The problems of sharing information

Today you can buy, for a couple of hundred pounds, a computer with enough memory to store every bryological record ever made in Britain or Ireland; and you can send huge chunks of data to other computers in seconds. It is possible, therefore, to envisage two extremes in the management of records: it would be perfectly feasible to hold all records in one central location (with adequate backup), with instant access to the data for anyone wishing to use it. Or anyone with an interest in bryology could hold the complete set of records on his/her own computer. I think the former is the ideal solution, and that we are very fortunate in having NBN Gateway as an approximation to it. Unfortunately (in my view), there is a strong tendency towards the second possibility: everyone (in particular local authorities, local record centres, conservation bodies etc.) seems to want records on their own computer. I have the impression that the accumulation of records becomes an end in itself; people are proud of the size of their database, just as a 19th century botanist might be proud of the size of his herbarium. And just as herbaria were swelled by 'duplicates' circulated by the botanical exchange club there is an active 'trade' in records. However, the only advantage of holding the same record on different computers is ease of access, an advantage which is effectively negated by the simplicity and speed of data transfer through the internet. I shall argue below that record duplication is a thoroughly bad thing and a source of chaos for the future.

The problem is that records are not fixed, they may change. First, there are simple mistakes like data entry errors and misidentifications. They should not happen, but in the real world they do. Many, we hope, will eventually be detected and records changed (where possible) or deleted. Then there are significant taxonomic changes – when *Hedwigia ciliata* was split into *H. ciliata s.s.* and *H. stellata*, databases suddenly acquired a rare taxon – until the appropriate changes were made. With the present protocol used by the BBS, changes are not a problem – the databases of vice-county recorders and CEH are tightly coupled and a recorder making a correction to her own database will always communicate the change to CEH. So one set of data is accurately mirrored by the other. But suppose that a recorder has responded to requests from several organizations (local record centres, local government, etc). Will he/she remember who has the data? If he sends the corrections, will they actually be made? And will the change be made incorrectly, compounding the error? If the organization has shared records with yet another, will the changes be communicated to them? It is important to remember that recipients will not always have expertise in bryophyte taxonomy and may be oblivious to anomalous records. In addition to these potential disasters, there is the nuisance value of duplication. Most of us have seen how Botanical Exchange Club specimens distributed to dozens of collectors gradually accumulate through legacies into big blocks in a few herbaria. This is an asset for taxonomy of course, but it will be nothing but a nuisance if in 20 years time there are 20 copies of the same record, some erroneous, at CEH. There are algorithms for removing duplicates, but do they always distinguish babies from bath water?

Of course we need to exchange data. But data transfer need not and should not involve copying *records* from one dataset to another. A database 'A' which routinely feeds records to 'B' (as v.-c. recorders feed CEH) should beware of communicating records to a third party 'C'. There is no harm in 'A' sending records to 'C' as long as they are *not incorporated into 'C's database*. It may be necessary for 'C' to create a temporary merger of the two sets in order to produce maps, etc., but this is no problem with the vast disk space available to everyone. This ensures that records (as opposed to information) flow in a hierarchical 'tree' structure rather than over an anarchic net. As long as the 'A'-derived dataset is kept separate from 'C's, it is easy for 'A' to ensure that the records at 'C' are properly maintained. As a touchstone, I would suggest 'A' responds to 'C's request for records by asking (1) what will you do when I send you an update in 2 years time, which includes all present records, and (2) what would you do if in 3 months time I told you that every single grid reference in the data sent to you is erroneous, and I am resending the data with the correct references inserted? Anyone who can face those prospects with equanimity, and who undertakes not to pass on your data but to refer requests to you is a satisfactory recipient.

John Lowell (e john.lowell@btinternet.com)

Thanks to M.O. Hill for comments on this topic.