On a prototype for a new distributed data base of volume data obtained by 3D imaging

Marabini, R. (a) , Vaquerizo, C. (a), Fernández, J.J. (a) and Carazo, J.M. (a,b)(*),
Ladjadj, M. (c), Odesanya, O.(c) and Frank, J. (c,d) (*)

(a) Centro Nacional de Biotecnología-CSIC, Univ.Autónoma, 28049 Madrid, Spain. (b) Department of Computer Architecture, Univ. of Malaga, Spain. (c) Wadsworth Center for Laboratories and Research, Empire State Plaza, Albany, New York, NY 12201-0509, USA. (d) Department of Biomedical Sciences, State University of New York at Albany, USA.

(*) Corresponding authors: carazo@samba.cnb.uam.es (Europe) and joachim@wadsworth.ph.albany.edu (USA)

ABSTRACT

In this communication we present a working prototype of a distributed data base of volume data from different forms of 3D imaging. Examples of three-dimensional information both from electron tomography of biological macromolecules and from medical imaging are already included in this development. This type of structural information is crosslinked, whenever applicable, to other sources of information, such as bibliographies. The solution we present here is sufficiently general to be applicable to data in a number of different fields of biomedical science.

1. INTRODUCTION

Nowadays a large number of tools is available which are able to produce three-dimensional images of an object of interest with resolutions ranging from milimeters to a few Angstroms. Among these we note the different forms of medical imaging, as well as electron tomography of biological macromolecules and various methods of 3D light and probe microscopy.

The final result of all these analyses is discrete data in the form of a volume. Other common features are the relative ease with which these results are obtained -due to the large number of instruments and research groups working in the field- and the fact that they frequently represent complementary information about the same object.

A further characteristic of these and most other efforts in modern biology, is that their full potential cannot be realized unless the specific information made available by these techniques is cross-linked to other sources of information. An obvious source of additional information is bibliography, which contains references by which the user of the new data base can get immediate access not only to the data themselves, but also to their complete description as originally reported by the authors. Other information crosslinks may be less obvious, since they may depend on the particular field of 3D imaging under consideration, and they will be discussed in the following sections.

These considerations have prompted us to propose the creation of a new data base of three-dimensional (3D) data from different forms of 3D imaging, whose first working prototype is described in this work. Although considerable effort is needed to set consensus formats in many areas, the development of this prototype already provides the structural biology research community with a test bed allowing a number of possible development

0-8194-1698-3/94/$6.00

venues to be checked. We hope that eventually, with the feedback from investigators engaged in all forms of biomedical 3D imaging, a data base system can be designed that is useful in many areas of application.

Since the authors of this work focus their research on one specific type of 3D imaging, namely 3D reconstruction of biological macromolecules, the examples that we will use to describe the organization of the data will be taken from this particular field.

2. PROTOTYPE DEVELOPMENT CONSIDERATIONS

The first and probably most important point to be considered when discussing possible ways of organizing information is the precise information content of the data. Although the final result of the different analytical approaches listed above is a volume, there are other important pieces of descriptive information that pertain to the experiment. We are referring to detailed information that describes the conditions under which the volume was obtained, as well as any pointers that allow the 3D volume to be linked to other sources of information.

In our case, that complementary information can be expressed in a textual form, and we have therefore decided to split each data set into two parts, one containing the discrete volume data and the other the descriptive text.

In the implementation described in this work, the textual and the volume information it refers to are separated into two files, with the link between them provided in the textual file itself. This is accomplished by introducing in the text file a number of especially coded sentences.

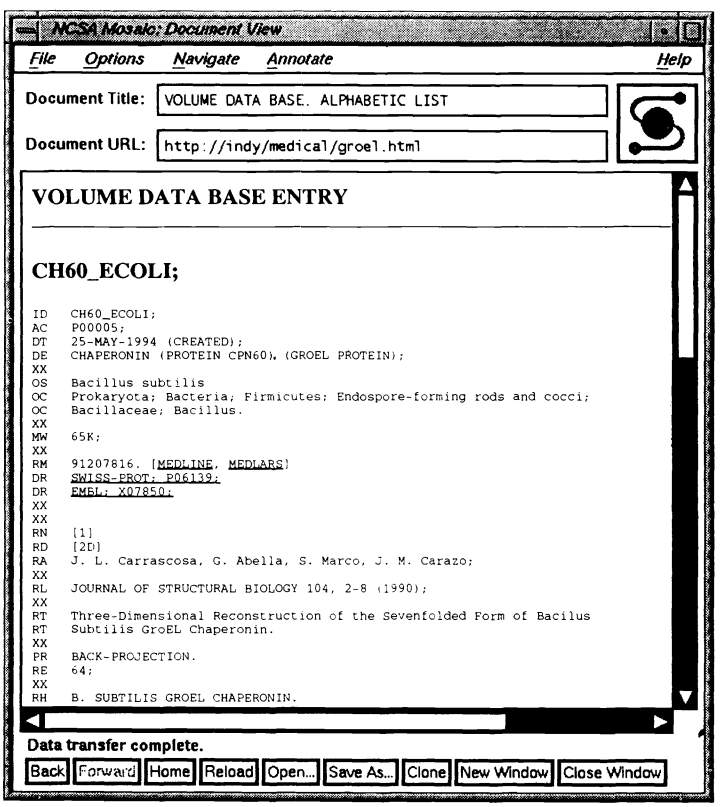

Figure 1: Textual section of a typical entry of our prototype data base. It corresponds to the GroEL macromolecule from *E.coli* and it shows information keyed by fields such as identifier (ID), creation date (DT), as well as crosslinks to other data bases, such as DR, pointing to other data bases.

As to the textual information, we have organized it into keyed fields. In this way it is possible to formulate complex queries to the database by interrelating the informational content of the different fields. Additionally, in an effort to integrate information from different sources, explicit links to other databases have also been provided.

An example of an actual text file of data obtained by electron tomography of macromolecules is shown in Figure 1, which corresponds to one of the structures we work with, the GroEL complex from *E.coli*. A close examination of the information contained in this text file illustrates the type of information that is being made accessible in this database. There are three important general organizational features that are readily noticed.

The first one is that each field starts with a two-letter code followed by a blank, and that the code is intended to be a short but

mnemotecnic description of the content of the field. This is a common practice in other biological data bases.

The second feature is that links to other sources of information are explicitly provided. This is accomplished through the specially coded "DR" field (for "Data Relations"), where the name of another database plus the entry number associated with that piece of information to be linked is provided. For the case of the 3D structure of GroEL, we are providing links to the sequence of the gene coding for each monomer (i.e., link to EMBL data base, entry number X07850), to the sequence of the protein itself (i.e., link to the SWISS-PROT data base, entry number P06139), and to the pertinent bibliographic source where the structure was first reported through MedLine data base. Certainly, links to other data bases could also be provided.

The third organizational feature is the presence, at the bottom part of the text file, of fields with an apparently "strange" syntax. These fields are not related to the data themselves, but to the precise data base querying and storing strategy that we have devised in this prototype. They will be described below.

Obviously, it is essential to conduct further research and discussions on the organization of the textual information in order to obtain a consensus on what those keyed fields should be and how their content should be defined. Flexibility has to be intrinsically provided, since there are many forms of 3D imaging and each of them will present specific needs. Parameters defining experimental conditions in one area might be entirely meaningless in another.

As to the precise data format in which the volume information is to be kept, the situation here is very complex, specially considering the potentially heterogeneous computing environment that might be typical of many applications. Two important considerations are in order in this respect.

The first one is that the data representation format itself has to be machine-independent, and that all sharable data have to be transformed to one such format. The machine-independent representation we have chosen for all original data contained in this prototype database is the well known hdf format (hdf stands for hierarchical data format[1] (However, images used for presentation pourposes are the so-called "in line" images, and they are stored either in the gif or jpeg formats.) In this way data can be safely treated and transmitted by a large number of applications.

The second important consideration is the internal structure of the data themselves. This is also a very difficult problem since a number of *de facto* standards coexist in this broad field, both among different application areas as well as within each one of them. In order to easily deal with this problem, we have simplified the data to the extent that, by default, only the pixel information itself, stored as one floating point number per pixel, together with the file dimensions, is kept. The problem of dealing with the different internal *standards* is then passed to the application layer, where a number of "translation tools" should be provided.

3. DATA DISTRIBUTION AND NAVIGATION STRATEGY

Considering the large number of 3D volume results obtained by the different kinds of 3D imaging, we have set means to allow for a distributed organization of the data. Our standpoint was that the data would finally be organized by stable "collection centers", which might be located anywhere in the world, but linked together via some form of data network.

From these concepts, a suitable implementation was then accomplished. Regarding networking, it was immediately clear that Internet was the obvious choice of network as it links most laboratories around the world. Another decision was the exclusive use of public domain software, since otherwise it might be difficult to impose specific products on different organizations. (Although we note that another consideration would come to the opposite conclusion: the data base management within each collection center, - depending upon the amount of data might be done more efficiently with commercial products).

With these considerations in mind, we have chosen WWW protocols to enable transparent data sharing (for a review see Obraczka et al., 1993[2]). Using WWW we allow information to be distributed over the whole of Internet yet to be retrieved transparently.

It was soon realized that the nature of the information provided in this data base was rather complex, mixing text with images and volumes. The need to provide the user with a powerful hypertext-based interface to help "navigating" through the data was then obvious. These considerations guided us to select for this implementation a hypertext-based WWW client such as X-Mosaic[2]

It is interesting to note that the apparent sophistication introduced by the use of tools such as WWW and X-Mosaic requires very little changes in the textual data entry. In fact, only a few lines at the bottom of this file are needed. Taking as example the text file corresponding to the entry of the structure of the GroEL macromolecule (Figure 1), these special directives are represented by expressions such as "http://INDY.CNB.UAM.ES/Base/AlphabeticList/GroEL_BSub_7/average.gif", or "ftp://INDY.CNB.UAM.ES/pub/DataBase/gro_EL_7/groe_7_vol.hdf"

These directives are especially handled and processed by X-Mosaic. The first type of expressions, those starting with "http" (from hypertext transfer protocol), are used to assemble the hypertext document, while those starting with "ftp" (from file transfer protocol) are used to establish direct links to the source where the data themselves are stored.

The task of selecting a particular piece of information that the user may be interested in within a data base with such diverse types of data as text, images and volumes, is certainly rather complex. For the sake of developing this first prototype, we have simplified this task by restricting the searching and inquiry possibilities to the information contained in the text file alone, that is, neither image nor volume information is "interpreted" in any way beyond the textual information provided by the authors.

Even with this restriction, it is not straightforward to provide text searching capabilities over distributed keyed files. The two most obvious candidates for public domain "searching engines" are WAIS (from Wide Area Information System; for a review see Obraczka et al., 1993) and SRS (from Sequence Retrieval System[3]. While WAIS does provide capabilities for searching over distributed text files, it does not handle keyed fields, that is, it provides only flat file searching. SRS, on the other hand, was especially designed to handle very efficiently keyed information; however, its present implementation did not allow for searching over distributed files.

Certainly, a possible solution would be to enforce the rule that all textual information should reside in all collection centers: this approach would eliminate the distributed organization at the expense of a certain complexity in maintenance (although data storage should not be a problem because of the limited size of text files). Still another possibility would be to develop some kind of searching engine especially suited for this task, although this solution does not seem to fit well within the world-wide effort to standarize information access tools. All these different possibilities are presently being considered.

4. EXAMPLE SESSION

In order to provide the reader with a clear understanding of the actual capabilities of our prototype, we are going to present in the following a possible session of data base access. We assume that the computer that is being used for this task is linked to Internet, and that an X-Mosaic client has been set up.

The database prototype can be reached just by starting X-Mosaic at the query laboratory, providing the address of the so-called "home page", which contains the specific directives to create the first "page" of the virtual document that is going to be assembled as an answer to the query. A suitable address or, more technically, a suitable URL (from "Universal Resource Locator") would be http://INDY.CNB.UAM.ES/Base.

The result of such a connection containing the prototype home page is presented in Figure 2, where the power of this interface in constructing a hypertext interface is evident, mixing images and text in a very informative document that is created dynamically each time the data base is accessed.

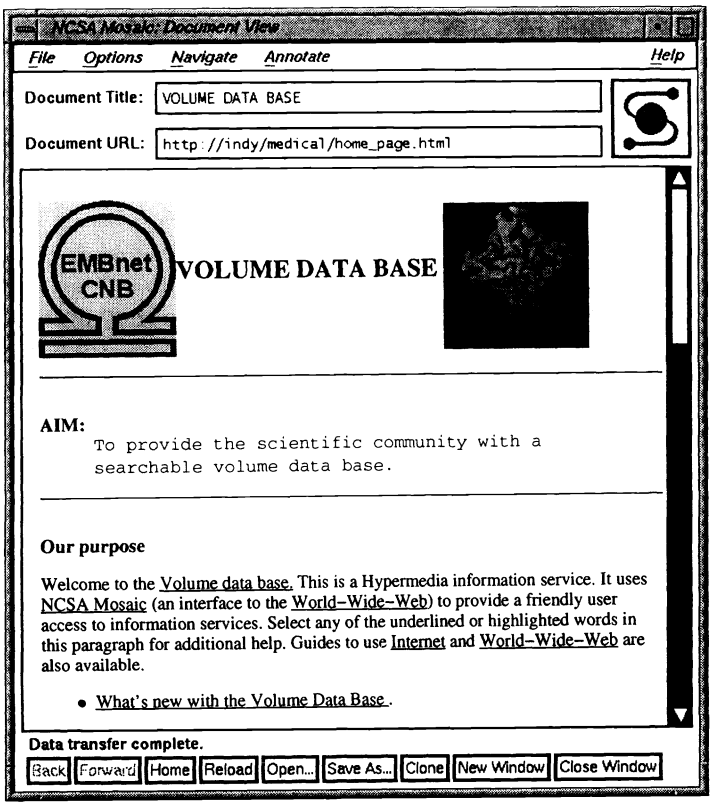

Figure 2: Home page of our WWW server. It states the content and aims of the prototype.

A number of data navigation options are then possible starting from this home page. As an example, we present in Figure 3 the result of a query. For each reference encountered in the database, either a 2D average, if the structural results only extend to two dimensions, or a 3D surface rendering of the volume are provided as icon image for further references. Clicking on any of the icon images will provide access to full scale images. A representative selection of sections through the volume is then shown. Once a given specimen has been selected, it is now possible either to interactively study the different sections of the volume data with tools such as "collage"[1], or to import the data to the local computer for further study. This import is accomplished via a direct link to the anonymous ftp server located at the collection center where the specific information required by the user is located.

5. WORK AHEAD

There are a number of specific problems that should be solved in order to make this prototype really functional. They involve both conceptual developments and implementation decisions. The first ones will insure the perdurance of the data base beyond technical possibilities at any given moment in time, while the latter ones will dictate what is the possible actual use of the data base and its general organizational principles.

A first and obvious conceptual development is the standarization of the "language" used in the textual files, that is, which fields are to be included and how their content is to be coded. Tools to "analyze" the 3D volume information are also needed, although their complexity is easily realized.

Another consideration is the format of the volume data themselves. Here we have followed the consideration that it might be difficult to really agree on a format, and we have then placed the format conversion problem right at the querying laboratory. Certainly, appropriate conversion tools should be produced and included with the prototype structure.

On the implementation side, actual tools such as WWW and X-Mosaic are already suitable for the main design goals of this prototype. Future enhancements on information sharing and presentation should therefore be easily accommodated. However, the problem of how to efficiently search over distributed keyed files should not be overlooked, since the appropiate tools are not yet there. This situation may become much more complex should 3D data analysis tools were already developed.

Last but not least, "political" considerations on data base maintenance and distribution policies are essential components of this project. They will require extensive international cooperation to set up and maintain a defined and stable framework that should allow this new data base project to prospere.

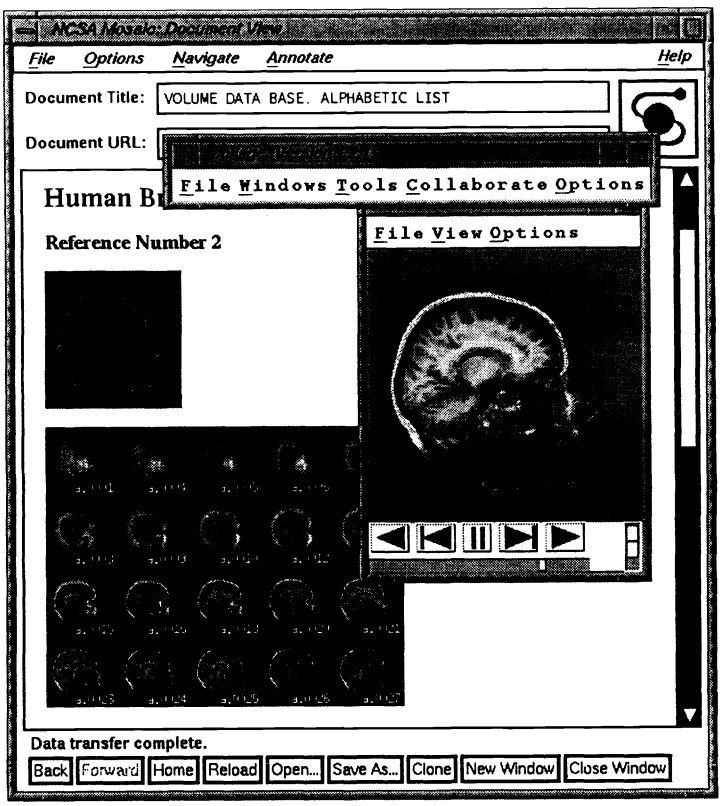

Figure 3: Typical result of a query to this prototype. Access to further information on the volume data is provided, as well as links to volume visualization tools such as collage and direct links to ftp servers.

6. CONCLUSIONS

We have presented in this work the prototype of a distributed data base of volume data obtained in different forms of 3D imaging. On the conceptual side, the information is organized into two parts, one containing the volume data themselves while the other contains the textual information as well as links to the volume data. Organizational principles allowing for a data distribution over a number of specialized "collection centers" have been devised. On the implementation side, only public domain tools such as X-Mosaic, NCSA Collage and SRS are being used.

This implementation is already operational at the two participating laboratories in Europe and the US on a trial base, providing the biology community with a framework within which future developments and enhancements may be tested.

7. ACKNOWLEDGMENTS

Supported, in part, by Spanish DGICYT Plan General de Promoción del Conocimiento, grant number PB91-0910 (to JMC), by the European Molecular Biology Network (EMBNet, European Union BRIDGE project number BIOT CT-910273) (to JMC), and by the National Science Foundation grant number BIR-9219043 (to JF). We acknowledge discussions on the general issue of 3D data bases with the participants of the 1993 Gordon Conference on 3D Electron Microscopy.

8. REFERENCES

[1]NCSA, HDF Calling Interfaces and Utilities, Version 3.0, National Center for Super Computing Applications, Univ. of Illinois at Urban-Champaign, Nov. 1989.

[2]K.Obraczka, P.B.Danzig and S.-H.Li, "Internet Resource Discovery Services", IEEE Computer, 26:8-22, 1993.

[3]T.Etzold and P.Argos , "SRS, an indexing and retrieval tool for flat file data libraries", CABIOS, 9:49-57, 1993.