INFORMATICS
the same old problems, just bigger

Today's information technologies profoundly affect the way we conduct research and development (R&D), and the last 20 years have seen great advances in information technology across both our work and home lives. From representing complex material structures to rapidly sequencing and annotating genomes, informatics continues to revolutionise the R&D arena.

By Glyn Williams and Dr Paul Denny-Gouldson

Informatics are used to drive productivity, efficiency, quality and innovation, and now they are set to help provide insights from oceans of disparate data – 'Big Data'.

However, it is not just the data – social media and mobile devices have transformed communications, data acquisition and information and data sharing. The mobile devices we carry at all times give us unprecedented access to information in real-time. Booking a restaurant, setting your TV, finding a car mechanic or calling a plumber can all be done in seconds. It seems in our working lives, we don't always enjoy the same level of access to information to help us make the correct decisions faster and be more productive. But things are changing, and the tools and way we use them in our home lives are starting to be transferred to our working lives.

This article looks at the data and informatics landscape in R&D and considers what and how these developments might impact the way we work and utilise data over the coming years.

Look back to go forward

To see how the future may look, it is often useful to reflect back on where we have been. Twenty years ago the pharmaceutical R&D IT landscape was relatively simple and mainly centred around cheminformatics. There were tools for drawing chemical structures, databases for registering those structures and systems for storing associated data. The challenges then were accessing data and how to cope with increasing data volumes.

Most processes were analogue, not digital, and the majority of decision-making happened in the labs or at project meetings around an overhead projector using acetate handwritten 'slides'. Fortunately, project teams and individual scientists were dealing with relatively small numbers of samples so it was relatively easy for team leaders to keep much of the data in their heads. Most of IT was writing large systems for corporate data or bespoke data warehouses. Local IT groups were populated by 'scientist developers' creating applications to solve point problems or to enable new techniques. For example, parallel synthesis (or combinatorial chemistry, as it was known then) on a worthwhile scale was not possible without software or for that matter automated preparation and storage. The same was true of many in-house High Throughput Screening (HTS) data systems.

Some things have moved a long way since those early days – but others not so much. In HTS and combinatorial chemistry, massive industrial change

was driven by instrument technology enabling everyone to create higher volumes of more complex data. This was supported by new software which was developed to analyse the data more creatively and quickly.
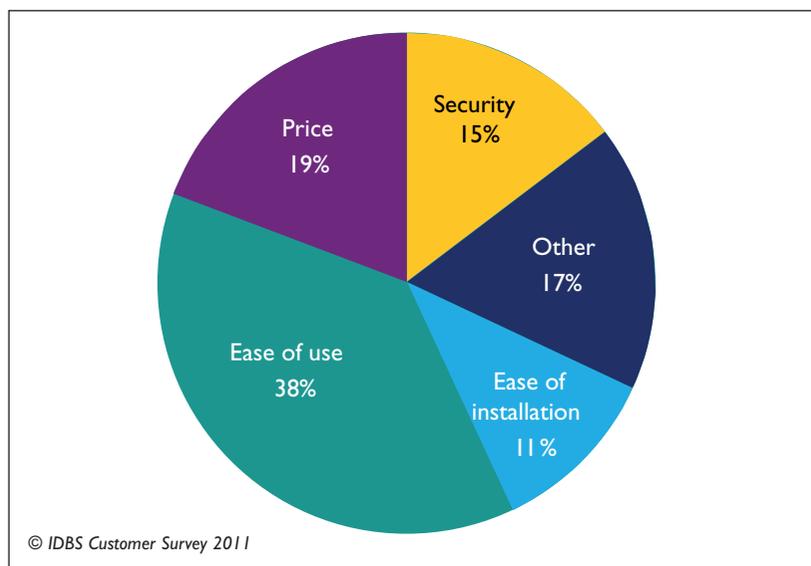
Here's how the HTS story went: advances in hardware and reagents drove increased throughput; these became so high that no-one could eyeball every result, so the industry began to employ techniques to analyse data quality (z' analysis of control deviation) and select hits (robust statistics of control and test populations).What was needed then was software to collect and perform the analysis. Then we needed new ways to visualise and consume the data that had been analysed. 96-well plates became 384 which became 1534, HTS became ultra-HTS and robotics drove the wheel even faster. Allied to this was the logistical need to order and create test plates ready for the robots, which became inventory-hungry monsters. All this drove the ability to automate by making the instruments talk to the software, and we started to shift the dynamic from having isolated pockets of data to a more joined-up world.

The human IT landscape has also changed dramatically. Pharmaceutical companies now have relatively few developers compared with the past and the modern philosophy is to buy as opposed to build. This is also helped by the number of software providers in this space.

Bigger and more varied challenges

The information challenges facing R&D now are still there but are far bigger and more varied. Data volumes and the rate they are produced are still increasing as in the days of HTS, and although now we call it Big Data, it is all relative. Yesterday's Big Data was a challenge because we were limited by processor power and storage capacity. We can now compute and store much more than we could have dreamed about in the pre-internet days but the same old problem still exists. Irrespective of the amount of data being generated, how do we make sense of it all?

We currently create more data than we can consume, and this is only set to increase. We need ways of triaging the information we receive to ensure what we get is what we need when we need it. In addition, high context data is essential to building a valuable knowledge ecosystem. It must be stored alongside its ontology and provenance. This is what enables it to be compared and used effectively, weighted against other competing data properly and quality controlled.

© IDBS Customer Survey 2011

Relative importance of attributes when selecting R&D informatics software

In science a lot of what we do is complex – it's a fact

Today we can do so much more. It is possible to have reams of information at our fingertips but this can easily lead to complexity and we can lose ourselves in a maelstrom of data. Software and infrastructure can help to control and avoid such confusion and help us better manage routine complexity.

There is also complexity in the number of different systems an organisation needs to use. These systems are useful in their own right, but when we can share and link data between different systems then the potential usefulness increases exponentially. This is particularly the case if systems can talk to each other in a way that enables you to stay in your application of choice and bring in data from other systems as and when needed.

Take the analogy of the iPhone. It is a telephone which also has a camera. Dependant on what people do, they may only require this camera and never use one that is 'phone-free'. The purist, however, wouldn't take a prized family portrait with their phone and, when buying their next high end SLR, wouldn't think of asking the salesperson if it comes with a jazzy MP3 ringtone. At the end of the day, it comes down to which tools you need to do the job you want to do.

The iPhone example has similarities to the R&D informatics arena. With the increased amount of software available, the lines between system capabilities have become blurred and the acronyms ever less relevant. Does anybody get really excited about whether they are using an ERP, PLM, ELN, LIMS, SDMS, MES, LES, CDS, CTMS, LIS,

Chemistry Registration or Biological Registration – so long as they can get their R&D done quickly, correctly and without a shelf-full of manuals?

However, the competition between vendors and the enterprise nature of some of these systems can create complications. For example, Laboratory Information Management Systems (LIMS) can be generic, specific or both, and Electronic Laboratory Notebooks (ELN) can integrate with LIMS or have LIMS features and can overlap with other acronymic systems such as SDMS, LES and CDS. This is often harder than it should be, as we all love to classify things – particularly as scientists – thanks Darwin! It is a good thing that the consumer phone world is not quite as confusing.

The underlying essence has to be: what do we need the software to do? It is very easy when you are looking for systems to do specific tasks, but when you are looking across the enterprise you have many more people to please. Compromise is essential but does that lead to a sub-optimal system? Possibly, but the Total Cost of Ownership (TCO) of a compromise solution can often outweigh the lower Return on Investment (ROI) for individual functions or groups.

Opting for a platform approach makes a lot of sense, but as we mention above it is vital that platforms can link and share data with other systems. Most modern software therefore now comes with comprehensive Application Programming Interfaces (APIs) and web services (RESTful and SOAP depending on the nature of the integration), and if they don't they should. These allow integration and configuration to cater for those specific edge cases and specialist systems to be integrated where applicable.

How can we bring our home life experiences into the lab?

In science a lot of what is done is complex, but we now have an expectation that these complex things can be done simply – search the web and find every place where the phrase 'regenerative medicine' exists in any language – and give me the results in less than a second and show me the context too and how it links to other things!

This expectation has been steadily rising and, having been conditioned by the likes of the iPhone, we now believe that we should be able to do everything from phone calls, email, photos, speed-dating and internet access all with the swipe of a finger.

Scientists are getting social too!

However, while most of these applications help us consume and use data really well, they only offer us extremely simplistic ways to capture and secure data and information.

Pick up the phone and ask for help – you must be joking!

Another factor in the expectation of users is that new technology has subtly changed the way we communicate. One of the important factors in the data-information-knowledge paradigm is how information flows. Applications such as Facebook have changed how information flows between people and how we share our information.

Recently I was sitting in a restaurant in Boston with colleagues and conversation got on to Fellini movies, so my colleague typed into Google 'La Dolce Vita' hoping to get a link to the film. However, the first page was filled with restaurants or establishments in the local area called La Dolce Vita. The name, which is pretty common in the North end of Boston, had to be prefixed with Fellini to get a more pertinent set of results. While it was impressive that Google knew where we were, due to GPS, it wrongly assumed what we were looking for. The phone was obviously not aware that we (well some of us) were a bunch of art house movie bores.

There is every expectation today that applications will 'know' such things by linking systems together – link your Facebook profile to your Google profile perhaps – or combining your newsfeeds to assess your interests or community. Similarly, the 'if you like this' feature in Amazon is another way in which we see things we may not have been looking for. What is of interest in our movie search story is that the software on the phone is using the data it has about where we are and modifying the search results accordingly. These 'm-apps' are already heavily used by retail marketers offering promotions and feedback from the stores you are walking towards or even out of, creating an 'always on' engagement with their customers. So, today's systems can estimate, based on algorithms, what we may be interested in rather than us having to look for it. They may not always be right, but by monitoring the choices we make, the systems can learn more about us and serve the data we need or in our personal lives – the adverts companies want us to see.

In essence the software deems what is appropriate or of particular interest to the phone's owner or user – herein lies another potential issue that is highlighted later. This is a great opportunity for these types of approaches in R&D, but the context

will need to be in terms of 'where you are scientifically in your work' not just where you are location wise. This is a far more complex problem to solve – a tractable one no doubt – for those clever people out there.

We all consume data and information differently

As we subtly change the type of data we can access, and subsequently how we communicate this then leads into changes in how we use the data. While the flow of information (communication) is important it is what people do with the information that potentially turns it into knowledge. As Peter Drucker said: "You can't manage knowledge. Knowledge is between two ears, and only between two ears."[1] Many of the social networking features we use help share information and make data more accessible, relevant, commentable and opinion-based. This is potentially very useful in the lab and in capturing the intellect of the R&D community. However, it is important that the right people are commenting or 'liking this' and have the context to make knowledgeable responses that lead to good decisions.

To take a non-science related example, through Facebook I learned that one of my friends was going to buy some medium density fibreboard (MDF), not something I would share publicly but there you go. What surprised me was the amount of people who commented on this, some with a semblance of knowledge and others without. What I now know is if I ever have the need to buy some MDF, I know who in my 'community' to ask. The

How we consume data is changing!

Informatics

References
1 Drucker, PF (1969). The age of discontinuity: guidelines to our changing society. New York, NY: Harper and Row.

likelihood is that I won't be buying MDF, which demonstrates the counterpoint to serving up knowledge in a work environment. You want people to have access to what is going to be useful to them to make the right decisions. They do not want to be bombarded with superfluous information that consumes their day.

Pharmaceutical companies employ clever people, with specific skills and domain expertise. If only two people out of 100 say they 'like this' on Facebook in our home lives, it may not be statistically valid based on a numbers game. But, in the lab it may be hugely scientifically valid if it is the right two people who comment. As well as providing productivity, efficiency and better context, integrated knowledge informatics could also bring people with complementary knowledge together. This is where the software having the capability to know what will be of interest, and more importantly, what won't become important. This is a fundamental requirement so people are spending more time on what will add value to their work.

Scientists and researchers have opinions about most things. While the social networking features may have many things to offer in the context of scientifically-based R&D, it will likely take some modification and control to apply them and to make them valuable tools in the day-to-day life of a scientist or engineer.

We are mobile scientists – but are we really touchy-feely scientists?

So, most of us have some device that allows us to be constantly connected to the world, and many companies see mobility and tablets as the way forward for their scientists. While there is no doubt that this has great promise, it is not necessarily the great panacea for science: well not yet anyway. These devices allow you to capture and consume very simple pieces of information.

The 'm-approach' consists of relatively small applications that talk to each other rather than an enterprise approach of large scale applications. This paradigm is effective when we may be just wanting to do specific tasks such as reviewing information, commenting, tagging or capturing a specific piece of data in a remote lab or place (eg balance reading). This can complement enterprise applications but does not necessarily replace them. We have already seen the use of print to/send to in most modern applications – now it needs to work with snippets of data rather than big lumps and be able to put it into the right place to make the information useful.

The opportunities are endless... so we need to be pragmatic

Undoubtedly we will see the rapid evolution of software and hardware usage in this space – look at how far we have already come in 20 years. Data-driven change is a fundamental part of life, both in the home and at work. It is a given for society and science to evolve.

Computer science has evolved to a point where we have the power and the accessibility to do most things we need and want to do. What is changing is the way we interact with hardware and software, and indeed information. It is a 'people thing'. It is how we access and consume information that is important. The information is out there and there is so much of it, that just asking the questions is not going to work, as this is based on the assumption that we are asking the right question in the first place. Therefore it is a combination of asking for the data we need, having relevant information served to us and enabling access to people with the right information or knowledge to lead us to better decisions. A great example of this stems from the personalised medicine world – where the vision is that based on data (your genome, proteome, family history, symptoms and lifestyle) preventative advice and specific treatments can be prescribed. This requires a whole host of data to come together, but the advances made already show its potential. Some cancer patients do not respond to certain drugs and this can be linked to the patient's genome or the cancer's genome – this is now known and used to help make clinical decisions in hospitals. Even the tools that remind you to take your medicines are changing outcomes.

Where will this all end up? Technology moves at such a pace but what is important is how it is applied in the context of what you are working on. NASA is looking at 3D printing to make pizzas in space. If they can do that then what does it mean for R&D companies making 'stuff'? Will it all be done on computers and 3D printers in the future? Maybe it is far-fetched, but then I never expected my computer to recommend 'Pretty Little Liars Season 3' to me (it doesn't know my daughter uses my Amazon account). Perhaps real time genetic fingerprinting does have its place! **DDW**

Glyn Williams is VP Product Delivery at IDBS.

Dr Paul Denny-Gouldson is VP of Translational Medicine at IDBS.