nature

www.nature.com/nature

Vol 455 | Issue no. 7209 | 4 September 2008

Community cleverness required

Researchers need to adapt their institutions and practices in response to torrents of new data — and need to complement smart science with smart searching.

The Internet search firm Google was incorporated just 10 years ago this week. Going from a collection of donated servers housed under a desk to a global network of dedicated data centres processing information by the petabyte, Google's growth mirrors that of the production and exploration of data in research. All of which makes this an apt moment for this special issue of *Nature*, which examines what big data sets mean for contemporary science.

'Big', of course, is a moving target. The portability of the tens of gigabytes we carry around on USB sticks would have seemed like fantasy a few years ago. But beyond a certain point, as an increasing number of research disciplines are discovering, the vast amounts of data are presenting fresh challenges that urgently need to be addressed.

The issue is partly a matter of the sheer scale of today's data sets. Managing this torrent of bits has forced more and more fields to move to industrial-scale data centres and cutting-edge networking technology (see page 16). But the data sets are also becoming increasingly complex. As researchers study the inner workings of the cell, for example, they now gather data on genomic sequences, protein sequences, protein structure and function, bimolecular interactions, signalling and metabolic pathways, regulatory motifs — on and on. No wonder even the smartest scientists turn with relief to advanced data-mining tools, online community collaborations (see page 22) and sophisticated visualization techniques (see page 30).

Sudden influxes of data have transformed researchers' understanding of nature before — even back in the days when 'computer' was still a job description (see page 36). Unfortunately, the institutions and culture of science remain rooted in that pre-electronic era. Taking full advantage of electronic data will require a great deal of additional infrastructure, both technical and cultural (see pages 8, 28 and 47).

The lack of standards, for instance, confounds many a researcher seeking to harness the diversity of knowledge now available on any chosen topic. All credit, then, to those in the vanguard of interoperability. In biology, for example, the Gene Ontology Consortium has spent the past decade devising consistent descriptions of gene products in different databases. Meanwhile, the Mouse Genome Informatics resource is a good demonstrator of complexity's challenges and solutions. Funding agencies have been slow to support data infrastructure and this is one cultural shift that needs to accelerate — although recent efforts by the US National Science Foundation and Germany's DFG are a good beginning. But above all, such standards require support from researchers, who should adopt them and deploy them consistently. This takes a degree of intellectual and practical commitment to what can seem like tedious bookkeeping.

Researchers need to be obliged to document and manage their data with as much professionalism as they devote to their experiments. And they should receive greater support in this endeavour than they are afforded at present. Those publicly funded databases that have taken on preservation responsibilities, such as GenBank and UniProt, are only a small part of the data landscape. Universities and funding agencies need to provide and support curation facilities, tools and training.

As is amply highlighted in this issue, all of these worthy aims require incentives. These include pressure from, and recognition through, journals. *Nature* and its sister publications have always worked closely with those developing databases and standards, and we remain committed to continuing such community collaborations. Incentives also include recognition of impactful informatics by peer committees and research-rating exercises.

Above all, data on today's scales require scientific and computational intelligence. Google may now have its critics, but no one can deny its impact, which ultimately stems from the cleverness of its informatics. The future of science depends in part on such cleverness again being applied to data for their own sake, complementing scientific hypotheses as a basis for exploring today's information cornucopia. ■

For podcast and more online extras see www.nature.com/news/specials/bigdata/