DOI: 10.1017/S1431927603447855
Microsc Microanal 9(Suppl 2), 2003

Image databases: What are they and what do they bring to microscopy

J.M.Carazo (1), P.de Alarcón (2) and M.Chagoyen (1)

(1) Centro Nacional de Biotecnología-CSIC, Campus Universidad Autónoma, 28049 Madrid, carazo@cnb.uam.es, (2) Integromics, Madrid Science Park, 28049 Madrid, www.integromics.com

Databases come into play when large amounts of data are to be organized in such a way that the retrieval of individual pieces of information becomes the main issue. In other words, databases are to simple "repositories" where the information is kept preserved, but complex structures where data items are organized such as they are accessed efficiently.

The field of databases is a classic one in computer science and informatics, and many paradigms on how to design these complex structures have been put forward over the years. In this way we encounter the so-called "relational databases", "object-oriented databases", and so forth. These terms refer to the basic manner the information is organized, as well as to the functionality that can be embedded within these structures. Of course, a basic understanding of the principle behind these terms is essential and they will be introduced in the tutorial through examples. Some recommended introductory readings can be found in www.cnb.uam.es/~carazo/msa2003, where a web site with links and references pertinent to the topic of this tutorial will be maintained.

Of course, a crucial point is "what kind of information can be properly stored"?. Traditionally, databases have contained text data. Typical examples are those databases containing information of the employees of a firm, or the ones containing our credit history. Even in the Life Science arena the best well known databases contain essentially text data, even if this "text" is somehow special and codifies for the sequence of bases of a gene, the sequence of amino acids of a protein, or triplets of (x, y, z) coordinates corresponding to atomic coordinates.

Obviously, in the microscopy field there is ample space for text-based databases. They provide, for instance, with ways to keep track of the experimental imaging conditions of a series of experiments, the person performing them, the project they are attached…. However, the real breakthrough comes when we also incorporate images into those databases, since they are the final microscopy information that needs to be organized. The problem appears, naturally, if that images are not structured entities such as a text, but essentially they are complex binary pieces of data. The traditional way to incorporate them is to define an entity called "blob", from binary large object, which essentially means to treat them as whole, blindly in a way. Of course, treated in this way images cannot be queried directly in a databases, and operations related to, for instance, comparing two images, are not trivial.

A further element of discussion appears when databases are tied into the daily operation of a resource in a way in which input to instruments and their outputs all follow a complex workflow in an automatized environment. Certainly, the need to manage complex workflows is not new, and a couple of decades ago the so-called LIMS (Laboratory Information Management Systems) were in use to help keep track of workflows in areas a diverse as the petrochemistry industry and centralized service resources. Still, the "type" of information that could be properly stored and analysed was basically alphanumeric. A new challenge appears when we want to organize the information contained in complex multidimensional images in such a way that not only they kept stored as "blobs", but also that they are analysed while being acquired and the result of this analysis is used as input to subsequent operations in the new automatized resource. Of course, this approach calls for a new modelling of the

image information such that a number of characteristics could be automatically extracted for later use. This area of research is usually referred to as "query by image content", and the performance of a number of such systems will be shown through examples.

Database systems, as well as LIMS, are complex and difficult to build pieces of software. They also tend to access the information in a very direct way, using mechanisms many time dependant of a particular hardware or a given operating system. Indeed, and specially referring to those cases in which the amount of data is very large and issues such as data integrity becomes an issue, there is ample room for sophisticated and expensive systems. However, with the current trend towards more open systems and public (although "academic") products, there are, on the one hand, a series of freely available database systems as well as LIMS, and, on the other, a clear desire of moving towards platform independent systems, mostly based on web. Some examples will be presented along the talk.

In the quest to keep combining more complex data types to move from information to knowledge, we are entering a new era of sharing data –or at least starting to have the capability to do so-, with the beginning of data grids initiatives that effectively extend the notion of "interoperability" to heterogeneous and distributed data sources, opening the way to complement in new ways diverse types of information.

In summary, databases are key elements in the process of data organization that is intrinsically associated with the new capabilities for mass production of data. Traditionally, text has been the single type of information normally stored in data bases, However, the capability to also organized the informational content of much complex data types such as multidimensional images is also being developed. Finally, databases of complex objects can be used now as part of LIMS systems to control and direct the workflow of the new automatized resources.

ACKNOWLEDGEMENTS

Our work on image databases is partly supported by the Spanish CICYT through project (P.N: Biotec., BIO 98-0761) , as well as the European Union through projects TEMBLOR (The European Molecular Biology Linked Original Resources) through grant EU (QLRI-CT-2001-00015) and IIMS through EU grant QLRI-CT-2000-31237.

REFERENCES:

For an updated list check at www.cnb.uam.es/ ~carazo/msa2003