Harvesting - from Data Collection to Deposition

Harold R. Powell[a], Andrew G.W. Leslie[a], T. Geoffrey Battye[a], Kim Henrick[b], [a]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge, UK.* [b]*EMBL Outstation Hinxton, European Bioinformatics Institute,Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.* E-mail: harry@mrc-lmb.cam.ac.uk

The best approach to data harvesting is undoubtedly to have all relevant information regarding the experiment recorded automatically as it is determined. Provided that this is done reliably (i.e. no data are missing and all items present are also correct), this can be used in an automated pipeline requiring minimal user intervention. By far the majority of errors in the harvesting process arise from manually entered user input.

While a standard format has been developed by the IUCr for area detector image data (imgCIF and its binary equivalent, the Crystallographic Binary File or CBF), it has not been adopted widely. In the absence of a commonly used format for the image data, it is important that files written using existing commercial formats contain the essential experimental information along with the intensity data. Further, the data processing programs should record these data in a format which is readily amenable to straightforward deposition in the appropriate database.

Keywords: data collection, data processing, data harvesting