nature neuroscience

Ensuring data integrity

A recent report by the National Academy of Science makes recommendations for ensuring the integrity of research data. Critically, it also highlights the urgent issues regarding the preservation of large datasets.

Reports of scientific misconduct often make for sensational media reports. High-profile cases of falsification inevitably call for a re-examination of whether and how fraud can be detected before publication. The growing ease and practice of digital data transformation make this issue ever thornier. Scientists can be genuinely misled into thinking that they have found a specific result, only to later discover that they've been fooled by an artifact created by their digital analysis methodology. Compounding such data integrity problems is the fact that researchers can now amass huge amounts of data with relatively little effort; consider, for example, the gene lists that are generated by a single microarray experiment or the new RNA-seq technology, or large-scale detailed brain maps that typically require tens of gigabytes of memory per image. Established guidelines on best practices for data analysis, integrity, accessibility and archiving have not kept pace with this data explosion. The problem appears particularly urgent in interdisciplinary fields such as the neurosciences, in which researchers often navigate multiple layers of resolution, including gene expression datasets, cellular imaging and physiology, functional imaging and clinical data, all of which may have their own standards to safeguard data accuracy.

Soon after the notorious Hwang stem cell fraud case, a group of scientific societies and publishers, including the Nature Publishing Group, approached the US National Academy of Science (NAS) to encourage a thorough study of data manipulation and preservation in the digital age and to recommend adequate best practices to ensure the accessibility of large datasets. The NAS committee, headed by cancer researcher Phillip Sharp and physicist Daniel Kleppner, published its report this July (http://books.nap.edu/openbook.php?record_id=12615).

The report's central tenet is that the individual scientists are responsible for the truth and accuracy of their data. Most of its recommendations for ensuring data integrity and combating fraud are common-sense guidelines that are followed in most laboratories. Research institutions need to ensure that appropriate tools for management of research data are available to their scientists. The report also emphasizes the obvious, that data and experimental details must be made accessible and archived to allow for replication and consequent studies. As expected, the panel found that different disciplines have rather diverse requirements regarding data quality.

Journals, as stakeholders in the research enterprise, also need to do their part, and many (including *Nature Neuroscience*) have taken steps to enhance the quality and reproducibility of published work. Journals may require detailed methods sections, mandate author contribution statements and many have published explicit policies on the manipulation of raw data. We, for example, ask authors to list all image-acquisition tools and image software packages, and if cropped electrophoretic gels are included in the paper, these must be indicated as such in the figure legend and uncropped gels and blots must be included in the Supplementary Information (http://www.nature.com/authors/editorial_policies/image.html).

Above and beyond recommending measures to combat mis-analysis and fraud, the NAS report also calls for an urgent evaluation of the provisions for long-term maintenance of research data. This involves the critical question of how the community (including individual scientists, universities, funding agencies and journals) can ensure that large datasets are appropriately stored, referenced and indexed for posterity. To achieve these goals, scientific disciplines and communities must first agree on the criteria as to what data should be retained, such as information about instrument calibration and proprietary tools, details of the data processing methodology, and similar nitty-gritty, but important, issues. Moreover, as the ultimate value of scientific datasets will depend on an interlinked database infrastructure, we must attempt to coordinate data standards between disciplines to ensure compatibility and avoid redundancy. In addition to agreeing on criteria for data annotation, communities must also agree on formal vocabularies for their data and concepts, to enable unambiguous description of data and to make it machine-readable.

Currently, researchers have few incentives to invest much time and energy into data preservation and annotation; a situation which has to be remedied if we are to ensure the integrity of digital data. Funding must be made available to achieve lasting conservation of and access to data. Funding bodies and scientists must also work toward developing metadata management tools that would help researchers annotate data more easily and creating software that would make it possible to track individual pieces of data so as to give credit where credit is due. The increasingly important role of data-processing professionals in all scientific endeavors must also be better recognized.

Better training and education of scientists in data stewardship issues is critical at this time. Many scientists have had little or no formal training in information management and are therefore simply ill-prepared to think intelligently about these matters. As the NAS panel points out, data management must start at the beginning of a project, not midway through it or as an afterthought. Institutions must put data stewardship policies in place and promote the necessary training of their employees. Such training would include an understanding of the storage and preservation of data, its annotation, some the central online databases and their organization, and an appreciation of the bioinformatic tools that are available.

The NAS panel correctly notes that maintaining the integrity and accessibility of research data in this evolving digital age requires the collective efforts of individual scientists, research institutions, funding agencies, universities and journals. We urgently need to invest into our bioinformatic infrastructure to create the framework necessary to ensure that data is stored, annotated and preserved in a way that will provide maximum benefit for future studies. ■