

## A TAXONOMY OF DATABASES: DATABASES FOR TAXONOMISTS

LAZARUS, David B., Museum für Naturkunde, Invalidenstrasse 43, D-10115 Berlin, Germany

'Database' is one of those wonderfully vague words which everyone is in favor of, but which can mean so many different things to different people that without considerable clarification, one generally has no idea what is specifically meant when it is used. This simply reflects the enormous variety of databases which are possible, a flexibility which is both a boon and a curse to anyone wishing to either create or use a database. The diversity of possible database types and designs makes common definitions and standards a necessity. Definitions and standards can be agreed upon at various levels - hardware and software development systems; data models; actual finished database systems, and common data exchange formats.

Although a great deal of discussion is generated by proponents of particular hardware and software development systems, these issues are probably the least important ones in developing a large database system. More important are the data model, which defines what type of information will be in the database; and common data exchange formats, which provide at least some degree of mobility and obsolescence protection for the actual data content of finished individual database systems.

Databases currently being developed that are of potential interest to paleontologists can be grouped into a few broad categories, each serving a different set of interests, users, and each having a dissimilar underlying data model.

'Taxonomic Catalog' databases are based on taxonomic descriptions, typically of species, and include detailed text descriptions, images of a small number of selected individuals, and short summaries of stratigraphic and geographic distributions.

'Research Taxonomic' databases are based on individual specimens or sample populations, and include many more images and sample related information than is typical for a taxonomic catalog. Their structure can be very complex, as they must incorporate to a greater degree the messy details of taxonomy, including partial synonymy, hierarchical nomenclature etc.

'Distributional' databases use the occurrences in samples or localities of named taxa as their primary data unit. Such databases can be used for stratigraphy, paleoceanography, or macroevolutionary studies. They are not taxonomic databases as such, although they can provide distributional information about taxa, and must incorporate at least some information about the taxonomic names in the database.

Lastly, 'Curatorial' databases are designed to provide storage and retrieval information about specimens, samples or other material, typically at a museum. They are something of a mixture of all the previous types of database, but generally without enough depth in any of these areas to serve as a replacement for them.

Given the multiplicity of database types and user groups, it is unlikely that any one database system can be developed which will meet the needs of all paleontologists. For this reason, defining common data exchange standards will be extremely important as databases of various sorts become increasingly common.

Another issue which has not received enough attention is the need for the long term maintenance of large database systems. Given the rapidity of change in the computer industry, virtually any large database system will become technologically obsolete soon after it is developed. Securing institutional mechanisms that will provide a constant stream (or at least a trickle) of support for technical upgrades over a period of many years are an essential component of any larger database project. Without such mechanisms, the often enormous effort that went into creating the database system in the first place will soon be lost.