Published: March 2012

The growing problem of data mortality

John Hargrove - Senior Research Fellow at SACEMA

Three of the articles in the current issue deal with the aggressive use of antiretroviral therapy (ART) as a means of reducing HIV incidence, and a fourth considers the problematic business of estimating HIV incidence. These matters put me in mind of several incidents that I have noted over the past 40-plus years of my scientific studies.

In the mid-1960s while at Pembroke College (Oxford) I used to spend time, rather too infrequently I confess, in the college library. It was a small space, made smaller by the fact that the bottom floor could not be used, for the good and sufficient reason that it was so full of ancient tomes that there was no room for people. And there was no way, either, of accessing any book further than one foot from the door. Some years later a wealthy American made a large bequest to Pembroke and a huge new library was built and suddenly there was room not only to store the old material but also for scholars to research material going back many centuries.

The problem of space to store, and also have accessible, hard-copy data has always been a problem. But as long as the data *are* stored there can come a time when they can be retrieved and used. Of course things don't always turn out that way …. In the 1970s, while working on tsetse flies in Rhodesia, we had occasion to move offices and questions arose about what to do with the mountains of old field records of fly catches, and laboratory records of blood meal analyses of what the flies had been feeding on. There was no sensible place to store these records and I confess to being party to the young and foolish decision that the methods used to collect the data were sufficiently spurious that the data were of dubious value: and we burned the lot. And I then almost immediately realised that I really needed those data. It was a harsh lesson, which has influenced my attitude to data ever since.

Of course starting in the 1980s it seemed that all of these problems had been solved. Suddenly we had computers with tape storage facilities, and then floppy disks, then "stiffies" (dread word) and CDs and then the web with gigabytes, terabytes and now, I see from Google, even yottabytes of storage space. Almost infinite amounts of data, it seemed, could be stored safely, for ever, and at no cost. Aaah, yes. Except that the tapes full of data that I brought back from California in the early 80s were unreadable only five years later, and the second generation of desk-top computers could no longer read floppies, and the next couldn't read stiffies, and even before increasing numbers of brave new storage devices became obsolete they, unaccountably, became corrupted. But never mind, onward and upward, we had hard disks that could hold 20K of data – imagine! And government departments with too much paper to store, copied the data onto their computers – and then followed my example and burned the paper.

Fires can of course happen by accident, but "book burning" is generally a conscious decision. To destroy all of the valuable data in the old Pembroke Library somebody would have needed to physically move all the books out of the basement, make the pile and light the mach. The destruction of data on computers, however, can happen much more insidiously. The person who collected the data moves on; the next person who takes the job has no idea what this, essentially "invisible", material is all about and either deletes it all at a single key-stroke or, more likely, doesn't think to make copies when the computer becomes obsolete and is tossed out.

Even the most modern web sites, which are in some senses independent of the individual hardware on which we collect the data, need to be maintained. Somebody has to take care of the site, somebody must pay, somebody must be paid, or the material simply vanishes – with no smoke and no flame to warn us of the crime that has just been committed. Perhaps in a matter of days, or months, or years, or perhaps, if we are lucky, a few decades later the material will be gone for ever. It seems we have no "Pembroke basement" where we can leave modern information, safe and undisturbed for centuries.

These are not simply theoretical worries. Certain death records in our own country were "computerised" some years back and the hard copies were about to be destroyed when a vigilant hoarder asked if she might take them. The custodians of the material perhaps thought her a little quaint but agreed to her request. Not many months later they came back to her, cap in hand, to ask if they might consult the hard copies. It turned out that somebody had hit the wrong key on a computer and all of the carefully entered electronic data had gone up in metaphorical smoke. In at least one neighbouring country the outcome was not so happy; when the computer records went that was the end and there is now no record of death rates in anything other than the immediate past.

This is one facet of the problem of information retention. There is another, and even more worrying, side to the problem and this concerns the collection of information which is simply never used. Such information might as well have been burned or, in fact, never collected since data collection is generally an even more expensive undertaking than data storage.

And this problem brings us back to the matter of ART and HIV incidence. The essential idea of Treatment as Prevention (TasP) is that, in settings where we have a high HIV prevalence, treatment should start immediately for those people who are HIV positive or, in a variant of the idea, should be offered as a prophylactic to those at highest perceived risk even before they become HIV positive.

The idea of all of this is to decrease HIV incidence. The question is how should we measure this incidence; how should we decide whether the increased use of TasP is indeed resulting in decreased HIV incidence? To this end there have been strenuous efforts made in recent years to come up with fancy ways of estimating HIV incidence from the more easily estimated prevalence levels, or death rates, or using bio-markers. In almost every case when workers have presented such approaches there has been a nod to the fact that there is a "gold standard" of incidence estimation which involves the follow-up of cohorts of individuals. Whereas this approach carries with it certain biases and assumptions, the major reasons given for not using the approach more frequently are generally cost and logistical difficulties.

But the common thread in the two variants of the TasP approach is that they will require vastly increased levels, and increased regularity, of HIV testing. This is already happening to a huge extent in countries such as Botswana, where persons attending clinics are tested on an opt-out basis. In other words, as part of the TasP process, it will be *necessary* to gather the very follow-up HIV testing history required for incidence estimation using the gold standard approach. Unfortunately nobody (yet) seems to be collating and using these data to see what incidence estimates emerge.

There will of course be problems because the samples of people are not randomly selected. On the other hand, the whole target of TasP, as applied in high prevalence settings, is that there should be an effort to test as large a proportion of the sexually active population as possible. The more closely this ideal is approached the better should be the resulting HIV incidence estimates. For example, if time-series information is available on the HIV status of every pregnant or postpartum woman every time she attends a pre- or post-natal clinic it would at least be possible to estimate accurately the incidence among pregnant and postpartum women at close to a population level.

What stands in the way of making sensible use of such data is the current obsession with secrecy when it comes to the collection of data on HIV, unlike the results on any other disease. Nonetheless it should be possible to devise ways whereby, without compromising the patient, serial HIV status records from individuals can be linked together and used to estimate HIV incidence.

If we can't link the records in this way then we might as well burn them, or bury them in the basement of the Pembroke library.

John Hargrove, Senior Research Fellow at SACEMA. Areas of interest: tsetse and trypanosomiasis biology and control; estimating HIV incidence. *jhargrove@sun.ac.za*