PALEONTOLOGICAL DATABASES AND PALEOINFOMATICS

MACLEOD, NORMAN Department of Palaeontology, The Natural History Museum, Cromwell Road, London, SW7 5BD, N.MacLeod@nhm.ac.uk.

Summary
Electronic databases represent practical, effective, and efficient means of accessing, combining, and controlling paleontological data. However, most databasing efforts currently underway are being organized in an ad hoc manner that will needlessly complicate more integrative phases of database-development. In order to address this issue there is a need for the paleontological community to develop databasing standards that will ensure the utility of current and planned databases for future generations of earth scientists. Aside from the immediate scientific benefits, such an standardization effort would also provide opportunities to forge necessary new internal and external organizational structures within the paleontological community as a whole.

Introduction
The past decade has borne witness to unprecedented levels of integration within the stratigraphical community, principally as a result of the electronic communications revolution. In addition, the continuing revolution in computing power has provided stratigraphers with the tools necessary to begin the task of migrating their data to the electronic repositories that will form the nodes of a distributed system for accessing and analysing stratigraphical research results. At the moment these tools are being employed in a haphazard manner by individual paleontologists/stratigraphers and small research groups around the world who need local databases to manage the copious amounts of data that (should) come together in any modern study. However, aside from being research and industrial tools in their own rights, such database systems also represent demonstrations of the adequacy of current computer technology to efficiently manipulate the types of data paleontologists and stratigraphers use, as well as tantalizing glimpses of what the future might hold for paleontological data management.

These developments come at a somewhat ironic time. Owing to a number of factors, the quality and depth of both systematic and stratigraphical expertise is in precipitous decline (Kaesler 1993; Flessa and Smith in press). Smaller numbers of graduate students are being trained and relatively few of those who do complete their degrees find employment in academia, industry, or museums as systematists-biostratigraphers. Moreover, many of the technical considerations that must be confronted in designing and maintaining a state-or-the-art paleontological database have little to do with systematics or stratigraphy per se. As a result, the diminished community of paleontologists is further hampered by a growing gap between computer-literate data analysts and programmers and more qualitatively-oriented, traditional paleontologists who have, up to now, dominated systematic paleontology.

In 1993 Richard Kaesler reviewed these arguments and concluded that, for the good of their science, paleontologists might want to consider making provision for the protection of their hard-won systematic and paleobiological insights by devoting some proportion of their activities to the developing and assembling expert systems of taxonomic information. Unfortunately, Kaelser's suggestions have not been taken up by the paleontological community. Instead, much recent work has focussed on the assembly of paleontological databases. While database assembly is a necessary precursor to the creation of an expert system, it does not involve the laborious—an often only marginally successful—programming required to quantatize the inherently qualitative judgements that systematists routinely make. Kaesler (1993) envisioned a system that would replace systematic paleontologists (albeit for an optimistically-envisioned short term), whereas databases keep paleontologists in the picture (though they do not solve the problem of the continuing decline in systematic expertise). Regardless, databases are seen by the paleontological community as practical and necessary; even urgent. Expert systems remain paleontological exotica.

Why has this change in attitude and direction come about? Certainly one factor has been paleontologists' growing appreciation that the technology needed to access databases of sufficient detail and scope to be practically useful in solving a host of data-related problems already exists on their desktops. The internet (via e-mail) has also provided this community with an opportunity to temporary offset the manpower problem (at least for many organismal groups) through the assembly of international partnerships that—if appropriately organized and administered—could ensure tight quality control for the initial phases of database creation. Last, but by no means least, recent developments in digital photography have provided paleontologists with the ability to significantly improve the manner in which fossil morphologies are illustrated. Since these advances are comparable to the introduction of scanning electron micrographical (SEM) methods for the illustration of fossils in the 1950's and 1960's, they can provide a focus for the descriptive effort that will inevitably need to accompany any comprehensive paleontological databasing effort.

One example of the type of database that future paleontologists-biosratigraphers might find particularly useful is *PaleoBase*. *PaleoBase* is a commercially-funded (Blackwell Science and The Natural History Museum [London]) relational database of macrofossil and microfossil genera designed for use in educational and research contexts. In addition to genus names, *PaleoBase* database records include full text-based descriptions of each genus, classification, ecology, biogeography, paleobiogeography, chronostratigraphical distribution, a morphological key, a extensive bibliography, and state-of-the art composite digital images of representative specimens (including many type specimens). These data are assembled into a fully-relational database (designed around the Compustrat-4D database engine) that allows extensive hard-coded and user-specified searches-sorts to be carried out. When completed the current *PaleoBase* system will include approximately 1000 macrofossil and 450 microfossil records.

In terms of organizational models for generalized databasing efforts, many of the problems associated with planning, organizing, and administering this type of international effort have been encountered and overcome by the *PaleoNet* (listserver), *PaleoBase* (electronic database), and *Paleontologia Electronica* (electronic publishing) initiatives I have been involved with over the past five years. Experiences with each of these has provided a series of

valuable object lessons in how to create and fund electronic paleontological products and services within both academic and corporate contexts. In addition, extra-paleontological initiatives such as GenBank point the way toward a possible future in which paleontologists are able to query metadatabases for data availability and have software automatically assemble datasets for their inspection/analysis for any fossil group, within any geological time interval, from any location in the world.

Considering that paleontological data remain (and will continue to remain) at the heart of such a wide variety of scientific disciplines, and considering the economic and intellectual implications such a "paleoinfomatics" system would have on the scientific community at large, the need for and economic viability of such a program seem obvious. Moreover, the limitations that would be imposed on any single institution or professional society that would attempt such a project ensure that this can only be contemplated as an international effort that cuts across traditional paleontological "business sectors" (academia, government, industry, museums). Even more importantly, though, such a project would serve several vitally-important functions within the paleontological community. First, it would act as a force for helping to unite what has become an increasingly disparate field in which practitioners who work on any one topic, with one group, or within one time interval feel they have little in common with other members of the community. Second, it would provide an avenue for improving the image of paleontology (as a rapidly-moving high-tech field on the cutting edge of research in the natural sciences and on top of the communications revolution). Thirdly—and most importantly—it would give provide an opportunity for paleontologists to use the communications revolution to gain control over its basic data. The work of the last quarter century has proven the value of the database approach to paleontological research beyond a shadow of a doubt (e.g., Sepkoski 1981, 1997). The ongoing phylogenetics revolution promises to dramatically improve the consistency of paleontological data (by improving classification and taxonomy, see Smith 1994) and the manner in which such data can be used in a variety of analytic contexts. A comprehensive "paleoinfomatics" system would consolidate these important advances and improve the quality of contemporary paleontology for both practitioners and clients. Moreover, all paleontologists could effectively participate in such a project and all would stand to benefit equally from its outcome.

Conclusions

Paleontologists no longer have any good reasons not to embark on a comprehensive, international database program and many, many good reasons to do so. Nevertheless, Kaesler's (1993) window metaphor remains valid in the context of such a project. There currently exists a window of opportunity for creating such a database that has remained open somewhat longer than originally expected thanks to innovations resulting from the development of the Internet. However, that window is closing as a result of the loss of individual expertise and the information represented by paleontological collections. Paleontologists of good will can disagree as to the duration of the time interval over which such a project could be started with some reasonable chance of success, but it is past the point of reasonable disagreement that such a project should be started as soon as possible. Local initiatives of this sort have already begun (e.g., *PaleoBase*, BIOLOG, the DSDP/ODP database system, PaleoBank, the electronic database programs of most major natural history museums that hold paleontological collections). What is needed is for the principal partners in such an effort (e.g., museums, systematists, computer communications specialists, professional societies, and users of paleontological data) to come together and begin discussing (1) how their information can be made available to their colleagues, (2) what type of infrastructure needs to be created to support common access, and (3) how such an international project can be co-ordinated and funded. It is only by conceiving of paleontological databases as a reflection of the paleontological community itself—composed of separate individuals, but united in a common purpose and willing to share information in an open and forthright manner—that we will succeed in surmounting the challenges and embracing the opportunities that are now before us.

References

Kaesler, R. L. 1993. A window of opportunity: peering into a new century of paleontology. Journal of Paleontology 67:329–333.

Flessa, K. W., and D. M. Smith. in press. Paleontology in academia: recent trends and future opportunities. In: Steininger, F. F., eds. Paleontology in the 21st Century. Verlag Waldemar Krammer, Frankfurt am Main.

Sepkoski, J. J., Jr. 1981. A factor analytic description of the Phanerozoic marine fossil record. Paleobiology 7:36–53.

Sepkoski, J. J., Jr. 1997. Biodiversity: past, present, future. Journal of Paleontology 71:533–539.

Smith, A. B. 1994. Systematics and the Fossil Record: Documenting Evolutionary Patterns. Blackwell, London.