can be useful, a full understanding of circuit function requires additional types of knowledge, for instance, about the neurotransmitters involved, the electrical properties of component neurons and the influence of modulatory systems.

Will BrainAligner become the software of choice for the community? History suggests that will depend on more than alignment quality and speed. BrainAligner is freeware, and it is well integrated with the V3D and AtlasViewer freeware developed by the same group[3]. Price is therefore not an issue, but documentation and technical support, platform dependence (BrainAligner is currently available in Macintosh and Linux formats) and the availability of updates could be. Furthermore, there are other promising ventures in various states of development such as Flybrain@Stanford[4], BrainGazer[5] and FlyCircuit[6] that have similar goals, so only time will tell. Let the bidding begin!

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Ito, K. *Front. Syst. Neurosci*. **4**, 26 (2010).
2. Peng, H. *et al. Nat. Methods* **8**, 493–498 (2011).
3. Peng, H., Ruan, Z., Long, F., Simpson, J.H. & Myers, E.W. *Nat. Biotechnol*. **28**, 348–353 (2010).
4. Jefferis, G.S. *et al. Cell* **128**, 1187–1203 (2007).
5. Bruckner, S. *et al. IEEE Trans. Vis. Comput. Graph*. **15**, 1497–1504 (2009).
6. Chiang, A.S. *et al. Curr. Biol*. **21**, 1–11 (2011).

Channeling the data deluge

Jason R Swedlow, Gianluigi Zanetti & Christoph Best

With vast increases in biological data generation, mechanisms for data storage and analysis have become limiting. A data structure, semantically typed data hypercubes (SDCubes), that combines hierarchical data format version 5 (HDF5) and extensible markup language (XML) file formats, now permits the flexible storage, annotation and retrieval of large and heterogenous datasets.

Biological research laboratories were once occupied by scientists whose main tools of the trade were the pipette, lab notebook, calculator and pen. Twenty-five years of automation and feats of engineering have revolutionized biology into a data-centric science, the best example being certainly the genome projects whose output is now the foundation of essentially all modern biological experiments. These projects were undertaken in a relatively few central facilities, which—after some negotiation—agreed to release their data within one day of collection using standardized formats. Today, most modern labs have access to sophisticated data generation and analysis systems that routinely generate similar amounts of data each day, all of which must be processed and analyzed to reveal biological understanding. In stark contrast to genomics, these data are produced locally by many individual scientists, but the overall scale and heterogeneity of these experimental efforts create a barrier to easy standardization: a data format that suffices for one lab will very likely only partially address the needs of another. When experimental design and outcome drive the data formats, straightforward standardization becomes nearly impossible. This priority is correct; scientific achievement should drive data formats and not vice versa.

Heterogeneity, however, comes with a considerable cost. Data generated in one lab cannot be analyzed by researchers in another, and data analyzed using one software tool often cannot be analyzed with another tool (even in a single lab). Reverse-engineering data formats is slow, time-consuming, error-prone and certainly scales poorly with the diversity of experiments. At the same time, although scientific data formats do not themselves enable discovery, they are a powerful enabling technology. Without the Genbank and Protein Data Bank (PDB) repositories, much of today's research would be impossible.

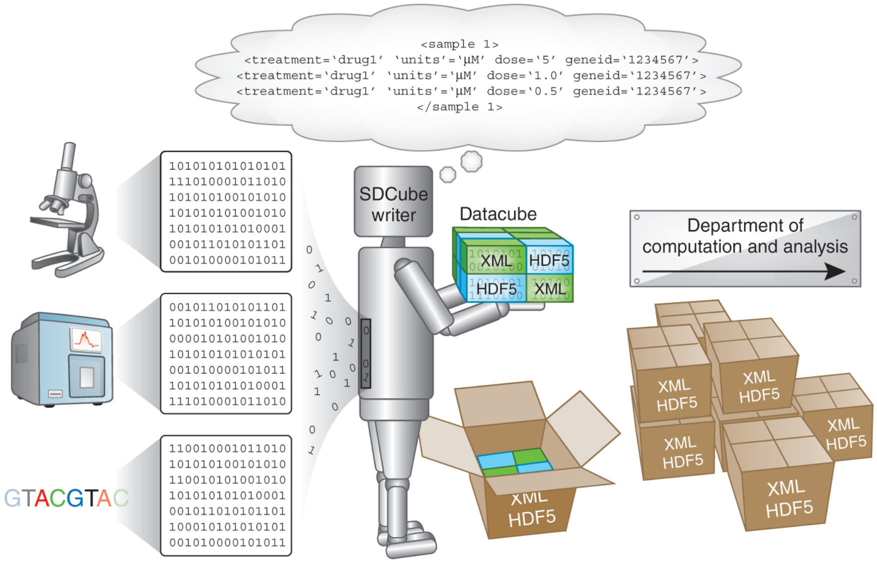
Seen in this context, weaning bench scientists from storing their data in randomly formatted spreadsheet files seems not only useful but scientifically valuable. This issue has not been lost on funding agencies such as the US National Science Foundation with its Office of Cyberinfrastructure, the UK Research Councils with their eScience program and the European Union, which funds several 'e-Infrastructures' in its FP7 program and recently commissioned a high-level expert group to report on the handling of scientific data[1].

Scientific data formats always involve a tradeoff between simplicity and flexibility. Some of the most useful formats, for example, the comma-separated values (CSV) spreadsheet or PDB and Genbank files, have a simple, line-oriented structure that is easy to process without extensive programming. But there are limits to these structures that force the use of awkward workarounds (for example, splitting a PDB file to accommodate more than 99,999 atoms in a ribosome). In this issue, Millard *et al.*[2] show that by leveraging well-established computer science tools and high-performance computing it is possible to build a simple data storage system that can efficiently and flexibly manage data coming from high-throughput imaging.

One tool they use is the hierarchical data format version 5 (HDF5), which works as a flexible vessel to efficiently store large arrays of numerical data along with textual metadata within a single file structure. HDF5 was first developed by the US National Center for Supercomputing Applications in the early 1990s as a flexible and efficient file format for large numerical datasets arising mainly in high-performance computing. An HDF5 file provides the flexibility of a file system: a single file can hold many different types of data, and arbitrary access to data elements within large matrices and datasets is supported. HDF5 is a sophisticated technology, but many open-source tools are now available that provide easy cross-platform access, making HDF5 a tool that can be used easily across scientific disciplines. As the needs for data have grown across the sciences, so has the readiness to accept the complexities of HDF5 for its flexibility and efficiency.

However, HDF5 has only limited capabilities to express nonnumerical information, such as metadata and experimental setups. Millard

JJason R. Swedlow is at the Wellcome Trust Centre for Gene Regulation and Expression, College of Life Sciences, University of Dundee, Dundee, Scotland, UK. Gianluigi Zanetti is at the Center for Advanced Studies, Research and Development in Sardinia, Pula, Italy. Christoph Best is at Google UK Ltd., London, UK.
e-mail: jason@lifesci.dundee.ac.uk



```
<sample 1>
<treatment='drug1' 'units'='μM' dose='5' geneid='1234567'>
<treatment='drug1' 'units'='μM' dose='1.0' geneid='1234567'>
<treatment='drug1' 'units'='μM' dose='0.5' geneid='1234567'>
</sample 1>
```

Katie Vicari

Figure 1 | A cartoon representation of data packaging into SDCubes.

et al.[2] therefore use another tool, the extensible markup language (XML), which has become the de facto modern standard for structured data interchange. XML files can easily describe what experiments have been performed and which columns in the numerical data blocks of the HDF5 relate to which experimental parameters. By pairing an XML with an HDF5 file, Millard *et al.*[2] create a data structure that combines the expressivity of XML with the efficiency of HDF5, allowing the heterogeneity and flexibility needed to support large, experimental datasets while still retaining a complete computer-readable description of experimental design and structure (**Fig. 1**).

Millard *et al.* refer to this mechanism as semantically typed data cubes (SDCubes). They apply this technique in a high-content screening experiment in which several different dose-response curves are generated in the exploration of phenotypic response of cells to several small molecule inhibitors. Although a new data storage mechanism by itself does not constitute a scientific discovery, the achievement of Millard *et al.*[2] is that they showed how a flexible and self-documented storage mechanism can be used in a highly complex and evolving biological experiment, and demonstrated how this can support a variety of experimental protocols and output data.

Data standardization linked to the appearance of new data-generation technologies is well-trodden ground. The 'minimum information about a microarray experiment' (MIAME)[3] was perhaps the most effective and has been followed by many similar efforts for other data types (microarray gene expression data (MGED)[4], open microscopy environment-XML (OME-XML)[5], minimial information about fluorescence *in situ* hybridization and immunocytochemistry experiment (MIFISHIE)[6], minimum information for biological and biomedical investigations (MIBBI)[7] and others). Much research in computer science has also been expended on how complex real-world situations such as biological experiments can be expressed in a way 'understandable' to a computer, in particular through the use of ontologies[8]. Projects such as Taverna[9], myExperiment[10] and Open Electronic Health Records Foundation (http://openehr.org/) use such concepts to provide ways of specifying, storing and sharing the data processing steps and workflows that underpin much biological analysis. However, a universal life sciences data format with enough flexibility to integrate these complex information items while ensuring that any system can read and compute on it has not been achieved.

SDCubes help fully describe experiments and their results in computer-readable fashion. What is the next step? At some point, single files stored on a local disk by themselves will become too cumbersome to use. Jim Gray and colleagues have referred to HDF5 and similar structured file formats as 'nascent database(s)'[11] and have predicted that large datasets will be stored in databases, possibly based on these file formats, but hosted on database servers and accessed through layers of software ('middleware') that hide their complexity from the user—in a concept very similar to what is today referred to as 'cloud computing'.

Finally, to make SDCubes heavily used, maintenance of open libraries that provide write and read access will be critical for the community. In our experience, standardization is provided not just by a technical specification, but by good examples, support and open-source reference implementations. Widespread adoption of a data format will also require involvement of the community at large before a standard can be finalized. With this we can look forward to the day when a standardized format for biological analytical output becomes linked to shared workflows so that real data sharing and analysis become possible.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details accompany the full-text HTML version of the paper at http://www.nature.com/naturemethods/.

1. High Level Expert Group on Scientific Data. Riding the wave: how Europe can gain from the rising tide of scientific data (http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf) (European Union, 2010).
2. Millard, B.L., Niepel, M., Menden, M.P., Muhlich, J. & Sorger, P.K. *Nat. Methods* **8**, 487–492 (2011).
3. Brazma, A. *et al. Nat. Genet.* **29**, 365–371 (2001).
4. Spellman, P.T. *et al. Genome Biol.* **3**, 0046 (2002).
5. Goldberg, I.G. *et al. Genome Biol.* **6**, R47 (2005).
6. Deutsch, E.W. *et al. Nat. Biotechnol.* **26**, 305–312 (2008).
7. Taylor, C.F. *et al. Nat. Biotechnol.* **26**, 889–896 (2008).
8. Ashburner, M. *et al. Nat. Genet.* **25**, 25–29 (2000).
9. Hull, D. *et al. Nucleic Acids Res.* **34**, W729–W732 (2006).
10. Goble, C.A. *et al. Nucleic Acids Res.* **38**, W677–82 (2010).
11. Gray, J. *et al. Proc. ACM SIGMOD Int. Conf. Manag. Data* **34**, 35–41 (2005).