STEWARDSHIP OF SCIENTIFIC DATA

Photo by UncaughtException via flickr.

Science magazine has an interesting special issue on scientific data, covering a variety of topics from data backup and data visualization to open data. It seems these contributions are free to access for registered user of their web site, and it certainly is worthwhile to have a look.

The editorial in particular lays out Science's policy on open data. Sharing scientific results is of course a motivation for publishing a paper in the first place. And to allow for independent verification of scientific results, the data contained in a publication has to be available and shared with other scientists. This sharing has to be done in a permanent way that guarantees access to archives also in future.

Is the data analysis traceable?

However, there is another point that hasn't come across that strong from this special issue, but one that I also consider to be very important. And that is that data processing itself needs to be tracked, by which I mean the steps from the raw scientific data as measured all the way to the plots in a scientific paper need to be traceable.

Logging data manipulation is important, not only to prevent fraud but also to re-analyse the data if needed, to uncover errors in the analysis for example. That involves of course that custom computer code is preserved. But it also means that any significant temporary data generated during the analysis is preserved. Much in the same way that any edits in Wikipedia can be tracked back step by step.

Practical issues

This kind of data archival, from preserving soft and hardware to multiple versions of large data sets, is probably something that can easily go beyond the capabilities of smaller research groups, and much more needs to be discussed how such archiving of data (and sharing it) can be facilitated. This raises questions such as whether commercial lab software could achieve this task, or whether central facilities on a university or a national level are better suited for this. And of course, also what the role of journals should be in that process.

Judging from my experience visiting research labs, data safeguarding and archiving might be on many researcher's minds, but clearly there is still a long way to go until we really are able to establish the kind of curation of scientific data that is fit for purpose. Putting the issue on the agenda is certainly an important step.

Source: http://allthatmatters.heber.org/2011/02/11/stewardship-of-scientific-data/