The search for data

The biological literature is vast and growing; tools to help researchers negotiate it are much needed. A new platform, SourceData, aims to help researchers hone in on the core of the research effort: the data.

In some respects, a scientific manuscript published today is a very different beast from one published two or three decades ago. Substantial material is presented separately from the main paper as part of a supplement. Data sets generated in a study are increasingly hosted on external sites, and the same is often true of software developed or used in the work. And yet, in other ways, scientific papers are still recognizably old fashioned. For one thing, it remains difficult to directly search for the data in published manuscripts.

To be sure, there is a growing consensus that certain data types should be freely available to help ensure both reproducibility of results and data reuse. Accordingly, publishers and funders now mandate availability of many, typically large-scale, data sets. But a good swathe of the biological literature consists of small-scale experiments—whether biochemical, genetic, biophysical or cell biological—that probe varied systems and questions. The data from these experiments, typically represented in figures, are used to build up the scientific argument in papers.

What if you had a way to directly query the figures in all published papers? For instance, what if you could, with the click of a search button, find all experiments in the literature in which the function of a particular gene had been tested in a particular signaling pathway, or in which the effect of a small molecule on a specific cell type had been studied? A platform announced in this issue (p1021) could, in principle, allow researchers to do just that.

The platform, SourceData (sourcedata.embo.org) is based on annotation of the figures in papers in a way that is machine readable and therefore searchable. In a proof-of-principle demonstration of this idea, the SourceData developers have manually annotated several hundreds of papers and about 18,000 experiments. Using a semantic system the researchers developed, human annotators describe individual figures based on information in the figure legend and in the labels of the figure itself. The platform in its present version is geared to the discovery of experiments where a perturbation or measurement is made on some scientific 'object' (which may be as varied as a cell line or a molecule or a mouse).

As such, this lays the groundwork for building a searchable database of figures in published papers. A query for experiments of interest then returns the specific figures describing those experiments as well as the DOIs of the papers of which they are a part. This changes the 'unit' of search: no longer keyword-defined manuscripts, but data in individual figures.

The SourceData platform should not be confused with the 'source data' that Nature Research journals have in recent years been encouraging authors to provide with their published papers. The latter are typically Excel spreadsheets that contain the numerical values underlying figures. In its present incarnation, SourceData calls up such spreadsheets when they are provided for a particular figure; however, annotation and search itself does not yet systematically extend to these files. No doubt this will be a potentially powerful extension for the future.

There will be challenges to incorporating such a platform into actual practice. First, who would annotate the figures in papers, and where would the resources for this additional effort come from? Annotation could be done by one (or more) of many parties: authors, journal staff, third-party annotators or even algorithms. Annotation could be a service provided by publishers—for example, Springer Nature now offers optional data annotation (http://www.springernature.com/gp/authors/research-data-policy/support-services/12327144) for some of its journals (not at present including the Nature Research journals). While this service is free in its present pilot form, providing a free service is likely to be more difficult at scale.

Indeed, the problem of scale goes beyond merely the question of who pays. Manual annotation is likely to be unfeasible at scale even prospectively, let alone retrospectively on the massive edifice of already published papers. Automated, computer-driven approaches will need to be developed. Furthermore, the quality of data description and of data itself is variable in published papers; whether this should be accounted for during data annotation, and how, remains to be seen. Finally, the performance of the annotation-search pipeline itself would need to be assessed.

But challenges notwithstanding, being able to search the literature based on data in figures is an exciting prospect. It could fundamentally shift how scientists interact with the collective body of knowledge and open up all data to synthesis, reassessment and reuse.