DATA MINING: MAKING THE MOST OF POLAR AND OCEANOGRAPHIC INFORMATION IN THE 21ST CENTURY

Ben Raymond
Department of the Environment and Heritage
Australian Antarctic Division
Channel Highway
Kingston 7050 Australia

ABSTRACT: Scientific data are routinely archived by many polar and oceanographic organisations. Data mining — the discovery of new information from existing data — is an increasingly popular means for data centres to add value to their holdings.

Polar and marine data pose various challenges to the mining process. The physical environment and high cost of data acquisition mean that data are generally sparse. Furthermore, the international and multidisciplinary nature of polar and marine scientific research activities has lead to data repositories that are diverse in subject, type, format, and access procedures.

Due to the sparsity of data, it is important that all available datasets be available for analyses. Data discovery and access across national and international organisations is critical. This is a fundamental problem for all science, and initiatives to provide improved data access are already in place (e.g. SCAR MarBIN and the International Oceanographic Data and Information Exchange). Technologies such as web services, which permit data exchange between applications, may help to make this process more efficient and transparent to the user.

Data mining also needs to be able to draw on information from a range of data types (e.g. image, video, text). Data can sometimes be found from sources not traditionally considered for scientific investigations (e.g. historical data from diaries and logs, or expeditioners' photographs). Such data can be useful, but their scientific value may be compromised by ad-hoc acquisition processes. Technical developments in data mining should therefore look to provide improved handling of issues such as data heterogeneity, incompleteness, and bias.

The long-term success of data mining will require its adoption as an integral part of the scientific process, rather than as a separate, add-on activity. It is clear that any contemporary data can be expected to be re-used in the future, and probably for purposes that cannot reasonably have been foreseen at the time of collection. Data acquisition and management strategies must therefore be designed to maximise future usability — for example, by providing detailed metadata, making data discoverable and accessible, and choosing data acquisition procedures wisely.

This integration of data mining and science will have many flow-on benefits — most obviously by raising the profile of data management, but also through feeding the results of the data mining process back into the scientific cycle. Examples of this feedback might include information useful to planning future experiments, such as holes in data coverage and regions in sample space where the acquisition of new data will have greatest scientific impact, or using data mining techniques to alert scientists to interesting patterns in real-time data, such as remotely sensed imagery or scientific journal papers.