TONY DAVIES COLUMN

Back to the future: total spectroscopic laboratory informatics

Tony Davies
External Professor, University of Glamorgan, UK
c/o Waters Informatics, Europaallee 27–29, 50226 Frechen, Germany

A.N. Davies

A.M.C. Davies

The start of the fifth year of the new millennium is a good point to look back over the recent past and see if we can establish any trends since the merger of *Spectroscopy International* and *Spectroscopy World* to form the new *Spectroscopy Europe*—with the emergence of a two "Tony Davies" column authorship team.

In line with our briefs, Tony "AMC" Davies has continued to concentrate on educating and advising in the more data processing and chemometrics applications of our field where I, "AN", have tended to deal with the more data handling and campaigning aspects of the column.

Considering the resources available, we have achieved a lot with your help in only a few years. We have established the use of international data standards and their availability in commercial spectroscopic data systems as the norm. If you as customers hadn't demanded this in your purchase orders and if our discussions with the various instrument vendors hadn't been persuasive enough, then this would never have taken place, so well done to all sides!

Chemometrics—from black magic to drug release

The use of chemometrics has advanced from the statisticians "black-box" which would deliver fantastically good calibrations and classifications which no honest spectroscopist really believed were true—to one of the normal features in the spectroscopist's toolbox, integrated in most standard spectroscopy data handling and instrument control packages. Trust in chemometric methods has even advanced to the point where the American Food and Drug Administration (US-FDA), within the bounds of their Process Analytical Technology initiative, doesn't see a problem in allowing the risk-based, real-time, on-line release of pharmaceutical active agents to the market based on numbers generated from chemometrics packages.

Just mentioning the FDA brings us to a clear change in the way we have been working in recent years. When we started writing these columns we were always working towards a "dream" spectroscopic laboratory with data and information flowing easily between different analytical systems and reporting packages—building up huge knowledge bases to provide expert resources, making your work simpler and more effective. With our vendor community, we have pushed for and been happy to promote these tools and services as they have become available. Many industrial and academic spectroscopic laboratories changed their working practices because they thought these ideas were good in principle.

Regulations, rules and compliance

More recently, the different government regulatory bodies, principally but not solely in the USA, have been calling the tune to an ever-greater extent—their influence and power extending well beyond the boundaries of the national sovereignties. The need to prove compliance with rules governing electronic record data storage, including the need for ensuring data longevity, are now standard texts for the regulated industries. Even some national research councils have realised the benefits of such approaches and have started demanding long-term secure storage of research data from projects they sponsor as an additional weapon against academic fraud.

So as you can see, the topics we covered have matured from being novel, interesting and potentially beneficial ideas to the hard-core of the mainstream analytical spectroscopic workplace.

Computing power and infrastructure

Over the years we have been covering this field, there has of course been a revolution in the computing power we are able to command. Interestingly, in some aspects, the limits that were placed on us in the cost of data storage capacity and computational speed in the past have been pushed back into irrelevancy, only to re-appear in a somewhat different form in current times.

Whereas, in the past, a single 30 minute 5000 point Raman measurement might generate a 40 Kbyte JCAMP-DX data file, you could still store 30-plus measurements on the standard data exchange media, a 3½ inch floppy disk—equivalent to the production of two full days solid measurement time.

Nowadays, the most modern hyphenated techniques such as UPLC/MS can complete what used to be a 30-minute HPLC experiment in three to

TONY DAVIES COLUMN

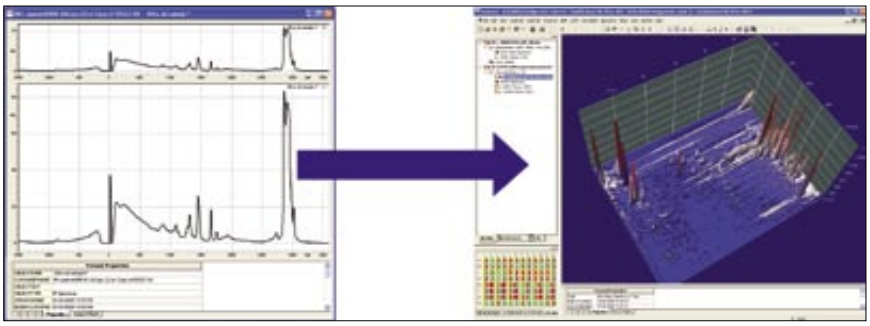

Figure 1. The data volume trend from single data sets of XY data pairs to hyphenated analytical methods with multiple spectroscopic detectors combined with chromatographic separation techniques measured by automated robotic systems from trays or racks.

four minutes, generating a two- or three-dimensional analytical data array of many hundreds of megabytes. This is often one measurement of a 96-well plate analysis containing blanks, repeats and reference samples for calibrations so that the entire output of a single days work on a single instrument could reach 40 GByte, at a time when the largest exchange media, a DVD, can accommodate only 10% of the daily data production volume. So, although we are happy to state that the cost of data storage and processing is no longer a barrier, we have succeeded in running up against other barriers.

The fastest networks currently found in most laboratories are 100 Mbit so-called Fast Ethernet, which are rarely dedicated to individual instruments. As long as the bandwidth is pretty much kept free within a particular network segment for the high data volume instruments' use, it should just about be able to cope with the load but does demand an efficient total informatics solution with networked storage area systems and secure archiving and retrieval if the best is to be made of the advanced spectrometer hardware now available. Simply buying another box of 3½ inch floppies for the week's work, unfortunately, is no longer enough!

Help is at hand. If you talk to your IT professionals, you will find that in many server rooms new standards have appeared for data transfer between high-load components—the so-called Gigabit LANs. These 1000 Megabit s^{-1} network communication technologies originally required the installation of optical fibre links but the newer Gigabit Ethernet standard 1000BaseT allows the use of traditional copper cabling. In fact the Gigabit Ethernet standard 1000BaseT has been specifically designed to be backward compatible, which should avoid having to re-cable your network. Where the 100 Mbit Fast Ethernet uses two pairs of cables, one for transmit and one for retrieve, the 1000BaseT uses all four pairs running at the same 125 Mbaud symbol rate but uses a more sophisticated coding scheme as well as sending and transmitting simultaneously on each pair.

You may well have come across the Gigabit Ethernet network cards as they are already being sold as standard components in many personal computers and even laptops on the market today.

Naturally to move up to these increased data transmission rates (1000BaseT ≈ 120 Mbyte s^{-1}) the appropriate network components must be installed all the way to your servers but at least the bottleneck will not be in the spectroscopy laboratory. There was even a white paper brought out as early as 1999, which has an interesting overview of the technology (see http://www.10gea.org/GEA_copper_0999_rev-wp.pdf).

Although you probably haven't had time to let the thought of 120 Mbytes s^{-1} sink in, it is worth at least knowing that 10 Gigabit Ethernet hardware is already available, although optical fibre based. Once you move to optical fibre systems, the introduction of multiple wavelength single fibre technologies adds further orders of magnitude to the transfer rates.

Reference data collections

One of the proudest moments I have had was when we switched the International Spectroscopic Database (IS-DB) systems live last year. For me, this was an excellent example of what the column and *Spectroscopy Europe*'s support can achieve by lobbying for advances and funding in areas where our readership, and the spectroscopic community in general, perceives a need. The long-term success of such systems relies completely on the willingness of the spectroscopic community to donate data. To date we have over 300 registered users and over 1600 electronic records collected. In addition, there are over 1300 spectra in chemometric data sets as well as pure compounds and structures. It will take time to generate a really useful resource but, the more data you are willing to donate, the faster we can grow and, the faster we grow, the more likely you will be able to find information which helps you in your day-to-day work!

Future directions

As with this column in the past, our success can only be measured by the responses we receive from our readership. So where do you want us to go? Are there any subject areas you would like to see covered which we have neglected in the past? Are there any topics that we ought to be covering in greater detail than we currently offer? Please let us know.

Either Ian Michael, myself or both of us will be attending various meeting in the next few months. If you are visiting Pittcon Exposition in Orlando between 28 February and 3 March come and let us know your needs—we can be contacted through the Wiley booth, #1784 (http://www.pittcon.org). We will also be available, as usual, at NIR-2005 held this year, 10–15 April, in Sky City, Auckland, New Zealand on the NIR Publications stand #10 (http://www.nir2005.com/) as well at the ASMS meeting, 5–9 June, in San Antonio, Texas (http://www.asms.org/Default.aspx?tabid=43).