Biology must develop its own big-data systems

Too many data-management projects fail because they ignore the changing nature of life-sciences data, argues John Boyle.

The last week of April was designated Big Data Week. But in modern biology, every week is big-data week: life-sciences research now routinely churns out more information than scientists can analyse without help. That help increasingly comes in the form of expensive data-management systems, but these are hard to design and most are even harder to use. As a result, a long line of data-management projects in the life sciences — many of which I have been involved with — have failed.

The size, complexity and heterogeneity of the data generated in labs across the world can only increase, and the introduction of cloud computing will encourage the same mistakes. Just a stone's throw from where I work, at least three computer companies are already touting cloud-based data-management systems for the life sciences. We need to find ways to manage and integrate data to make discoveries in fields such as genomics, and we need to do this quickly.

At their most basic, data-management systems allow people to organize and share information. In the case of small amounts of uniform data from a single experiment, this can be done with a spreadsheet. But with multiple experiments that produce diverse data — on gene expression, metabolites and protein abundance, for example — we need something more sophisticated.

An ideal data-management system would store data, provide common and secure access methods, and allow for linking, annotation and a way to query and retrieve information. It would be able to cope with data in different locations — on remote servers, on desktops, in a database or spread across different machines — and formats, including spreadsheets, badly named files, blogs or even scanned-in notebooks.

That ideal system does not exist. Most academic organizations have, through trial and error, developed their own in-house systems that work — or just about. The systems have limited functionality and cannot be connected, which makes collaboration difficult. The situation is as unworkable as if every lab in the country had decided to devise its own (poor) document-editing software.

Efforts to introduce overarching data-management systems, to which any and all scientists in a particular field could plug in, have failed for two main reasons. Either they demand that scientists change the format of their data, to allow information to be entered into the system, or they demand that scientists change the way they work, to generate standardized sets of results. The systems are thrust on scientists who are then expected to change, rather than taking the work of scientists as a starting point. It should not be scientists who are required to be flexible; it should be the system that they are being asked to use.

IT SHOULD NOT BE
THE SCIENTISTS WHO
ARE REQUIRED TO BE
FLEXIBLE;
IT SHOULD BE THE
SYSTEM THAT THEY
ARE BEING
ASKED TO USE.

These problems are exemplified by the expensive flop that was the US National Cancer Institute's caBIG data-integration project, scrapped last year after almost a decade and tens or even hundreds of millions of dollars. It had admirable goals and seemed workable in theory, but in the end it was too complicated to use. Crucially, caBIG relied on standardized data formats, which called for standardized experiments. Its one-size-fits-all approach fit nearly nobody.

There have been some successes. A widely used system called SRS allows the linking of data held in separate well-structured repositories. And the Biomart project joins up specially designed databases. But these were both fairly bespoke research applications; computer giants Microsoft and IBM are among the commercial firms that have introduced systems that aimed at a wider reach but had little impact.

To be useful to the life-sciences community, a data-management system probably needs to be devised and developed by the life-sciences community. The US National Institutes of Health has a 'Big Data' initiative, and agency head Francis Collins has spoken many times of the need to address the problem. Now is the time for researchers to plan an open data-management system that scientists will want to adopt. Many of the software pieces are already available.

As a starting point, here are three lessons from the successes and failures of the past.

First, the data are going to change. Biological information will always come in varied formats, and these formats cannot be defined in advance. Software engineers hate this. But a useful system must be flexible and updatable.

Second, people are not going to change. Busy scientists will adopt a new system only if it offers substantial benefit and is painless. Many commercial systems are unpopular because they make simple steps such as data retrieval complicated, to stop scientists using several (rival) systems at once.

Third, the problem is not technical. Although the latest kit is always alluring to funders, today's cutting-edge devices will be blunt tomorrow. Data-management systems must be driven by the need to find a workable solution to the problem, not by a desire to make the problem fit the latest fashionable technology.

Development of a biology-friendly system is possible, but it will require a change in mentality. As a useful test, a good data-management system should cost more to maintain, update and change with the times than it does to develop. Otherwise the price is too high. ■

John Boyle will shortly become senior director of bioinformatics at Kymab in Cambridge, UK.
e-mail: john.boyle@kymab.com

➔ **NATURE.COM**
Discuss this article
online at:
go.nature.com/zyllsz