

# Effective Access and Retrieval of Laboratory Data to Enable Knowledge Management

by Mark Fish

Laboratory services in both the research and production environments are increasingly under the spotlight as science-based organizations focus on ways to improve efficiency and productivity. In particular, pressure is on the world's research and development laboratories to produce new and enhanced products at unprecedented rates and with improved predictability. The key goals for these laboratories are to reduce the time scales for product development, improve the chances of product success, and reduce discovery and development costs. The ability to access and build on prior research and development data is critical to this objective.

Drug development is a good way to illustrate the reasons behind the motivation to improve productivity and increase efficiency in the laboratory. Leading pharmaceutical companies are expected to submit between four and six New Drug Applications (NDAs) per year<sup>1</sup>. The FDA estimates that it takes approx. 8.5 years to study and test a new drug before it can be approved for the general public.<sup>2</sup> Other sources estimate 10–15 years, with the cost of developing a new drug averaging

more than \$800 million,<sup>3</sup> and attrition rates as high as 95% in early stage research and discovery, 98% in preclinical, and 80% in clinical trials.<sup>4</sup> From these findings, the chance that the next compound to be synthesized on the bench will be a blockbuster drug is around 1 in 50,000.

With costs and risks on this scale, the incentive to improve productivity and increase efficiency in the pharmaceutical industry is clear. All over the world, similar pressures exist in research and development laboratories irrespective of the industry. In many cases, it is all the more challenging for the scientist to make important technical advances in an extensively researched and developed marketplace. Any technologies that can help reduce the time and resources required to produce new products can have a significant impact on the bottom line.

Drawing again on the area of drug discovery, advances in genomics, proteomics, high-throughput screening, combinatorial synthesis, and in silico testing have been made in an attempt to optimize and automate the processes of target compound generation and lead com-

pound identification. Ironically, in many cases the high attrition rates in these phases of research and early discovery are desirable.

The closer a compound gets to submission to the FDA, the more the cost of development increases throughout the various stages of a clinical trial. The elimination of poor candidate compounds earlier in the development process can result in enormous savings in the long run. These highly automated techniques have proven successful and have generated huge volumes of invaluable data. The problem that organizations now face is how to use, manage, and apply the data to maximum benefit. Despite more sophisticated data management tools, perhaps now more than ever the laboratory is drowning in a sea of data. Reaping the benefits from the knowledge obtained from these data has been compared to trying to take a sip from a fire hose.

Data volume is not the only challenge to overcome if laboratories are to move toward genuine knowledge management. Effective technical integration, data retrieval, and presentation are all considerable obstacles. Integration (not only of data) is a pervasive theme

in today's science-based industry. It is, however, useful to consider why laboratories need to integrate different software applications.

Scientists are primarily concerned with taking raw observation and data and distilling them into information and knowledge. This is the overriding reason that organizations invest heavily in integrating systems. Systems integration costs alone reportedly account for around 30% of IT budgets.

It is not uncommon to find different vendors' information management applications in use for the same function at various locations within the same organization. Industry trends toward globalization and consolidation have exacerbated the situation. In highly automated and integrated laboratories, it is reasonable to expect that the information gathered from past projects should be available to help accelerate new product development.

In the drive to more effectively leverage information throughout the enterprise, industry is increasingly faced with the problem of integrating data sources built on different technologies with diverse data formats. The existence of departmental, geographical, and organizational barriers can present difficulties. Unfortunately, often the biggest hurdle is the technology employed to help capture and manage the data.

Discrete software applications

tend to be good at the jobs they do. In fact, these systems are valuable because they simplify the retrieval and collection of the data they contain for a specific function. There are instrument data systems to manage instrument data, bioinformatics for biological data, LIMS to manage samples, chemical information management systems (CIMS) to register structures, and document management systems (DMS) to manage standard operating procedures (SOPs), among many others. The best examples of each are effective at their specific function.

The real problems arise when sharing information between systems. This is where the traditional challenges of technical systems integration come into play. The problem is magnified when the wealth of information within these systems is required to be available throughout the entire organiza-

tion. This is the goal of knowledge management. In addition to the technical challenge of integration, issues of retrieval and presentation must also be addressed.

Traditionally, systems integration has focused on enabling communication between two discrete systems. This is analogous to building a bridge between two different islands. In knowledge management, however, the task is to enable many different systems in an organization to communicate effectively via a common interface. This process has been likened to walking on water.

## Conventional methods of systems integration

Many well-established methods enable different applications to communicate. These include such technologies as application programmable interfaces (API), object

**Table 1 Key characteristics of conventional methods of integration\***

API	SQL	ORB	Files
Powerful	Fastest	Flexible	Proven reliability
Nonportable	Unlimited power	Complex	Platform neutral
Product/version specific	Can circumvent security	Platform issues	Easy to test debut
	Can violate business rules		

\*API, application programmable interfaces; SQL, standard query language; ORB, object request brokers.

request brokers (ORB), direct database access, reporting technologies, and, of course, paper. All of these technologies have their place when integrating systems, and each can be used to permit effective point-to-point integration. The benefits and drawbacks are listed in *Table 1*.

The biggest disadvantage of these technologies is that, like building bridges, they tend to be expensive and very difficult to change once they have been constructed. This is a fundamental problem with a point-to-point approach to systems integration, and one of the biggest impediments to allowing successful knowledge management (see *Table 2*).

## Emerging methods based on established technologies

Undoubtedly, the best example of a major scale integrated system is the Internet. The Internet is a powerful and robust tool that enables individuals to perform research, manage finances, shop, and reserve airline flights, etc., through a common interface. It is interesting to reflect on the technology that has been fundamental to its success.

### Internet communications protocols

It is commonly accepted that the Internet was created by Tim Berners Lee, who brought together three discrete technolo-

**Table 2 Systems integration dilemmas**

#### Paper

A great deal of systems integration can be achieved with paper. In many ways paper is the ultimate integration tool (cut and paste).

#### Rocks

Standard interfaces can provide “out-of-the-box” integration solutions but, like rocks, they may not be flexible enough to meet all needs.

#### Glue

In software integration, some “glueware” is often needed. The custom code needed to enable systems to communicate is a potential maintenance nightmare.

#### String

Point-to-point integration can be expensive to build and maintain.

gies to create the Internet as it is known today:

- HTML—A markup language that allows one to format the content of a document
- HTTP—A protocol that permits one to transfer documents via standard network technology over a wide area network
- URL—A facility that enables documents to be indexed and retrieved using a universal resource locator.

The combination of these three elements fostered the creation of the first Web browser. It is remarkable that these underlying technologies have remained stable for over 13 years. However, adding Web interfaces to existing applications does not provide the solution to problems of integration. A Web browser can be thought of as a

“dumb” terminal of old (as part of a mainframe computing architecture). It may mean that applications on one network can be accessed via a common interface without having to tolerate a green and black screen, but it does not solve integration and communication problems. This is because a Web browser, like a “dumb” terminal, addresses presentation only, not access to the underlying data themselves.

### XML format

eXtensible Markup Language (XML) is a subset of the HTML format. HTML can be used to format text, but it cannot interpret data. The need to send data over the Internet was one of the primary motivations for the conception of XML. While they are both sophisticated text files, HTML

handles format and XML manages the structure and content of the data. An important attribute of XML is that it can be used to convey the content value within the information, which permits standard formats to be developed and defined, and allows developers to create applications using a simple guide called a document type definition (DTD) document. Data can be exchanged without an intimate knowledge of the application that initially produced them.

### **Web services architecture**

Web services are cited as the set of technologies used to facilitate easier integration and data exchange, simplified application access, and more effective information and knowledge management throughout the organization. To make informed decisions about scientific data, there remains a strong reliance on expensive point-to-point integration for collation, and the interpretation and processing of complex data formats in order to reach the actual data required. Web services are expected to address these types of issues and provide real benefits through their dual qualities of accessibility and integration.

A Web service is an application that can be delivered as a network service and integrated using standard Internet technologies. The basic platform for a Web service is a combination of XML and HTTP technology. XML is used to describe the functions of the appli-

cation and for communication through the transfer of XML documents. The method employed to transfer these documents is HTTP, the protocol used to transfer Web pages from the Internet to the Web browser on the desktop. Much of the strength of the Web services concept lies in the fact that it combines an easy-to-understand format with a proven and well-established communications technology.

The union of XML and HTTP, as a platform-independent method for which applications can communicate, is known as simple object access protocol (SOAP). It is widely speculated in the Information Services (IS) industry that Web services and SOAP could supersede previous methods for exchanging business data such as electronic data interchange (EDI). With the support of influential organizations, these claims are proving correct.

XML provides accessibility since it can be easily organized, programmed, edited, and exchanged. It is a future-proof, self-describing, plain text format that can be validated using a schema that offers rules of use in different environments. The generalized analytical markup language (GAML) schema was recently proposed as a means with which to handle analytical data.

Integration is provided by tried and tested Internet communication protocols. TCP/IP, HTTP, and

SMTP can all be used and are the backbone technologies of the Internet. The fact that these technologies have long been available and are so widely supported helps integration and platform independence considerably.

Forward-thinking laboratory informatics vendors are looking toward delivering Web services-based solutions based on, for example, LIMS, using a secure SOAP-based architecture. Web services are viewed as one of the fundamental tools that can help reduce the cost of system integration, enable data sharing, and increase interoperability. In this era of information explosion, a commitment to such goals should be very good news to laboratory managers and bioinformaticians.

### **The need for a fully integrated solution**

Particularly in research and development, where the scientist works with a wide variety of different data types and applications, there is a genuine need to develop effective integrated solutions that provide the scientist with a workspace that can be used to capture and share data and information as transparently as possible. Ideally, such a system would assist the scientist in all facets of the scientific process, from planning, through experimentation, through capturing raw data, to the analysis and conclusions that constitute all of the important intellectual property that can be used to support development of a

new product. Due to the sheer diversity of data and information used by the research scientist throughout the course of the day, it is not surprising that paper is still the primary record-keeping tool used in scientific research. Paper, with all of its limitations, remains the lowest common denominator in terms of format and flexibility. Almost without exception, any applications utilized by the scientist can be translated into common paper format, and paper can be easily bound, which is a very cost-effective form of integration.

This is by no means a new argument, and the quest for a searchable, on-line version of the traditional tool of the research chemist—the paper laboratory notebook—has been a topic of study for some time. In fact, there are organizations dedicated to enabling this objective.<sup>5</sup> It is recognized, however, that it is impractical to find or develop a single software application that can incorporate all of the functionality we take for granted in a paper scientific notebook. Many systems specialize in isolation in the management of different data types (chemical, analytical, spectral, pictorial, etc.) and provide essential functions (planning, tracking, archiving, and searching). However, it would be both infeasible and undesirable to create a single application to do it all. The real challenge is to integrate different systems to create a practical workspace that allows the scientist to share and collate information most effectively. This is where for-

mat like XML and technologies such as Web services play a vital role.

## R&D and production

Although the focus of this article has been the research and development environment, the advantages of more effective systems integration and knowledge management are just as important in the production environment. In some ways, the production environment has been the forerunner in terms of systems integration, for it is here that interfaces to enterprise resource planning (ERP) and process information management systems (PIMS) are commonplace. The advantages of instrument integration have long been recognized in automating routine analyses in this setting. The production laboratory can also realize the benefits of knowledge management to attain more efficient problem resolution, more consistent business metrics, and reliable forecasting based on data available across different systems throughout the organization.

## Summary

Those who work for companies that actively promote public-domain XML data formats for instrument data<sup>6</sup> technology to allow more cost-effective systems integration and tools that enable the conversion of data into common formats are sometimes asked why vendors do not make more of an effort to promote the standardization of laboratory data. Herein lies an interesting dilemma: The

creation of standards for data exchange and communication cannot be the responsibility of the vendor community or international standards bodies. The key to enabling standards for communication and eventually knowledge management is adoption. Unless standard formats and protocols for simplified communication and systems integration make it onto the list of essential purchasing requirements, it is unlikely that the open standards and technologies to allow effective laboratory management will become a priority.

## References

1. *R&D and the Internet: Opportunities to profit from Web-enabled pharmaceutical research and development*, Nov 2000, Andersen Consulting, London, U.K.
2. *The new drug development process: Steps from test tube to new drug application review*, US FDA, Center for Drug Evaluation and Research, Washington D.C., [www.fda.gov/cder/handbook/develop.htm](http://www.fda.gov/cder/handbook/develop.htm).
3. *Outlook 2002*, Tufts Center for the Study of Drug Development, Boston, MA, <http://csdd.tufts.edu/InfoServices/OutlookReports.asp>.
4. Bolten BM, DeGregorio T, *Nature Reviews, Drug Discovery*. Trends in development cycles, May 2002. ( See [url http://www.nature.com/cgi-taf/DynaPage.taf?file=/nrd/journal/v1/n5/full/nrd805\\_fs.html](http://www.nature.com/cgi-taf/DynaPage.taf?file=/nrd/journal/v1/n5/full/nrd805_fs.html).)
5. *The Collaborative Electronic Notebook Systems Association (CENSA)*, Woburn, MA, [www.censa.org](http://www.censa.org).
6. *Generalized Analytical Markup Language (www.gaml.org)*.

Mr. Fish is a Services Product Manager, **Thermo Electron**, Informatics and Services, Hanover Business Park, Altrincham, U.K., tel.: +44 161 942 3000; fax: +44 161 942 3001; e-mail: [mark.fish@thermo.com](mailto:mark.fish@thermo.com).