Agsandrew/Shutterstock

WHEN DATA COMES TOGETHER

Peter Boogaard explores the process of laboratory data integration

When considering data integration, we must first stop thinking 'technology' – integration is not just about instruments or other software platforms. Instead, it is about integrating processes, accelerating ideas and facilitating mandatory compliance requirements more economically. Cross-functional collaboration between research, development, quality assurance and manufacturing is all about optimising and integrating multi-discipline distributed processes; all of which require significant amounts of data. By integrating this data into the scientific workflow, its availability and quality within the entire scientific and business community will increase.

The adoption of data integration has been accepted more within GxP-regulated laboratories than in R&D; however, pharmaceutical companies are still predominantly deploying traditional paper-based solutions. Despite the enormous potential for compliance and efficiency gains, significant barriers to successful paperless lab implementations remain. In cost-sensitive industry segments, such as health and patient care, automated workflow integration, including automated science-based data integration, is commonly accepted. Having said that, today's analytical laboratories are using the most advanced and sophisticated scientific instruments, while their data acquisition and storage methods are somewhat old fashioned.

Change is on the horizon, however, as the adoption of new mainstream technologies, like cloud, service-oriented-services, and mobile devices, becomes more prevalent. Economic pressures are accelerating an overall change in the industry's mindset and companies now have to rethink how to facilitate cross-departmental knowledge sharing in a truly global, multi-discipline collaboration environment. Data integration is the first piece of the puzzle – and that's the good news!

COMMON GOAL

Data integration is not a goal in itself. It has been an industry buzzword for quite some time now, and not only has it come to mean many things to many people, but the term often hides the complexity surrounding what it actually comprises. It may sound obvious, but people should look before they leap when saying: 'Yes, we can integrate a LIMS or ELN to instrument xyz or a corporate computer system.' Failing to define clear objectives, measurable metrics, technical implications and corporate benefits usually results in a project disappointment, due to lack of understanding of what should have been included in that implied integration process.

Industry studies show that, on average, each single batch in a pharmaceutical process requires hundreds of manual data transcriptions during the life cycle of a product. It also shows that in many cases, more than a dozen different documents are being processed. A significant reduction in transcription errors will increase the quality, reduce costs and cut non-labour related activities. But the question is, when it comes to achieving successful data integration, where does the process begin?

The first step should be the formation of an overall plan that takes the following points into consideration:

- **What is expected from the data integration?** The essence of integration is to share and merge data between parties and systems. Simply mentioning that the data needs to be exchanged is not good enough; the format of the data or object has to be clear. Consensus and acceptance about how the information is transported across the IT infrastructure is critical.
- **Who is involved?** By definition, integration is between at least two parties

or systems, but often includes many more. It is essential to know who all of these parties are and understand their goals and individual wins.

- **What is each party's contribution?** Each party in the integration has a role to play and these roles can range from being aware of data changes for reporting to participating in the system. Identify what each party is expecting from the process, define project ownership and clearly articulate an escalation plan in the event of difficulties.
- **Avoid customisation or programming at the interface level between systems.** Most data needs to be changed or reformatted prior to, or just after, transferring it. Spend a significant amount of time when debating this subject with vendors at RFI or implementation level.
- **Define internal and external communication strategies to ensure that expectations are set**. When data integration projects are failing, it is often due to a lack of clear understanding of the final goal. Don't making assumptions – ask for details in the process.

STANDARDS – FIRST THINGS FIRST

Data integration in laboratories is not straightforward. It may seem a boring topic these days, but the need for standardisation in our industry, has never been higher. Without these standards, automating data capture from instruments or data systems can be challenging. Several initiatives are working hard to address these badly-needed common standards. The Pistoia Alliance aims to lower barriers to innovation by improving the interoperability of R&D business processes through precompetitive collaboration. The alliance was conceived by informatics experts at AstraZeneca, GSK, Novartis and Pfizer who were all attending a meeting in Pistoia, Italy.

Pistoia's founders realised that their organisations were all tackling the same precompetitive problems – issues around aggregating, accessing and sharing data that are essential to innovation, but provide little competitive advantage. They realised that working together to solve these common problems would free their organisations to innovate by enabling them to cut costs and repurpose precious resources to projects with more strategic, competitive impact.

In June, an industry-sponsored initiative to promote open information standards for the analytical laboratory was formed. The Allotrope Foundation, sponsored by Abbott, Amgen, Baxter, BI, BMS, Merck, GSK and others, is addressing the lack of common metadata repository formats. The proposed framework will consist of (a) open document standards based upon XML and JSON, (b) open metadata repositories to provide accurate input from numerous data sources, and (c) open source class libraries to support these components.

A standard is only a standard if organisations adopt it. The AnIML standard supports a full audit trail capability, digital signatures and validation for regulatory compliance and is gaining acceptance. AnIML is a standardised data format that allows for the storing and sharing of experiment data. It is suitable for a wide range of analytical measurement techniques. AnIML documents can capture laboratory workflows and results, no matter the instruments or techniques used. AnIML is based on XML, which has two consequences: first, many tools for XML manipulation are readily available off-the-shelf, making implementation easier. Second, as XML is a text-based format, AnIML documents are human-readable – an important aspect for long-term storage. AnIML is being developed by the ASTM E13.15 subcommittee on analytical data, comprising volunteers from industrial, academic, government and vendor communities.

Other standards which have a significant impact on how data integration can be successfully implemented include ISA, ASTM and IEEE. These are multi-industry, globally accepted standards. For example, the ISA standard consists of models and terminology for structuring the production process and for developing the control of equipment (ISA-88) and for production, maintenance and quality (ISA-95). Centocor Ortho Biotech used S88 principles to develop a system-independent, recipe-based ELN/LES system based upon Accelrys' ELN technologies. Using S88 resulted in structural processes and provided a zero-day release and zero-day transfer to and from external contract labs and between internal groups including the laboratory, pilot plant scale-up and production facilities.

CONTEXT IS KING

Overall, there are three basic operating principles to optimise data integration. First of all, capture the data at the point of origin to eliminate human error and reduce system complexity. Smartlab from VelQuest, LIMSLink from Labtronics, and LabWare all began with instrument data integration in mind. Their original products were designed to capture laboratory data at the data source. Secondly, simplify and implement self-documenting processes to eliminate transcription errors and avoid unnecessary retyping of data. In a recent survey, 32 per cent of people voted that data integration in a paperless laboratory will eliminate manual entries and data transfer.

> **'Change is on the horizon, however, as the adoption of new mainstream technologies like cloud, service-oriented-services, and mobile devices, becomes more prevalent'**

Finally, ensure that metadata is captured in a structured way. Raw data represents a set of unorganised and unprocessed facts (e.g. collection of numbers) and is usually static in nature. A data file without context or metadata information is meaningless. The scientist is no longer in the laboratory, but integrated in the overall quality process. To ensure tacit knowledge is maintained in computerised systems, context information must form the foundation for integrated scientific analysis and interpretation.

Context is the organisation of related elements that makes analysis and interpretation possible:

- **Data type context** – enables specific types of data analyses.
- **Batch context** – enables batch-to-batch comparisons.
- **Process context** – enables process-to-process comparisons.
- **Site context** – enables site-to-site comparisons.
- **Genealogy context** – enables upstream/downstream correlations.

Thermo Scientific Integration Manager for the Paperless Lab provides bridges between the islands of data generated in the lab and transforms that data into information that can be used across the enterprise. It provides access to all instrument data via a single interface and enables the real-time investigation of results. The technology converts raw data to an XML storage format to ensure future-proof data archiving and to facilitate data and information sharing across the organisation, without having to rely on access to the original application.

PerkinElmer offers an iLAB solution, which is part of its Ensemble for QA/QC

➤

suite and like Thermo Scientific's product, provides access to all instrument data via a single interface, as well as real-time investigation of results. Again, raw data is converted to an XML storage format for future-proof data archiving and the sharing of data and information without the need to access the original application. Accelrys offers Discoverant, which is a validatable environment that delivers on-demand manufacturing process intelligence by aggregating and contextualising different data types and sources, to create an end-to-end view of the manufacturing process.

CHANGING TECHNOLOGIES

Technologies supporting conceptual modelling through ontologies and data standards are maturing rapidly. Although they remain in the early stages of adoption, ontology mediated system interoperability and formal metadata management have substantial potential to facilitate a best of breed or Service-Oriented-Architecture (SOA) strategy. The ontology-based approach will allow the user to integrate existing database sources and achieve interoperability between different data formats and applications. Key to the success of the development of such applications is the need to make existing content in organisational data warehouses or siloed data stores available to ontologies. Both Google and Microsoft are investing here.

Balance and titrator vendors are increasing the value of their instruments by implementing approved and pre-validated methods in their firmware. For example, Sartorius allows methods to be implemented directly in its balances. Mettler-Toledo is deploying LabX middleware to realise that functionality across its balance, titrator and other LabX-supported instruments. This could have an impact on validation efforts in lab and manufacturing operations, such as fewer points of failure during operation, reduced customisation of software, and better documentation.

'Tacit knowledge is based on common sense, while explicit knowledge is based on academic accomplishment – both are underutilised'

Data-intensive science is becoming far more mainstream. Research is increasingly collaborative and complex, leveraging multiple technologies to get a systems level understanding of diseases and organisms. Data integration is crucial for enabling virtual knowledge sharing and the exponential rise in the scale of (big) data being generated, combined with increased collaboration, has resulted in the need to rethink how data is cost-effectively stored, analysed and shared. Communication is a common dominator. Tacit knowledge is based on common sense, while explicit knowledge is based on academic accomplishment – both are underutilised. Combining explicit information, stored in computer systems, with tacit information is where inventions and knowledge are created and shared.

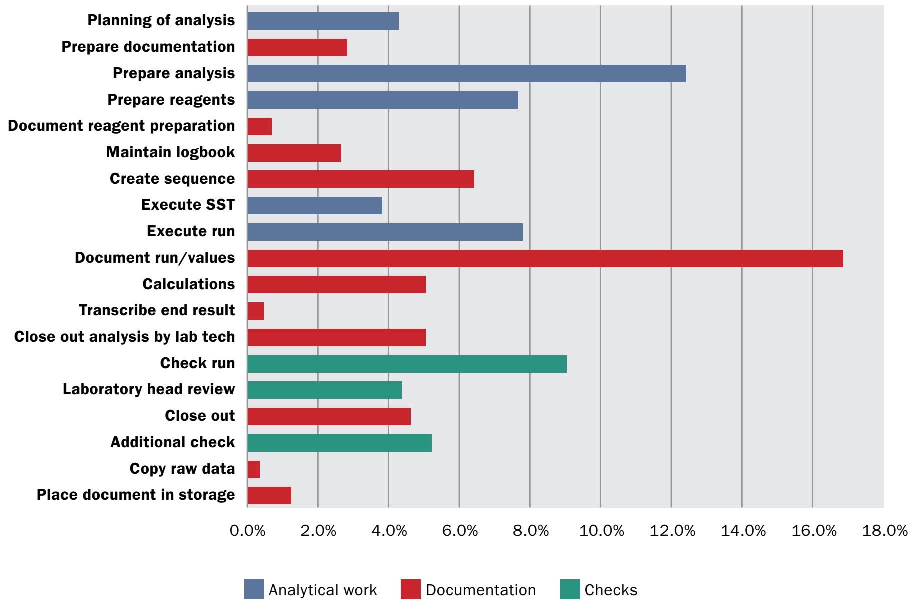
Technology is set to change the dynamics of how scientists work together. The Cloud, for example, is not just an IT initiative; it really changes the ways in which people and science can work together. For example, it eliminates the need to wait for months for a particular scientific paper to be published. Building trust within relationships in order to create these teams remains a people issue and when considering where active communication occurs in science, thoughts may lean towards scientific presentations and great papers.

As John Trigg points out (p.7), Steven Johnson reported at TED.com that he had conducted research into where scientific innovation really takes place. What he discovered was that most innovation actually occurred through social interaction at regular face-to-face lab meetings. There, ideas are shared, data challenged, and concepts rallied. Ideas truly become innovation when combined with others or added to existing facts. Towards the end of the scientific process, conclusions are written as a one-sided conversation with an imaginary colleague, anticipating questions, challenging and stimulating debate. In order to support a strong scientific discussion, people not only need access to relevant facts, but to a collaboration platform which allows virtual real-time interaction with all pertinent information. This forms the basis for capturing the discussion, decision and opinion of an integrated set of people. One example is Shire, a global specialty biopharmaceutical company, which uses IDBS' E-WorkBook platform to enable virtual laboratory meetings and to streamline and facilitate distributed drug research.

CHEATING IS ALLOWED

Those laboratories yet to deploy an informatics system shouldn't worry – being late in the adoption of lab data integration does have one great advantage. Healthcare, banking and the consumer industry have all adopted paperless and electronic data integration approaches, which means that many accepted technologies are now at the disposal of laboratories. This is an exciting time as cross industry best practices can be used to create start-to-finish knowledge management repositories and enable cross-functional collaboration between internal information silos.

Data integration facilitates self-documenting processes



Peter Boogaard is an independent LIMS consultant: peterboogaard@industriallabautomation.com