Graellsia, 59(1): 15-27 (2003)

WHAT CAN BIOINFORMATICS DO FOR NATURAL HISTORY MUSEUMS?

J. M. Becerra*

ABSTRACT

We propose the founding of a Natural History bioinformatics framework, which would solve one of the main problems in Natural History: data which is scattered around in many incompatible systems (not only computer systems, but also paper ones). This framework consists of computer resources (hardware and software), methodologies that ease the circulation of data, and staff expert in dealing with computers, who will develop software solutions to the problems encountered by naturalists.

This system is organized in three layers: acquisition, data and analysis. Each layer is described, and an account of the elements that constitute it given.

Key words: Taxonomy, Systematic, Bioinformatics.

RESUMEN

¿Cómo puede ayudar la bioinformática a los Museos de Historia Natural?

Se presentan las bases de una estructura bioinformática para Historia Natural, que trata de resolver uno de los principales problemas en ésta: la presencia de datos distribuidos a lo largo de muchos sistemas incompatibles entre sí (y no sólo hablamos de sistemas informáticos, sino también en papel). Esta estructura se sustenta en recursos informáticos (en sus dos vertientes: *hardware* y *software*), en metodologías que permitan la fácil circulación de los datos, y personal experto en el uso de ordenadores que se encargue de desarrollar soluciones *software* a los problemas que plantean los naturalistas. Este sistema estaría organizado en tres capas: de adquisición, de datos y de análisis. Cada una de estas capas se describe, indicando los elementos que la componen.

Palabras clave: Taxonomía, Sistema´tica, Bioinformática.

Taxonomy, the science of naming and classifying organisms, is the original bioinformatics and a fundamental basis for all biology. (Mallet and Willmott, 2003)

Introduction

Natural History can be regarded as a broad term that brings together two classical scientific activities: Taxonomy and Systematics. These activities treat a rich set of multimedia data: alphanumeric, for quantities (lengths, weights, physical measures of the environment…) and names (scientific names, locations, etc.); images, drawings of specimens or maps; sounds; and lately, video. Also, the origin of this data is mixed: obtained from sampling in the field, from already established scientific collections, genetic data from

* Dirección General de Universidades. c/ Serrano, 150. Madrid – 28006 (Spain). e-mail: josem.becerra@univ.mecd.es

laboratories, data from publications in journals, etc. And all this information is stored in many different ways and formats: isolated computers with data managed by "homemade" programs, large or medium database systems (Sarasan & Neuer, 1983), on plain paper (old labels in collections, for example), video tapes or audio cassettes, the specimens themselves, etc.

This high level of variation in data source poses a real difficulty when dealing with all this data as a whole in order to obtain integrated results. It is also very difficult to access it, even when it is stored in computers.

It was supposed that the advent of the Internet era would make all this information easily accessible to anyone (Godfray, 2002), but, in the end, the use of this resource by naturalists is plagued with the same problems associated with the computer initiatives in Natural History (Mallet & Willmott, 2003): many projects are completely unrelated, and always neglect the "difficult" basic data, specially that in antique collections. They focuse all their efforts on giving textual results, and forget to improve the activities and methods that feed the data into Natural History, methods that usually are very slow and manually oriented. For an introduction to the use of computers in Natural History, see Bello (1996).

The Human Genome project has made one term very well know: Bioinformatics. Bioinformatics addresses problems related to the storage, retrieval and analysis of information about biological structure, sequence and function (Altman, 1998). If we look at many other definitions, we see that bioinformatics is usually associated with the treatment of DNA data exclusively. We are against this restrictive view of the term bioinformatics, but it can be considered as a reflection of the situation of informatics in Natural History; that is, there is neither unified treatment of the data, nor generic automatic methods which can alleviate the problems of data acquisition and treatment in Natural History. In the field of genetic bioinformatics a real effort has been made to integrate the data at their disposal, and projects for making the information from very different sources easily accessible, as if they came from only one source, are being set up (Anonymous, 2002; Wilkinson & Links, 2002; Siepel *et al.*, 2001). There is no reason why Natural History cannot achieve these goals.

The Natural History bioinformatics framework

We propose here the establishment of a Natural History bioinformatics framework, consisting of (Fig. 1):

- Hardware, not only computers, but also interfaces among computers and the regular equipment used by naturalist in their daily work.
- Software, that, on the one hand eliminates the need for repetitive tasks to be performed manually, and on the other, allows naturalists to pose questions in a natural way; integrates the data from very different sources analysing it to give **multimedia results** (that is, in the format that they appear in life) to answer not only research questions but also environmental problems and public interest.
- Computer scientists, who integrate the two elements above, and give a framework and not just a collection of isolated pieces, as well as a set of methodologies that can be used by the naturalist to accomplish all these tasks.

This system could use many different techniques and hardware which is well proven in the business field as in other research fields with many years of experience in the treatment of information by computer. The framework could be organized in three completely interrelated layers, with the bioinformatics department permeating all three (Figure 1):

- the acquisition layer.
- the core layer.
- the analysis layer, where much of the work remains to be done.

Acquisition layer: Obtaining the data

The first step in any of the tasks in Natural History is obtaining data. And in this step is where problems arise. The sources for data in Natural History Museums can be, amongst others:

1. Samples
2. Scientific collections
3. Publications
4. Work in the laboratory
5. Other sources

At the same time, one source can be the origin for other source, as in the case of sampling and scientific collections. Let's see what the problems and data are and what solutions bioinformatics can provide.

Sampling in the field. Figure 2 shows a hypothetical flowchart of work for a researcher in the field. When working in the field, a researcher must absorb a lot of information and manage a lot of devices:

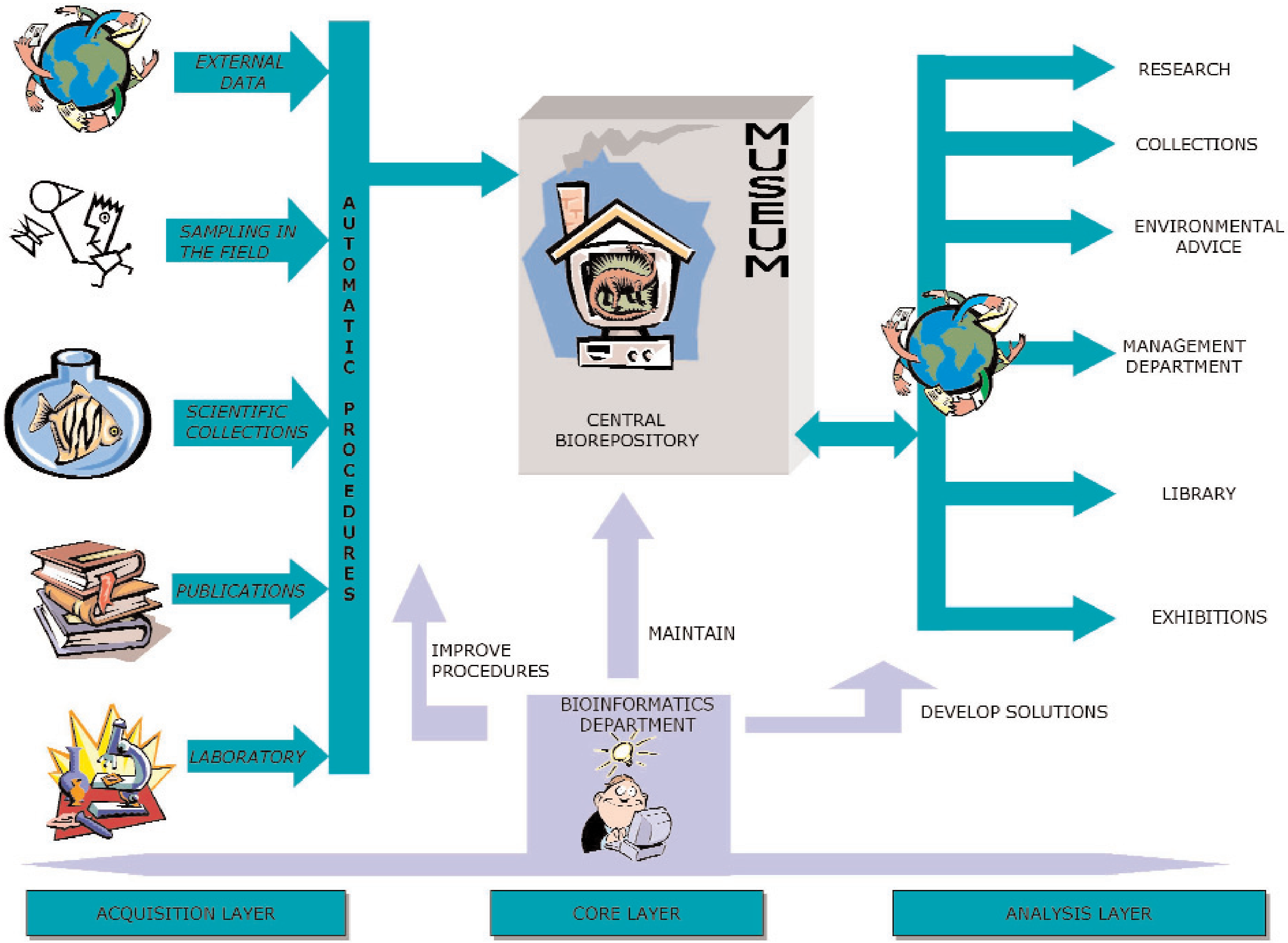

Fig. 1.— Hypothetical structure of the Natural History bioinformatics framework, with the three layers shown.

Fig. 1.— Estructura hipotética de la estructura bioinformática de Historia Natural, mostrando las tres capas.

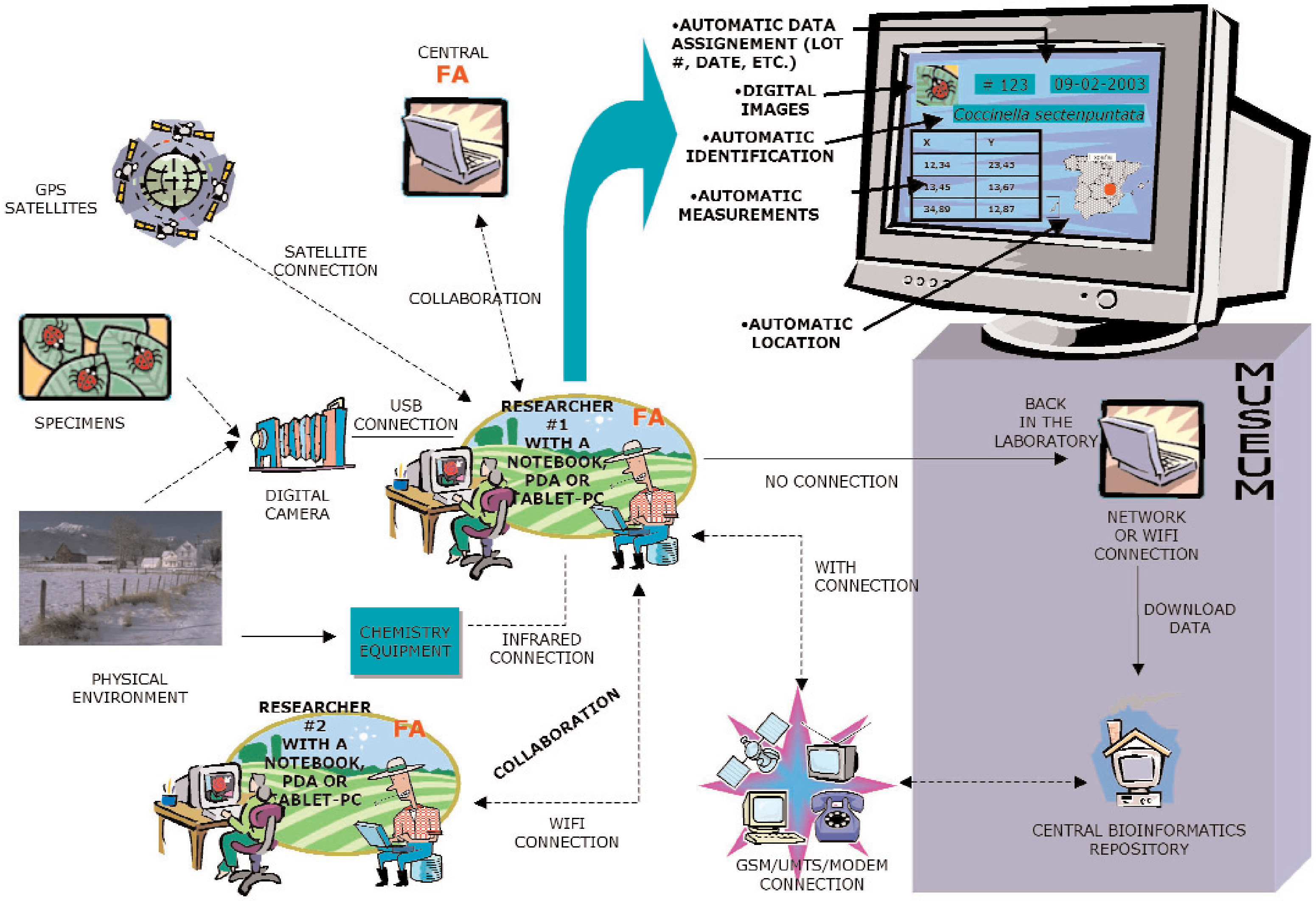

Fig. 2.— Hypothetical flow of data, methods and devices in a field sampling work. FA: Field Assistant. Wi-Fi: Wireless Fidelity.

Fig. 2.— Flujo de datos hipotético, métodos y dispositivos en un trabajo de muestreo en el campo. FA: Asistente de campo. Wi-Fi: Conexión inalámbrica.

- management data: location, date, and lot number.
- environmental data: physical and chemical, etc.

One of the big problems in sampling is the great quantity of data obtained, which usually is written down on pages of notebooks, or left pending arrival at the laboratory. This data can also be obtained from many devices carried by the naturalist. Bioinformatics can help in this tasks developing the "**Field Assistant**" (FA from now on); this FA would be organized around a computer device, like a notebook, a PDA or a tablet PC, in general any device with the capacity to store software modules and human-computer interfaces developed in some cross-platform language like Java. These devices would also be tested to make sure they can withstand for hard-environmental conditions, usually found on sampling campaigns. This FA would have a Wireless connection with other FAs, and a modem card or a card to connect a mobile phone to send and receive information from the central repository. Furthermore, this FA would have physical interfaces for the field equipment, like physic-chemical devices, printers, GPS devices, digital cameras, etc. This connection would ideally be carried out by a WIFI connection, or an infrared connection, or in the worst situation, with an USB connection. One of these FAs (a powerful notebook, for example) could be used as a central field station, acting as a central repository for the information collected by the scientist, while the scientist themselves use a thin FAs (like a PDA). The software interface of the FAs would be the same as in the central repository. As much information as possible would be collected and assigned automatically, such as dates, lot numbers, location (UTM and name of places), etc. A standard set of modules in the FAs would work on the data collected to generate derived information: from example, from digital images, the modules would be capable of identifying organisms, obtaining automatic measures, giving clues where to continue with the sampling, all of this based on the information collected and the experience already stored in the central repository. The information collected could be downloaded to the central repository wherever a connection is at one's disposal, and where more processing power is available, to render 3-D images or extract patterns from the data. When it is clear a connection will not be available, a medium for storing the information would have to be provided: a CD-RW device or magnetic tape, or a high capacity hard disk (although in this case, a backup device is advisable). When the sampling team (the researcher and the FAs) arrive at the Museum, they simply connect the FAs to the central network (via physical connection or using a WIFI connection) and run the corresponding modules, and the data will be integrated into the central repository.

Scientific collections. This is one of the most troublesome sources of data. Apart from some recent collections, almost all institutions are starting the automatization of collections with a certain quantity of material already stored, with documentation in several places, from collectors' notebooks to simple cards; the information is also different from source to source, always in the need of previous normalization and study, in order to treat old data, like locality data.

Data here can be stored on paper, already in computer format, and in the specimen itself.

In the case of paper, one solution is to scan the paper (notebook or label) to an image, and then, try to apply OCR over the image. Depending on whether or not the data is handwritten or typed and the degree of conservation, the results will be very variable. Anyway, many studies are being conducted today to improve the efficiency of OCR processing of handwritten texts (Hahn *et al.*, 1999) and degraded documents (Natarajan *et al.,* 1999). Furthermore, there actually are high capacity scanners with trays that can feed the scanner with a great amount of cards or papers, reducing the time needed to acquire images for the OCR. Paper can also contain drawings of animals or plants. This information can be scanned (as has been done with rare books in "El Archivo General de las Indias", in Sevilla) and the images obtained stored in the repository.

The other source of data in collections is the specimen itself. Very often, the specimens are irreplaceable. Taxonomy is very much about phenotype data, so the visual information is very important. In this field, computers can help a lot. Ranging from the microscopic to the macroscopic world, there are many tools for obtaining visual information from objects (in this case, organisms), both 2-D to 3-D. This information must be of good enough quality to serve for comparison purposes with real specimens or images of them, and for extracting information, like measurements and colour.

In recent years, many collections have undergone a process of computerizing the available information. This information can be taken advantage of the repository. One important task here is the development of modules that allow the downloading of information to the central repository, and integrate it in the new structure.

The management of the collection itself would improve with the use of computers. Basically, collections are warehouses of objects, with the same processes as in any big market store: new entries, loaning of objects, replacing of spoiled objects, etc.

The flow of data from collections to exhibitions could be facilitated with the right design and software modules; knowing what specimens are available and in what degree of conservation would help us decide what to include in an exhibition, as well as this, all the information associated in the repository to the specimen would improve the production of the presentations.

Publications. There is a great deal of data and information in publications. With the advent of the computer era to the scientific journal world, many publications have digitised their issues. The information can be classified in two kinds: data about the reference (author/s, title, date, pages, etc.) and the text of the publication. The first kind of data can be easily downloaded into databases, while the second can be stored in PDF format, and indexed for keywords search, linking the text to the reference. Even more, developing personal electronic libraries is a significant step ahead, only hampered by the fact that many vital taxonomic publications are very old and not digitized yet.

Work in the laboratory. With the arrival of the DNA era, data obtained from this source is now not only quite common but essential (Mallet & Willmott, 2003). The devices used in these methods are usually connected to computers and only modules that interface with the central repository to store the information in the right format need to be developed. As in the case of collections, not only textual data will be stored, but also sequences, images and the like obtained from experiments and analysis.

Other sources. There is plenty of information out there: the Internet. There are many sites devoted to showing Natural History information. In the beginning, the method to recover all this information was known as **web mining**, an extension to **data mining** (Kohavi *et al.*, 2002; Yi & Sundaresan, 2000) applied to web. This technique allows us to search the Internet with a series of criteria, using agents or bots, and the information obtained integrated into existing databases. With human interaction, the agent can learn what the necessities are and improve the results on every new “trip”. This vision is an ideal vision, as the format of the data in Internet is so variable and makes it very difficult to retrieve this information, with no structure underlying it. **Web mining** gave way to **Web farming**, the systematic refining for Web-based information for business intelligence (Hackathorn, 1999). Web farming assumes that obtaining information from the data present in Internet is not an easy task, and this work needs a structured and well-planned approach. But the rewards would be great. And with the introduction of XML, databases and, in the end, real structure to the information contained in the web, the benefits will also be undoubtedly important. For example, accessing scientific names databases like the Fauna Europea or Directories of Taxonomists, which are accessible on the web at the moment.

Standard processes and programs need to be developed in this layer, in order to permit an easy integration of the data obtained everywhere into the central repository, as well as loading the FAs with the information needed for a sampling campaign: maps, identification systems, statistics of previous campaigns, in order to detect important variations, etc.

Core layer: Biological Information Repository

Until now, we have dealt with many kinds of data sources, represented as text, images, sounds, etc. But where do we store all this data? In the Biological Information Repository (BIR from now on). This is the heart of the bioinformatics framework system and what gives coherence to the whole. If this layer did not exist, we would only have more information to use than before, but with the same problems as formerly: information scattered around, with no inter-connections and no interrelations. So, one vital task is the development of flows and methodologies that permit the integration, and specifically normalization of the acquired data, in order to give consistency and which can easily be entered into the BIR. To achieve this goal, it is very important to conform to open standards.

BIR is a concept based on three principles: hardware, software and people.

Hardware. Figure 3 shows the proposed structure of the BIR: the core of this system is the database repository, a data warehouse where all the data is stored, and from where the processed information is obtained. This database repository must be based in a central computer, with enough processing power to give answers to all the elements of the subsystems; we must take into account that we are going to manage gigabytes of data (text, images,

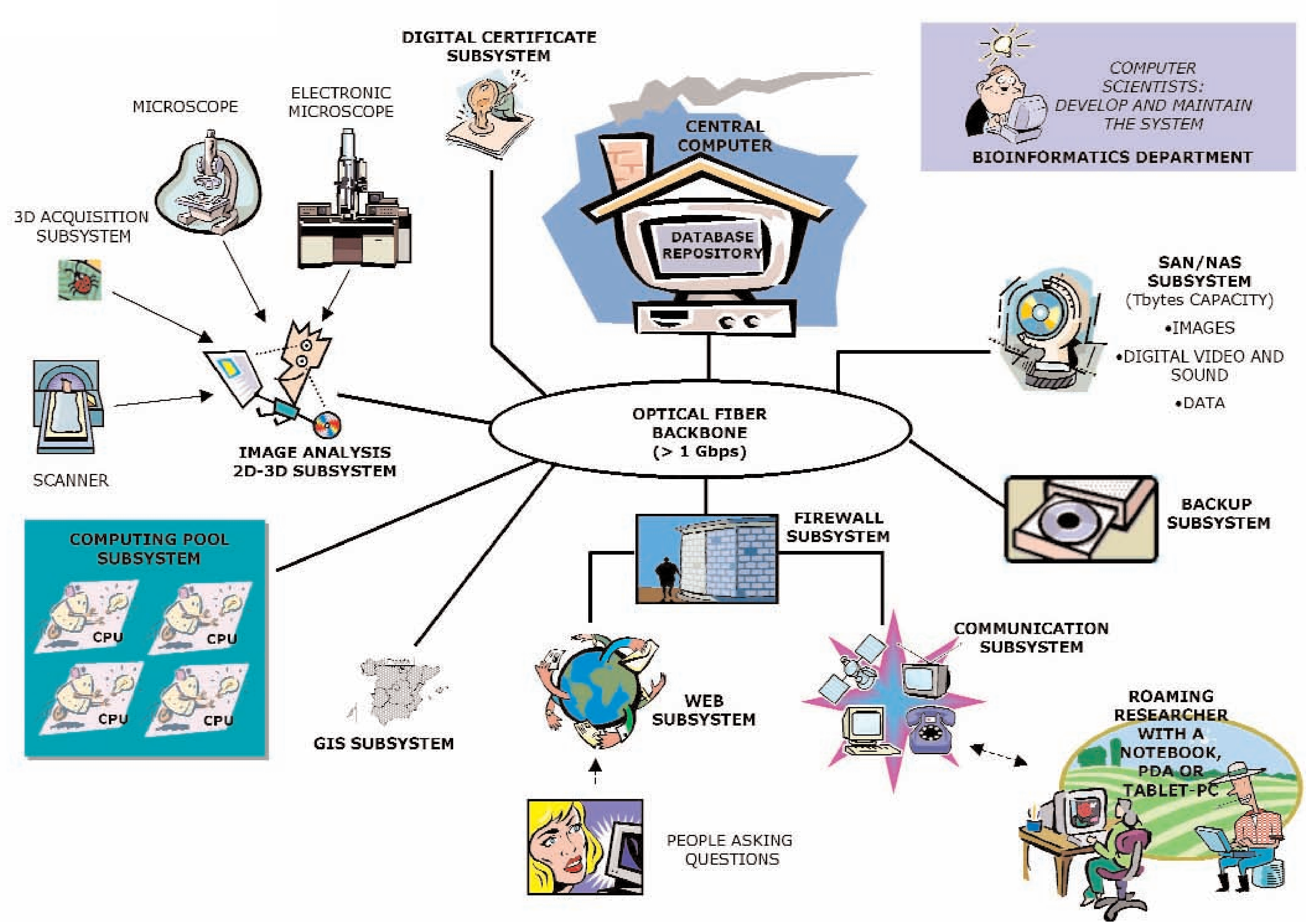

Fig. 3.— Proposed structure of the data layer, showing the subsystems associated to the Biological Information Repository (BIR).

Fig. 3.— Estructura propuesta de la capa de datos, en la que se muestra los subsistemas asociados al Repositorio de Información Biológica (RIB).

sound, video, etc.). To physically store this data, the best solution is to use a SAN/NAS system, that is a rack with an array of disks, which allow for a capacity of Terabytes of information, directly connected to the network, and which can grow easily by adding more disks to the rack, or by adding more racks; this kind of storing device has several benefits, among them, that it is transparent to the operating system of the machines. A good, powerful backup system is also needed, as we are going to store a great amount of data, and all of it is vital, as it is data that is difficult to obtain again (the results from a sampling campaign are unique for that campaign and can not be reproduced again). As the amount of data that is going to be managed is great, the processes for generating derived information will need high processing power. This will be accomplished with the so-called computing pool subsystem. This subsystem will benefit from developments in parallel computing improvements. This subsystem could be established around a pool of computers, whic can process the data in parallel, running in a coordinated way among them, or it could be established around a real multiprocessor machine. Right now hundreds of processors can be installed in a computer to work in SMP (Symmetric Multi-Processing) way. The software would, of course, be ready to exploit this feature. In this way, very big 3-D images can be rendered, or patterns can be extracted from huge amounts of data. The experiences in the genetic area could be taken into account (Carina, 2002).

It is very important to have a image analysis subsystem available. As commented above, there are now available many devices for acquiring images, ranging from the microscopic scale (electron microscope or light microscope with the use of digital cameras) up to the macroscopic scale (3-D acquisition systems like those used in Medical imaging – Duncan, 2000). All these images would be directly stored in the database repository, using standard processes to improve the speed of capture, and the quality of the images. Also included in this subsystem would be the scanners necessary for capturing the images of documents mentioned above.

Other subsystems included in the BIR are:

- a digital certificate subsystem in order to permit the secure exchange of information, which the naturalist could access to level of security clearance she/he has, a PKI (Public Key Infrastructure) must be setup. This infrastructure would identify the naturalists, and encrypt the information that is circulating on the system or going out of it.
- a GIS (Geographical Information System) subsystem, as the geographical information and distribution is very important for Natural History work.
- a web subsystem, where all the relations with the outer world (extranet and internet) will take place.
- a communication subsystem, to allow the researchers using their FAs to retrieve or download information into the system.
- a firewall subsystem, to protect the inner area from the dangers of the outer world.

All these subsystems must be connected by a high-speed line, using optical fibre, with speeds starting from 1 Gbps, as the traffic is going to be very high, due to the great amount of data used.

An important point to bear in mind at the time of choosing a computer system to hold the repository is the possibility of adding more function by just buying and inserting new cards; this implies that the system should be based on a rack, with the components being just option cards inserted into this rack.

All the infrastructure must be placed in an adequate environment: a room with air-conditioning to dissipate the heat generated by all the devices, as well as a UPS system that guaranties the continuous work of the system and protects the machines from the fluctuation in the electrical current.

Software. The core of the BIR consists of two kind of modules:

- the database management software
- the analysis software.

We will deal with the second kind of modules in the analysis layer section.

The database management software (DBMS) must be based on open standards. Must be able to work with languages like Java and must be multiplatform. It should support multimedia types of data such as native data that can be stored directly into the database. The interrogation language must be standard SQL. We could consider two possibilities in this regard:

- freeware software, like MySQL (www.mysql.com)
- commercial software, like Oracle (www.oracle.com) or DB2 (www.ibm.com).

Each has its benefits and drawbacks, and an in depth study must be conducted in order to make an informed decision.

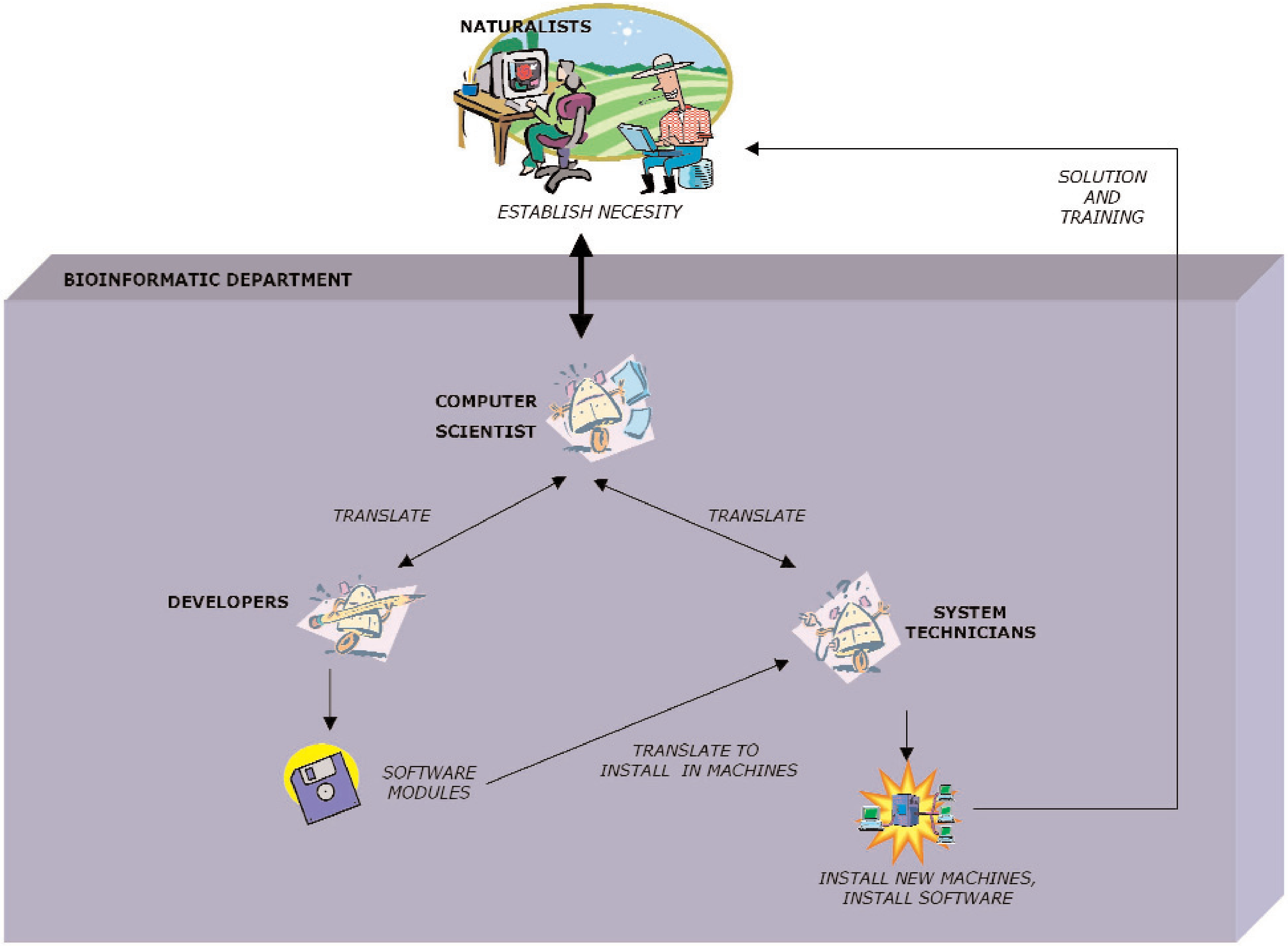

Fig. 4.— Structure and flow of tasks and information in the Bioinformatics department in the Natural History Museum.

Fig. 4.— Estructura y flujo de tareas y de información en el departamento de Bioinformática del Museo de Historia Natural.

People. A bioinformatics department could be based on the following staff (Figure 4):

- Analysts and programmers, to develop all the software interfaces (for human and machines). Our experience shows that it is better to have people in-house for these tasks, than to establish closed projects, as the experience and knowledge of the working environment do not need to be relearned for every project. They should be fluent with open standard, cross-platform, network ready languages like Java.
- System technicians, in charge of building the system and maintaining it. The people in this area would be responsible for installing new machines, establishing specification for new devices, and maintaining the daily functioning of the system. Part of the support in this area could be outsourced (broken down machines, problems with installation, etc.).
- Computer scientists, who act as a interface between the naturalist and the computer experts. They will be in charge of coordinating the work of the two other groups in the bioinformatic department, developing new methods and ideas for facilitating naturalists tasks, etc.

Analysis layer

This layer is where much work is to be done by the bioinformatic department, especially by the group of programmers, working together with the scientist. The main idea in this layer is that the software developed must mimic the way naturalists work, i.e., the system must allow questions in a natural way and the answers must be multimedia, and of course, adapted to the audience seeking the information. Based on the kind of access and the information searched for, we can distinguish three different approaches to the analysis:

- intranet: in this area we include the naturalists, the curators, exhibition staff, librarians and management staff. These groups of people will have total access to the information stored in the BIR, usually locally or when in the field. The local access allows for high-speed connections to the BIR, through optical fiber. In this way, the management of high quantities of data at real-time is ensured.

Most of the work to be done here is on the naturalist-BIR interfaces. The main idea behind the development of analysis software is that the naturalist must be freed from “fighting against computers”. We propose the development of a research environment, where the naturalists have small generic pieces of software, resembling the objects used in the real life (for example, a generic object could be a specimen, on which the naturalist works, or the data of a sample in the field, etc.). On these “objects”, the naturalist will perform actions, like obtaining images, measurements, etc. The naturalist will only need to drop objects onto a desktop and then subject them to actions to obtain results. All these modules could be shared among scientists scattered through the system. They will only need to worry about choosing the right objects, and filling data into the BIR through objects, and later, making analyses (actions) on the data (objects) introduced in the BIR (Figure 5).

Areas where research must be done and resources must be invested are:

- algorithms for searching for images, for automatic identification of specimens (Lew, 2000).
- automatic 3-D coordinate acquisition for morphometric analysis.
- automatic sequence identification, by accessing the main sequence databases located around the world.
- integration of very different types of data for phylogenetic analysis.
- 3-D rendering of specimen images, with the idea of creating virtual types that can be rotated and used to extract derived information.
- Better image and video compression algorithms and methodologies (Wactlar *et al.*, 1999; Talagala *et al.*, 2000), in order to construct Terabyte visual databases of Natural History images and videos.

This schema can be applied to all the categories we have dealt with before: curators, exhibition staff and management staff. As all the information is stored in the BIR, everyone can share the information, and take decisions, be it assigning money to exhibitions, choosing the right resource for an exhibition, deciding the next sampling campaign to fill gaps in the collection, or making decisions on the policies of the Museum. Other departments in the Museum, like the photography department, should have a very close relationship with the bioinformatics department, as photography is more digital nowadays.

To ensure that the user that is accessing the resources has the rights to do so, a digital certificate policy is a must.

The close collaboration among the development team and the scientists is very important for the

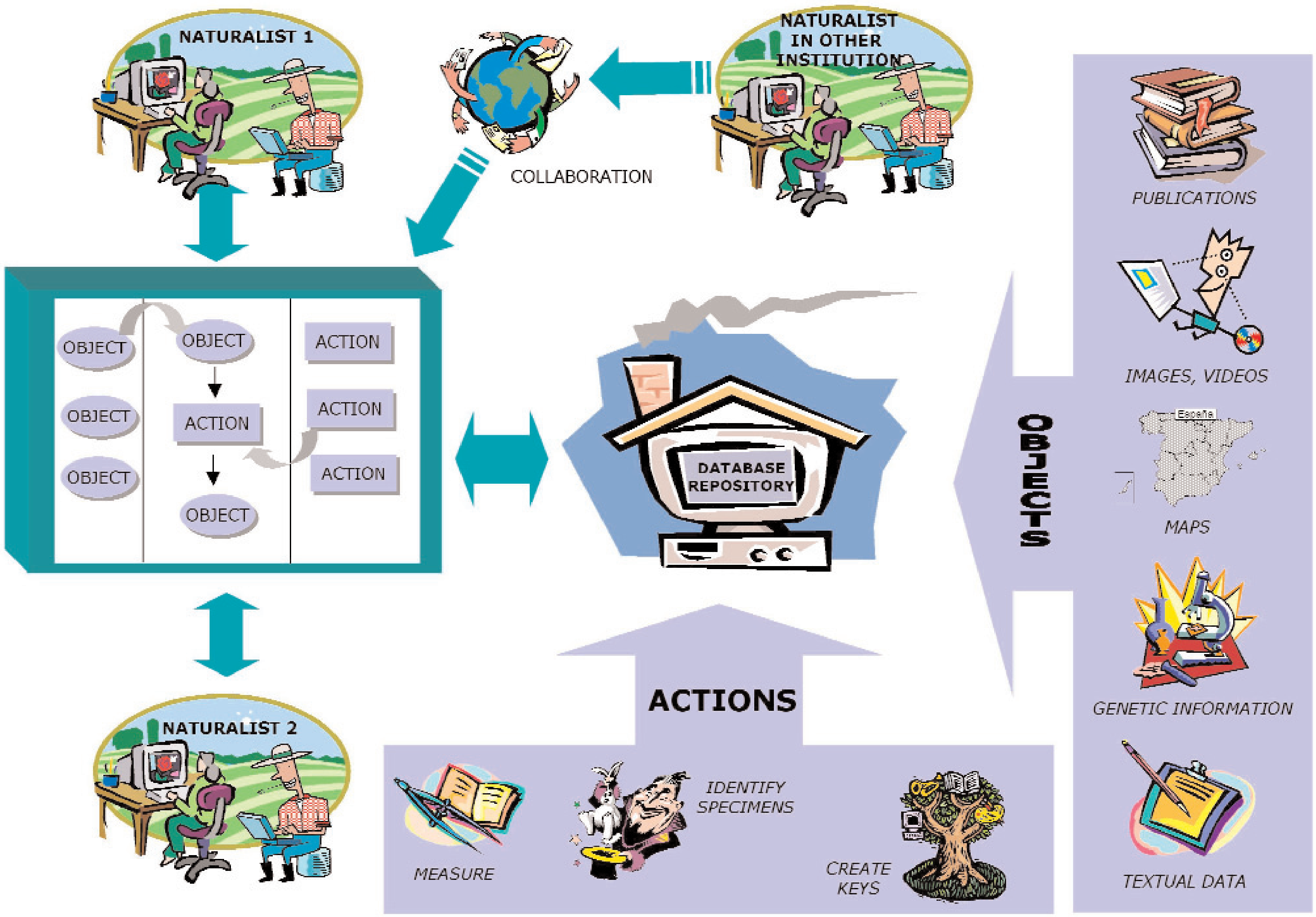

Fig. 5.— Structure of the intranet area of the Natural History bioinformatics framework, showing the concept of human – BIR interface, with the use of objects and actions.

Fig. 5.— Estructura del área de intranet de la estructura bioinformática de Historia Natural, en el que se muestra el concepto de interfaz hombre – RIB, con el uso de objetos y acciones.

success of this layer. Differences will arise, as in any collaborative project, and in these cases, the role of the bioinformatic/computer scientis is vital in order to translate the needs of the researchers to the development team.

Another possibility, exploiting high-speed communications (not only inside the Museum, but also outside with other institutions, using the Internet-2, capable of reaching 10-100 times the current speed) is real-time collaboration among scientists, and the possibility of using expensive resources located in other places (institutions, cities, even countries). Even when working inside the Museum, the use of instant messaging and multiuser videoconference can upgrade the level of collaboration.

- **extranet**: governments and companies that need information to make decisions or fill contracts or do business integrate this area. They will have an intermediate access to the information, with tools that allow for a broader range of questions than in the case of the Internet users, or they will order studies, and will obtain the results in a way they can manipulate. Their access will always be remote. People asking information from collections fall into this category. This area is very important nowadays, as government agencies are realizing that environmental care is a priority nowadays. It is also a good way of obtaining funds for studies. Good directories of taxonomists will allow environmental agencies to find specialists to solve specific questions or identify specimens. Also search in images databases for the automatic identification of specimens (to find out whether a plague is attacking, for example) will be very useful for farmers.

- **internet**: In this area we include the exhibitions at the Museum as well as the information laid in educational webs. This information is fixed and could no be changed by this kind of user. Usually, the questions that they can ask are a subset of the ones stored in the BIR and these questions are more elaborated. Their access would be local when the users are present in the exhibitions, and when they are dealing with web sites would be remote.

One important task of this area is the educational one. Schools are nowadays getting connected to the Internet through high-speed lines. The Internet area of a Natural History Museum must act as a very important educational resource for the students, and help to form them a conscience that the environment must be respected and must be care of. The level of knowledge put in this web can be graded taking into account the age of the student who would access to the Museum internet: ranging from games that allows identification of animals and plants, to high quality and detailed identification keys for high-school students. Using chat technology the researchers at the Museum could also give conferences to students, and answer questions related to their work at the Museum. The Internet area of the Museum would also acts as a Natural History library that allows the student to seek information for their school works. The publications, reports, etc. created by the Museum scientific staff would also be accessible to the University students, so they would use these in the studies and works, learning in the process the methods and techniques used by researchers. In this way, these students would become new researchers, attracted by the new look these old disciplines are changing into.

The Museum web site would have virtual tours, which allows the visitor to virtually work on the Museum specimens, with high resolution 3-D images, that can be rotated and with information about every specimen associated to the images:the environment where it lives, interrelations with other species, etc.

Conclusions

Natural History is in the need of new and well-established computer methods and techniques, that allow to alleviate many manual and routine tasks, and all the data that now are disperse in many computer systems, should be integrated in one unique bio data system. This will allow to concentrate resources, and the presence of bioinformatics (in the sense of a person that understand the biological problem and translate it to a computer solution) will make this system possible. As has been done in the Genetic area, resources (in the form of money, mainly) must be set, and standard (in methodology, software, and devices) must be established and disseminate among the Natural History community, to make exchange of data an easy task, and form the public opinion and governments with objective data. The researcher should be freed from dealing with computers and cumbersome software, and concentrate in studying data and collaborating with other researchers in order to solve scientific problems, and that these results could be easily published and disseminated in the scientific community. And the data generated by one researcher in their analysis could be straighforwarg used by other researcher, even working in a different discipline.

ACKNOWLEDGMENT

Ana Camacho and Antonio G. Valdecasas for their useful comments and discussion on this paper. Mark Creber made a revision of the English.

References

ALTMAN, R. B., 1998. A curriculum for Bioinformatics: The time is ripe. *Bioinformatics*, 14(7): 549-550.

ANONYMOUS, 2002. Bioinformatics: Bringing it all together. *Nature*, 419: 751-757.

BELLO, E., 1996. *Herramientas taxonómicas por ordenador. Los Halacáridos subterráneos continentales en la Península Ibérica*. Tesis Doctoral. Universidad Autónoma de Madrid. 510 pp.

CARINA, D., 2002. Information overload. News feature. *Nature*, 417: 14.

DUNCAN, J. S. 2000. Medical image analysis: Progress over two decades and the challenges ahead. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1): 85-106.

GODFRAY, H. C. J., 2002. Challenges for Taxonomy. *Nature*, 417: 17-19.

HAHN, S.-H., LEE, J. H. & KIM, J.-H., 1999. A study on utilizing OCR technology in building text databases. *10th International Workshop on Database Expert Systems Applications*, Florence: 5 pages.

HACKATHORN, R., 1999. Web Farming. *DB2 Magazine.* 1-10.

KOHAVI, R., MASAND, B., SPILIOPOULOU, M. & SRIVASTAVA, J. 2002. Web mining. *Data Mining and Knowledge Discovery*, 6: 5-8.

LEW, M. 2000. Next-generation web searches for visual content. *Computer*, 33(11): 46-53.

MALLET, J. & WILLMOTT, K., 2003 Taxonomy: renaissance or Towel of Babel? *Trends in Ecology and Evolution*, 18(2): 57-59.

NATARAJAN, P., BAZZI, I., LU, Z., MAKHOUT, J. & SCHWARTZ, R. 1999. Robust OCR of degraded documents. *Fifth International Conference on Document Analysis and Recognition*, pages 357-361.

SARASAN, L. & NEUNER, A., 1983. *Museum collections and computers*. Association of Systematic Collections Publication. Lawrence. 292 pp.

SIEPEL, A., FARMER, A., TOLOPKO, A., ZHUANG, M., MENDES, P., BEAVIS, W. & SOBRAL, B., 2001. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics*, 17(1): 83-94.

TALAGALA, N., ASAMI, S., PATTERSON, D., FUTERNICK, B. & HART, D. 2000. The art of massive storage: a Web image archive. *Computer*, 33(11): 22-28.

WACTLAR, H., CHRISTEL, M., GONG, Y. & HAUPTMANN, A. 1999. Lessons learned from building a Terabyte digital video library. *Computer*, 32(2): 66-73.

WILKINSON, M. & LINKS, M., 2002. BioMOBY: An open source biological web services proposal. *Briefings in Bioinformatics*, 3(4): 331-341.

YI, J. & SUNDARESAN, N. 2000. Metadata Based Web Mining for Relevance. 2000 *International Database Engineering and Applications Symposium*, Yokohama: 113-121.

Recibido, 22-II-2003
Aceptado, 6-IX-2003
Publicado, 15-XII-2003