

## **Databases, data sets, and data accessibility – views and practices of scholarly publishers**

### **A statement by the International Association of Scientific, Technical and Medical Publishers (STM) and the Association of Learned and Professional Society Publishers (ALPSP)**

Publishers recognise that in many disciplines data itself, in various forms, is now a key output of research. Data searching and mining tools permit increasingly sophisticated use of raw data. Of course, journal articles provide one ‘view’ of the significance and interpretation of that data – and conference presentations and informal exchanges may provide other ‘views’ – but data itself is an increasingly important community resource.

Science is best advanced by allowing as many scientists as possible to have access to as much prior data as possible; this avoids costly repetition of work, and allows creative new integration and reworking of existing data.

There is considerable controversy in the scholarly community about ‘ownership’ of and access to data, some of which arises because of the difficulty in distinguishing between information products created for the specific display and retrieval of data (‘databases’) and sets or collections of raw relevant data captured in the course of research or other efforts (‘data sets’). Another point of difficulty is that in many cases data sets or even smaller sub-sets of data are also provided as an electronic adjunct to a paper submitted to a scholarly journal, either for online publication or simply to allow the referees to verify the conclusions.

We believe that, as a general principle, data sets, the raw data outputs of research, and sets or sub-sets of that data which are submitted with a paper to a journal, should wherever possible be made freely accessible to other scholars. We believe that the best practice for scholarly journal publishers is to separate supporting data from the article itself, and not to require any transfer of or ownership in such data or data sets as a condition of publication of the article in question. Further, we believe that when articles are published that have associated data files, it would be highly desirable, whenever feasible, to provide free access to that data, immediately or shortly after publication, whether the data is hosted on the publisher’s own site or elsewhere (even when the article itself is published under a business model which does not make it immediately free to all).

We recognise, however, that hosting, maintaining and preserving raw data or data sets, and continuing to make such data available over the long term, has a cost which, in certain circumstances, the host site may need to recover. We also recognize that on occasion the generation of data has been privately funded, and the funding entity may have a particular reason for restricting access to the data (either temporarily or even permanently), but we believe these should be limited exceptions, and that journal publishers themselves should claim no ownership interest in such data. The academic and publishing communities should discuss further (in the context of the debate on the public funding of research) whether more reliable and more permanent sites should be established to host research data.

None of this means, however, that databases themselves – collections of data specifically organised and presented, often at considerable cost, for the ease of viewing, retrieval and analysis – do not merit intellectual property protection, under copyright or database protection principles. Such databases are often characterized by the sophistication of their data field structuring, searchability tools, and the like, and scholarly publishers are often involved in producing and marketing databases that contain valuable and useful information for scholarly research. The research interest and value of raw research data sets and individual data points is entirely different, and serves different purposes, from that of specific databases that have been organised and compiled for particular research needs.

There is sometimes confusion about whether the use of individual ‘facts’ and data points extracted from a database is permitted under law. Facts themselves are not copyrightable, but only the way in which information is expressed – this is fundamental in copyright law. In the EU, the use of ‘insubstantial’ parts of a database, provided it is not systematic and repeated, does not infringe the database maker’s rights.

Articles published in scholarly journals often include tables and charts in which certain data points are included or expressed. Journal publishers often do seek the transfer of or ownership of the publishing rights in such illustrations (as they might do with respect to an author’s photograph), but this does not amount to a claim to the underlying data itself.

We hope that this statement is helpful in clarifying the views of publishers concerning raw data, data sets and databases, and that the statement will serve as useful guidance for publishers in their policies concerning data sets submitted with papers. Scholarly and scientific publishers share the view that research data should be as widely available as possible.

June 2006