MSVision

Dedicated to Mass Spectrometry

MS VISION TECH TIPPS SERIES
PART II — BACKUP OF DATA

How old are your current mass spectrometers?

1 year? 5 years? 10 years? Even older?

Do you know where the images from your installation are? And on which media?

Why am I asking this? Because the storage of data, whether installation data or measurements is a difficult topic in mass spec, particularly for smaller labs. Typical storage media have restricted lifetimes, not so much from the physical point of view but from the point that they lose the information stored upon them over time. Or the media don´t exist nowadays anymore. Two examples: when I made my PhD, data was typically stored on QIC tapes, 120 or 250 Mbytes. Endless space for data when using a Sun Sparc 5 workstation. I was very happy at the end of my PhD that I had no paperless office and printouts of all the spectra as the tapes were stored in a shelf in the mass spec lab. Where also two sector field instruments were located. Magnetic tapes and magnetic sector fields are an unholy combination! Almost all tapes were unreadable after 1-2 years… After my PhD I became head of a core facility lab and bought a modern QTOF system. These systems, already in the very beginning, were producing hundreds of Mbytes of data. Floppy disks were unable to handle the data generated by this, the hard disk drives had capacities in the range of 2-300 Mbyte, CD's were not suitable as well and DVD's or BluRay did not exist. At that time I decided for an Iomega Jazz-Drive for data backup. Worst decision I ever made… The media were extremely expensive per Mbyte and the product disappeared quickly.

Back then, it was seriously discussed in the community to make the data backup in a way that you have removable hard drive slots, push in a hard disk, fill it, put it into the shelf, fill the next hard disk and so on. For those which assume we are talking about stone age – no, this was around 2000, just 20 years ago.

Luckily, hard disk capacities have dramatically changed since then but the fundamental problem persists:

How to handle long time storage of data? In particular of a lot of data…

Fundamentally there are a few ways to address the problem:

- Don't care. Just fill the hard disk drive and delete old stuff once in a while. A fundamentally bad idea! Hard disks can crash (they typically have failure rates between 1 and 10%! For details see here: https://computerbase.de/2014-09/ausfallraten-von-38.000-festplatten-dokumentiert/) and all the data is gone. From your thesis, your co-workers thesis, your customers…

- USB sticks. USB sticks are nowadays available at significant sizes and could be used to save data as well. I personally use this for my laptop. I have a set of 10x 32Gbyte USB sticks and save all the important data every Friday completely on a stick. So the maximum I can lose is one weeks work. Additionally, we use net drives for specific data. For me this is a pretty safe strategy. Regarding the storage of the data, manufacturers typically specify 10 years lifetime for the data and 100.000 write cycles per cell. Thus, lifetime is limited as well, yet I will never reach the 10 years (every stick will be rewritten every 10 weeks) nor 100.000 cycles (100.000 cycles, 1 cycle per 10 weeks = 19.230 years…). And they are not susceptible to magnetic fields such as in FT-ICR or sector field instrument labs…

- Professional backup solutions. These can be very costly but provide utmost security.

- Backup on CD/DVD/BluRay. Forget CD when you think about LC-MS data from proteomics experiments. Not even a single run will fit onto it.

DVD or BluRay might work. But it's slow and much more important: the lifetime of self burned CD/DVD/BluRay is approximately 10 years only. Then you can expect a steep increase in the error rate. An alternative to DVD's are DVD-RAM drives and disks. These are comparable to DVD's (4.7 GByte per disk), but have a much longer MTBF (mean time between failure) in comparison to DVD's. However, they require special drives and again: will these still be available in 10 years at all? Nowadays they have been largely replaced by flash memory (USB sticks, SSD's) which hold up to several hundred GByte easily and have much higher read/write rates.

- RAID systems. RAID stands for Redundant Array of Independent Disks. These can be conventional hard disks or solid state disks (SSD). The basic idea of RAID systems is to ensure data security when one component fails. There are several levels of RAID systems. The most common ones nowadays are RAID 0, RAID 1 and RAID 5.

  All systems have in common that they work with several hard disks and may store the information in a redundant way. RAID systems are supported by all modern operation systems. Strictly speaking, RAID 0 is an exception as it uses multiple disks but no redundant information. But that way always only a part of the data can be lost (to bad when it was the important part). RAID 1 duplicates the entire data and spreads it over two or more disks. Thus, if one disk fails, the data will still be there on another one. RAID 5 uses the so called parity information to increase the security while also increasing storage capacity.

  These three levels differ by the degree of security as well as by the available storage capacity. Let's assume a system with 4 hard disk drives of 1 TB. Let's also assume that a single hard disk has an independent failure rate of 1% (if you would like to see some real data on this have a look at the link above):

RAID level	Storage capacity [TB]	Overall failure probability
0	4	3.94%
1	2	0.000001%
5	3	0.0592%

Thus RAID level 1 has the highest security but at the price of the lowest capacity. Level 5 is somewhere in between regarding capacity and significantly better than level 0 in security.

There is also an additional consideration: hard disks are not the only component which can fail in a computer (we will cover this in another article separately). Controllers can do as well. So, if the controller fails, access is also impossible. For this reason traditionally RAID 1 was subdivided into mirroring (multiple disks on one controller) and duplexing (each disk has it's own controller). However, failure rates of controllers are much lower than of hard disks, thus nowadays the most common mode of operation is mirroring.

But to make it clear: strictly speaking a RAID system or mirrored disks are no replacement for proper backup! A fire, flooding or a virus might attack the entire system still.

While I was researching for this article, I found the quote: "Only copied data is safe data." Thus a backup is something you should think about. Probably for a small to medium sized lab a server with RAID 5 is the best option in terms of performance and price for medium time data storage when it is decoupled from the internet and located in a separate room or building. But still, you should consider how to make a proper backup of your data. In academic institutions you should check with your central IT facilities. They might be able to provide support or have capacities available or are able to organize these for you.

Coming back to my entry question: how old is your mass spec? Traditionally, when the systems are installed, after the installation an image of the acquisition PC is burned and stored on a CD-ROM. As mentioned above, these have typical lifetimes of 10 years. Is your mass spec 10 years or even older? Are you starting to get nervous? Perhaps it's about time to make a copy… As only copied data is safe data!