EDITORIAL

nature cell biology

Accurately reporting research

The cell biology literature contains manipulated data that distort findings, usually in an attempt to 'beautify' and, rarely, to commit fraud. A new National Academy of Sciences (NAS) report considers data integrity, as well as accessibility and archiving. However, the scientific record can also be distorted through miscitation.

Ever since Woo Suk Hwang's fraudulent and ethically compromised stem cell paper in 2005, data manipulation has been a favourite with the media. Although the news coverage has dented the reputation of science, it has also led to an increase in the teaching of research ethics, better supervision and the establishment of independent institutions that monitor, advise and arbitrate. The apparent rise in data misrepresentation in the past few years is probably a consequence of the relative ease of digital image manipulation, but also of the increased pressure to publish and publish fast — rising scientific output and speed of data acquisition certainly conspire to increase the likelihood of being scooped. In our experience, data manipulation increases when data is added in revision, undoubtedly because the pressure and desire to achieve publication rise with each revision, and some researchers feel experiments requested by referees and editors are less important.

We have commented previously (*Nature Cell Biol.* **8,** 101–102; 2006) on steps we have taken to address data manipulation, notably the publication of explicit policies, the requirement for uncropped gels/blots (*Nature Cell Biol.* **8,** 203, 2006 and *Nature Cell Biol.* **6,** 275; 2004) and the forensic checking of a subset of our accepted papers (*Nature Cell Biol.* **9,** 355, 2007). In the wake of the Hwang case, a number of journals, including this one, approached the NAS to encourage study of digital data manipulation.

The resulting report, *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*, arrives at no hard and fast rules; the panel found that different fields have quite different requirements. In the words of panel chairs Phillip Sharp and Daniel Kleppner, "the report provides a framework for dealing with the challenges to the community generated by the onrush of digital technology." Nevertheless, the key tenets that researchers are responsible for ensuring the integrity and accuracy of their data and appropriate training in the management of research data, that all data and experimental details from papers be publicly accessible and carefully archived to allow verification and to facilitate future discoveries, and that field-specific standards have to be developed by researchers, funders, societies and journals, benefit from being spelled out in one document. We have previously discussed these and related issues, such as the need for detailed methods (including reagent information) to allow data reproducibility, and we have discouraged 'data not shown' (*Nature Cell Biol.* **11,** 667; 2009, *Nature Cell Biol.* **9,** 481; 2007 and *Nature Cell Biol.* **8,** 541; 2006).

Importantly, the issue of data stewardship goes beyond keeping all relevant raw data at hand; it highlights the question of how journals, funding agencies and the community can ensure large datasets and associated metadata (such as instrument settings and procedures) can be appropriately preserved in an accessible form. Many types of high-throughput or memory-intensive data cannot yet be captured in public databases and journals struggle to host them. In some cases, databases suffer from insecure long-term funding (*Nature Cell Biol.* **8,** 425, 2006). The NAS report notes, "the questions of who is responsible for storing research data and who pays for maintaining the archive are urgent" and is right to recommend that funding be earmarked for data access and stewardship.

Whereas some new types of data, such as deep sequencing runs, are absorbed by existing databases, key areas such as microscopy data remain 'orphans'. There are significant ongoing efforts to standardize microscopy data to allow inclusion in databases, notably the Open Microscopy Environment (*Nature Cell Biol.* **6,** 909; 2004) and the Interactive Scientific Publishing Project of the Optical Society. This journal is actively engaged in proposals to set up a community database with visualization tools for minimally processed imaging data and associated metadata. It is essential that funding agencies address the need to invest in the development and, importantly, the maintenance of publicly accessible databases to archive complex datasets, and to make such data accessible efficiently to human and machine interrogation (systems biology).

Whereas data manipulation is a recent phenomenon facilitated by the digital data revolution, allegations of intellectual property theft and plagiarism stretch all the way back to Leibnitz and Newton, via Faraday and Davy, and Montagnier and Gallo. Several systematic analyses and questionnaire-based studies indicate that the literature in most subject areas is rife with text duplications (for example, *Nature* **435,** 737–738, 2005 and *Nature* **451,** 397–399, 2008). In an effort to monitor this form of intellectual property theft, we routinely spot check accepted manuscripts with semantic text comparison tools. It is worth reiterating that re-publication of ones own writing and data constitutes self-plagiarism.

Importantly, many still seem unaware that plagiarism extends to concepts, in particular, inappropriate attribution of prior knowledge by overlooking citations, misciting or even scooping projects based on prepublication information from conferences. The boom in reviews can also easily lead to misinformation and the misattribution of academic credit. To redress this, we have increased our reference limits and encourage primary literature citation (*Nature Cell Biol.* **11,** 1; 2009). Moreover, we reiterate our call to disambiguate journal impact factors by separating primary from review impacts (*Nature Cell Biol.* **7,** 925; 2005 and *Nature Cell Biol.* **5,** 681; 2003).

In a series of striking papers, Simkin and Roychowdhury have presented evidence that the majority of citations may not actually be based on reading the original reference, but rather on copying citations from the reference lists of other papers. They conclude that certain papers thus accumulate citations irrespective of quality, 'of course great scientists do exist. It is just that even if they would not exist, the citation data would look the same', (*Complex Syst.* **14,** 269–274; 2003, *Scientometrics* **62,** 367–384; 2005 and *Annals Improb. Res.* **11,** 24–27; 2005). Recently, a network analysis of 242 papers and 675 citations on a controversial role of β-amyloid in skeletal muscle concluded that citation distortions led to unfounded claims due to bias against papers that refuted or weakened the claims and amplification of the claims through a few reviews and various forms of miscitation (*BMJ* **339,** 2680; 2009). Interestingly, the author noted that negative data is inherently less cited and therefore tends to fail to spread through citation networks. These studies underscore the importance of accurate, informed and honest citation to ensure academic credit goes where it is due and to avoid leading the community up the garden path..