Maximising long-term value

CORRECT STORAGE AND USE OF DATA IS THE KEY TO KEEPING DATA 'CURRENT' AND RELEVANT IN MODERN LABORATORIES, WRITES SOPHIA KTORI

Data is a company's biggest asset, yet for any organisation, keeping a handle on the potentially vast volumes and diversity of data that are generated can represent a considerable issue.

Burkhard Schaefer, BSSN Software head of product and technology, feels that pitfalls occur when companies become blinkered to reaching their short-term endpoints and don't co-ordinate their broader goals and expectations.

'Organisations will commonly buy informatics platforms and dedicated pieces of software with a focus on solving one problem, or achieving one business or scientific aim. They buy their instruments, are allocated space on their network for the software, and away they go.'

It's a very opportunistic approach that just introduces more data generators into an ecosystem that may already be chaotic, Schaefer suggested. 'When an organisation reaches the sort of size where it has to start segregating work between departments, countries or regions, data control starts to become a major issue.'

Bottom up, or top down?

There are two approaches to handling the problem, Schaefer suggests. 'Either we continue with your 'best of breed' approach to purchasing and deploying software based on an up-front need, and then try and bring it all together at a later stage, or we try and deal with things top-down. Taking this approach means taking a step back to look at existing systems and data management practices enterprise-wide, to understand whether any new systems can feasibly be integrated into that existing framework without impacting on the ability to manage and control new and existing data, and its associated metadata.'

BSSN Software has developed a two-part approach to data husbanding and control. The concept involves creating a data lake as a structured file store that contains all of the key data, in parallel with a metadata repository that contains pointers to all of the key and relational data and signposts to where that data resides. 'In this way users can easily find and extract the information they need, and they can search through all of the data and metadata using key fields that will help guide them through associated information, such as how the study was carried out, by whom and using which instruments, and indicate where they can find additional underlying data.'

Navigating large-scale data ecosystems

It's a user-friendly way of navigating large-scale data ecosystems without trying to squeeze everything into one place, Schaefer explained. Try and do that and you end up storing representations of results, rather than data in its original form. 'If you stick with a centralised approach, then every data format that you have in a central repository constitutes a liability. That's because you need to put an infrastructure in place that can read and present that data to the user in a human-friendly format. Whatever format the file is in will require the software that can decode it. This can become costly and may engender access problems. And if you aggregate your data to reduce complexity, it will hold less value because you have no way of looking at it in its native format.'

Standardised communication languages for instrument interfacing, such as SiLA (Standardisation in Lab Automation), and data format standards, such as AnIML (Analytical Information Markup Language), reduce ambiguity, and ensure that all data is usable, irrespective of where or how it was derived, Schaefer notes. 'What people typically have to consider is, what will it cost me to bring that data into this open format, and what does it cost me to get that data out again, in front of the scientists or decision makers who need to use it?'

A global philosophy of data control

AnIML is an ideal format because it's XML-based, and so immediately human

WhiteMocca/Shutterstock.com

"Users can easily find and extract the information they need, and they can search through all of the data and metadata using key fields that will help guide them through associated information"

readable. 'Adopt AnIML and even tools that haven't been built specifically to support the format will be able to work with AnIML, as long as they support XML. The XML ecosystem now includes possibly thousands of relevant tools, and this will then drive down the cost of both data access and control, because it means you don't have to custom-build everything for reading your data from scratch. For the user, this translates to not having to turn every search into a major IT project.'

AnIML ticks all the boxes in terms of data accessibility, and also fits in with the principals of FAIR data, Schaefer said. 'If you can fulfil these basic principles and address accessibility and reusability, then you are working towards a more global philosophy of data control.'

That ability to access, search and understand data in context can be particularly important when something goes wrong and you want to find out why. For example, if there is a purity problem on a production line, Schaefer suggests. Having immediate access to all the data linked with the affected batches can speed identification of where the problem may have arisen, and how to deal with it. 'You can then compare every factor, including operating parameters, sources of materials, instruments used and operators working those machines, between batches, to identify where the problem may have arisen.' This isn't a validated system, Schaefer stresses.

'But what it does let you do is feed GMP data in and use it to compare with other data. It helps you see the bigger picture. What you can't do is make GMP decisions without a validated process. The area of data analytics and regulated environments is another topic – and something that people tend to shy away from because there are few tools and processes.'

Putting in place an infrastructure that will give an organisation control and access to a complete breadth and depth of data will almost inevitably mean working with legacy systems and legacy data, Schaefer acknowledges.

Proof in the field

Even today, companies' ability to find and have confidence in a platform that will facilitate data control without hindering permission-based access at a granular level is held back, so that it can take five to ten years for software – and particularly for those platforms that have to accomplish such a lot with vast amounts of data – to prove themselves in the field, suggests Jeff Carter, co-founder and COO at Arxspan, which was acquired by Bruker in March. 'The accumulation of all that data makes performance and responsiveness a real challenge, and companies want proof that a platform can cope with today's data, and also equally manage accumulated data over subsequent years, whether from existing or new sources.'

Yet modern-day technology runs in relatively short development cycles, whereas the pharma industry, for example, runs in really long cycles, Carter suggested, so that in five or so years technology can almost become obsolete. We just have to look at mobile phone technology to appreciate that speed of development. 'While pharma may not be ready to adopt platforms that are being released today for another five or more years down the line – when they've been proven – in that timeframe technology may already have moved through two development cycles and brought out new generations of software.'

But it is doable, Carter notes. 'You just have to look at Google and Facebook.' The differentiating factor is that Google and Facebook can build their own hardware from the ground up and run their own, massive data centres. That's not part of what pharma wants to have to manage routinely. Rather, pharma companies are increasingly looking to get out of having to run data centres, and use one of the big data cloud hosts.'

An issue of vision and foresight

Carter maintains that implementing a sustainable infrastructure for managing and controlling data may thus not be so much an informatics technology issue, as it is an issue of vision and foresight. Companies, labs and individual scientists

→

→ are under pressure to complete short-term projects, rather than think about the overarching business objectives, and 'what it is we are trying to solve in the long-term,' Carter adds, concurring with Schaefer.

Organisations should try to turn their thinking around and, instead of focusing on a short-term answer to an immediate question, look more globally at the overall problem, what are the available solutions, how long will they take to implement, and how that can be worked into a business model. 'It's a case of letting the timelines be dictated by the solution, rather than the solution being dictated by the timeline.'

Shifting regulatory sands

The constantly changing regulatory landscape is also dictating the direction of platform development. There are tools emerging now that can overlay data and ensure it complies with regulations when applied in a regulated context, Carter notes. 'The vision is that we will be able to separate regulatory elements of data management from the technological aspects of data collection, storage, interrogation and analysis. Companies like Arxspan want to be able to present the ideal software capabilities, user interfaces and security for our customers, knowing that there are third-party platforms that can be layered on top of these solutions to manage regulatory compliance aspects to data utility and control.' One example, Carter notes, is Tranquil Data, a 2018 venture-backed startup that is developing software that is claimed to help firms transform and scale by addressing the challenge of implementing transparent methods for governing how data is used.

Whatever the approach to data control – top-down or bottom up in Schaefer's words – solutions will inevitably be cloud-based, suggests Carter. Trying to shoehorn data into a legacy SDMS system doesn't make sense, nor does moving it in this form into the cloud, which may not be feasible. 'As one consultant in the SDMS space pointed out to me recently, this approach is effectively just transferring your data control issues from inside your data centre, to outside of your data centre, which can be cost-prohibitive. It also doesn't address another key issue, which isn't so much about where to put it, but how to put it there. Legacy systems may commonly limit the ability of administrators to put data in the cloud in the most cost-effective way.'

Making data intimately accessible while still under control will always be linked with the ability to keep that data secure, Carter notes. 'Most commercial data systems operate at record level security, but the Arxspan ELN and suite of cloud-hosted registration, inventory and assay management tools has been developed as a fully integrated data management and search platform that addresses security at the field level, and so provides an extra layer of confidence for collaborative research.'

Sakkmesterke/Shutterstock.com

> "What will it cost me to bring that data into this open format, and what does it cost me to get that data out again in front of the scientists or decision makers who need it"

Security at the most granular level

What that means is that the system gives customers the option to set in place security permissions and access control at the most fundamental level, he continued. 'By allowing view, and edit privileges at a field level, multiple scientists can collaborate on a scientific process or workflow without exposing sensitive information. Field security works with field state to limit or expose access to data in a just-in-time fashion for a user.' Organising data security at this level takes a lot more thought to instigate and maintain, but can save problems downstream, Carter notes. 'We offer that level of security, but the discipline of maintaining that security level and understanding how to document data is critical.'

The Arxspan platform has a number of differentiating features that give customers better control of their data and how that data is used, he claims. 'Our architecture gives users the ability to run proprietary algorithms on-premise and still interface with the cloud solution, while the architecture's RESTful API set give companies complete flexibility to carry out functions such as creating new forms, using their own tools.'

Two key benefits are scaling of individual processes and future-proofing of the system. 'Managing the system by API means Arxspan can monitor which processes are under high demand and scale the resources for these processes, to improve end-user performance and eliminate potential system failures. Using RESTful API for the platform allows new capabilities to be built and plugged into the system, with limited impact on platform operation.'

This allows for more dynamic updates to individual components of the system, without delays that might be caused by a major release cycle, Carter notes. 'Customers can also use the API hooks to write, maintain and host proprietary web services in their own data centre or virtual private cloud, and still have the processes executed as part of normal user experience.

Synchronous / asynchronous processing

Uniquely, Arxspan architecture allows users to separate operations into synchronous processing – in which case the operation must finish before the user has a response – and asynchronous processing, which can include long-running processes such as analytics or data warehouse population. 'This allows for optimisation of the end-user experience, while still providing API hooks that allow longer-running activities without negatively impacting the user,' Carter said.