The potential business value and scientific impact of data integration in drug discovery is substantial, particularly with the advent of numerous platform technologies and data resources that have even inspired the coining of the term ‘integromics’1. By some accounts, ‘omic’ data will soon exceed exabyte (10^18 bytes) quantities2, which approximates the content of all the words ever uttered by humans3. However, concurrent with the explosion of ‘omic’ data, the capabilities of computer hardware and storage costs have kept pace thanks to the venerable MOORE’S LAW⁴, and it is now widely recognized that the substantive challenges in fruitfully combining data stem in large measure from the diversity of data, and not just the sheer quantity. Meanwhile, the changing nature of the pharmaceutical and biotechnology industries, which is characterized by mergers, acquisitions and exchanges of assets, creates a need to maintain comprehensive data ‘packages’ within ever larger and more dynamic portfolios5. These are just a few of the drivers demanding a capability to assemble multifarious information relevant to drug discovery and development in a coherent and accessible fashion.

In approaching data integration, the first question to ask is where such integration should occur along the path from PRIMARY DATA, which is stored in operational or ‘working’ databases, to DERIVED DATA, which is refined and presented at a higher level, where it is aggregated, visualized, statistically characterized, interpreted and used to drive decision-making. The integration of derived data — what can be termed high-level data integration — involves the collation of presumed facts and conclusions, rather than the raw supporting data. At the highest level, this bears more on issues of ‘organizational memory’, institutional communication and the proper juxtaposition of related information than on data analysis proper.

The character of data

In approaching data integration, the first question to ask is where such integration should occur along the path from PRIMARY DATA, which is stored in operational or ‘working’ databases, to DERIVED DATA, which is refined and presented at a higher level, where it is aggregated, visualized, statistically characterized, interpreted and used to drive decision-making. The integration of derived data — what can be termed high-level data integration — involves the collation of presumed facts and conclusions, rather than the raw supporting data. At the highest level, this bears more on issues of ‘organizational memory’, institutional communication and the proper juxtaposition of related information than on data analysis proper.