Life and death of a data set: a forensic investigation

Antony N. Davies

SERC, Sustainable Environment Research Centre, Faculty of Computing, Engineering and Science, University of South Wales, UK

I recently reflected that life can be extremely unFAIR, especially if you are a spectrum. I had looked down on the body of this dead spectrometer and reflected how many keen enthusiastic young researchers had tussled with the complexities of what had once been a state-of-the-art scientific wonder which was now reduced to a problem of recycling, disposal and potential contamination risk.

How many now established scientists had benefitted from the children of this spectrometer, the excellent spectroscopic data sets that it was capable of generating in its heyday. They had crafted from these data sets—sometimes with a little help from their supervisors (and no doubt sometimes subtle data processing), their theses and early publications—the passports to their now established careers. But where is all that data now? (Figure 1).

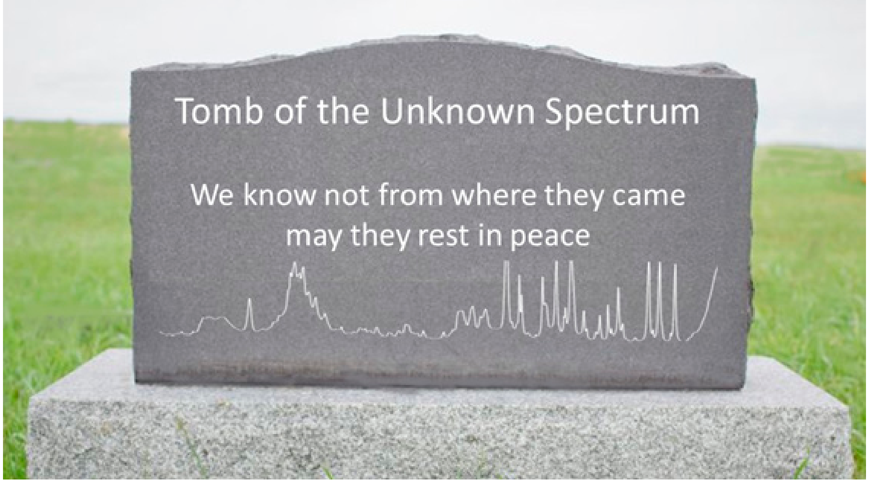
Tomb of the Unknown Spectrum
We know not from where they came
may they rest in peace

Figure 1. There are many ways a spectroscopic data set can "die".

Oddly enough we were lucky enough to attend a Rick Wakeman concert in London which included a very well received rendition of his famous "Six Wives of Henry the Eighth". This reminded me of my daughters' favourite mnemonic chant of "*Divorced*, *Beheaded*, *Died*, *Divorced*, *Beheaded*, *Survived*" normally used to remember the fate of Henry VIII's six queens—Catherine of Aragon, Anne Boleyn, Jane Seymour, Anne of Cleves, Catherine Howard and Katherine Parr. But as you all know conversations sometimes take bizarre turns, this same mnemonic suddenly seemed very relevant to the various common fates of our spectrometer's data sets.

DOI: 10.1255/sew.2023.a3

Figure 2. Does this remind you of your research supervisor? If so, look out! Photo 173782772 © Spiroview Inc. | Dreamstime.com

Birth of a spectroscopic data set

As originally delivered the spectrometer was capable of generating some of the best data sets we had ever been able to measure. Excellent signal-to-noise ratios and extremely stable calibrations. The associated computer hardware was somewhat behind the state-of-the-art, but this was quite normal due to the long development times for the instrument hardware. And so, in the hands of many expert, and some less expert, scientists, this wonderful spectrometer gave birth to many spectroscopic data sets—reinforcing theories and dispelling some myths. Over its lifespan serving as an excellent measurement platform allowing many modifications to the original basic equipment. So, while our data was young there were no problems. All the measurement parameters were stored with the data set so we could check the instrument had been set up correctly. Where the appropriate comments had been written to the data file, we could even see how the background compensation had been carried out. However, life became more complicated when more advanced data processing was required which was not possible on the original spectrometer control computer.

Divorced

Early attempts to carry out more adventurous data processing required the data to be moved off the spectrometer computer. Where we were lucky, there was a data export function built into the spectrometer software and we were able to get at least the X-Y data points across on to the second computer. Even where there was a data exchange format deployed to the spectrometer such as the JCAMP-DX standard, we did not get all the information onto the new computer. Since the standard only required the minimum amount of information required for accurate data processing to be exported as a requirement to be compliant. All the additional metadata from the originating spectrometer system could be transferred in a defined compliant manner, but only if the vendors believed it was a good idea. In most cases this left the spectra data divorced from much of the metadata about how the spectra had been originally measured and with the death of the spectrometer any hope of recovery from this divorce was gone.

Beheaded

We were successful in setting up workflows to transfer the children off the birthing spectrometer and on to computer hardware capable of more advanced data processing. However, most of the commercial data processing packages were, and still are, not capable of maintaining the integrity of the spectroscopic data set during the import into, for example, a chemometrics data package. Even where the chemometrics software vendor has implemented file filters for direct import from the spectrometer's native file format the header information is very frequently left behind. So, the spectra are essentially beheaded of their supporting metadata.

Died

Often the divorced or beheaded children of the spectrometer were the lucky ones. Many precious spectra have died, not only when the instrument which measured them was retired, but also when the vendor enforced a systems upgrade—sometime completely replacing the control computer for one working on a completely different operating system. This meant the original native binary format files of the earlier work were no longer readable. On one instrument an early attempt to meet FDA guidelines on data integrity, we experienced original data files being embedded within another binary wrapper to provide electronic signatures capability proving no data manipulation had taken place. The only issue there was that the fully validated FDA-compliant ready data migration software for this spectrometer type knew nothing of the home-made additional wrappers and simply failed to read the spectra. In this specific case we were fortunate enough to still have access to some of the IT team who has dreamt up this “cost saving” one-off solution and could reanimate the “Dead” spectra with probably more effort than had been originally expended to measure them!

Divorced

Now looking further down the publication and exploitation pathway of the data, we can see that some journals are providing authors with the opportunity to supply the relevant spectroscopic data along with the submitted manuscripts. These ground-breaking publications have unfortunately little or no guidance on how the spectroscopic data is to be presented or uploaded. Indeed, Professor Robert Lancashire recently came across some guidance by one journal that limited the amount of data that could be deposited to a few MB.

During a recent meeting of the IUPAC FAIRSpec project team, they also suggested that the enforced limited upload of only a few selected example spectra (usually the best measured rather than “typical” for any given experiment) could also fall under the category of Divorced or even having suffered the medieval torture/execution category of Dismemberment of the complete experimental data set with the accompanying critical loss of context.[1]

Beheaded

In the second Beheaded category I have decided to mention an example of users or vendors carrying out what initially looks like a useful fix to a software issue which has unexpected consequences. Having worked in the spectrometer software industry I have experienced the pain of salespeople selling software features which only exist on paper or maybe have only been discussed in long-term planning meetings and haven't yet reached the stage of being put down on paper. If the sales effort is successful these projects usually result in very rushed implementations of the absolute minimal number of feature improvements to meet the contract obligations. This is never a great way to develop robust software and often suffers from the law of unexpected consequences! Fortunately, this is not the norm.

One example, not from anyone I have worked for, involved an instrumentation engineer who was having difficulty fixing some relatively minor problems with an instrument at an important customer site. They decided to try using a version of the spectrometer software that they had been given to test and potentially glean customer feedback. However, the new version was never intended to get in the hands of real, live customers! Somewhat surprisingly this fixed the specific instrument problem the key account was having, and all were happy… for a while.

When the next regular update was due the key account customer obviously received the upgrade and were horrified to find that all the spectroscopic data they had measured in the last twelve months would no longer load. Unfortunately for the engineer and the important customer, the unreleased review copy of the software had also included an experimental, innovative new data storage concept which had failed at the pre-release testing phase and the vendor had reverted to the older tried and trusted storage file format. This truly beheaded the archived data and again cost a significant amount of money to recover from. I am not even going to attempt to discuss the compliance and data integrity issues this sort of mistake raises.

Survived

So how can the children of our spectrometer emulate Katherine Parr and survive all the potential pitfalls in their expected lifetimes? Well, many of the answers still lie in the FAIR principles and how to implement them in an analytical laboratory. One of the starting points would certainly be to assign all spectroscopic data sets a persistent unique identifier at birth [*FAIR Principle F1. (Meta)data are assigned a globally unique and persistent identifier and F3. Metadata clearly and explicitly include the identifier of the data they describe*]. This would make life much easier for researchers, supervisors, principal investigators, publishers and regulators alike. Even if, during the lifetime of the data set it, was separated from some of the critical metadata, so long as the unique identifier remained intact data archaeologists could always reinstate relevant metadata.

Now it may be wishful thinking at present, but if we look at many of the issues highlighted in the column above, keeping the metadata available would solve many of the problems commonly associated with the premature death of a data set—Accessibility Principal *A2. Metadata are accessible, even when the data are no longer available*. If correctly implemented would mean that it would always be possible to envisage metadata reinstatement as discussed above.

During a dataset's lifetime it will pass through many different software systems and rather than each stripping away metadata that it does not require for its specific operation—such as chemometric software only importing X-Y data pairs and ignoring the rest of the metadata—it should be possible that the original metadata is preserved with its own provenance and new processing should only add new metadata fully describing, maybe like a compliant audit trail, the actions that have been undertaken with the spectroscopic data. In this way the ability of subsequent researchers to reproduce a piece of scientific work published in the literature will be enhanced. We all need to stand on the shoulders of giants as Newton's famous metaphor confesses,[2] but there is still far too much published which belongs in the *Journal of Irreproducible Results*!

Finally, it is obvious that all we have discussed above, into the forensic investigation of the premature demise of spectroscopic data sets, revolves around misuse, mis-design, mis-deployment and mishandling of spectroscopic and general scientific data handling software. So, it was interesting to see the publication of FAIR principles for Scientific Software (FAIR4RS) at the end of 2022.[3] Unfortunately, there were no data sets or software code published with the paper although it does cite three software examples claiming to follow the FAIR4RS principles which could be worth following should you be interested.[3]

Where better to end than a quote purported to be from the great king himself, "*Of all losses, time is the most irrecuperable for it can never be redeemed*". So, let's not waste time generating spectra with artificially short lifetimes, let's get this FAIRification of scientific data done!

References

1. https://www.oxfordlearnersdictionaries.com/definition/english/dismemberment
2. I. Newton, *Letter from Sir Isaac Newton to Robert Hooke*. Historical Society of Pennsylvania. https://digitallibrary.hsp.org/index.php/Detail/objects/9792
3. M. Barker, N.P. Chue Hong, D.S. Katz, A.-L. Lamprecht, C. Martinez-Ortiz, F. Psomopoulos, J. Harrow, L.J. Castro, M. Gruenpeter, P.A. Martinez and T. Honeyman, "Introducing the FAIR Principles for research software", *Sci. Data* **9,** 622 (2022). https://doi.org/10.1038/s41597-022-01710-x

Tony Davies is a long-standing *Spectroscopy Europe* column editor and recognised thought leader on standardisation and regulatory compliance with a foot in both industrial and academic camps. He spent most of his working life in Germany and the Netherlands, most recently as Lead Scientist, Strategic Research Group – Measurement and Analytical Science at AkzoNobel/Nouryon Chemicals BV in the Netherlands. A strong advocate of the correct use of Open Innovation.

https://orcid.org/0000-0002-3119-4202

antony.n.davies@gmail.com