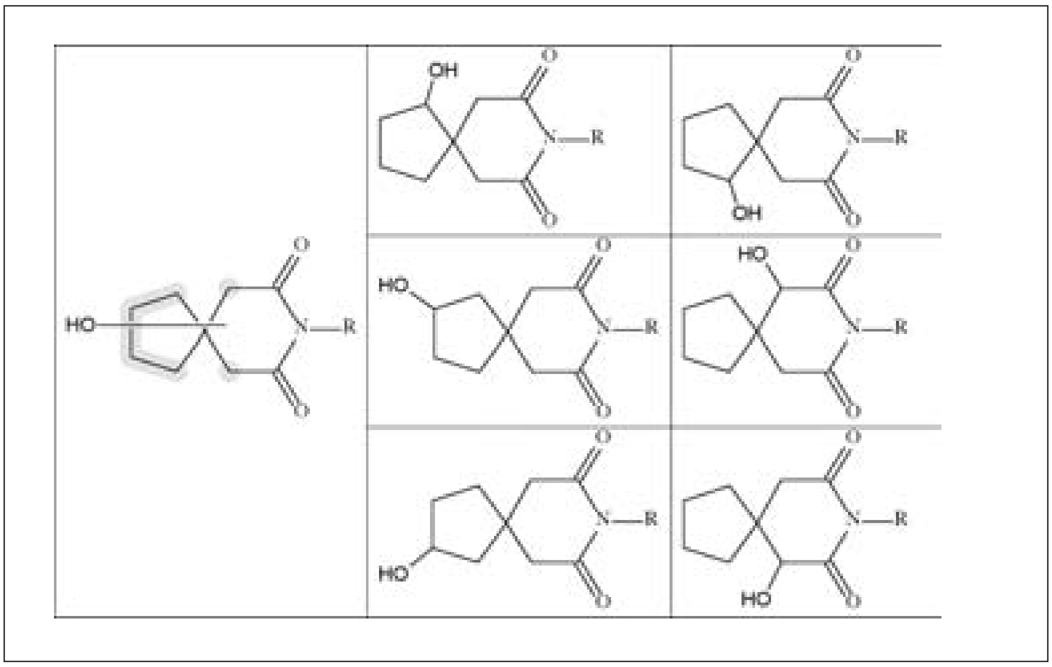

Figure 1
Top left is a Markush structure representation, which is a single structure representation that is able to encompass all of the structures shown to the right. In early chemical R&D, this form of structure representation is suitable because additional investment of time and effort, using different analytical techniques, is required to finalise the structure. It may be numerous months or even years before the determination of the final structure. In many cases, this form of structure is not supported directly by the structure registration systems. Organisationally, companies may have issues when defining the first inception date of such a structure

that is expected to deliver organisational benefit. Finally, we should discuss how analytical knowledge can be inter-related with other organisational infrastructure knowledge management systems to produce something larger than the constituent parts.

Corporate-wide resource, material and knowledge management, when properly implemented, has boosted productivity and resulted in immediate returns, as demonstrated by many success stories[1-6]. Classical LIMS systems have done a fine job in supporting optimal processes and efficient, flexible workflows. E-archiving and storage facilities are a must in regulated environments. ELNs, the latest must-have in modern R&D[7], help make handwritten records legible and searchable, and address a number of legal issues. However, some of the general efficiency goals involving analytical information have generally been difficult to meet within traditional ELN, LIMS and e-archiving approaches. Research calls for frequent access to previous results to help the evaluation of the new data, and for close collaboration between scientists to provide breakthrough in the discovery process. Linearity of the data and sample management systems, and lack of analytical and chemical 'intelligence' make even the best system design rigid and inflexible for the task.

There really is no single point where knowledge creation starts which makes it rather difficult to conceptualise and create an all encompassing knowledge management system. Sometimes it can be merely the jottings of thoughts that our structural design experts have; it can be the deductions made by analytical scientists relating to ideas of a potential structure or scaffold; it can come from the chemists working to synthesise new chemistries. All in all we quickly find that no single infrastructural system is really appropriate to capture what we might refer to at the highest level as 'Organisational Knowledge'. Just consider the most prized and protected piece of intellectual property, the structure – how many places can a new structure be conceived? In the defence of a patent case, would the earliest date of conception be the date of registration in the corporate database? And is this truly the earliest date of the structure being defined? Would the determination of a structure during the analytical process related to a similar compound not also be a potential date of conception of a novel structure, and could this not be a potential new and more powerful drug? It is a complex question and thus there is no simple and single answer. One may never know when a piece of analytical or other knowledge is required, but like insurance our organisations are tasked with ensuring the safety of this information for the 'just in case scenarios' which recoup the investment by allowing rapid solutions to our complex problems.

Structure confirmation or even *de novo* structure determinations by analytical measurement will typically be derived from multiple analytical techniques and will often be the result of many hundreds and thousands of megabytes of raw data files and that may just be for a single structure. The need to capture these raw data files and have them such that the data be easily and quickly retrieved is extremely important. Numerous systems exist within the marketplace for this, with examples including Waters Corporation 'NuGenesis SDMS', Agilent Corporation 'Cyberlab', plus of course the numerous in-house developed solutions. Archive in this sense is meant to infer that the physical files that are produced by analytical instrumentation, whether as raw data or processed data, are maintained in a secure electronic environment that allows recovery when required for further interrogation or processing. However, can we consider that simply retaining just the physical data files in a repository fully encompasses what we might call an end-to-end Knowledge Management Solution? While it is certainly a part, it is not the complete solution.

For many years, LIMS systems have acted as the archive of the results of the analytical process. However, they are predominantly only stored as alpha-numeric values and meta data, often with the additional advantage of powerful workflow management pieces. LIMS systems, however, are typically not structurally enabled and rarely support the complex multi-dimensional data types such as LC/MS, LC/MS/MS, 2D NMR and the like. Similarly to the archive system, a LIMS is just a component of the end-to-end knowledge management system.

Structure registration systems are not analytical data aware and in many cases only support finalised structures, when in reality a structure may be conceived from data and may often be only partially defined in terms of the finalised structure. These indeterminate structure representations are termed 'Markush' structures (**Figure 1**) and are most often found in close proximity to the impurity and metabolism parts of the R&D process. Thus in the case of structure inception dates, would a Markush structure be admissible in a legal defence, when attached to the analytical data and stored in a secure Analytical Knowledge Management System?

Analytical Knowledge Management Systems as in the case of the above mentioned systems offer a domain specific solution that play a contributing and equivalently important role in the larger organisational knowledge management solution infrastructure. Providers of such systems, including Waters Corporation and Advanced Chemistry Development, Inc, (ACD/Labs) are faced with a huge challenge due to the massive complexity of analytical data and the multiplicity of analytical techniques and supporting instrument and vendors for such instrumentation. Today's research analytical laboratory supports a myriad of applications and special considerations must come into play[8].

For example, in Mass Spectrometry alone there are nearly 10 major vendors, namely the Waters Corporation, Agilent Corporation, ThermoFisher Scientific, MDS Sciex, Shimadzu Corporation, Leco Corporation, Jeol, Hitachi, Varian,and PerkinElmer. From each of these vendors it is typical to have both Liquid Chromatographic (LC) and Gas Chromatographic (GC) Mass Spectrometers.

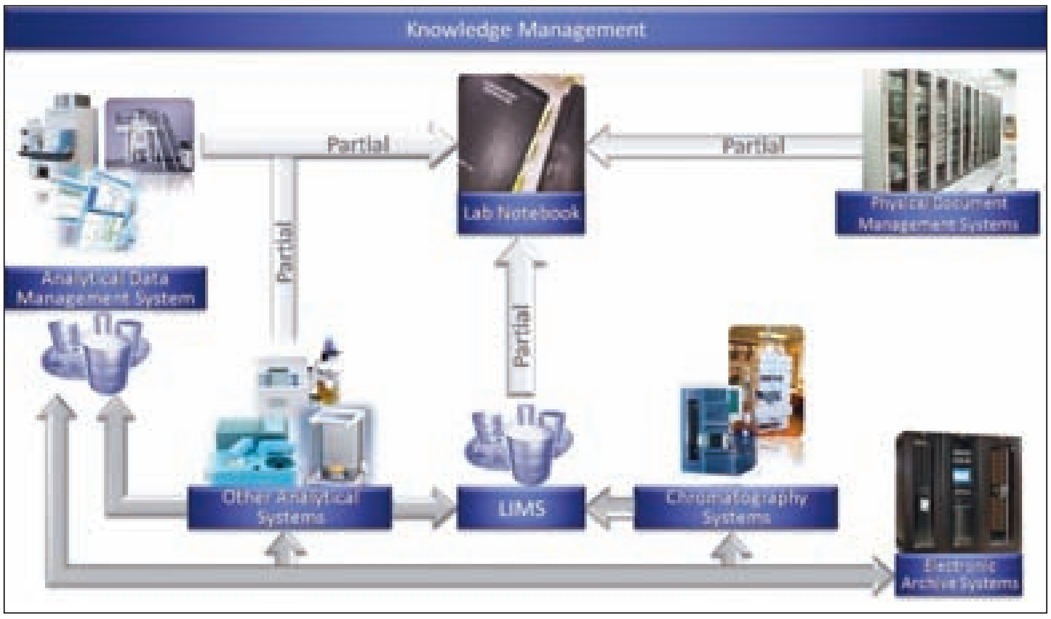

Figure 2
The current situation of Knowledge Management related to the Analytical Process. Currently the output from each of the components of the system end up in paper format and pasted into Laboratory Notebooks which are then filed in document management systems. However, in many cases, only a part of the captured knowledge may finally end up in a laboratory notebook for many reasons which are not included in this discussion

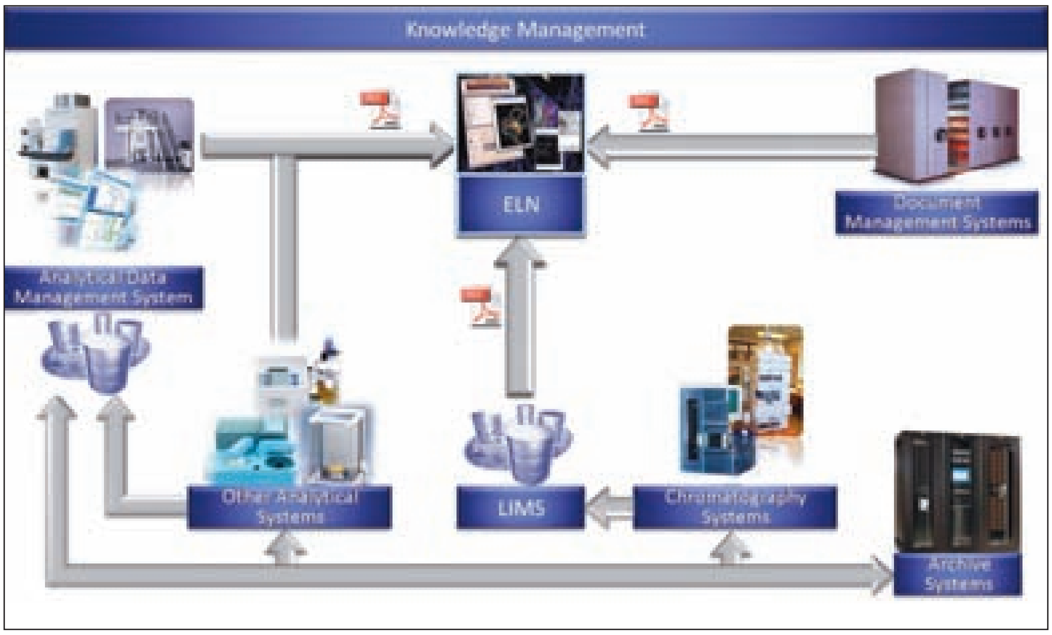

Figure 3
Proposed integration of different systems that captures the contents of analytical processes. Each individual system offers a level of granularity to individuals within organisations based on their job function. The ALN or, as shown here, Analytical Data Management System, is designed to capture all forms of analytical data including complex multidimensional data types

References

1 Kihlen, Mats. Electronic Lab Notebooks – do they work in reality? Drug Discovery Today (DDT), Vol 10, Number 18, Sept 2005, p.1205–1207.

2 Stamatiadis, Dimitri. Electronic Archiving: A New Paradigm. American Pharmaceutical Review, Vol 8, Issue 5, Sept/Oct 2005, p. 10–16.

3 McDowall, RD. Future Trends in LIMS. American Pharmaceutical Review, Vol 8, Issue 6, Nov/Dec 2005 p.10–15.

4 Taylor, KT. The status of electronic laboratory notebooks for chemistry and biology. Curr Opin Drug Discov Devel. 2006 May, 9(3), p.348–53.

5 Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology: Report of a Workshop (1999). http://books.nap.edu/openbook.php?record_id=9591&page=154.

6 Brown, Douglas, Williams, Antony and McLaughlin, David. Web-Based Information Management System. TrAC: Trends Anal. Chem., 16 (1997), p.370 at www1.elsevier.com/homepage/saa/trac/wimsarti.htm.

Continued on page 38

Within each category there may be multiple individual instrument types including, but not limited to, Quadrupoles (Single and Triple), Ion Traps, Time Of Flight, Magnetic Sector, Fourier Transform Mass Spectrometers (FTMS) and a myriad of hybrid systems. Within the MS realm alone, a company has to handle a wide range of data formats (**Figure 2**).

In early R&D, the translation of data into knowledge is built on the backs of scientific domain experts who are the go-to people when a complex problem requires historical knowledge. One reason for this is that historical findings are stored as hard copies or electronic reports, either completely unsearchable electronically or with limited text or metadata-based search abilities. For example, how would a chemist perform a structure search of a Word document stored in a long term file storage system? It is suggested that viable and useful searches should include structures (including Markush structures) as exact structure, structure similarity and sub-structure; metadata; spectral searches; and chromatographic data searches. Indeed, access to vital knowledge may require more than a single search type in order to get to the final limited subset of returned records. In more recent times, the rapid move to Electronic Laboratory Notebooks (ELNs) has improved this situation somewhat in that older, paper-based notebooks are replaced with an electronic archive of documents typically in Adobe PDF™ format. At least now individual notebook references can be rapidly recovered and reviewed. However, again these systems can be limited in their searchability, especially when considering that searching for structures which are embedded in PDF documents depends on whether the PDF generator has been extended to allow structure indexing. Analytical data of course remains outside the domain of searchability in these systems.

Different roles within organisations are looking for different ways to access the knowledge stored within their organisational systems, and thus knowledge tailoring should be considered. As the primary focus of this article is to consider the needs of the Analytics Process, we should define a series of expectations that may drive such a knowledge silo. The introduction of the concept of the Analytical Laboratory Notebook (ALN) is considered appropriate at this stage and is proposed to allow historical access to capture knowledge through a wide series of search types including: Spectral, Sub-spectral, Chromatographic, textual, metadata, structural, including Markush, sub-structural and structure similarity. A number of such systems are available from providers such as Waters Corporation, ThermoFisher Scientific, Bio-Rad and ACD/Labs. It is suggested that ALNs are not a substitute for ELNs, rather they are an adjunct offering a level of knowledge granularity that is not necessarily organisationally at the level of the ELN. Information from an ALN is summarised in PDF format and stored within the ELN. The integration of these different electronic systems is summarised in **Figure 3**. An example of an integrated system that brings together the benefits of ELN and an analytical management system is demonstrated in **Figure 4**.

For this type of system to be effective, it should be able to accept data entry from all instrument vendors and all instrument types. Data from these disparate instrument types should be normalised to enable globalisation of metadata, the reasoning being that data is typically stored in different ways for what is in essence the same thing. Thus, when it comes to searching this information, it is necessary that it be captured and stored in the same way. Instrument vendors for the most part, do enable access to their raw data through what is often termed an Application Program Interface (API) that allows third parties to integrate into their systems. These APIs help insulate independent vendors from changes that are made by the original instrument vendor to their internal data structures, which are often subject to regular change. In principle, the majority of requirements of such an ALN can be delivered using existing and available technologies. However, one challenge that remains is the extraction of knowledge captured in the original vendor higher-order processing systems. This situation certainly may change with the encouragement of vendors to provide APIs that are compatible with these higher-order data processing environments. In the case of report generation and the need to integrate to ELNs, the ability to output in PDF enables capture of salient information. In most cases vendors providing Analytical Data Management Systems offer integration to these systems through their own APIs or integration toolboxes (**Figure 5**). In the case of the web services layer from ACD/Labs, it is reported[9] that any data element stored in the Analytical Data Management System may be extracted including even the stored spectral data.

Conclusions

Is the analytical laboratory well served by the existing IT data management solutions? While the ELNs certainly can deliver, especially in some areas of R&D, they still have not fulfilled the task in the analytical lab. Among the missing pieces is the one related to the handling of live analytical data. Searchability of the analytical data by specific chemical and spectral parameters that can offer an analysis of trends and an inspiration for a researcher is also amiss. Collaborative nature of the research can be enabled by better uniformity and accessibility of the analytical tools and results.

There have been efforts in integrating CDS,

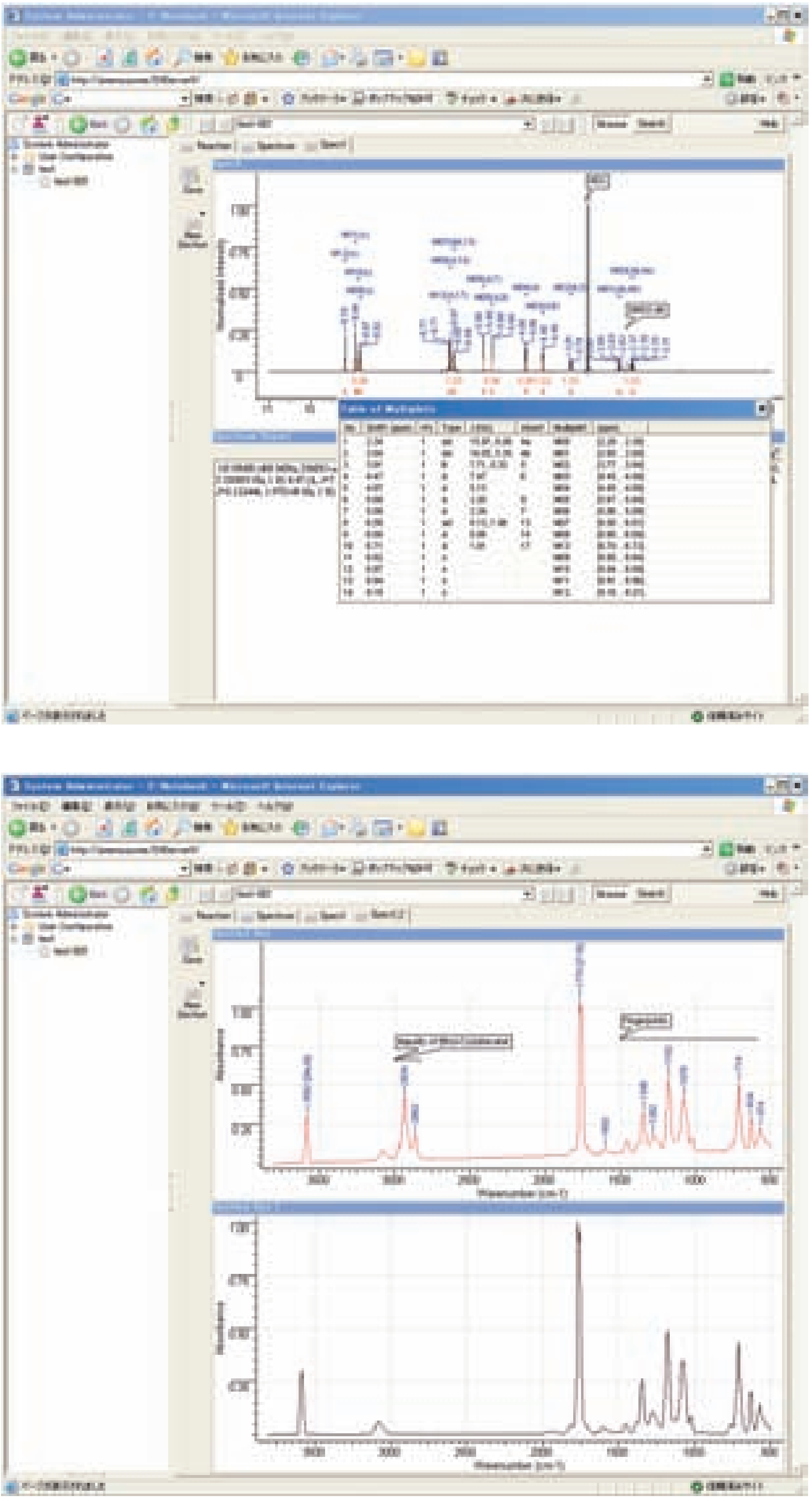

Figure 4: Analytical results opened by ACS/SpecManager are incorporated into the CambridgeSoft Electronic Notebook. The integration shown is enabled and distributed in Japan by Fujitsu

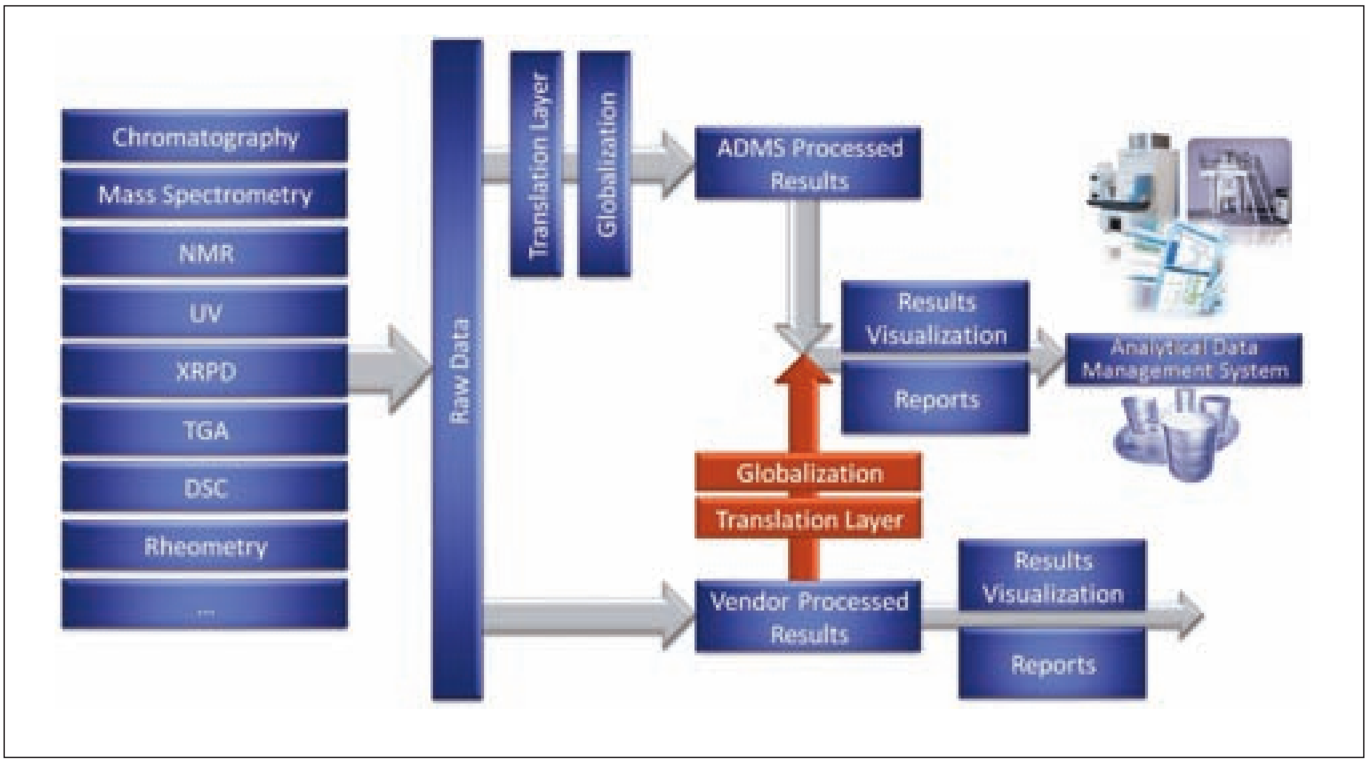

Figure 5: Proposed Analytical Laboratory Notebook System allowing for extraction of knowledge from the Analytics part of our organisational processes

SDM, LIMS and ELNs but the overall analytical integration is rarely available. However, current scientific data management packages offer promise and feature advanced integration that can address the needs of an analytical laboratory and its users, and progress to offer an ultimate Analytical Laboratory Notebook. DDW

With background in mass spectrometry technologies and software, Mark Bayliss is the Director of Analytical Informatics at Advanced Chemistry Development, Inc (ACD/Labs). Prior to joining ACD/Labs in 2002, Mark worked as the Product Manager at Thermo Finnigan, and earlier at Micromass, covering its quadrupole-based technologies. Mark came into analytical chemistry through his Doctoral research at the University of Wales in Swansea.

Continued from page 36

7 Elliott, Michael H. The state of the ELN Market. Scientific Computing World: December 2006/January 2007, p.53.
8 McLaughlin, David R, Williams, Antony J in Lindon, JC, Tranter, GE and Holmes, JL (ed). Laboratory Information Management Systems (LIMS) Encyclopedia of spectroscopy and spectrometry. Academic Press, 1999, p. 1105–8.
9 ACD/Web Librarian web interface is used to grant public assess to the ACD/Chromatography Applications Database at www.chromdb.com/. ACD/Web Librarian Web Services is described at www.acdlabs.com/wlws/.