FINAL REPORT OF THE

OECD MEGASCIENCE FORUM

WORKING GROUP ON BIOLOGICAL INFORMATICS

January, 1999

Contents

Executive Summary

FINDINGS AND RECOMMENDATIONS

1. **Biological informatics**—a rapidly growing, interdisciplinary scientific area that brings the advantages of computational science, networking capabilities, and information science and technology to bear on biological data—**is an enabling discipline for all of modern biology.** The databases, computational tools and search engines that form the basis for biological informatics are located in various places around the world, but at present are especially concentrated in OECD countries. It is the need to link these informatics resources (and the people who use them) into a synergistic, interoperable whole that makes biological informatics a megascience endeavor.

General Findings and Recommendations

2. **The governments of OECD countries have the opportunity to play a crucial role in fostering biological informatics by eliminating the barriers that prevent cooperation and by providing incentives to potential participants.** Actions that could be taken include, *inter alia*:

   - Funding development of software or hardware capabilities that enable biological informatics;
   - Protecting the needs of scientific research and education when making international agreements about intellectual property and biological information;
   - Making databasing and information provision a condition of government funding of biodiversity and neuroscience research projects;
   - Participating in international efforts to establish standards in several areas (including data validation and description); and
   - Encouraging, through tax concessions, contracts, or other incentives, participation in the worldwide biological informatics effort by academia, private companies, the publishing industry, and other segments of the private sector.

3. Two areas of biological science where the need for biological informatics is especially crucial are **biodiversity informatics** and **neuroinformatics**.

Findings and Recommendations for Biodiversity Informatics

1. An international mechanism is needed to make biodiversity data and information accessible worldwide. The existence of such a mechanism will produce many economic and social benefits. For example, the Convention on Biological Diversity (CBD) obligates nations to implement provisions relating to conservation, use, and equitable sharing of biodiversity. A scientific information resource that could facilitate fulfillment of these obligations is greatly needed. Such a resource will also contribute to biotechnology and bioengineering, and therefore will be a central element in sustainable development. Because biodiversity is one of the primary measures of environmental impact, sound scientific information about it provides a way of determining whether development in a particular region is actually sustainable.

2. **The Subgroup on Biodiversity Informatics recommends that the governments of OECD countries establish and support a distributed system of interlinked and interoperable modules (databases, software and networking tools, search engines, analytical algorithms, etc.) that together will form a Global Biodiversity Information Facility (GBIF).** This facility will enable users to navigate and put to use vast quantities of biodiversity information, thereby

   - advancing *scientific research in areas such as agriculture, biomedicine, biotechnology, environmental management, pest control, health, education, and conservation, among others;*

   - serving *the economic and quality-of-life interests of society;* and

- providing *a basis from which our knowledge of the natural world can grow rapidly and in a manner that avoids duplication of effort and expenditure.*

3. A GBIF Secretariat (Director, plus three to six program officers and appropriate staff, with input from legal and technical consultants as necessary) will build coalitions among ongoing efforts, encourage new developments, and provide mechanisms for coordinating separate national investments and forging international agreements. The Secretariat will be responsible to a Governing Board composed of delegates from countries that elect to support GBIF, and be advised by *ad hoc* Scientific and Technical Advisory Groups.
4. Funding for individual projects will continue to be provided through existing national and regional mechanisms. The GBIF secretariat will provide, among many other services, a clearinghouse for information about past, present, and proposed projects. GBIF personnel, through their activities (conduct of studies, facilitation of workshops, coordination and networking, etc.) will promote more interoperability between databases, more coordination between independently-funded programs, more standards and protocols for linking databases, and more practical applications that demonstrate the utility of biodiversity data for addressing critical social concerns.
5. GBIF will be closely linked with established programs and organizations that compile, maintain and use biological information resources such as Diversitas, Species 2000, the Integrated Taxonomic Information System (of US agencies), and the Clearinghouse Mechanism of the CBD.
6. **The Subgroup also recommends that governments accelerate efforts to compile electronically available data about living organisms and ecosystems,** especially those whose existence is threatened and those of potential economic importance, and to enter these data (as well as existing biodiversity and ecosystems data) into databases that are interlinked via the GBIF.

Findings and Recommendations for Neuroinformatics

1. Neuroinformatics combines neuroscience and informatics research to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain. Neuroinformatics research is uniquely placed at the intersections of medical and behavioral sciences, biology, physical and mathematical sciences, computer science and engineering. The synergy from combining these approaches will accelerate scientific and technological progress, resulting in major medical, social, and economic benefits.
2. Indeed, neuroscience has so far been dominated by the acquisition of experimental data, and the time is propitious to facilitate the development of theoretical models and tools to help manage and use the data to yield new knowledge and understanding.
3. The scientific goals of Neuroinformatics are to accelerate the progress of neuroscience and informatics by:

    - Making better and more efficient use of neuroscience data using informatics-based, including computational, approaches;
    - Generating and evaluating new hypotheses and computational theories about brain function to drive further experiments;
    - Developing and applying new tools and methods for acquiring, visualizing, and analyzing data important for understanding how the brain functions;
    - Enabling the more efficient application of the accumulating knowledge of how the brain functions to be applied to understanding its dysfunction in disease; and
    - Developing computer systems and technological applications which simulate or emulate specific aspects of brain function.

4. The Subgroup on Neuroinformatics recommends that the Megascience Forum support the establishment of a global neuroinformatics capability. This capability needs to be developed as a network of neuroinformatics facilities and approaches, distributed across many research centers around the world. This network of neuroinformatics facilities

will be diverse, with major foci representing:

- Databases, increasingly capable of handling the full complexity and organization of the nervous system, from molecular to behavioral levels;
- Powerful new tools for data-acquisition, analysis, visualization and distribution; and
- Theoretical, computational and simulation environments for modeling and understanding the brain.

5. International coordination, as well as national efforts, are needed to assure that these steps are properly implemented and sustained. An international scientific coordinating body, the International Neuroinformatics Committee (INC), and an associated secretariat should be established through the support of the participating countries.

BACKGROUND

The Biological Informatics Working Group was established by the Megascience Forum in January 1996. Its goals are:

- To promote international cooperation in the development and implementation of federated, interoperable databases and other informatics resources related to biological diversity, especially at the specimen, species and ecosystems levels, and to foster the rapid development and general distribution of informatics tools for the field of biological systematics;

- To strengthen international cooperation in neuroinformatics, the area of science that combines brain and behavioral (clinical and basic) research with informatics research, with the intent of developing advanced tools for better utilizing brain and behavioral data; and

- To identify the intellectual property rights issues related to biodiversity and neuroinformatics databases as well as the conditions required to maintain an open accessibility to these universal resources.

Working Group member countries include:

Australia	Belgium	Canada	Denmark
Finland	France	Germany	Israel
Italy	Japan	Korea	Mexico
Netherlands	Norway	New Zealand	Poland
Portugal	Russia	Sweden	Switzerland
United Kingdom	United States	Commission of the European Community	

What is Informatics?

Informatics is an emerging area of science and technology that combines the advantages of computational science, networking capabilities, and information science and technology, and may be described as

> Research on, development of, and use of technological, sociological, and organizational tools and approaches for the dynamic acquisition, indexing, modeling, dissemination, storage, querying, retrieval, visualization, integration, analysis, synthesis, sharing (including electronic research collaborations), and publication of data and information such that economic and other benefits may be derived from them by users in all sectors of society.

"Informatics" is a term originally coined in Russian (as "informatika") in 1967 to refer to the dissemination of electronic information via networks. Since that time, the development and rapid spread of the Internet, as well as the recognition that informatics also encompasses sociological issues, have expanded the concept. Indeed, the Turing Award winning computer scientist Robin Milner recently stated that "networking allows computer science (theory of calculation) to grow into informatics (theory of interaction)". "Informatics" has come to the forefront in some scientific disciplines, though in others there is much yet to be done. The life sciences are uniquely placed for coupling with informatics because these two areas share the study of communication, and the employment of mobile and complex agents. Using informatics in biological research thus fuels a "virtuous circle" in which informatics supports the life scientists, whose results and insights can then be fed into informatics tools.

Why is Biological Informatics Important and Needed?

The biological information contained in print media, in outmoded electronic form, and in modern databases constitutes an intellectual wealth produced by decades and centuries of research and considerable societal investment. If significant further advances in scientific understanding of biodiversity at the gene, organism (including neurological), population, species, ecological community and landscape, and global levels are to be made, the results of the work of the predecessors as well as contemporaries of the world's biological scientists should, using the technologies now at our disposal, be made readily and comprehensively available to the current generation of researchers, no matter where they reside. This same information is needed by persons with policy- and decision-making responsibilities, and there are applications in education, both formal and informal, and industry to which the information could contribute.

With the proper investments in infrastructural and software developments, the advantages of modern informatics techniques can be employed to exploit this intellectual wealth — with great benefit not only to biological research, but to decision- and policy-makers, education, and society at large. However, advancements in informatics capabilities for the biological sciences (in data management, in network connections, and in data content) are still needed. Meeting these needs will benefit researchers and society in general by, *inter alia*,

- Making information from all biological disciplines readily available worldwide;
- Managing masses of data to reduce them to the important kernels of information;
- Correlating information from disparate sectors of knowledge;
- Enabling analysis and synthesis of volumes of data too great to be handled by the individual human mind; and
- Providing a complete biological context for information from molecular and other databases that are already available.

The science of informatics itself has needs that may be addressed in the longer term by developments in biological informatics. The mobile, complex, intercommunicating agents with which informatics is concerned can be either software agents that traverse networks, or hardware agents such as robots on a factory shop floor, or prosthetic devices. In the near future, individuals will be launching hundreds of software agents a day from their PCs into the Internet; thus, there is an obvious need arising to understand how to build and organize such societies of agents. With all the advances in robotics, there remains a long way to go before they can navigate in, and interact with, their environments as well as living organisms do. Biological systems can inspire solutions to problems the informaticians are meeting (witness the interest in "ant algorithms" for routing phone calls, based on observations of the behavior of ants).

Rationale for Focus on Biodiversity Informatics and Neuroinformatics

Current Status

Currently, "bioinformatics" is for the most part concentrated on the molecular and genetic subdisciplines of biology, primarily gene sequencing and other genetic information ("genomics"). The overwhelming majority of the employment positions in biological informatics are in these subdisciplines, and the majority of software developments are geared directly to handling these sorts of data (one important exception can be found in software developments for geographic information systems [GIS]). Because the molecular subdiscipline has received so much informatics attention, this Working Group does not include a Subgroup on this area.

However, the full value of molecular biological information cannot be realized until it is possible to correlate genetic information with data on (for example) the native habitat, neurobiology, physiology, or genealogical relationships of the species from which the genes were derived. At the same time, both biodiversity informatics and neuroinformatics would greatly benefit from intercompatibility with molecular-level datasets.

Scientific Needs

In the last 20 years, brain and behavioral research has experienced explosive growth because conceptual links have been made across different species, different levels of biological organization, and different experimental and theoretical approaches. The dramatic increase in the amount of information generated has caused neuroscientists, of necessity, to increasingly narrow their areas of specialty, just to be able to keep up with publications most relevant to their own research. The cost of such specialization is a decrease in the development of new conceptual linkages. Thus, the amount of information generated by the engine of interlinked research threatens to choke the engine itself. However, advances in informatics focused on brain and behavioral research information can prevent the stifling of this success.

A major scientific consideration in biodiversity science is the need to bring 25 decades worth of accumulated information into an electronically available format. Unlike some other subdisciplines of biology, biodiversity (primarily taxonomic and ecological) research results do not rapidly go out of date. In fact, many such results probably cannot be replicated because of anthropogenic habitat modifications that have occurred since the research was done. In addition, new kinds of data are being generated by satellite imagery and other measures of non-biological, global phenomena—phenomena that have significant influence upon biodiversity. Great forward strides could be made in the understanding of the biological world, for instance, if informatics techniques were developed to make it possible to correlate historical information with newly collected satellite data; if molecular genetic datasets could be linked to species-documentation datasets such as those held by natural history collections; and if neurobiological, physiological, chemical, and other sorts of datasets could be correlated with taxonomic and ecological ones.

Societal Needs

In order to comprehend and sustainably utilize the biodiversity resources of the world, humankind must learn how to exploit massive data sets, learn how to store and access them for analytic purposes, and develop methods to cope with growth and change in data. The informatics developments that are recommended here will be an enabling framework that could unlock the knowledge and economic power lying dormant in the masses of biodiversity data that we have on hand.

Advanced informatics solutions will accelerate understanding of the brain and lead to a better understanding of ourselves and the nervous system, thus enabling translation of basic research into better means to diagnose, monitor, treat, and prevent brain disorders. Conversely, understanding of the biological mechanisms that acquire, store, retrieve, analyze, synthesize and visualize data and information will reciprocally illuminate informatics techniques, such that, over time, computers will be better able to emulate the workings of the human brain, which is still (even given the remarkable advances made in recent years) the ultimate in computers.

Opportunities in Biological Informatics for OECD Countries

Economic Opportunities

As molecular biological informatics has demonstrated, there are economic gains to be made when information from multiple sources is readily searchable and can be easily correlated. Less than $10 million per year (globally) maintains the three major regional nucleotide sequence databases, but these are essential to the several billions of dollars grossed yearly by the pharmaceutical, agrobusiness, and health industries.

This will be no less true for the rest of biological information; in fact, the broader the range of data types available, the wider the range of applications to which those data may be put. For example, agriculture will benefit from information on the habitat and evolutionary relationships of the wild relatives of crop species; health-related research will benefit from correlations among many types of neurological data (images, numerical results of surveys, etc.); and the pharmaceutical industry will benefit from access to information from biological collections.

There are also economic opportunities for software and hardware developers, researchers, and technicians ("informaticians") associated with biological informatics:

- There is high demand for individuals who have the training to contribute to information management in the biological sciences.
- There will be many occasions for public-private partnerships in the development of hardware and software for biological informatics in the broad sense.
- The science of informatics will itself benefit from feedback from biological systems, which will lead to new advancements that will iteratively improve biological informatics.

Research and Development Opportunities

The exciting aspect of the biological informatics needs described in this document is that the methods and technologies required to meet the challenges are conceivable and could be within reach—all that is needed is that these methods and technologies be applied in a focused manner to the special challenges presented by biological data and information.

Furthermore, the innovative capabilities and facilities that are needed to accomplish these tasks by and large reside in countries that are members of the OECD. It is the special privilege of OECD countries to be in a position to apply the talents of their citizens, through investments that direct the attention of those talents, to the achievement of a goal that is particularly important both to science and to society. These opportunities include, *inter alia*:

- Supporting the development of software that will make possible the broad-scale correlations and computation that will lead to paradigmatic advances;
- Encouraging development and installation of high bandwidth networks; and
- Funding research in information science that will better enable modeling of biological information, which in turn will improve database structure and function.

International Benefits

OECD countries can be of great assistance to developing countries by supporting the actions recommended here and in the Subgroup reports. Particularly in the area of biodiversity, but to a certain extent also in neuroinformatics, the issue of "repatriation of data" holds great importance. Institutions located in OECD countries hold the vast majority of the biological data content that needs to be made fully available to scientists in other OECD countries and in other parts of the world. Importantly, OECD countries hold at least 75% of the approximately two to three billion biological specimens in the world's natural history museums, and a great percentage of these specimens were collected in countries which are not OECD countries. It is incumbent upon the OECD countries who hold the information associated with these specimens to make it possible for the scientists and citizens of non-OECD countries (as well as their own scientists and citizens) to have ready and easy access to this information, especially in the light of provisions of the Convention on Biological Diversity. To achieve this goal, OECD countries could chose to:

- Assure that biological research institutions have Internet access and appropriate computer equipment;
- Enable data-providing institutions by funding the tasks of bringing their datasets online; and
- Develop informatics capabilities that are congruent with the goals of international efforts such as the Clearing House Mechanism of the Convention on Biological Diversity.

Scientific and Infrastructural Challenges and Issues

The scientific and infrastructural challenges that accompany the vision of true biological informatics are several, stemming from network and hardware needs, required software developments, sociological adjustments, and interactions between providers and users of the information.

- ***Network Needs—***The basis of the network that will be required is in place (the Internet as it is known today). However, there are many regions of the world and many institutions that have yet to be provided with network connections. In addition, as improvements in network capacity (increased bandwidth, etc.) occur, they need to be provided to biological research institutions as a high priority. Both a fuller implementation of current technologies—such as

digital signatures and public-key infrastructure for managing cryptographic key distribution—and a consideration of tools and services in a broader context related to use are needed.

- ***Hardware Requirements—***

  Computation—It has become apparent that the gathering of biological information into a single centralized database is not only impractical but also antithetical to scientific advance. Rather, the computer servers required to house and provide data should be maintained by the institutions that own and provide the data. These institutions may require financial assistance to guarantee the continuity and provision of those data.

  At the same time, many biological informatics operations will be computationally intensive because they call for the aggregation and combination of information from large numbers of autonomously managed resources and their presentation to the user as a coherent whole. This fact generates a need for several “nodes” that would provide high-capacity, high-speed computational ability specialized to the biological disciplines. Funding for these “nodes” should be assured by appropriate consortia among industry and governmental agencies at both national and international levels. Interlinking these nodes would produce a “decentralized distributed centers system.”

  Storage—As research is conducted to devise new ways to manipulate huge datasets, massive storage capabilities will be needed.

- ***Software Requirements—***

  Information management—Major advances are needed in methods for knowledge representation and interchange, database management and federation, navigation, modeling, and data-driven simulation; in effective approaches to describing large complex networked information resources; and in techniques to support networked information discovery and retrieval in extremely large-scale distributed systems. We need to preserve and support the knowledge of library and information science researchers, and help scale up the skills of knowledge organization and information retrieval.

  Tools for visualizing data—The value of raw data is typically predicated on our ability to extract higher-order understanding from those data. Wherever possible, tools will be adopted and adapted from other arenas, such as defense, intelligence, and industry. A reciprocal relationship among partners in these developments will provide the most rapid progress and best results. Among the most important issues are content-based analysis, data integration, automatic indexing on multiple levels (of content within databases, of content and quality of databases across disciplines and networks, of compilations of data made in the process of research, etc.), and data cleansing.

- ***Information Resource Issues***—Generally, data owners adhere to the scientific principles that would lead them to make data available. However, at present, there are barriers to making their data electronically available for use by others. Means must be found to enable institutions to add the providing of electronic data to their missions. Internet connections must become one of the basic requirements of a biological research institute, and employee positions must be allocated to system and information management. And, international mechanisms must be designed to assure that attribution will be appropriately reported.

- ***Data Validation and Assessment***—Global information systems in biology are of greatest value when they provide data of reliable and known quality. Mechanisms to monitor and document data quality are essential for the efficient exchange of data in these systems and are an essential part of interoperability of data and databases. There is an apparent conflict between the objective to speed up the global availability of biological information and the necessity to improve the quality of that information, which may slow down the process. Data quality assurance and documentation are essential parts of data management and stewardship. Data quality is generally agreed to be related to "Fitness for Use." Data documentation (often called metadata) describes a dataset so that users may determine the fitness of the dataset for the use to which they wish to put the information. Issues that affect the validation and assessment of data include:

Characteristics of Biological Data—Databases in biological sciences are diverse, complex, heterogeneous and distributed throughout all areas of the world.

Indexing—Domain taxonomies for both biodiversity science and the neurosciences are still being developed; these should incorporate full flexibility to account for scientific and technological progress while accounting for historical variability.

Quality of Biological Data—It is important that the scientific quality of the data is identified and assessed by generally accepted procedures in scientific research. Quality description and validation must be applied not only to the data, but also to software systems used for the storage and validation of data.

Validation of Biological Data—The biological sciences have lagged behind other fields in the adoption of standardized methods for data validation, and this tardiness should be corrected.

Compatibility—It is no longer necessary for software or hardware to be compatible to participate in distributed systems; however, if databases from different sources are to be integrated, then it is essential that the data fields be defined and described. Standards for this within the biological sciences are being developed, and OECD countries are important partners in this development.

Documentation—Using databases without documentation is like using a library without a catalogue. Full documentation is needed both for data and for database tools.

Ownership of Data —Any information system should protect the rights of the database providers. These may relate to copyrights or the right of admission to, or use of, the data.

Distributed Nature of Databases—None of the above factors is inconsistent with the distributed nature of databases in the biological sciences or their use by scientists and the general public. Local, national, and international groups should all be strongly encouraged to make their data available and to present the data with full and consistent metadata documentation.

- ***Training and Infrastructure—***The next decade will bring about a rapid and increasing expansion of biological informatics tasks, not only in research and academia, but also in industry and commerce and in the government sector, where a growing number of departments and agencies will need bioinformatics capabilities. One of the keys to rapid progress lies in the area of training and education of the "informaticians" that will be needed to carry out these tasks. Issues that affect such training and education include:

  Non-recognition of biological informatics as an autonomous field—At present, training of specialized personnel is often strictly task-oriented (and frequently rather autodidactic).

  Scarcity of specialized staff—Increasing demand on the public sector for coordinating measures and in the academic sector for instructors will have to cope with competition for expertise caused by the increasing needs of the private sector.

  Missing technical standards—There are too few (and poorly funded) public bodies with the task to formulate technical standards and guidelines for the optimal modeling, archiving, storage and access of biological data.

  Poor coordination of international infrastructures—Biological informatics activities have benefited from projects aimed at networking institutions and research teams. However, consistent funding for the necessary basic infrastructure remains a problem.

  Lack of access—Availability of communications and information technology is often inadequate in areas with a high intensity of biological knowledge and representation of biodiversity, such as museums, botanical gardens, zoological gardens, natural history museums, and biological research institutes.

To alleviate these problems, and increase the number of broadly trained informaticians, OECD countries could choose to implement policies that would

- lead to the official and academic recognition of biological informatics as an autonomous, interdisciplinary field;
- develop the necessary standards and guidelines;
- give career credit for database development and information provision;
- provide the necessary administrative backbone for cooperative projects; and
- diminish costly duplication of efforts.

With such policies in place, and as the demand in the job market grows, academic institutions will likely establish chairs, faculties and institutes in Biological Informatics. These should, from the start, be organized into an international, coordinated network having the aim to create an effective infrastructure for the training of specialists for research in the areas of optimal processing of biological data and related IPR issues. Governments of OECD countries could choose to play a role in the development and maintenance of such a network. In so doing, they would support the training of the informaticians needed in the public sector, and promote the activities of the private sector. In partnership, the public and private sectors could greatly assist the academic sector to produce new research products and qualified students.

Intellectual Property Rights

The provisions of the Convention on Biological Diversity, specifically in Articles 15, 16 and 17, present several IPR issues to biological informatics that require resolution, including:

- Does the scope of sovereignty cover biological samples collected before the coming into force of the Convention?
- What is the connection between sovereignty on physical specimens and access to data about them?
- Are human genetic resources included in the field of application of the Convention?

Modern juridical systems establish a distinction between intellectual and physical property whereby, in principle, the ownership of biological samples or materials does not rule out the existence of a possible intellectual property on the discovered findings and/or on creations based on such materials. The ownership of the source of the information does not entail the property of the information itself. In practice, however, the following situations occur that prevent a thorough application of this general principle:

- The actual holding of biological materials clearly leads to the control of the access to related information and, in practice, also to the control of any possible invention that could derive from those materials. In this perspective, the problem of intellectual property of biological information is also connected to the problem of ownership of the media where that information is stored.
- In the framework of contractual terms, the owner of biological material may also keep the ownership of the results of research carried out using that material. It is a common practice in cooperative contracts in the field of biotechnology that the terms of such contracts binds the parties to the sharing of the commercial or other benefits arising from utilization of such resources. In the context of biodiversity, this has been the basic guiding principle of the Convention on Biological Diversity confirming the sovereignty over genetic resources

Raw data, in and of themselves, may not be protected by copyright. Elements in databases of biological information that can be protected under IPR include software developments that may be defined as literary works under the Berne Convention and the TRIPS agreement, database structures insofar as they are original and creative, and (in Europe) the “whole or a significant part” of the content of a database. The “denial right” of database authors potentially may be seen as a barrier to the free flow of information, although the WIPO Treaty on Intellectual Property in Respect of Non Original Databases was not adopted by the WIPO Diplomatic Conference on Certain Copyright and Neighboring Rights Questions held in Geneva in December 1996, because this issue had only recently been added to the international agenda. Since then, the issue has been the subject of an "Information meeting" in Geneva and various regional consultation exercises with a view to determining the scope for international law making concerning databases. Of most concern to this Working Group is the lack of explicit exception, for research purposes, to the “sui generis” right to forbid extraction from databases in the Draft Convention (Article 5).

At this time, the Working Group simply wishes to raise these issues as ones of concern. In no way does it wish to prejudice ongoing efforts in other fora to develop policies regarding intellectual property rights for biodiversity data.

Support and Funding

Enabling conditions

Success in implementing a megascience effort in Biological Informatics will depend upon the dedicated provision of necessary financial, technical and human resources and upon the optimal use of those assets in enabling a set of coordinated, global actions. There are certain conditions which, if established and maintained, will facilitate the developments that are identified as needed in this Report. Among these are:

- Assurance of support for electronic data provision, without diminishing support for traditional functions of the providing institutions;
- Creation of positions dedicated to the maintenance of computer systems and information management within (or shared between) data-provider institutions; and
- Assistance by users of data (industries, etc.) directly to data providers.

Funding

This document aims at identifying the sources and targets of the funding requested, as well as the mechanisms that will have a multiplying effect on the devoted resources in terms of savings and improvement of results. For this purpose, the document considers those aspects that can be common to Biological Informatics. Specific actions to be supported and detailed funding schemes are given in the individual Subgroup reports.

Leverage and Evolution Curve—The cost of the Biological Informatics activities proposed here, before being judged as too high, must be compared with the value added by accomplishing the goals, both in terms of accessibility of the information itself and in terms of other tangible and intangible benefits.

Governments are already spending significantly on biological informatics through traditional resources, although this funding is not usually visibly targeted as such. Examination of current funding levels and determination of what is required in additional or reallocated, targeted resources (or incentives) will result in significant synergies with existing programs and expenditures. Every country is already investing, but in an uncoordinated fashion. What is needed is a concerted, concentrated and coordinated investment that will result in major economies of scale and reduce wastage that comes from poor integration of existing programs.

In general, information technology-related costs occupy between 10% and 15% of the total costs of most research projects conducted within government agencies. This percentage can reasonably, and definitely should, be expected to be higher (up to 25%), at the initial stages of system design and data collection, and to rapidly drop to normal values once the information system becomes stable and fully operational.

Examples of Value-Addition in the Informatics Arena

There are a number of lessons that can be taken from the experiences of the molecular biology community. Less than $10 million per year (globally) maintains the three major regional nucleotide sequence databases, which are doubling in size every 18 months. A 100% increase in volume of information requires only a 10% increase in maintenance costs. Re-usability of programs and interfaces, agreement on standards and a modest international management structure assure the provision of a core information resource for research at a more than reasonable cost compared to the benefits it enables. Probably, several billions of dollars per year in pharmaceutical, biomedical and agroindustrial research and development activities in both the governmental and industrial domains are dependent upon the presence of the genomic informatics infrastructure.

Another outstanding example of the leverage effect of investments in information technology-related activities is provided by the Internet itself. For the Internet, the simple registration of Internet Protocol (IP)

addresses and domain names has been an essential enabler of the total suite of Internet functionalities. At an estimated cumulative cost of approximately $100 million, this activity represents only approximately one-tenth of one percent (0.01%) of the total cost of today's Internet, which is estimated at a cumulative $100 billion. This helps to demonstrate that, for networking-type efforts, relatively small investments can be leveraged to gain extraordinarily high payoffs if those investments provide critical enabling capabilities.

Direct governmental investments could be made in several ways, such as:

- New funding allocations (grants, contracts) to software developers and information providers;
- Re-allocation of effort within/among agencies to assure 15% minimum concentration of funding on information technology and information management (IT/IM));
- Directions to agencies that they share IT/IM resources and developments for benefits of scale and cost-efficiency;
- National and international public-private working groups focused on identifying information sources and prioritizing information needs; and
- Provision for maintenance (storage facilities or funding) and curation of data (migration to new platforms, software, etc.) into the future, after the original provider (researcher or institution) has retired or gone out of business.

When weighing the pros and cons of such investments, governments should consider the costs of *not* having strong biological informatics research and infrastructure, e.g. environmental destruction and loss of biodiversity resources, and lost biomedical and economic opportunities, as a consequence of poor data collection, management, analysis, coordination and distribution.

Thanks to its investment in informatics through CONABIO, the government of Mexico has very recently been able to quickly and reliably assess the impact by Hurricane Pauline on the biodiversity and biodiversity-based industries (forestry, agriculture, etc.) of the state of Oaxaca, a feat that few other countries could (at present) match because they have not made a concerted investment in building their biological informatics capacities.

Relatively small investments can be leveraged to gain extraordinarily high payoffs, if those investments provide critical enabling capabilities. Therefore, in order to assure that kick-off investments enable future developments, the identification of appropriate first initiatives is of crucial importance. If this plan is followed, there will be benefits immediately, as the system is being built. Modular design will allow production of results from the very moment the investment starts. It will not be necessary to wait for benefits until the total system is built, as happens with other types of physical facilities.

***Leadership, Partnership and Sponsorship*—**Governments, academia, industry and social groups will play complementary roles in this Megascience project. Leadership, partnership and sponsorship represent ways to channel this participation while responding to diverse interests of the various stakeholders.

Most of the activities of the Megascience endeavor that will result in the availability of information content (data collection and organization, networking and provision of access to the databases) will be performed by participating nodes that may be expected to be academic and/or dedicated research institutions. Therefore, incentives for these data providers should be generated.

However, companies and non-governmental organizations may also carry out specific tasks, and governmental agencies often host important databases of relevant information. In order to encourage private sector participation, if not sponsorship, governments could consider creating incentives for industry, such as:

- Assurances of government contracts for certain software if they will turn their expertise to developing it;
- Tax benefits for making formerly "proprietary" data available; and
- Tax concessions for industries that have funded in-house or collaborative biological informatics research.

REPORT OF THE SUBGROUP ON BIODIVERSITY INFORMATICS of the WORKING GROUP ON BIOLOGICAL INFORMATICS

SUMMARY

An international mechanism is needed to make biodiversity data and information accessible world-wide. Attempts to integrate parts of the biological and ecological data matrix are occurring in a handful of projects (e.g., INBio in Costa Rica, Diversitas [an ICSU-UNESCO program for biodiversity research], and Species 2000), but these efforts need to be augmented and coordinated.

The existence of such a mechanism will produce many economic and social benefits. For example, the Convention on Biological Diversity (CBD) obligates nations to implement provisions relating to conservation, use and equitable sharing of biodiversity. A scientific information resource that could facilitate fulfillment of these obligations is greatly needed. Such a resource will also contribute to biotechnology and bioengineering, and therefore will be a central element in sustainable development. Because biodiversity is one of the primary measures of environmental impact, sound scientific information about it provides a way of determining whether development in a particular region is actually sustainable.

The Subgroup on Biodiversity Informatics recommends that the governments of OECD Member countries establish and support a distributed system of interlinked and interoperable modules (databases, software and networking tools, search engines, analytical algorithms, etc.) that together will form a Global Biodiversity Information Facility (GBIF). This facility will enable users to navigate and put to use vast quantities of biodiversity information, thereby advancing scientific research in areas such as agriculture, biomedicine, biotechnology, environmental management, pest control, health, education, and conservation, among others; serving the economic and quality-of-life interests of society; and providing a basis from which knowledge of the natural world can grow rapidly, in a manner that avoids duplication of effort and expenditure.

This Facility will be ***distributed***, while encouraging co-operation and coherence; ***global*** in scale, though implemented nationally and regionally; and ***open*** to participation and benefit by all countries, while having the majority of its support provided by those countries that have the greatest financial, scientific, and technical capacities to do so.

The Subgroup also recommends that governments accelerate efforts to compile data about living organisms and ecosystems, especially those whose existence is threatened and those of potential economic importance, and to enter these data (as well as existing biodiversity and ecosystems data) into databases that are interlinked via the GBIF.

This Report reaches beyond the obvious need to conserve biodiversity to include other major policy objectives of the OECD and particularly its Committee for Scientific and Technological Policy. These are:

1) ***the advancement of science—***Biological informatics is fundamental to the future development of all life sciences. Depending upon country and definition, probably half of all scientific research is devoted to the life sciences and associated fields, such as health, agriculture and food, ecology and environment. The GBIF, in part by linking to existing molecular and genetic databases, will give a gradually increasing and ultimately very substantial boost to all research endeavors directly or indirectly related to living things.
2) ***greater efficiency and economies in R&D spending—*** Joint implementation of GBIF will be less costly than multiple efforts undertaken independently by individual governments.
3) ***technological applications and economic opportunities—***GBIF will facilitate application of life science data in industry, agriculture, conservation and health. The exploration of science and technology for economic benefit is a goal shared by all countries. GBIF will be a resource that will stimulate development of new commercial products and informatics tools and aid in preserving biodiversity. Databases that are essential to achieving ecological and economic compatibility, responsible resource management, and sustainable development will be interlinked and made accessible by GBIF.

The Subgroup recommends that OECD member countries:

- ***Establish*** a Global Biodiversity Information Facility (GBIF).
- ***Support*** the GBIF Secretariat and GBIF programs financially, and ***appoint*** a representative to the GBIF Governing Board.
- ***Invest*** in related national and international activities that further the goals of the GBIF, such as:

    1. Contributing data, information, and capabilities to GBIF. More specifically,
        a. data about whole organisms
        b. specimen data from biological collections
        c. environmental and remote sensing data
        d. molecular, gene and genome data
        e. new information and communication software and tools
        f. laboratory, computing, and training facilities.
    2. Promoting national involvement in GBIF by
        a. developing national nodes of GBIF
        b. coordinating and synchronizing funding activities with respect to GBIF priorities
        c. expanding databases with data from collections, organisms, ecosystems, genes, etc.
        d. providing access to databases
        e. developing software and tools for information and communications technology
        f. adopting international standards
        g. improving high-speed networking and computation infrastructures

h. enhancing infrastructures for
    1) data providers
    2) database custodians
    3) expert centers

i. supporting training

3. Sharing
    a. computational facilities
    b. storage capacity to house major databases or operate database mirroring sites
4. Hosting
    a. the GBIF Secretariat (in whole or in part)
    b. specific GBIF projects

INTRODUCTION

The Subgroup on Biodiversity Informatics was constituted as one of the subgroups of the Working Group on Biological Informatics of the OECD Megascience Forum. Its charge was to produce a report on the state of biodiversity informatics in OECD Member countries and to evaluate the opportunities for collective progress. The Subgroup also examined possibilities for filling gaps in biodiversity information content on the Internet, and explored linkages with other kinds of informatics resources. Members of the Subgroup were representatives appointed by the governments of the OECD Member countries that chose to participate in the Working Group. The Subgroup met seven times between June 1996 and September 1998, and communicated frequently via a listserv between meetings. Several task groups were constituted to address particular issues as these arose; the white papers produced by these task groups were incorporated into this Report as appropriate. The Subgroup also consulted frequently with existing efforts and projects (e.g., European Molecular Biological Laboratory, the Clearinghouse Mechanism of the Convention on Biological Diversity, Species 2000, Diversitas).

There is increasing awareness that responsible environmental stewardship is not only compatible with developing economies, but is imperative for the long term survival of our species. In order to increase stewardship capabilities, knowledge and information about biodiversity must be available to policy-makers, voters, and the scientists who will continue to increase the store of knowledge. To make this knowledge and information available to those who need it, using the most up-to-date computational, networking and information management technologies, a global biodiversity information infrastructure that supports a common, global, updateable, electronically accessible knowledge base is required.

Box 1: Case example - Threatened Species in Relation to Forest Practices

In Sweden there is an ongoing discussion concerning the preservation of virgin and so-called natural forests. Both reserves and different methods of forestry are used to prevent further diminishment of biodiversity in those forests. The preservation of red-listed species is used as a measurement of the success of reaching the goal of preserving biodiversity. Pulling together geographical and ecological data on Swedish forests and data on the occurrences of threatened species can give us good guidance on the size of the areas that should be strictly preserved, and on the size of the areas with highly modified forestry that is required to preserve biodiversity.

Box 2: Application of Biodiversity Informatics in Australia

Many agencies in Australia are providing biodiversity information via the WWW, such as the CSIRO, the parks departments of New South Wales and Tasmania, and the Environmental Resources Information Network (ERIN). ERIN (part of the Department of the Environment) uses computing technology to provide access to information that is not held by ERIN but rather is drawn from distributed sources, which are maintained by many agencies and institutions. The parks departments are accumulating information and then making it available from more centralized sites. CSIRO is providing results of projects undertaken by its researchers.

Use of environmental information is contributing to the development of environmental policies, assessment of environmental impact, and developments leading to sustainability. For example, information about plant and animal species has been integrated with climate models to predict the distribution of a range of species under a number of climate change scenarios developed from Global Climate Models. This information has then been integrated with other information about vegetation and soils for use in the development of species management plans.

Box 3: Application of Biodiversity Informatics in Mexico

In April, 1996, CONABIO had enough information in its databases to begin to respond to requests for information on a regular basis. At present, CONABIO receives between five and ten formal requests for information per week. About 50 per cent of these come from government agencies, mainly of the environment portfolio. The rest of the information requests come from scientists (about 40 per cent) and private companies (10 per cent).

The government questions are often of the nature of “what protected species occur in a given place?” and are posed in relation to protected areas or zones where environmental impact assessments are being contested. In other examples, state or municipal governments require lists of their endangered or protected species.

A recurrent set of questions concerning suitable areas for reforestation with a given tree species led to the development of a database and GIS package containing the correct scientific names, a catalogue of common names, and museum information on the presence of the main 300 species used for reforestation, along with their ecological profiles (ranges of altitude, latitude, temperature etc.) and text information about seed production, phenology, etc. This package has been tested for the state of Morelos and will be released for the whole country at the end of the year. It is expected that the final package will include modeling tools to correct automatically the likely areas of success as new data is added to the database. A similar package has been requested for the non-timber forest products, of which good data is available for around 500 species.

CONABIO´s databases are also used to prioritize areas for conservation and for the Country Study and National Strategy of Biodiversity requested by the Convention on Biological Diversity.

RECOMMENDATIONS

Biodiversity itself is distributed all over the Earth, with concentrations primarily in developing countries. In contrast, scientific biodiversity knowledge is concentrated in major centers in developed countries. To be useful in management and use of biodiversity, biodiversity information should be available when and where it is needed. At present, it is more likely that information on the plants of a particular part of Africa is stored in an herbarium in Europe, for example. Because it is not immediately at hand, biodiversity information is often not applied in policy or management decisions that affect the organisms involved. At present, scientists and others find it difficult to discover, access, and use biodiversity data that have already been collected, and to synthesize information from disparate sources. This is difficult to do because of the long history of "bottom-up" evolution of scientific biodiversity information and the mismatch between the distribution of biodiversity itself and the distribution of the data about it. In contrast, in disciplines that have emerged very recently, such as genomics (which has a history measured in mere decades in comparison to the centuries of history of biodiversity science), researchers have been able to capitalize on modern information technology to capture the data in digital form and make the data more readily accessible from the very beginning of their science.

An international mechanism is needed to make biodiversity data and information accessible world-wide. Attempts to integrate parts of the biological and ecological data matrix are occurring in a handful of projects (e.g., INBio in Costa Rica, Diversitas [an ICSU-UNESCO program for biodiversity research], and Species 2000), but these efforts need to be augmented and coordinated.

The existence of such a mechanism will produce many economic and social benefits. For example, the Convention on Biological Diversity (CBD) obligates nations to implement provisions relating to conservation, use and equitable sharing of biodiversity. A scientific information resource that could facilitate fulfillment of these obligations is greatly needed. Such a resource will also contribute to biotechnology and bioengineering, and therefore will be a central element in sustainable development. Because biodiversity is one of the primary measures of environmental impact, sound scientific information about it provides a way of determining whether development in a particular region is actually sustainable.

The Subgroup on Biodiversity Informatics recommends that the governments of OECD Member countries establish and support a distributed system of interlinked and interoperable modules (databases, software and networking tools, search engines, analytical algorithms, etc.) that together will form a Global Biodiversity Information Facility (GBIF). This facility will enable users to navigate and put to use vast quantities of biodiversity information, thereby advancing scientific research in areas such as agriculture, biomedicine, biotechnology, environmental management, pest control, health, education, and conservation, among others; serving the economic and quality-of-life interests of society; and providing a basis from which knowledge of the natural world can grow rapidly, in a manner that avoids duplication of effort and expenditure.

This Facility will be ***distributed***, while encouraging co-operation and coherence; ***global*** in scale, though implemented nationally and regionally; and ***open*** to participation and benefit by all countries, while having the majority of its support provided by those countries that have the greatest financial, scientific, and technical capacities to do so.

The Subgroup also recommends that governments accelerate efforts to compile data about living organisms and ecosystems, especially those whose existence is threatened and those of potential economic importance, and to enter these data (as well as existing biodiversity and ecosystems data) into databases that are interlinked via the GBIF.

The GBIF knowledge base and informatics tools will provide infrastructure support to information networking efforts such as, *inter alia*, the Clearing House Mechanism of the Conference of Parties to the Convention on Biological Diversity, the Environmental Information Organisation network of the European Commission, and the North American and Inter-American Biodiversity Information networks. The GBIF is complimentary to the "geographical-environmental locator service (GELOS)" activity of the Committee on Natural Resource Management of the G7, and will provide support for Diversitas activities (especially Core Program Element 3) in both OECD and non-OECD Member countries. Documents about initiatives such as these often presume that the informatics capabilities requisite to their missions already exist. In many cases, these capabilities are insufficiently developed to accomplish the goals of the initiatives. One purpose of the GBIF would be to promote the development of needed capacities so that goals of multiple organizations can be achieved without duplication of effort.

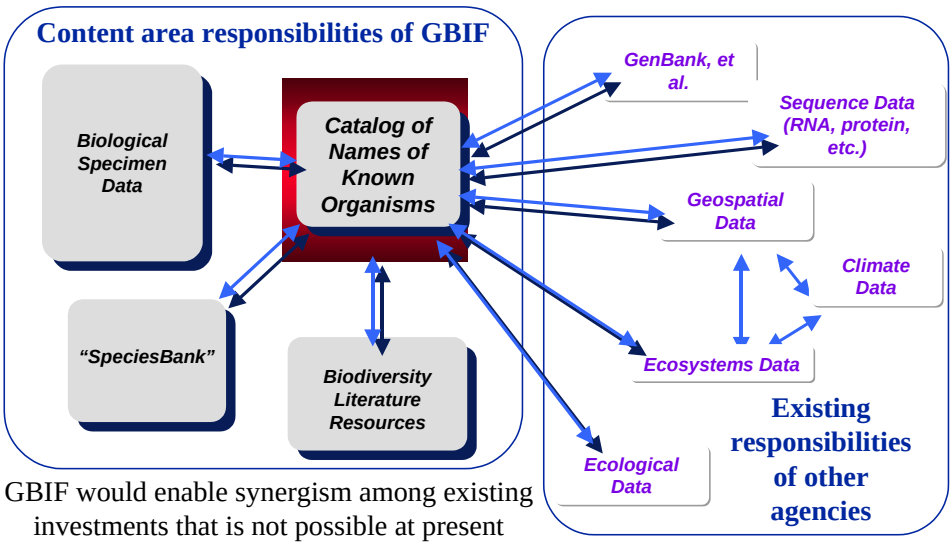

Figure 1. An Electronic Catalogue of Names of Organisms Known to Science will make linkages among many types of biological and non-biological databases possible. These linkages will enable "data-mining" that cannot be imagined today because at present it is difficult if not impossible to discover correlations among different data sets.

The GBIF will utilize Internet connections and other tools that are applicable to storage, dissemination, and sharing across networks of any type of information. Certain parts of the needed hardware and software technologies have been or are already being constructed for multiple uses, and the GBIF only needs to adapt them. However, there are unique, highly complex characteristics of biodiversity information that require the specific and concerted attention of information modelers, network specialists, computation technicians, and others. Distinct needs exist for co-ordination and prioritization of informatics software development and other projects to maximize global synergy and avoid duplication of effort. To do this, the GBIF stakeholders might identify, promote, and support best practices, standards, quality control and validation, metadata, cataloguing, development of authority files, and other resources that will allow the biodiversity community to link its data and information with digital spatial libraries, genome databases, and other major digital resources.

To enable GBIF, the Subgroup recommends that OECD member countries:

- ***Establish*** a Global Biodiversity Information Facility (GBIF).
- ***Support*** the GBIF Secretariat and GBIF programs financially, and ***appoint*** a representative to the GBIF Governing Board.
- ***Invest*** in related national and international activities that further the goals of the GBIF, such as:

    1. Contributing data, information, and capabilities to the GBIF. More specifically,
        a. data about whole organisms
        b. specimen data from biological collections
        c. environmental and remote sensing data
        d. molecular, gene and genome data
        e. new information and communication software and tools
        f. laboratory, computing, and training facilities.
    2. Promoting national involvement in GBIF by
        a. developing national nodes of the GBIF
        b. coordinating and synchronizing funding activities with respect to GBIF priorities
        c. expanding databases with data from collections, organisms, ecosystems, genes, etc.
        d. providing access to databases
        e. developing software and tools for information and communications technology
        f. adopting international standards
        g. improving high-speed networking and computation infrastructures
        h. enhancing infrastructures for
            1) data providers
            2) database custodians
            3) expert centers
        i. supporting training

3. Sharing
    a. computational facilities
    b. storage capacity to house major databases or operate database mirroring sites
4. Hosting
    a. the GBIF Secretariat (in whole or in part)
    b. specific GBIF projects

FUNDING

Within its agreed, segmented architecture, there are many viable options for investment in the GBIF, many of which will provide more than one return on the investment made. GBIF exploits the fact that big, monolithic, physical facilities are not necessary to improve biodiversity information provision or management. For example: OECD Member countries can elect to support the GBIF Secretariat monetarily and/or in kind (by hosting offices, etc.). Some countries may contribute to the infrastructure of the GBIF by providing incentives for software developments. Other countries may contribute to the knowledge base itself by providing a high degree of support for increasing the content of the world's data supply (particularly through digitization of museum specimen labels or library materials). Yet others may contribute to the "global" nature of GBIF by funding Internet connections for biodiversity institutions in currently under-supplied countries. Investors could opt to underwrite specific projects, such as a specific part of the Catalog of Names of Known Organisms (e.g. the International Plant Names Index).

A number of key initiatives exist that can be partnered with, built upon, expanded, or improved—e. g. Diversitas, Species 2000, the Integrated Taxonomic Information System, the Convention on Biological Diversity Clearing House Mechanism, the International Organization for Plant Information, the World Micro-organism Data Centre, etc. Critical information processing capabilities and tools exist or are near realization. Information management techniques (software developed to enable interoperability, analysis and synthesis of multiple data sets) are evolving to facilitate scientific conceptualization that includes knowledge from many information domains or scientific disciplines.

Another goal would be to move toward full and open provision of biodiversity data to all peoples to the maximum extent possible, while protecting intellectual property rights via carefully crafted legislation and regulation (current copyright and patent laws do not deal adequately with the intricacies of electronic information resources.) In addition, incentives might be designed to stimulate software developers, network entities, and information content providers to address the particular needs of biodiversity and ecosystem information provision.

Action Plan for
Global Biodiversity Information Facility (GBIF)

This Action Plan is a product of the Biodiversity Informatics Subgroup (BIS) of the Organization for Economic Co-operation and Development (OECD) Megascience Forum's Working Group on Biological Informatics. This document provides a provisional Action Plan for implementing the GBIF. Once initial governance and staffing for the effort are in place, this plan will be further improved and refined.

The GBIF will be established under the aegis of the OECD. Its Governing Board will be constituted by those countries that choose to support the GBIF. The Governing Board will be responsible for the selection and hiring of the Director and staff of the GBIF Secretariat, and for deciding among tenders for the siting of the GBIF Secretariat. The Secretariat staff will be accountable to the Governing Board, and will be advised as necessary by *ad hoc* Scientific and Technical Advisory Groups. The initial term of service of the Director and other Secretariat staff will be approximately 5 years. Scientific and Technical Advisory Groups will serve for only the amount of time needed to produce their reports, as requested by the Governing Board and the Secretariat.

Once five or more countries have elected to participate in the Governing Board and have appointed their individual delegates, the Governing Board can begin to function on an initial basis. The target is for the Governing Board to hold its first meeting before January 2000. Countries may elect to support GBIF at any time, and send delegates to future meetings of the Governing Board.

The GBIF Secretariat will work internationally to co-ordinate national and regional efforts. In addition, it will manage (through a competitive granting mechanism) a small amount of seed money (that is, a small percentage of the total funds necessary for the activities that it will encourage) to be used for leveraging activities being conducted by other agencies/countries.

The major programmatic areas/activities that the GBIF will co-ordinate and encourage include:

- **Access/interoperability/search capabilities** (the system that allows things to come together and makes links with and services existing regional and national networks)

  Deliverables:

  - Software/hardware infrastructure to enable

    1. linkages among/with molecular databases
    2. linkages between biological and non-biological information
    3. algorithms to search multiple databases simultaneously
    4. more rapid advance of scientific investigation

- Development of data and software standards

- Analysis of mechanisms for connections with existing databases

Estimated Costs:

- World-wide per-year average: $3.0 M. Of this, the majority will be in-country national investments. Approximately 5 per cent of this total should be delegated to the GBIF Secretariat so that it can encourage particular developments via peer-reviewed proposals from qualified projects. Ten-year, cumulative: $30.0 M. Suggested continuing investment yet to be estimated.

- Development of appropriate hardware and software tools will depend on focusing attention of software developers and others on this information domain. This is already under way in some countries, but can be enhanced by the coordination activities of the GBIF Secretariat.

- Such focusing of attention would be enhanced if the GBIF Secretariat has "seed money" that it can use to build partnerships among appropriate groups (biologists, information technologists and computer programmers, as well as information scientists).

- Many countries that choose to support the GBIF may have ongoing projects or current funding that already contributes in this area.

- **Electronic Catalogue of Names of Known Organisms** (encouragement and speeding up of processes already in place, such as Species 2000, to produce a reasonably complete catalogue within 10 years)

Deliverables:

- Content infrastructure to enable

    1. linkages among/with molecular databases
    2. linkages between biological and non-biological information
    3. more rapid advance of scientific investigation

- Co-ordination of various taxonomic reference files (e.g. NCBI list, Gray Card Index, et al.) and development of those that are at present unavailable, incomplete, or not digitized

- Analysis of status of biodiversity knowledge, which will facilitate more rapid scientific progress

Estimated Costs: Total cost of the electronic catalogue is estimated to be $280M, but it is also estimated that about $112M has already been spent or committed to this effort.

- World-wide per-year average (total among all countries contributing to the GBIF) over the next 10 years : $16.8 M. Ten-year cumulative (post GBIF-start): $168 M. Of this, the majority should be in-country national or regional investments, but approximately 5 per cent of the total should be delegated per year to the GBIF Secretariat for peer-reviewed dispersal among qualified projects.

- Ongoing (after the first 10 years) world-wide investment (to account for new species descriptions after the Catalogue is established): approximately $4 M per year.

- Development of the Electronic Catalogue of Names of Known Organisms is already well under way (approximately 40 per cent of species names are included in current projects), and certain of its components are already being funded at the national or regional levels.

- Encouragement and coordination of the Electronic Catalogue by the GBIF Secretariat would be enhanced if it had a small fund that it could use to leverage funding from other sources for those aspects of the Catalogue that are not yet started.

- **Digitization of natural history collections data, and provision of access to those databases**

Deliverables:

- Repatriation of data from developed to developing world

- Advancement of biodiversity science

- Analysis of status of biodiversity knowledge, which will enable prioritization of research

Estimated Costs:

- Per-year average: $20 M. (Of this, the majority will be in-country national investments, but approximately 5 per cent should be delegated to the GBIF Secretariat for peer-reviewed dispersal among qualified projects). Ten-year cumulative: $200 M.

- Ongoing world-wide investment: to be decided.

- Digitization of natural history collection data is already under way in many institutions, funded at institutional, national or regional levels.

- The rate of data input needs to be significantly enhanced if the repatriation of biodiversity data is to be accomplished relatively quickly. This can be achieved, at least in part, by

- "cooperative specialization" among institutions, which could be coordinated by the GBIF Secretariat.

- Encouragement and co-ordination of specimen-data digitization would be enhanced if the GBIF Secretariat had a small funding source that it could use to leverage funding from other sources for digitization of data from, for example, specimens of economically important species, invasive species, or endangered species.

- **"Species Bank"** (a means of access to information of all sorts about species, both known and new)

  Deliverables:

  - Automated searching of existing species descriptions to facilitate description of new species and automatic digital deposition of information as new species are described

  - Analysis of user needs for information about species ("users" includes individual scientists, Clearing House Mechanism, and national environmental agencies, for examples)

  - Understanding of gaps in knowledge about species that must be prioritized for research

  Estimated Costs:

  - Most of this work will be carried out by institutions funded at the national or regional level (actual total expenditures are extremely difficult to estimate). However, to encourage coordinated effort and leverage partnership initiatives, the GBIF Secretariat should have approximately $0.5 M per year for peer-reviewed dispersal among qualified projects. Ten-year cumulative: $5.0 M. Continuing investment: to be decided.

  - Development of SpeciesBank can be coordinated among various ongoing or new national, institutional, and industry efforts by the GBIF Secretariat.

- **Digital biodiversity literature resources** (essentially a digital library of biodiversity information compiled from information available in print libraries, prioritized according to quality, need, etc.)

  Deliverables:

  - Information on useful substances; for CITES and co-ordination of Red Lists; about invasive species; on species useful for detoxification; etc.,

  - Collation and correlation of various kinds of biological information

- Response to user needs for information

Estimated Costs:

- Most of this work will be carried out by institutions funded at the national or regional level (actual total expenditures are extremely difficult to estimate). However, to encourage coordinated effort and leverage partnership initiatives, the GBIF Secretariat should have approximately $0.5 M per year for peer-reviewed dispersal among qualified projects. Ten-year cumulative: $5.0 M. Continuing investment to be decided.

- Development of digital biodiversity libraries can be coordinated among various national, institutional, and industry efforts by the GBIF Secretariat.

- Encouragement and co-ordination by the GBIF Secretariat would be enhanced if it had "seed money" that it could use to leverage funding from other sources.

- **Training** (not only programs to produce informaticists/informaticians, but also providing assistance to developing world, to biologists about new informatics techniques, to computer scientists regarding needs of biodiversity and ecological sciences, etc.)

Deliverables:

- Scientists and resource managers in the developing world who have access to information and knowledge that has been unavailable to them.

- Experts who can advance the science and technology of biodiversity informatics.

- Biologists who will be able to work more efficiently because they can use new informatics tools, and computer scientists and information technologists who can deal with complex challenges presented by biological information.

Estimated Costs:

- Per-year average for the GBIF to promote and encourage the development of programs: $1 M. Ten-year cumulative: $10 M. Continuing investment: to be decided.

- The GBIF Secretariat will coordinate workshops that will provide instruction and provide travel funds for persons from institutions and countries that would be unable to support participation by their representatives.

- In addition, the Secretariat will encourage the establishment of permanent training programs in appropriate institutions, advise such programs as to curriculum and content, and other such activities as appropriate.

- Some projects of this sort are already underway; the GBIF Secretariat will work to enhance their effectiveness and increase their numbers.

- **Outreach** (working with other organizations to provide connectivity and digitization to the developing world, sharing technological developments, etc.)

  Deliverables:

  - Identification of other efforts with which the GBIF can interact and for which it can provide leverage

  - Good will resulting from data repatriation

  - Advancement of science and economy in developed and developing world

  Estimated Costs:

  - Per-year average for the GBIF to promote and encourage the development of programs: $1 M. Ten-year cumulative: $10 M. Continuing investment: to be decided.

  - The GBIF Secretariat would coordinate several workshops that would develop means of interactivity and compatibility among existing databases and provide travel funds for persons from institutions and countries that would be unable to support participation.

  - Establish connections among ongoing projects (benefits would include: avoidance of duplication of efforts, synergistic effects of cooperation, support for "information networks" at the national and regional levels, etc.).

  - Support workshops and other forms of communication to make such cooperative ventures happen.

  - Also, demonstrations of advantages will be necessary; it may also be necessary to work with other agencies to make equipment purchases and establish Internet connections to encourage participation, particularly in under-funded institutions in developing countries.

GBIF Secretariat Personnel and Budget:

1. Basics, such as office space, telephone, computer and networking resources, payroll and benefits accounting, etc., should be included in the bid package provided by the country(ies)/institution(s) that offer to house the GBIF Secretariat.

2. Personnel (salaries and benefits: ca. $1.5 M per year)
    - 1 Director
    - 1 Assistant Director
    - 3 Senior Program Managers
    - 3 Junior Program Managers
    - 2 secretaries (one of whom is also in part a public relations manager)

3. Travel expenses (ca. $1 M per year)
    - Director and Senior Program Managers (in their functions as international promoters and coordinators of GBIF-related activities)
    - Members of Scientific and Technical Advisory Groups (travel and expenses for meetings, but not salaries)
    - Members of the GBIF Governing Board (travel and expenses for meetings, but not salaries)

4. Other costs (ca. $0.5 M per year)
    - Legal, Intellectual Property Rights and other consultants, as necessary
    - Access to adequate computing and communications resources, document delivery and distribution, etc.

5. "Seed money" for leveraging the main activities of the Secretariat (co-ordination, encouragement, etc., of informatics activities on a global basis; to be dispersed via a peer-reviewed mechanism): Five to ten per cent of the total amount to be spent world-wide through national and regional programs according to the subsidiarity principle.

Memberships and Mandates:

1. Membership of Governing Board:
    - Countries that provide financial support
    - Countries that provide "in-kind" support, as approved by the Board
    - Ex officio: Executive Director (or other representative appointed by the Secretariat of the Convention on Biological Diversity) of the Clearing House Mechanism
    - Observer status (potential): NGOs, private foundations, industries, non-contributing countries
2. GBIF Governing Board mandate:
    - Global policy-setting for GBIF activities
    - Represent interests of their countries
    - Political brokering

- Assist in garnering resources
- Monitoring performance of the GBIF Secretariat
- Request bids for hosting of the GBIF Secretariat
- Appoint Director of the GBIF Secretariat

3. *Ad hoc* Scientific and Technical Advisory Groups:
   - Will be created as needed on an ad hoc basis
   - Specific tasks
   - Limited lifetime
   - Personnel who are acknowledged experts
   - Travel and subsistence but no salary
   - Number of members will vary according to task (typically 3-5)
4. GBIF Secretariat:
   - Report to/communicate with the Governing Board
   - Co-ordination, ambassadorship, workshops, etc.
   - Management of programs
   - Fund-raising
   - Liaisons with:
     - International standards bodies
     - Clearing House Mechanism
     - National contact points
     - Major institutions
   - Contracts (e.g. legal and intellectual property rights advice, etc.) as appropriate

ANNEX I: GBIF Action Plan: Programmatic Areas / Projects

Project / Program Area	Value Added to Existing Databases	Linkages Established	Utility Outside Science	Avoidance of Duplication	Contribution to Developing World	Performance Measures	Time frame
Linking / Interopera-bility Projects	• Correlation of information from a variety of sources will yield new scientific, economic, and resource management insights	• Linkages with geo-spatial, chemical, biochemical, molecular biological, genetic (etc.) databases will be enabled	• Correlations of information from various biotic and abiotic data sources will facilitate new scientific, technological and economic advances	• No other existing entity is specifically dedicated to the formation of such linkages; once the links are made, duplication will be unnecessary because GBIF will be accessible to all	• Scientific correlations and insights enabled by GBIF-sponsored interconnections will facilitate biodiversity discovery and science in developing countries	• Seamless integration of databases • Capability exists to search several databases simultaneously and report the combined results	• First 12 months of GBIF: Workshops held on developing standards for database interoperability. • Lifetime of GBIF: achieving database interoperability with robust search capabilities
Electronic Catalogue of Names of Known Organisms	• Essential infrastructure to enable full use of digital libraries and collections data • Essential to making linkages between biological and non-biological databases	• Will enable links among existing sequence databases because names are the unique identifiers included in those sources • Will enable links between taxonomic and ecological databases not at present possible • Will provide access to SpeciesBank , the digital biodiversity information library	• Is the unifying linkage among biological information sources (that in databases, physical libraries, and in natural history collections) • Will provide access to correlated biodiversity information for non-scientist users	• GBIF will enable the co-ordination of efforts already begun to concatenate the databases produced by various experts • The Catalogue, once completed, will obviate the need for individual researchers or organizations to create "taxonomies" for specialized purposes	• Is the unifying linkage among biological information sources (that in databases, physical libraries, and in natural history collections) • Will provide access to collections data, digital library information, etc.	• Biodiversity data providers are using the GBIF-encouraged Catalogue as the "authority file" for scientific names • Synonyms will be automatically handled by the information system	• By Year 4 of GBIF: 40 per cent of scientific names (including synonyms) electronically available • By Year 10 of GBIF: 90 % of scientific names (including synonyms) electronically available

Natural History Collection Data Digitization	• Will enable repatriation of biodiversity data to developing world • Will facilitate improvements in ecological and systematic research	• Links will be established between natural history museums to provide more thorough coverage of the globe for known species distributions • Digitized specimen data can be correlated with other biotic and abiotic information sources	• Natural resource management will be greatly enhanced through use of digitized specimen data combined with other ecological information • Economic uses of individual species can be better developed using the information stored in specimen data	• Co-ordination among natural history museums can reduce workload and increase coverage of distribution data • Thorough digitization of specimen data will enable prioritization of future collections research	• Digitization of the 75% of biodiversity specimen data held by institutions in the developed world allows these data to be repatriated to the developing world	• Pace of digitization of natural history collection data has increased markedly • The digitized data are used for cutting-edge research and to solve real-world problems	• First 12 months of GBIF: A prioritization plan is developed for dealing with digitizing of specimen data • By year 5 of GBIF: GBIF members have digitized at least 35% of their specimen data
Species Bank	• No database currently exists specifically to facilitate new species discovery and description	• Will link to any accessible existing database that holds either biotic or abiotic information about species	• Will facilitate searching of Internet resources by non-specialists (students, resource managers, et al.)	• Will assist taxonomists to avoid re-naming already-described species • Will facilitate rapid dissemination of information on newly discovered species	• Will repatriate information about species native to developing world	• Over time, increase in rate at which new species are described • Accessibility of species information to users (scientific and non-scientific)	• First 12 months of GBIF: Initial workshops to design SpeciesBank database and linkage structures • Lifetime of GBIF and beyond: Building of SpeciesBank information sets
Literature Resources	• Digital libraries are more accessible and more useful to more people than are physical libraries	• Digitization is needed to gain the ability to readily correlate data from multiple branches of biology	• Library search capabilities make it easier for non-specialists to gain needed knowledge	• Biological publications that have already been made available electronically need not be digitized again	• Physical libraries are difficult to duplicate, but digitized libraries can be accessed by either developed or developing world	• Accessibility of biodiversity information of all sorts to users (scientific and non-scientific) • Gradual appearance of a digital "knowledge base" drawn from physical library resources	• First 12 months of GBIF: Initial "networking" by Secretariat staff to identify partnerships; workshops to prioritize information sources for digitization • Lifetime of GBIF and beyond: Building of digital biodiversity information libraries

Training	• Will open new avenues of research both in computer science and information technology, and in biological sciences • Will expand user-base for GBIF information	• Interactions among the biological, computer science, information science, and information technology communities will be forged in training programs better than in any other GBIF activity	• Will provide an expert group of users and researchers who are fully able to take advantage of GBIF • Experts will be available to handle real-world biodiversity problems	• Model curricula will be of great value to universities and other training institutions • Experts with a common set of skills will be able to work co-operatively and effectively	• Research techniques and capacities will be transmitted to the developing world	• A cadre of biodiversity informaticians who can "mine" biodiversity databases for economically, politically and scientifically useful results	• By year 2 of GBIF: model curricula for biodiversity informatics training are developed • Lifetime of GBIF: experts who can fully utilize GBIF are trained in all countries
Outreach	• Will advance science, technology and economy in both developed and developing world	• GBIF Secretariat personnel will work constantly to form linkages wherever possible	• Will inform potential users of utility of the GBIF • Will expand the GBIF provider and user bases • Will engender good will in OECD and other countries • Will encourage scientists to provide data usable by non-scientist audience	• Every connection made will avoid duplication of effort because the Internet allows sharing and reorganization of the same information for different purposes	• Efforts constantly will be made to assure that developing countries receive GBIF benefits	• Internet linkages are developed everywhere • GBIF concept is accepted and adopted universally	• First 12 months of GBIF: plan is developed for catalyzing development of Internet connections in all countries • Lifetime of GBIF: GBIF, in partnership with other organizations, implements Internet connections and promotes GBIF concept

GBIF Action Plan: Management Structure

Time	Personnel	Actions
Prior to "Time Zero"	Subgroup members	• Draft position descriptions for GBIF Secretariat Director and staff • Draft tender for hosting GBIF Secretariat • Lay out initial ideas for first feasibility studies, descriptors for initial Scientific and Technical Advisory Groups, etc.
"Time Zero" (OECD Ministerial meeting)	OECD Ministers	• Recognize need for GBIF
	Five OECD countries invest in GBIF	• Initial funding for GBIF Secretariat secured
0 – 6 months	Representatives of countries that invested in GBIF	• GBIF Governing Board established
	GBIF Governing Board	• Select host country/institution for GBIF Secretariat • Hire GBIF Secretariat Director
	GBIF Secretariat Director (in consultation with Governing Board)	• Hire Program Managers, secretarial staff • Establish office for GBIF Secretariat • Initial activities (feasibility studies, task analyses, etc.)
6 – 18 months	GBIF Secretariat (Director and Program Mangers) and consultants as necessary; appropriate Scientific and Technical Advisory Groups	• Feasibility studies (technologies, IPR, etc.) • User needs surveys • Initial workshops (data standards for interoperability; prioritization of specimen digitization, structure of SpeciesBank, prioritization of digital library resources) • Initial "networking" by GBIF Secretariat to identify and establish partnerships • Plan for catalyzing development of Internet connections in all countries

2 years	• Researchers at appropriate institutions (efforts coordinated by GBIF Secretariat) ♦ Database owners and providers	• Model curricula for training in biodiversity informatics developed • Emerging standards for interoperability • Specimen digitization efforts underway • Host institutions for SpeciesBank identified, initial development of database underway • Internet connections installed in countries/institutions with greatest need • Continuation of "networking", workshops as needed ♦ Increase in rate of data entry of specimen data ♦ Initial digitization of library resources ♦ More databases available and inter-operable
4 years	• Researchers at appropriate institutions (efforts coordinated by GBIF Secretariat); database owners and providers ♦ GBIF Secretariat	• Electronic Catalogue of Names of Known Organisms contains 40 per cent of all scientific names • Significant increase in interoperability of biodiversity databases • Significant increases in amount of data available via GBIF ♦ Continuing and ongoing coordinating function: User needs analyses; networking; workshops (as appropriate)
5 years	• Researchers at appropriate institutions (efforts coordinated by GBIF Secretariat); database owners and providers ♦ GBIF Secretariat	• 35% of natural history specimen data digitized and available via the Internet • Significant increase in interoperability of biodiversity databases • Significant increases in amount of data available via GBIF ♦ Continuing and ongoing coordinating function: User needs analyses; networking; workshops (as appropriate) ♦ Internet connections are available in all countries

10 years	• Researchers at appropriate institutions (efforts coordinated by GBIF Secretariat); database owners and providers ♦ GBIF Secretariat	• 85% of natural history specimen data digitized and available via the Internet • Electronic Catalogue of Names of Known Organisms contains 90% of all scientific names • Biodiversity databases are inter-operable • Information content of GBIF has captured most important of the information in physical libraries ♦ Continuing and ongoing coordinating function: User needs analyses; networking; workshops (as appropriate)
Lifetime of GBIF	• GBIF Secretariat, researchers, co-operating agencies, users, providers, et al.	• Database interoperability with robust search capabilities • SpeciesBank capabilities increase • Biodiversity information digital libraries grow • Personnel trained to use all GBIF capabilities found in all countries • Internet connections and GBIF concept accepted in all countries

ANNEX II: FINDINGS

This Annex is organized into three Findings, each of which is organized into three sections: a Situation Analysis of the Social and Scientific Context, Priority Policy Adjustments, and Role of the OECD.

FINDING 1: THE BIODIVERSITY INFORMATION DOMAIN IS VAST AND COMPLEX, BUT CRITICALLY IMPORTANT TO SOCIETY.

Societal and Scientific Context: Situation Analysis

Our knowledge of biodiversity, even though incomplete, is a vast and complex information domain. The complexity arises from two sources:

1. The underlying biological complexity of the organisms themselves: There are millions of species, each of which is highly variable across individual organisms and populations. These species each have complex chemistries, physiologies, developmental cycles and behaviors, all resulting from more than three billion years of evolution. There are hundreds, if not thousands, of ecosystems, each comprising complex interactions among large numbers of species, and between those species and multiple abiotic factors.

2. The overlying complexity introduced by variations in the data about them: This is engendered by differences (historical, philosophical, educational, etc.) among the people who collected the data. The manner and mechanisms that have been employed in biodiversity data collection and storage are almost as varied as the natural world the datasets document. The range of biodiversity data types includes not only text and numerical measurements, but also images, sound, and video.

The variability and structural complexity of biodiversity information constitutes a set of challenges within both the realm of management of and research on biodiversity itself, and within the realm of complex knowledge engineering. In order to exploit what is known, and expand that knowledge through research, it is important that these challenges be met.

Examples of major scientific and biotechnological questions in biodiversity include:

- Exactly how diverse is life on Earth? That is, how many species exist? (According to the *Global Biodiversity Assessment*, science has discovered and described about 1. 5 to 1. 75 million species out of 12 million or more, or roughly 1 out of 8).
- How do organisms function within their environments so that life sustains life?
- What are the opportunities to sustainably use biodiversity?

- How can we learn from nature and apply what we learn to aid sustainable development?
- What is the complete suite of attributes of each organism of interest — biochemistry, genome, physiology, reproductive cycle, behavior, and so forth? How do these attributes affect other organisms, including humankind?

Such research questions increasingly demand that

- A much greater proportion of the global portfolio of biodiversity information be captured in digital form, and that
- Biodiversity databases interoperate with a broad range of other kinds of datasets: molecular, genetic, geographical, meteorological, geological, chemical, etc.

Biodiversity is an extraordinarily challenging field for the application of scientific conceptualization and research, but unifying principles do exist (from Linnaean taxonomy to genomics). These principles do and will allow for development and progress in managing biodiversity information. Among OECD Member countries and elsewhere, leading practitioners have demonstrated the potential for biodiversity informatics to make tremendous progress in the Information Age.

Priority Policy Adjustments

Governments, in formulating economic and environmental policies, should be cognizant that biodiversity information is a very important input needed to enable a faster shift toward sustainable development and to meet international obligations under the Convention on Biological Diversity and other such instruments. To properly exploit this information, policies should be put in place that encourage the application of the tools of modern informatics to biodiversity information. The logical outcome of such policies would be the construction of an international, cooperative framework that will allow the acceleration of and focus for biological exploration and discovery of new information, while at the same time maintaining support for retroactive electronic capture of valuable, but currently archived, data.

It is particularly in OECD Member countries that biodiversity information needs can be met with new technological potentials. These can be used to enhance the value of existing information assets and generate efficiencies that will make information useable for many purposes, including both informed decision-making and scientific research that will generate new information. An integrated global project is needed so that questions that range 1) from molecular levels to ecosystem levels, 2) from individual researchers and institutions to the global enterprise, and 3) from local or regional to national governmental levels may address the issues of complexity and scaling found in each of these areas of endeavor.

Role of the OECD

OECD Member countries would greatly benefit from creating and sustaining the Global Biodiversity Information Facility (GBIF). Other countries could easily be associated with this effort.

Creating and sustaining the GBIF will require international leadership, a clear focus, and targeted, leveraged monetary investments in catalytic resources to hire and house personnel to guide the growth and evolution of the GBIF.

FINDING 2: AT PRESENT, EXISTING BIODIVERSITY AND ECOSYSTEMS INFORMATION IS NEITHER READILY ACCESSIBLE NOR FULLY USEFUL.

Societal and Scientific Context: Situation Analysis

The western scientific tradition has had biodiversity and ecology at its core from its beginnings as “natural philosophy ” over two thousand years ago. Many of the earliest surviving ancient books contain inventories of plants and animals and commentary on the effects that certain species had on their surroundings. Our legacy of information on biodiversity exists in distributed and manual (non-digital) form as books, journals, card files, and notebooks, stored in libraries around the world, that have accumulated primarily over the last 250 years. There is also an undocumented but highly valuable store of knowledge in the brains of specialists who are alive now but may not be for long. In addition, there is an unmeasured store of indigenous knowledge of biodiversity all over the globe that has as yet gone unrecorded and is in danger of being lost. Thus, though there exists a huge accumulation of knowledge, the corpus of that knowledge is fragmented and difficult to access.

The global biodiversity informatics enterprise, at present, lacks the infrastructure that would achieve ready accessibility of information so that it can be compiled, organized, coordinated, and made amenable to analysis in a timely fashion. There is no facility that uses the power of computers and networks to allow researchers to draw on distributed information in a manner that nets the given user all of the information needed but only the information needed—that is, to take a sip but not be overwhelmed by an ocean wave. An important issue is to create agreed reference-coordinate systems. Many aspects of biodiversity information management could be improved dramatically if such systems were put in place.

Priority Policy Adjustments

Governmental (national, international and regional) funding decisions should reflect recognition of the need to convert legacy biodiversity information to digital form. The biodiversity research community is at present moving data into digital form in a slow, haphazard and uncoordinated fashion. Current intellectual and monetary investments are not optimally exploited to obtain synergies and accelerate progress. However, biodiversity information needs to be accessible world-wide, for legitimate use by various sectors of society.

It will be necessary to

a) help provide and support practical tools for computerization, networking, modeling, geographic information systems (GIS), remote sensing, and other related projects that are well-matched to operational needs. OECD Member countries should promote use of biodiversity informatics resources in environmental assessment and monitoring, and by field biologists for local assessment, exploration and other purposes; and

b) encourage electronic capture of data as research is conducted, and provision of those data over the Internet. A possible model is the requirement by many journals and employers that gene sequences be uploaded to GenBank prior to publication of papers that discuss those sequences.

In order to encourage active participation by biologists of all disciplines in the GBIF effort, it would be an advantage to modify and adapt the culture of academic and career credit to reflect current best practices and promote the "publication" of research results as electronic databases. This is already underway in some disciplines: Hubble telescope data and GenBank sequence data are immediately available to all researchers via the Internet. Further, the contribution of sequence data into molecular databases is recognized as equivalent to publication of those data.

Role of the OECD

Countries that participate in the GBIF would benefit by establishing and supporting a coherent national informatics structure or organization that will link to and support the international efforts of the GBIF.

Participating GBIF countries should be aware of how their individual actions fit the global architecture and provide appropriate levels of investment to accelerate the digitization of legacy data, networking, synthesis and analysis within a scalable structure that can also accommodate future additions and changing needs.

GBIF participant countries should work towards eliminating barriers to the full and open sharing of biodiversity information across political boundaries by adhering to and promoting principles of “fair use” of such information in research and education.

FINDING 3: RECENT TECHNOLOGICAL AND POLITICAL DEVELOPMENTS PRESENT LEADERSHIP OPPORTUNITIES FOR OECD COUNTRIES

Societal and Scientific Context: Situation Analysis

The Internet, World Wide Web, and other key information technologies have enormous potential to accelerate research and technological developments in biodiversity science, as they have already done for other disciplines. New scientific ideas and results can be developed from existing data (e. g., “macro-ecological modeling”) as new informatics tools and capabilities are developed and used (iterative cycling), allowing the comparison and synthesis of information from many sources and disciplines. OECD Member countries hold a large percentage of the global inventory of biodiversity information that is needed world-wide. The circumstances of the acquisition of these data and the necessarily global nature of the effort to sustain biodiversity make it imminently desirable

to provide and enable access to these data for all countries who might employ them. OECD Member countries are well positioned to create a political and legal framework and a flexible network that can provide comprehensive information to satisfy the evolving needs of stakeholders.

A number of key initiatives exist that can be partnered with, built upon, expanded, or improved—e. g., Diversitas, Species 2000, the Integrated Taxonomic Information System, the Convention on Biological Diversity Clearing House Mechanism, the International Organization for Plant Information, the World Microorganism Data Centre, etc. Critical information-processing capabilities and tools exist or are near realization. Information management techniques (software developed to enable interoperability, analysis and synthesis of multiple datasets) are evolving to facilitate scientific conceptualization that includes knowledge from many information domains or scientific disciplines.

Many tools are in place, but there are certain critical areas in which additional developments, tailored to the characteristics of biodiversity information, are important to progress: broadband network connectivity, intelligent optical character recognition, intelligent routing systems, rapid indexing of large scale full-text and image files; agentry ("know-bots"), and OPM or CORBA object brokering are primary examples.

Because they do not perceive potential for economic gain, software developers and information companies have not focused on the needs of biodiversity informatics although it is a highly challenging area for research endeavor.

Priority Policy Adjustments

Software developers should be encouraged, through incentives put in place by governments, to create tools for biodiversity informatics. These incentives could include governmental, industrial, scientific, and other markets for the products of their development efforts in this area.

Distinct needs exist for coordination and prioritization of informatics software tools development and other projects to maximize global synergies and avoid duplication of effort. To do this, GBIF countries might identify, promote, and support best practices, standards, quality control and validation, metadata, cataloguing, development of authority files, and other resources that will allow the biodiversity community to link its data and information with digital spatial libraries, genome databases, and other major digital resources.

Role of the OECD

If this effort is to global impact, it is the OECD Member countries which must take the initiative. Among the priority targets is the completion and maintenance of globally ubiquitous Internet connectivity (actual hardware connections) so that there can be information interchange among all countries in a manner that takes maximum advantage of technological advances.

Another goal would be to move toward full and open provision of biodiversity data and information to all peoples to the maximum extent possible while protecting intellectual property rights via carefully crafted new legislation and regulation (current copyright and patent laws, established for print and other physical media, do not deal adequately with the intricacies of electronic information resources).

In addition, tax incentives might be designed to stimulate software developers, network entities, and information content providers to address the particular needs of biodiversity and ecosystem information provision.

ANNEX III: DEFINITIONS

1. **Informatics**: Research on, development of, and use of technological, sociological, and organizational tools and approaches for the dynamic acquisition, indexing, modeling, dissemination, storage, querying, retrieval, visualization, integration, analysis, synthesis, sharing (including electronic research collaborations), and publication of data and information such that economic and other benefits may be derived from them by users in all sectors of society.

2. **Biodiversity** (short form for "biological diversity"), as defined in the Convention on Biological Diversity, is: The variability among living organisms from all sources including, *inter alia*, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which they are a part; this includes diversity within species, between species and of ecosystems. This Subgroup of the Working Group on Biological Informatics would include "diversity among genes, individuals, and populations within and between species" in the definition.

3. **Biodiversity Informatics**: The application of informatics to recorded and yet-to-be-discovered information specifically about biodiversity, and the linking of this information with genomic, geospatial and other biological and non-biological datasets.

4. **Facility**: Something created to facilitate ease in acting, processing, or working such that value is added to a particular function or functions. A facility may be unitary, such as a large telescope, or may comprise a number of elements that are distributed but interlinked, such as numerous computer servers connected via a network.

5. **Legacy Data:** Information that is stored in a manner that is not electronically accessible by current methods, such as paper, analogue tapes, printed photographs, etc. Even digitized data that is stored on outmoded media may be regarded as legacy data. Such data may be invaluable, but the mode of storage presents challenges to making them accessible.

6. **Taxonomy:** The science of classification of organisms in an ordered system that indicates natural relationships.

7. **Taxon:** A general term for any taxonomic category or group, such as a phylum, order, family, genus, or species.

OECD Megascience Forum
Biological Informatics Working Group
Neuroinformatics Subgroup

Executive Summary

This Working Group was proposed and accepted at the January 1996 OECD Megascience meeting in Paris. Within this working group two subgroups were initiated, one on Biological Diversity Informatics and the second on Neuroinformatics. This report reflects the consensus of the Neuroinformatics Subgroup which met a number of times over a 24-month period to consider how best to achieve the goals of the Subgroup.

The goals of the Neuroinformatics Subgroup are to establish ways to develop the international capability for the acquisition, storage, analysis, understanding and sharing of data on the nervous system. The report reflects the added value of creating this capability and provides recommendations of actions needed to achieve the goals of the Subgroup.

Neuroinformatics is a new research area that will help to accelerate progress in understanding brain function in the 21st century. Neuroinformatics is interdisciplinary, combining research in neuroscience and informatics (including computation) to develop and apply advanced tools and approaches needed for understanding the brain. In its study of the competence and flexibility of the brain, neuroinformatics research is uniquely placed at the intersection of medical, biological, and behavioral science, the physical sciences, computer science, mathematics and engineering. The resultant synergy from combining these approaches will accelerate scientific and technological progress resulting in major medical, social and economic benefits.

Understanding the structure, function, and development of the human brain in health and disease is perhaps the great challenge in science for the 21st century, and is of interest to all members of society. Our brains mediate our perceptions of the world around us, generating the myriad thoughts, memories, and emotions that make us uniquely human, and control our immediate and long-term behavioral responses to the environment. Understanding our brains is essential for alleviating enormous health problems such as mental illness, memory loss and aging, drug abuse, and other brain diseases. It also has important applications for advancing Information Technology, for instance, in developing artificial systems implementing the types of processing and natural computation found in the brain. It may be said that neuroinformatics research helps to understand the brain; to heal the brain; and to create brain-like computer and robotics applications.

Understanding the brain is a major scientific challenge, because of the complexity of the brain. The human brain contains roughly 100 billion nerve cells, 3.2 million kilometers of "wires," a million billion connections, all packed in a volume of 1.5 liters weighing a

bit more than 1.5 kilograms, and consuming about 10 Watts of energy. **The Challenge is** to understand the brain through research on its structure and connections; on the operation of its neurons and genetic regulation and determination; on its pharmacology and molecular mechanisms; on the effects of damage to each part; on the activity found in each brain area during the performance of different tasks; and on how the networks of neurons actually operate to perform computational, cognitive, behavioral and maintenance functions. The data obtained from each of these areas must be, and for the first time can now be, combined in order to produce a computational model for how each part of the brain operates, and thus ultimately how the brain works as a system.

Indeed, neuroscience has so far been dominated by the acquisition of experimental data, and the time is propitious to facilitate the development of theoretical models, and tools to help manage and use the data to yield new knowledge and understanding. The **scientific goals of Neuroinformatics** are to accelerate the progress of neuroscience and informatics by:

- Making better and more efficient use of neuroscience data using informatics-based, including computational, approaches;
- Generating and evaluating new hypotheses and computational theories about brain function to drive further experiments;
- Developing and applying new tools and methods for acquiring, visualizing, and analyzing data important for understanding how the brain functions;
- Enabling the more efficient application of the accumulating knowledge of how the brain functions to be applied to understanding its dysfunction in disease; and
- Developing computer systems and technological applications that simulate or emulate specific aspects of brain function.

The **Lessons from Bioinformatics** must be applied to neuroscience through Neuroinformatics to create an information management system. When the Human Genome Project began in 1990, it was recognized that the information to be generated by the planned sequencing and mapping efforts would quickly overwhelm investigators, and that a concerted effort was needed to organize this information. In the Human Genome Project, research and development of bioinformatics tools were developed hand-in-hand with the ongoing biological research as the base-pair sequences were identified and mapped onto locations along the chromosomes. Today, bioinformatics tools allow scientists access to this genomic data, and also permit scientists to manipulate, visualize, and analyze it, as well as explore theoretical issues. Neuroinformatics will provide these capabilities to neuroscientists.

Some of the areas in which Neuroinformatics will have a major impact in the ways listed above are highlighted in the following paragraphs, and elucidated in the Report.

1. Neuroinformatics will enhance our **understanding of the brain** and the operation of its circuits. An overarching objective of systems neuroscience is to decipher how information is processed, stored, and translated into action by the intricate neural

circuitry of the brain. For example, specific cortical neurons are involved in pattern recognition and perception. There are neurons specialized for the analysis of spatially localized features contained in visual images, such as orientation, color, direction of motion, and stereoscopic depth. To understand how fundamental properties such as these arise within the intricate circuitry of the cortex, evidence from thousands of experiments on neuronal responses, connections, connection modifiability, and biophysical properties must be combined with computational approaches to network operations that can for example extract perceptually useful information from natural images. Neuroinformatics will play an important role here, by facilitating bringing together all the relevant evidence from different disciplines, and by bringing together expertise in empirical studies of the brain with that in information-processing systems.

2. Another area is in **understanding Cortical Maps**. Recent Neuroinformatics-based advances in computational neuroanatomy are providing new insights here. Cortical cartographers have charted the layout of different areas across the cortical sheet with greatly improved resolution and visualization capabilities, including, for example, digital reconstruction of the cortical surface. Many different types of information can be displayed on this surface-based cortical atlas, including the type of function such as vision, touch, hearing, etc.; major neuroanatomical subdivisions; and the mapping of neurotransmitters, their receptors and genetic regulatory mechanisms. Such maps, in a standard framework, can be continually revised and updated by emerging discoveries, with the relations between the mapped variables being made apparent.
3. Related to this, Neuroinformatics will facilitate the **Understanding of Brain Structure, Function, and Variability,** by providing new integrating tools. These tools are needed to deal with the integration of the many diverse types of data, including results obtained with different neuroimaging methods that are complementary in their spatial and temporal resolution, and in revealing the functional, structural, and biochemical machinery of the brain. The establishment of sophisticated dynamic digital brain atlases, and databases for each major species, including humans, will provide increasingly powerful capabilities for interactive querying, with flexible selection from a variety of volume-based and surface-based visualization formats. For any given location or region of interest, these atlases will provide probability distributions and confidence limits for the assignments of structure, function, behavioral characteristics, and (for the human brain) associated clinical measures.
4. The Neuroinformatics contributions towards **Understanding Cellular and Subcellular Level** processes involved in brain function are threefold: first, the modeling of cellular and subcellular processes through theoretical and computational approaches; second, the description, storage, and handling of empirical findings; and third, the advancement of methods and approaches of data acquisition. Nerve cells are specialized for different functions. The properties of each nerve cell are genetically regulated and result from the combination of its morphological properties and the vast amounts of ion channel and neurotransmitter receptor molecules that the cell is expressing and has anchored in its membrane. It is also the combination of these factors that is considered the neuronal substrate for learning and memory. Hence, in

order to understand the function of an individual nerve cell, multiple scales need to be considered, ranging from millimeters to Angstroms. In addition, effects described in electrical, biochemical, genetic, and spatio-temporal terms need to be taken into account and appropriately modeled for full understanding and the advancement of knowledge of brain function.

5. Neuroinformatics will **Impact Information, Communication, and Computing Technology** due to its focus on the most robust and efficient real-world information-processing device known: the brain. The bi-directional flow of information will influence the products and thinking in hardware technology as well as software technology, and will strongly influence both the fields of robotics and intelligent machines. There is a potential to create a positive feedback loop between the domains of Neuroinformatics and Computing and Information Technology, creating a unique synergy. Neuroinformatics will also stimulate developments in the field of neuromorphic engineering or bionics. In this domain, methods and approaches are developed that use alternative computational methods, for instance silicon analog Very Large Scale Integrated circuits (aVLSI) which can provide novel approaches towards emulating neural function.
6. It is projected that brain disorders will be the foremost source of disabilities and associated costs in the next millennium. From the **Health and Treatment** perspectives, Neuroinformatics will facilitate the understanding and, ultimately, diagnosis and treatment of brain disorders more efficiently and rapidly to meet the projected health crisis in brain disorders. This information will be utilized for example by the pharmaceutical industry to produce more effective therapeutic interventions.

To obtain all the evidence required to understand the operation of each part of the brain is beyond the capacity of any one country. This is due partly to the complexity of the brain and the fact that there are tens of thousands of neuroscientists throughout the world studying brain function. But it is also due to the fact that evidence relevant to understanding how the brain functions must be drawn from the many different methodologies and disciplines described above, including those used to study brain connectivity, neuron function, the computational properties of neuronal networks, and clinical disorders. It is for these reasons that Neuroinformatics is a Megascience issue.

Neuroinformatics will provide the tools and capacity to create a global information management system for the enormous and ever-increasing quantity of diverse data on the structure and function of the nervous system. Once available, it will accelerate our understanding of brain function in health and disease. The Three Key Challenges and Implications of Neuroinformatics are:

1. ***To provide better ways to draw upon and access all the data and findings obtained with these different methodologies in order to bring together all the evidence to understand how each part of the brain functions. Meeting this challenge will require global cooperation. Neuroinformatics will need to utilize current and future technological capabilities and advances to permit the appropriate creation of an information management system for the vast quantity and variety of neuroscience***

data that have been and will continue to be created at an increasing rate and greater levels of complexity.

2. ***To provide ways to enable the expertise needed to understand the brain, which comes from disciplines as far apart as anatomy, genetics and computational approaches to brain function, to be brought together and applied to the problem.***
3. ***To provide ways for our developing understanding of brain function to be applied to the information technology domain. Bringing together the fields of neuroscience and informatics will benefit each field, because of the unique way in which neuroscience and the understanding of brain function can stimulate and lead to the development of novel information technologies, while these novel information technologies have direct application to neuroscientific problems. Hence, the development of research at this interface will accelerate progress in both fields simultaneously.***

These challenges should be addressed at both the national and the international levels in order to develop a Global Neuroinformatics Capability. The steps include:

1. Improved integration, coordination, and standardization efforts in Neuroinformatics.
2. The development of new tools and approaches for data acquisition and dissemination.
3. The development and application of new tools, methods and techniques for the theoretical and computational study of the nervous system.
4. The establishment and promotion of National Neuroinformatics Nodes. These centers will be both real or virtual to link groups of different scientists. These centers will function as a national resource for technology and for coordination within each country and will serve as the link to the international community. The need is to link closely in interdisciplinary research programs experts trained in different disciplines, including neuroscience, computer science, mathematics and theoretical physics, and information technology, to enable the rapid advances in understanding how the brain functions that are now shown to be possible, to be actually realized, and to enable those advances to feed back into information technology.
5. Support for international collaboration in interdisciplinary Neuroinformatics research.
6. Enhanced global communication of research results through common distributed repositories of key neuroscience data. This will be facilitated by promoting appropriate software and hardware for different but interoperable distributed databases of information with access through smart retrieval systems.
7. Interdisciplinary international education and training initiatives should be promoted to train future generations of Neuroinformatics researchers.

These considerations lead to the following Recommendations.

RECOMMENDATIONS:

Neuroinformatics combines neuroscience and informatics research to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain. Neuroinformatics research is uniquely placed at the intersections of medical and behavioral sciences, biology, physical and mathematical sciences, computer science and engineering. The synergy from combining these approaches will accelerate scientific and technological progress, resulting in major medical, social and economic benefits.

The main recommendations of the Neuroinformatics subgroup are:

I. Establish a Global Neuroinformatics Capability

This capability needs to be developed as a network of Neuroinformatics facilities and approaches, distributed across many research centers around the world. This network of Neuroinformatics facilities will be diverse, with major foci representing the development and application of:

- Databases, increasingly capable of handling the full complexity and organization of the nervous system, from molecular to behavioral levels;

- Powerful new tools for data-acquisition, analysis, visualization and distribution; and,

- Theoretical, computational and simulation approaches, methods, and environments for modeling and understanding the brain.

II. Establish a Global Coordinating Capability

International coordination as well as national efforts are needed to assure success and proper implementation and a sustained capability. An international scientific coordinating body, the International Neuroinformatics Committee (INC), and an associated secretariat should be established through the support of the participating countries.

The implementation steps needed to achieve this Global Neuroinformatics Capability are:

- Establish the coordination, standardization and interoperability requirements needed for successful application, integration, stabilization and quality assessment of the distributed and local Neuroinformatics facilities.

- Enhance collaborative opportunities in Neuroinformatics, both nationally and internationally. Such collaborations should be highly interdisciplinary and can range from as few as two laboratories to large-scale centers, either real or virtual.

- Develop national and international opportunities for recruitment, education and training in Neuroinformatics and for stable career pathways in this emerging interdisciplinary area.

- Enhance technology transfer to industrial and clinical arenas (e.g., electronics, computers, robotics, health care, pharmaceuticals and educational areas).

To this end, during its mandate, the OECD Megascience Neuroinformatics Subgroup will continue to:

Phase 1: (i) work on the objectives and a set of specific early tasks for the coordinating body; (ii) initiate national and international Neuroinformatics activities; and (iii) broaden the awareness in the community and prepare an operational plan for the next Phase.

Phase 2: after approval the INC needs to be established as well as an operational plan for the program.

Phase 3: will include a feasibility assessment on the basis of pilot projects, while

Phase 4: is devoted to the main realization of the program.

OECD Megascience Forum
Biological Informatics Working Group
Neuroinformatics Subgroup

NEUROINFORMATICS: a developing field uniting neuroscience and interfacing it to the medical and information technology domains.

I. INTRODUCTION

The human brain is a marvelously complex and sophisticated organ. It mediates our perceptions of the world around us, generates the myriad thoughts, memories, and emotions that make us uniquely human, and controls our immediate and long-term behavioral responses to the environment. The effort to understand the structure, function, and development of the brain in health and disease represents one of the great scientific challenges of the 21st century. Meeting this challenge will require powerful new approaches for acquiring, representing, modeling, and communicating information about the nervous system of many different species. The interdisciplinary field of Neuroinformatics is poised to be an increasingly important driving force in this endeavor.

Neuroinformatics combines neuroscience and informatics research to develop and apply the advanced tools and approaches that are essential for major advances in understanding the structure and function of the brain. Neuroinformatics research is uniquely placed at the intersections of medical and behavioral sciences, biology, physical and mathematical sciences, computer science, and engineering. There is a strong synergy in these interactions, including a positive feedback loop from the interactions between informatics and neuroscience. This synergy can lead to a rapid acceleration of scientific and technological progress, which in turn will have major medical, social, and economic impacts.

Nearly all people are afflicted with a disease or disorder of the nervous system at least once in their lifetime, and often the affliction is seriously debilitating or even life threatening. While many important clinical advances have occurred in recent years, adequate treatments or cures are still lacking for the great majority of diseases and disorders of the nervous system. Strategic investment in Neuroinformatics research can have a strong leveraging role, accelerating progress in basic neuroscientific research and its translation to improved diagnosis and treatment in the clinical arena, in areas ranging from new pharmaco-therapeutic approaches to new prosthetic devices for restoring brain function. Success in this endeavor will markedly reduce the enormous social and psychological costs caused by pain and suffering of afflicted individuals and their families. Likewise, it will commensurately reduce the staggering economic costs caused by reduced productivity of the workforce and massive expenditures for medical care.

On a parallel front, Neuroinformatics can contribute to major technological advances in the areas of information and communication technology, robotics and intelligent machines, and the interface between machines and humans. Computers and robots of the 21st century will be increasingly reliant on strategies of information processing that are analogous to those used by the brain. Neuroinformatics provides the conduit for elucidating these information processing strategies and translating them to the industrial arena. Accordingly, judicious investments in Neuroinformatics research offer the prospect of greatly enhanced competitiveness in the industrial and economic climate of the 21st century.

This report aims to convey the scope and excitement of current efforts in the emerging area of Neuroinformatics research and to illustrate the potential for rapid progress on many promising fronts, leading to a truly global Neuroinformatics capability. We recommend that a number of steps be taken at an international level as well as at national levels, in order to develop a global Neuroinformatics capability. In brief, these steps include improved integration, coordination, and standardization efforts in Neuroinformatics; the development of new tools and approaches for data acquisition and dissemination; development and application of new tools, methods and techniques for the theoretical and computational study of the nervous system; establishment and promotion of Neuroinformatics centers; support for international collaboration in interdisciplinary Neuroinformatics research; and enhanced communication of research results through common repositories of key neuroscience data. To train future generations of Neuroinformatics researchers, interdisciplinary international education and training initiatives should be promoted. Finally, to provide continuity and vision during a period of rapid technological development, an appropriately constituted international body, the International Neuroinformatics Committee, should provide sustained oversight.

II. THE CHALLENGE

The human brain contains roughly 100 billion nerve cells, 3.2 million kilometers of “wires,” a million billion connections, all packed in a volume of 1.5 liters weighing a bit more than 1.5 kilograms, and consuming about 10 Watts of energy. In order to understand its functions and properties, we must bridge many levels of description from molecule, cell and synapse to perception, cognition and behavior. The nervous system contains circuits which solve specific tasks - whether detecting particular features in the environment, coordinating a group of muscles to produce a given type of movement, or storing different types of information. Some of these circuits mature as the nervous system develops, and become modified by experience - learning. The total sensory, experiential, mental, and behavioral repertoire of each individual, whether mouse, man or mollusk, is represented in the nervous system by the total number of neural circuits available at any given instant, and their potential for interaction.

Over the last century, neuroscience has matured as an empirical science. Individual neuroscientists focus on particular aspects of neural function, from the intrinsic function

of single molecules like ion channels, to global aspects like cognition, Alzheimer's disease or schizophrenia. Facts about different aspects of the vast area of neuroscience accumulate at an increasing pace, but the insights gained about the actual principles underlying the operation of the nervous system have accumulated much more slowly. One reason for this is that a great number of parallel, dynamic and interactive processes occur at the single nerve cell level, as well as on the circuit and systems level. The sum of these processes will result in some fragment of behavior. It is therefore difficult, often impossible, to deduce intuitively the net result of the different observations made experimentally. This problem is presently a central challenge of neuroscience.

The different levels of organization displayed by the nervous system are paralleled by the specialization of individual neuroscientists, each studying the brain at a particular level of description, tied to the methods and techniques available. This specialization has strongly contributed to the growth of the field. It has, however, also contributed to its fragmentation, which hampers further growth since it obstructs meaningful interaction between the numerous subdisciplines. Specialization, then, can both drive and impede progress. For example, a team of molecular neurobiologists might use the sophisticated tools of their discipline to identify a mutation in a gene that codes for a particular receptor and determine that this mutation alters the properties of the receptor. Such an observation can have important ramifications for fields outside of molecular neurobiology, such as neuropharmacology, behavioral psychology, and psychopathology. In many cases, however, practitioners of these disciplines would not come in contact with these new findings, or would not be able to integrate the findings into their current theories, since this type of data is so remote from current thinking in their own subdiscipline.

If specialization is impeding the dissemination of new information among brain and behavioral researchers, so too is the diversity and quantity of the data generated by these same scientists. This diversity derives from the wide range of species studied, from invertebrates to humans, as well as from the spectrum of levels of biological organizations addressed, including molecules, cells, systems of cells, brain regions, and behaviors. Additional diversity in the data pool is introduced by the many different methodologies used and by the interest of brain and behavioral researchers in understanding normal, genetically modified, and diseased states throughout the life span of a myriad selection of species. Next to the diversity of the data obtained in brain and behavioral research, its sheer amount is staggering, being generated by tens of thousands of researchers working around the world, and being reported in hundreds of journals and conferences.

In summary, despite the enormous growth in the number of facts and regularities observed in neuroscience, the current state of the field does not allow an efficient dissemination of these observations and especially their integration in coherent theoretical frameworks. The first challenge needs to be addressed by a large-scale effort directed at optimizing the ways in which neuroscience acquires, describes, stores, and communicates its findings. Although this challenge seems of a practical nature, its

solution implies a megascience effort focused at a further integration with modern information technology.

The second challenge, in contrast, is more fundamental and its solution requires at least a paradigm expansion, or shift. In order to explore the relationships between the different levels of description which constitute neuroscience, as opposed to providing further descriptions of observations at isolated levels, the methods and approaches employed in neuroscience need to be extended. For instance, in the last decade theoretical approaches, such as simulation studies and formal mathematical analysis, have become immensely powerful methods to bridge the gaps between the disciplines of neuroscience. It constitutes a complementary methodology for investigating the nervous system, integrating experimental observations on the molecular, cellular, network and systems levels, in order to formulate experimentally testable models of aspects of neural function. So far, however, these methods have only been accessible to a small group of researchers due to the specialized knowledge required, ranging from mathematics to computer science and electrical engineering. A concentrated, internationally coordinated effort is required to increase the availability and usability of these Neuroinformatics methods and approaches. Presently, only the large-scale application of these approaches will facilitate the further intellectual growth of neuroscience in order for it to meet the challenges of the 21st century.

Not only neuroscience can profit from the expanding field of information technology. Information technology, in turn, will profit strongly from an increased understanding of the feats of information processing accomplished by even modest invertebrate nervous systems, let alone those demonstrated by more advanced vertebrate nervous systems. For instance, one of the major technological accomplishments of the last decade of the 20th century is the successful landing and use of a robotic device on the surface of Mars. The control of this device, however, was fully dependent on earth based teleoperation. In this case, every step of the Sojourner needed to be planned in advance on earth, by a human operator, and subsequently transmitted to Mars. This example demonstrates that despite the enormous advances in our information and computational technology, it is still far removed from the perceptual and behavioral abilities displayed by the terrestrial nervous systems studied in neuroscience. Other areas where this understanding can be expected to have an impact is, for instance, in the development of compact, low power, “smart” devices needed to support the “portability revolution” in IT and communication technology. Hence, advances in our understanding of “natural computation,” supported through a large-scale effort in Neuroinformatics, will have immediate relevance to the information technology of the future. The interaction between neuroscience and informatics, through Neuroinformatics, is expected to provide a positive feedback between the two domains. This “Neuroinformatics loop” will not only transfer results between them, but in turn sustain and further enhance this interaction.

III. LESSONS FROM BIOINFORMATICS

When the Human Genome Project began in 1990, it was recognized that the information to be generated by the planned sequencing and mapping efforts would quickly overwhelm investigators, and that a concerted effort was needed to organize this information. In the Human Genome Project, research and development of bioinformatics tools were

developed hand-in-hand with the ongoing biological research as the base-pair sequences were identified and mapped onto locations along the chromosomes. Today, bioinformatics tools not only allow scientist's access to this genomic data, but also permit scientists to manipulate, visualize, and analyze it, as well as explore theoretical issues. While bioinformatics research and development funding comprises less than 5 percent of the total budget of the Human Genome Project, it is patently unimaginable to see how this project could succeed, and how that success could be translated into new economic, public health, and societal benefits without this effort, which only takes a fraction of the total investment. In this light, it is clear that scientific informatics and computing adds great value and vigor to the scientific domain for which it is elaborated.

The data and theories of neuroscience are again more complex than those presented by nucleotide sequences; indeed, such genetic aspects are but one of the many facets of neuroscience. The human genome project does provide an effective demonstration, however, of how a scientific field can profit from a close integration with informatics. A megascience Neuroinformatics effort to accelerate the growth of neuroscience is facing challenges that by far exceed those addressed in any previous bioinformatics project.

Finally, while Neuroinformatics research is just forming and such efforts are few compared to the whole of bioinformatics research, there are important areas where Neuroinformatics is clearly in a good position compared to the human genome project. For example, the research community interested in genetic sequence data is now beginning to develop the mathematical and computational models for exploring the functional implications of the information gathered. Thus far, in this arena, Neuroinformatics already has a strong tradition that is directly contributing to the advancement of neuroscience. In addition, the multidisciplinary interactions which the human genome project fosters are already deeply ingrained in the developing field of Neuroinformatics. Its current challenge, however, is the need for a dramatic enhancement of its scale of application.

IV. IMPACT ON UNDERSTANDING THE BRAIN

This section will review the contribution of Neuroinformatics to understanding the structure and function of the brain. The examples will focus on specific brain structures and approaches. They are chosen from a large set and therefore cannot be taken as a complete and exclusive description of Neuroinformatics and its role in neuroscience. They do provide a representative overview of the general areas of application and the available Neuroinformatics methods and approaches. Particular emphasis is placed on the future challenges facing Neuroinformatics and neuroscience.

Neural Systems and Circuits. An overarching objective of systems neuroscience is to decipher how information is processed, stored, and translated into action by the intricate neural circuitry of the brain. The sections below illustrate the pivotal role of Neuroinformatics in elucidating basic principles of information processing by single cells

and distributed networks of cells, in mapping the cerebral cortex, and in understanding the functional organization of the brain.

Neural circuits for pattern recognition and perception. The primary visual cortex (visual area VI) is the largest and most intensively studied subdivision of the cerebral cortex, which itself is the dominant structure of the mammalian brain. Neurons in area VI are specialized for the analysis of spatially localized features contained in visual images, such as orientation, color, direction of motion, and stereoscopic depth. To elucidate how these fundamental properties arise within the intricate circuitry of the cortex, many mechanisms have been hypothesized, and thousands of experimental studies have been published, often with seemingly conflicting findings. In recent years, Neuroinformatics research has addressed many of these controversies and helped focus attention on hybrid, cooperative mechanisms that are well suited for extracting perceptually useful information from natural visual images. Computational modeling efforts have also provided key insights concerning mechanisms of visual development and plasticity. For example, computational studies of learning rules that describe how the strength of neural connections are adjusted based on neural activity can account for many of the feature-analyzing capabilities of neurons that emerge as a result of visual experience.

The results of the local feature analysis carried out within area VI are transmitted through many stages of processing to areas that mediate a variety of high-level tasks. A particularly impressive capability is object recognition, the process that allows us to recognize countless patterns and objects, such as a familiar face, independent of the distance or perspective from which they are viewed. One prominent hypothesis is that the perception of an object as a coherent entity occurs because its component features are bound together through synchronized activity across a distributed population of neurons. However, alternative models and mechanisms have been proposed, and the nature of temporal coding and the dynamic routing of information in distributed ensembles of neurons remains an extremely challenging question.

Progress in attacking this and related questions will benefit from Neuroinformatics contributions on a number of fronts. These include advances in single-cell recording methods that will allow simultaneous recordings from large numbers of individual neurons (dozens or even hundreds) in freely behaving animals, plus new approaches to how information is encoded and processed by large populations of neurons arranged in widely distributed circuits.

From perception to actions and memory. Sensory processing and perception affect behavior in a variety of ways, both immediate and long-term. Many of the immediate effects involve the role of sensory signals in guiding and modulating the movements of our eyes, limbs, and bodies as we interact with the environment. The coordinated yet flexible nature of sensory-motor processing involves a variety of complex computations and transformations. Experimentalists and computational neuroscientists working together have made great strides in recent years in attacking these challenges. This has

also been the case in the study of memory, the process that extracts information from our ongoing perceptions and stores this information by changing the strength of connections within specific ensembles of neurons. Theoretical studies of how information can be stored and retrieved in networks of artificial neurons have been highly influential in guiding the formulation of increasingly sophisticated computational models of memory mechanisms in the nervous system.

Contributions to robotics and machine vision. Despite the blinding speed of modern digital computers, current robotics devices are typically inflexible and poorly coordinated, and machine-vision systems are generally unreliable at recognizing complex natural objects and patterns in real-world environments. Insights gained from studying biological systems are already having an impact on the design of robots and artificial vision systems, and there is good reason to believe that this impact will be far greater in the coming decade. The symbiotic nature of this interaction is also evident, because many current computational models of specific neural functions have been directly inspired by research in robotics and machine vision.

Understanding Cortical Maps. The cortical sheet contains a complex mosaic of perhaps 100 or more distinct areas that differ from one another in their structure, connections, and functional specialization. Although some cortical areas, such as area VI, are large and easily identified, in many regions the distinctions between cortical areas are very difficult to discern. This problem is compounded by the fact that in many species (especially humans) the cortex is folded into extensive convolutions that allows a large surface area, that contains billions of neurons, to fit into a compact cranial volume.

Recent Neuroinformatics-based advances in computational neuroanatomy allow cortical cartographers to chart the layout of different areas across the cortical sheet with greatly improved resolution and visualization capabilities. One set of tools involves digital reconstruction of the cortical surface onto a flat map. This flat cortical map allows the entire surface (including the 70 percent that is buried within the cortical convolutions) to be seen in a single relatively undistorted view (as with maps of the earth, cuts are made in the cortical surface to reduce distortions). Many different types of information can be displayed on this surface-based cortical atlas, e.g., regions associated with different functional modalities (vision, touch, hearing, etc.) can be simultaneously represented by shading in different colors. Simultaneously, the map can also show major subdivisions from a partitioning scheme generated by classical neuroanatomists at the beginning of this century.

This flat cortical map is not a static entity, but rather one snapshot from an ongoing progress report that can be revised and updated by discoveries emerging from recent advances in neuroimaging capabilities. Future advances, particularly those involving simultaneous application of multiple techniques, such as functional Magnetic Resonance Imaging (fMRI) and structural MRI combined with magnetoencephalography (MEG), electroencephalography (EEG), Event Related Brain Potentials (ERP), Positron Emission

Tomography (PET) or optical imaging, will improve even further the resolution with which brain function can be mapped onto brain structure.

V. UNDERSTANDING BRAIN STRUCTURE, FUNCTION, AND VARIABILITY

To capitalize fully on the advances in neuroimaging techniques, several major technical challenges must be addressed. One challenge is to cope with the staggering amounts of experimental data continuously emerging. A single fMRI experiment may generate more than 1 gigabyte of associated data; the total number of such experiments carried out per year is in the thousands and is rapidly increasing. Another challenge is to integrate the many diverse types of information routinely attained using methods that are complementary in their spatial and temporal resolution and in the types of functional, structural, and biochemical information they provide. A third challenge is related to the high degree of individual variability in brain structure and function. Variability is particularly pronounced for the cerebral cortex, the convolutions of which differ as much from one person to the next as do their fingerprints. Variability in the functional organization of the cortex is presumed to account for many of the differences in behavior and capabilities that define our unique personalities.

These factors motivate the establishment of sophisticated digital brain atlases and databases for each major species, including humans. This has been aided by the development of probabilistic atlases based on a large population of subjects. Future atlases will provide increasingly powerful capabilities for interactive querying, with flexible selection from a variety of volume-based and surface-based visualization formats. For any given location or region of interest, these atlases will provide probability distributions and confidence limits for the assignments of structure, function, behavioral characteristics, and (for the human brain) associated clinical measures. These atlases will be highly dynamic, allowing efficient incorporation of data from a rapidly increasing population of experimental subjects. They will use advanced brain-warping methods to improve registration of data obtained from different individuals. This will not only reduce experimental uncertainties but also allow a more precise assessment of genuine individual differences. Brain warping will also be applied to comparisons across species, thereby enhancing the relevance of laboratory animal studies in efforts to understand the human brain.

Probabilistic brain atlases will greatly increase the opportunity for fundamental hypothesis-driven studies about brain function and dysfunction. These studies will allow for syntheses from the combination of mathematical, computational, and statistical approaches designed by Neuroinformatics experts with anatomical and physiological approaches designed by experimental neuroscientists. They will lead to new insights and perspectives on the principles underlying the organization of the human brain in health and disease.

VI. UNDERSTANDING THE CELLULAR AND SUBCELLULAR LEVEL

The Neuroinformatics contributions towards understanding the properties of cellular and subcellular processes considered are threefold: first, the modeling of cellular and subcellular processes through theoretical and computational approaches; second, the description, storage, and handling of empirical findings; and, third, the advancement of methods and approaches of data acquisition. Each of these contributions will be discussed. The examples used should be considered as representative of Neuroinformatics activities in the field and not as an exhaustive description.

The richness of neuroscience information at the cellular and subcellular levels. Neural function ultimately depends on the rapid conduction of information along the fine processes (axons) of the nerve cells through electric signals (action potentials), and chemical transmission between nerve cells at their contact points (synapses). Nerve cells are specialized for different functions. Some classes of neurons are faithful relay interneurons, others generate spontaneous patterns as pacemaker neurons, and yet other classes of neurons have complex dendritic (receiving) areas in which very intricate processing takes place involving thousands of different synaptic inputs. The properties of each nerve cell result from the combination of its morphological properties and the vast amounts of ion channel and neurotransmitter receptor molecules that the cell is expressing and has anchored in its membrane. Ion channels determine the behavior of the cell and its dendritic processes by balancing a number of inward and outward currents across the cellular membrane. Synapses may excite, inhibit or modify the properties of the ion channels present in some membrane compartment of the target cell, and induce a change in the local balance of currents. In other words, synaptic input can significantly alter the properties of the target nerve cell. The modification of cellular functions can also result from the activity of additional neurotransmitter receptors that can act through complicated signaling cascades. These effects can range from the modification of ion channel properties, for instance, closing or opening specific ion channels, to the transcription of particular genes inducing, for instance, morphological changes to the cell. It is also the combination of these factors that is considered the neuronal substrate for learning and memory. Hence, in order to understand the function of an individual nerve cell, multiple scales need to be considered, ranging from millimeters to Angstroms. In addition, effects described in electrical, biochemical, genetic, and spatio-temporal terms need to be taken into account.

Theoretical and computational approaches. One important area of Neuroinformatics research is to develop models of the large variety of known cells that form the brain. These approaches allow the exploration of the computational functions performed by neurons including the role and interaction of the different complex cellular morphologies, their blend of ion channels and receptors, and the subcellular events of, for instance, transmitter release, binding, and reuptake. Over the last several years, several physiologically based mathematical and computational methods have been developed which have generated powerful heuristics in the study of cellular and subcellular processes. The application of these methods has generated insights in a wide variety of

cellular and subcellular phenomena. For instance, important questions regarding the signal transduction performed by a nerve cell focus on the effect of synaptic events on the ability of the cell to generate an action potential. The actual electrical state of the membrane, at the moment such an event occurs, can play a decisive role in the contribution of a synaptic event to the subsequent generation of an action potential by the cell. In addition, the conduction of a synaptic event to the cell's soma will depend on the dynamic balance of the different inward and outward currents. Hypotheses on how the interaction between these different processes affect the behavior of a cell are difficult to address directly using current empirical methods, given the many spatial and temporal levels of interaction involved. It is exactly in this realm that theoretical and computational efforts have made their contribution by first translating the available data in testable model assumptions and subsequently providing hypotheses that can be tested with available experimental methods. One example is the interaction between neuromodulators and signal propagation in dendritic trees. Synapses can be placed at varying positions along a dendrite of a receiving cell, from very near to the soma to very distant. Dependent on the actual balance of the inward and outward currents in the dendrite, however, the signals of only a specific subset of synapses can effectively propagate to the soma, where ultimately the action potential is generated. This raises the question of how the cell can select this effective set of synapses. By means of modeling studies researchers have evaluated novel approaches to this issue, exploring the possible role neuromodulators could play, given their ability to change the dynamic balance of inward and outward currents. This, in turn, has created a set of hypotheses on dendritic function which have been introduced into the empirical domain. The above example illustrates how Neuroinformatics approaches can facilitate empirical research by providing efficient and rapid exploration of scenarios on cell function, spanning several levels of description, and, in addition, how these approaches can be used to summarize a large set of otherwise unrelated empirical findings.

Synthetic approaches. Another arena where neuroscience will profit from present developments in Neuroinformatics is the field of neuromorphic engineering or bionics. In this domain, methods and approaches are developed which use silicon analog Very Large Scale Integrated circuits (aVLSI) which can provide novel approaches towards emulating neural function. For instance, in modeling the nervous system, a practical problem is the demand for close to real-time performance in order to simulate large neural systems, while including sufficient detail at the level of the biophysics of single neurons. With the present technology, this is not feasible. Modeling a neuron, including a detailed description of its intricate morphology, and all the different channels, receptors, and currents, will necessarily limit the computational performance of the model. Presently, in the field of aVLSI design, technology starts to emerge which allows the development of special-purpose hardware that implements neurons with a sufficient degree of realism. The behavior of these silicon neurons is practically faster than real-time. This technology could prove to be an important component of the Neuroinformatics approaches developed and applied in neuroscience. Effort should be placed in further expanding this technology and in the construction of hybrid digital-analog systems focused on simulating large-scale neuronal systems interfaced to behaving devices.

With the increased possibility to construct so called nanomachines, alternative ways to define probes for neuroscientific research become available. Possibilities could be the construction of very small integrated electrode-amplifier systems, or the construction of "autonomous" probes that would navigate along neuronal processes using known biochemical markers. This type of technology could, on the one hand, increase the number of recording sites and, on the other, add a higher precision of targeting.

Description, storage, and handling of empirical findings. The sheer amount of empirical findings relevant to the understanding of cellular and subcellular processes is staggering. It is the result of decades of intense research by a large number of scientists using a wide range of methods. To address this problem, a number of groups have started Neuroinformatics projects that allow a more efficient storage and use of this data. One example is the construction of a database that contains the basic properties of a large set of identified neurons, such as their membrane channels, neurotransmitter receptors, and expressed neurotransmitter substances. This database, in turn, serves as the basis for constructing models of these neurons. In this way each model constitutes a computational tool for analyzing the functional properties of a given neuron in parallel with experimental analysis of that neuron. This greatly enhances the insights gained from experiments, and puts the functional interpretations on a firm theoretical foundation. In this project, the system for storage, analysis and synthesis is tightly coupled to the modeling environment that makes the step from the empirical to the computational realm much more efficient.

Next to integrating data derived from different levels of description for a given neuron, neuroscientists need to compare data across different neurons and species in order to determine general principles underlying their function. Some investigators have approached this problem by constructing special-purpose Neuroinformatics tools that establish relationships between the properties of different neurons. The essence of these tools is that from any component of any neuron in a database, a researchers can select a particular property, such as an ion channel, neurotransmitter receptor or neurotransmitter, and query the database for the presence of that property in any other described neuron. For example, a query for voltage-gated sodium (Na) channels in the soma brings up the soma compartments of all the matching neurons in the database. Currently, the distribution of voltage-gated Na channels in dendrites is of considerable experimental interest. A query on these channels in distal dendrites brings up a range of neuron types including the described variations in channel density; links to citation databases enable the user to assess immediately the experimental data, and links to the modeling environment enable the user to assess immediately the functional consequences of differing channel densities and other properties. These Neuroinformatics tools will become indispensable as neuroscience data continues its growth beyond the abilities of single researchers or laboratories to keep up with work that may be critically relevant. But, especially they will reestablish the necessary links between the different disciplines of neuroscience which study cellular and subcellular processes.

Advancement of methods and approaches to data acquisition. Central to the scientific method is the ability to observe and measure. In the study of cellular and subcellular processes, traditionally physiological techniques were applied, sometimes *in vivo* but more generally *in vitro*, using electrodes complemented with pharmacological methods. Despite the wealth of information these approaches have generated on the function of neurons, they have only provided a limited insight into the behavior of neurons in *in vivo* situations, such as their activation during particular behaviors. Recently, new visualization methods have been developed (through the close collaboration of neuroscientists, physicists, mathematicians, and computer scientists) that have extended the available physiological methods to image neurons in the intact brain of behaving animals. These novel imaging methods, using a combination of fluorescent indicators, such as calcium or voltage sensitive dyes, and confocal or two-photon microscopy, have provided the spatio-temporal resolution necessary to analyze cellular and subcellular processes under *in vivo* conditions. In certain applications, these methods allowed the direct evaluation of the hypothesized role of particular neurons in the generation of behavior.

In other applications, the activity of second messengers and their effect on the behavior of neurons or the release and fusion of neurotransmitters, or the interaction of particular ions, could be visualized. These new methods will have a profound influence on the experimental methods of neuroscience, allowing the experimentalist to pose and explore questions that go beyond the limits of single subfields. In addition, the field will benefit from new powerful technology from the post-genomic era that is expected to accelerate the identification and functional characterization of expressed brain genes both in health and disease. For example, using high throughput mRNA (cDNA arrays) and protein (two-dimensional gel electrophoresis in combination with mass spectrometry) based technology one can readily identify novel genes and proteins that are differentially expressed under various physiological conditions. Novel genes can be characterized using the green fluorescence protein (GFP) expression system, which uses GFP as a marker to follow protein localization *in vivo*. Both the visualization of cells as well as the cellular localization of novel antigens will be greatly facilitated in the future thanks to the recent development of large phage antibody libraries, which in theory allow the isolation of antibodies against virtually any antigen.

In summary, it can be stated that it would not be possible to understand the complex and dynamic interaction that occurs in the living nerve cell and the circuits they form without an interaction with mathematical, computational, and synthetic approaches. These approaches are used to study the brain at many different levels, ranging from behavior and neuron systems to structure and function at the molecular level (e.g., receptor structures, ion channels and transduction processes). So far, however, these methods have not been sufficiently integrated in the neuroscience community. The available simulation environments, for instance, have been the results of the efforts of small groups of researchers. These environments have not reached the quality and portability of commercial software products. We foresee, however, a development in which practically

every researcher working in neuroscience will be able to utilize these Neuroinformatics methods and approaches in interaction with experiments to interpret, integrate, view, communicate, and derive underlying principles from their findings. Although new software has been developed at this level, there is a need for substantial improvements.

VII. IMPACT ON INFORMATION, COMMUNICATION, AND COMPUTING TECHNOLOGY

Information Technology (IT). The world IT (hardware, software, and computer services) market grows at an annual rate of about 10 percent, twice that of world Gross Domestic Product (GDP), and revenues in 1995 exceeded $500 billion. The role of Information and Communication technology (ICT) in international trade (11 percent of world-wide merchandise growing with 22 percent annually since 1993) and Research and Development (R&D) (a quarter of all business enterprise R&D) is especially significant (data from Organization for Economic Cooperation and Development report on IT, July 1997). IT, so far, is based on computing and communication methods that are static and brittle in interacting with the real world. There is a great interest in IT to address these problems, but the pertinent technology is presently not available and only actively researched in a small number of groups worldwide. Neuroinformatics can contribute to this development due to its focus on the most robust and efficient real-world device known—the brain. This section will list a number of representative examples of fruitful present and future interactions between Neuroinformatics and IT.

Database and communication technology. Building federations of heterogeneous, distributed databases is not only a matter of immediate interest to the neuroscience communities, but is also a recognizable subfield of informatics. At issue is more than just storing and retrieving data, but also quality control and consistency of the data, and deduction from the data, learning of new data, and reflection upon the contents of the database system itself.

Relatedly, in the last decade, an area has emerged called “data warehousing,” closely associated with “data mining,” worth $13 billion annually in the business-management information-systems communities alone. Many organizations obtain vast quantities of data, stored in a non-systematic fashion in a number of different databases, from which valuable deductions could be made if only they could actually access that data. The idea of data warehousing is to merge and clean up the “legacy data,” to make sure it is reliable and consistent, and insert it into the carefully prepared data warehouses. Conclusions for business decision support can then be derived by exploring the data in the data warehouse and extracting “nuggets” of information from it – hence, the term data mining. This technique distinguishes itself from traditional approaches in two ways. First, it focuses on the creation of new relationships in data sets. Second, it deals with data sets which, in size, dwarf those analyzed with traditional methods—typically billions of data records.

Neuroinformatics constitutes a medium for a fruitful two-way interaction between neuroscience and information technology in this area. Setting up the heterogeneous, distributed databases in which to house the data on the brain is a challenging application area for current informatics. It is an application of such complexity (in terms of the huge amounts of data, and its internal complexity, as well as the global distribution of the relevant databases) that informatics will immediately gain from the experience. It will also provide new challenging domains in which emerging IT technologies, such as augmented reality and immersive collaboration (the use of virtual technology to find new forms of interaction with large datasets), software agents (autonomous software entities which explore datasets and extract information), the grid (a U.S. initiative to pool the computing power of a large numbers of computers to achieve model-based integration of large, distributed, heterogeneous databases, presently applied to chemistry, biology, and cosmology), new communication protocols and specification languages can be applied, tested, and developed. In the second instance, neuroscience will benefit from this enhanced informatics support and the enhanced access and interaction it will provide with the knowledge base of neuroscience.

In a next step, the applied IT technology can be improved by a further understanding of the brain. For instance, advanced search strategies rely on so-called software agents. These agents are supposed to autonomously explore, interact, and learn from their information “environment” in a way similar to how an animal would interact and learn from its physical environment. Presently, the abilities of software agents are relatively restricted, partly due to our limited understanding of learning in biological systems. For instance, software agents are strongly dependent on the user or programmer to define their tasks and search strategies, reducing their autonomy. Biological systems, however, show a great flexibility in defining and solving the tasks they face. A better understanding of these abilities will automatically translate into optimized search abilities of software agents. The above argument also holds for the emerging trend of so called “component software,” which can be mixed and matched across different hardware platforms, raising important questions regarding interface and communication standards.

Another area relates to the way in which IT handles communication, from the level of the physical bus, connecting the components of a computer, to software-defined protocols. Current IT depends on sequential, reliable, broad-band, large-package communication. These forms of communication quickly lose efficiency in the case of noisy transmission or the use of heterogeneous components. The brain has adopted a communication strategy that seems to be of a very different nature: parallel, noisy, low bandwidth, small packages. This implies that the way information is stored and retrieved in the nervous system will obey different rules. Understanding these rules will strongly facilitate the development of novel information and communication technologies.

The above set of issues and approaches will lead to a continuous interaction between neuroscience and IT through Neuroinformatics. Neuroscience will benefit through increasing our understanding of how the brain stores and retrieves information, maintains

its quality and consistency, makes deductions, learns, and reflects upon its own behavior. From this understanding, Neuroinformaticians will abstract principles for building better heterogeneous, distributed databases. Informaticians, in turn, will set up better systems, based upon these principles, for the neuroscientists.

It is important to emphasize that the two-way interaction of the Neuroinformatics loop can only be fueled if the Neuroinformatics program is balanced, focusing sufficient attention to the fundamental questions of neuroscience, as opposed to a single-sided direction towards applied research. Without fundamental understanding, there is not much to apply beyond *ad hoc* solutions.

Hardware technology. Present silicon technology allows the placement of about one billion transistors on a single chip. This density will further increase in the near future. The exploitation of this technology, however, is hampered by the methods to design and understand the circuits it can implement. An additional obstacle is the need to find circuits that have only limited power needs. This is especially pressing in the application of this technology in portable devices, which is a strongly developing market. In the field of neuromorphic engineering, a subdiscipline of Neuroinformatics, researchers attempt to extract principles of "natural computing" from our understanding of the brain in order to face these challenges. This has already created new technology, such as artificial retinas and cochleae, which have become successful products. For instance, the Logitech trackball uses a neuromorphic imager that has the advantage of being independent of any mechanical device to detect motion. Since its introduction three years ago, about 4,000,000 units have been sold. Other applications have been in the control of measurement equipment and power cells in space probes. The application of this technology to machine vision is now pending. This form of Neuroinformatics research will bring about a revolution in the design of advanced high-density processes needed to face the IT challenges of the future.

Robotics and intelligent machines. In 1996, there were 680,000 robots active in the world. Of these, 80,500 were installed during that year, representing an increase of 11 percent over 1995, and a market of $5.3 billion. This market will show further progression of the order of 13 percent per annum from now until the year 2000. Robots are mainly to be found in the automobile industry. For instance, in Japan, there are 830 robots for every 10,000 persons employed in the sector. To the extent that a robot is generally capable of replacing at least two persons on a production line, one can consider that they represent more than 4 percent of the workforce in the automobile industry in Western Europe. (Data from "World Industrial Robots: Statistics, Analysis and Forecasts to 2000," 1997, United Nations). In the U.S., robotic technology affects about 15 percent of GDP, and is expected to grow to 40 percent. In the U.S., robot sales have tripled since 1991 (data from white paper of the Robotics and Intelligent Machines Cooperative Council-RIMCC, 1997). Next to the automotive industry, another rapidly developing field of robot applications is in the management of hazardous environments such as dismantling nuclear, chemical, and conventional weapons (i.e., landmine clearing).

Current robot technology is far removed from the forms of behavior and adaptation displayed by biological systems. For instance, industrial robots, e.g., welding robots, are most often static devices that perform highly stereotypic actions in a strongly predictable and controlled environment using a minimum number of sensors. Current trends in this field move towards applying robots to more real-world tasks. Developing domains are cleaning, delivery, and maintenance robots. Current trends are towards adaptive, rapidly reconfigurable systems, human-like dexterity for assembly, and autonomous navigation. Applications in the area of entertainment are also under development. The required technology emphasizes machines which can sense and reason on physical objects, respond rapidly and autonomously to change, and enhance safety. It is predicted (RIMCC) that the realization of these goals will fuel a revolution as profound as the computer revolution. These developments pose two opportunities for current Neuroinformatics research. On one hand the type of sensing and control sought closely resemble those observed in biological systems. The field of robotics and machine learning, however, does not have the methods and technology to realize this level of competence. Hence, Neuroinformatics research can have a strong impact on this field. Again, by creating the kind of two-way interaction delineated earlier, hypotheses about brain function can be tested using robot technology, and robot technology can profit from the gained insights. This will in turn lead to new questions regarding brain function and behavior, opening up new vistas for Neuroinformatics research. A second demand stems from the need for devices that are not only autonomous in their behavior but also physically independent. This will create a strong demand for cheap, low-power, but intelligent solutions that are developed within Neuroinformatics (see previous section).

A particularly promising area of neuroscience research is the study of learning and memory. In the domain of robotics and intelligent machines, there is a strong demand for methods that allow these devices to adapt to the environment and learn from their experience. In ongoing Neuroinformatics research, advanced models of learning and memory are being developed that reflect pertinent aspects of the underlying substrate. These models are already successfully applied to robots. In addition, in Neuroinformatics research, robots are equipped with neuromorphic sensors. Next to their importance as synthetic approaches towards understanding brain function, they constitute an excellent starting point to establish a Neuroinformatics loop with this field.

In summary, Neuroinformatics is an interface between neuroscience and the domain of information and computing technology, and robotics and intelligent machines. It has the potential to create a positive feedback loop between these domains fully realizing the synergy between them.

VIII. IMPACT ON HEALTH BURDEN AND THE TREATMENT AND DIAGNOSIS OF BRAIN DISORDERS

Health Burden. All of humanity will benefit by the formation of the field of Neuroinformatics. The knowledge created through Neuroinformatics research will be used to help cure, prevent, and diagnose mental, neurological, and developmental brain disorders. The 1996 World Health Organization report "Global Burden of Disease" showed that in 1990, neuropsychiatric conditions ranked first in the list of diseases affecting life expectancy and quality of life. This report uses the so-called Disability Adjusted Life Year (DALY) measure, which is a composite measure of time lost due to premature mortality and time lived with disability. Neuropsychiatric conditions listed include: unipolar major depression (rank 1), alcohol use (rank 4), bipolar disorder (rank 6), and schizophrenia (rank 9). It also projects that brain disorders will be the greatest contributors to the Disability Adjusted Life Year (DALY) in 2020.

The primary epidemiological indicator of disability for non-fatal outcomes is Years Lived with Disabilities (YLD). In 1990, neuropsychiatric conditions directly related to the brain and central nervous system accounted for a majority of years lived with disabilities. Worldwide, of the top ten causes of YLDs, five are brain disorders. These five are: unipolar major depression (1), alcohol use (4), bipolar disorder (6), schizophrenia (9), and obsessive-compulsive disorders (10). Especially with the increase of life expectancy, partially due to the development of novel longevity drugs, the total incidence of neurodegenerative diseases and stroke will have a stronger impact on the budget for medical care. For instance, Parkinson's disease affects approximately one percent of the ever-growing world population.

We must meet the challenge of having a healthy population and reducing the human suffering and extreme medical-care costs of these illnesses. This can occur in the area of neuroscience through the opportunity created by Neuroinformatics, which facilitates the translation of our expanding knowledge of the brain into effective therapeutic interventions and diagnosis.

Impact on Treatment. Interventions for the prevention and treatment of brain disorders are developed through the use of scientific data available around the globe. This data serves as the major resource for the pharmaceutical industry in creating new therapeutic agents. The decision to develop a new agent is a function of both the market place (health burden) and the availability of potential target sites for a new medication. The accelerated rate of growth of knowledge in the area of neuroscience has already developed a significant knowledge base for the development of therapeutic agents affecting the central nervous system. As table 1 shows, drugs affecting the central nervous system generated over $8.5 billion in worldwide sales in 1996. This places central nervous system therapies in fourth place in sales worldwide.

Therapeutic category	$ Value in Millions
Cardiovascular Drugs	12,990
Infection Fighters	12,562
Gastrointestinal Products	10,409
Central Nervous System Drugs	8,541
Respiratory Therapies	5,526
Cholesterol Reducing Drugs	5,494
Cancer and Related Treatments	4,660
Women's Health Products	2,328
Erythropoiesis Enhancers	2,145
Arthritis Remedies	1,924

Table 1: Market value of the 10 leading therapeutic categories of the applications of pharmaceuticals (in millions of U.S. dollars). *1996 Data (Source: PharmaBusiness: The International Magazine of Pharmaceutical Business. July/August 1997, No. 16, pp. 31-32, Engel Publishing Partners, West Trenton, NJ.)

The richness of the current Neuroscience data is also reflected in the inventory of drugs in development. A total of 44 therapeutic categories are reported; of these, 31 categories registered fewer than 50 drugs in development. Only 6 categories (Table 2) claimed over 100 drugs in development; second among these, by a large margin, is the category of central nervous system disorders.

Therapeutic Category	Number of Drugs in Development
Cancer and Related Treatments	394
Central Nervous System Disorders	245
Analgesics & Anesthetics	50
Cardiovascular Products	184
Infection Fighters	127
Respiratory Products	125
AIDS and Related Treatments	102

Table 2: What's in the Pipeline: the July 1997 Special Issue of MED AD NEWS (the magazine of pharmaceutical business and marketing) reports Pipeline Products by Therapeutic Category (pp. 88-127): *Data reported are for calendar year 1996.)

If the information in Table 2 holds true, the therapeutic and economic gain from central nervous system drugs should increase dramatically. Future development is primarily a function of new discoveries of targets in the central nervous system and new knowledge of the functional systems in the brain. This capability promised by Neuroinformatics should greatly facilitate the development of new therapeutics interventions through the pharmaceutical industry.

In summary, the field of Neuroscience is poised to make major advances in understanding brain function and unraveling the mysteries of the workings of our brains in health and disease. It is clear that with time even more data will be obtained, not only with the continued efforts, but also with new tools and knowledge created through technological advances and capabilities. Many of the approaches currently being used were unknown as little as twenty years ago. Neuroinformatics will spur new initiatives in viewing and integrating data, and, in novel dimensions, enhance the effective and efficient communication and sharing of data, minimize duplication of effort, allow for more theoretical considerations of integrative brain function, and create new "principles" describing brain function. This information will be rapidly used and assimilated into the pharmaceutical industry to produce more effective therapeutic interventions. These interventions will be needed to battle the growing impact of brain disorders both in terms of minimizing the human suffering as well as reducing the cost to society, both direct and indirect. Neuroinformatics will ensure that, in the 21st century, the human energy and capital expended to date will realize its full potential, serving the overall scientific enterprise and society.

IX. IMPLICATIONS OF NEUROINFORMATICS

Neuroinformatics will become a strategic domain accelerating the growth of our understanding of the brain and the translation of this increased understanding to the medical arena and the areas of information and communication technology. In order to realize this potential, a number of challenges need to be faced. The first challenge is how Neuroinformatics will facilitate the integration of the many different levels of observation employed in studying the brain, and the resulting fragmentation of neuroscience into many practically unrelated subdisciplines. Answering this challenge will foster a change in the practice of neuroscience research, from adding more and more data to the brain data puzzle to extracting and evaluating underlying principles. A second challenge results from the strong multi-disciplinary nature of Neuroinformatics. In order for Neuroinformatics to mature and establish a positive feedback Neuroinformatics loop, it must foster close ties between its many constituent disciplines. In order to achieve this goal, Neuroinformatics needs to develop a common language and find a common ground in the highly variable set of available methods and approaches. These issues will be amplified by the lack of an appropriate organizational infrastructure at the level of academic institutions and funding mechanisms. The third challenge Neuroinformatics faces is due to the strong international nature of its research activities. The knowledge base of neuroscience is distributed over the planet and stored in many different formats (from lab notes to databases) and media (from paper to CD ROMs). Despite the present efforts to streamline the storage and communication of the elements of this knowledge base, the enormous scale of this endeavor can hardly be overestimated. A global Neuroinformatics effort will, however, guarantee that the neuroscience knowledge base will be more efficiently managed, accessed, and expanded. This will increase the return of the investments made in neuroscience.

This report attempts to convey the excitement and potential of Neuroinformatics. The realization of this potential constitutes a megascience challenge. This is due to the complexity of the domain, its highly multi-disciplinary nature, and the international effort required. No single country will be able to fully address or develop such an effort in isolation.

The recommendations emphasize the need for the establishment of a global Neuroinformatics capability involving many different research groups distributed over the planet. This implies that the solution sought has a different structure than those traditionally pursued through an OECD framework, such as the dedicated physical facilities constructed and operated with much success for the physical sciences. Hence, a megascience effort in Neuroinformatics constitutes an alternative perspective on large-scale, multi-national scientific projects and facilities, using modern information, computing, and communication technology, which seems especially appropriate for the life sciences.

The realization of a global Neuroinformatics capability requires inter-governmental coordination, since it involves not only scientific development and cooperation, but also has economical and societal implications. Hence, the OECD constitutes the appropriate body to mediate such an effort. In addition, the OECD would provide the means to nucleate such a field and facilitate the transfer of its results to non-OECD countries, enhancing international cooperation between the developed and developing countries.

X. RECOMMENDATIONS

Neuroinformatics combines neuroscience and informatics research to develop and apply advanced tools and approaches essential for a major advancement in understanding the structure and function of the brain. Neuroinformatics research is uniquely placed at the intersections of medical and behavioral sciences, biology, physical and mathematical sciences, computer science and engineering. The synergy from combining these approaches will accelerate scientific and technological progress, resulting in major medical, social, and economic benefits.

The main recommendations of the Neuroinformatics subgroup are:

I. **Establish a Global Neuroinformatics Capability**

This capability needs to be developed as a network of Neuroinformatics facilities and approaches, distributed across many research centers around the world. This network of Neuroinformatics facilities will be diverse, with major foci representing the development and application of:

- Databases, increasingly capable of handling the full complexity and organization of the nervous system, from molecular to behavioral levels;

- Powerful new tools for data-acquisition, analysis, visualization and distribution; and

- Theoretical, computational and simulation approaches, methods, and environments for modeling and understanding the brain.

II. **Establish a Global Coordinating Capability**

International coordination, as well as national efforts, are needed to assure success, proper implementation, and a sustained capability. An international scientific coordinating body, the International Neuroinformatics Committee (INC), and an associated secretariat should be established through the support of the participating countries.

The implementation steps needed to achieve this Global Neuroinformatics Capability are:

- Establish the coordination, standardization and interoperability requirements needed for successful application, integration, stabilization and quality assessment of the distributed and local Neuroinformatics facilities.

- Enhance collaborative opportunities in Neuroinformatics, both nationally and internationally. Such collaborations should be highly interdisciplinary and can range from as few as two laboratories to large-scale centers, either real or virtual.

- Develop national and international opportunities for recruitment, education and training in Neuroinformatics and for stable career pathways in this emerging interdisciplinary area.

- Enhance technology transfer to industrial and clinical arenas (e.g., electronics, computers, robotics, health care, pharmaceuticals and educational areas).

To this end, during its mandate, the OECD Megascience Neuroinformatics Subgroup will continue to:

Phase 1: (i) Work on the objectives and a set of specific early tasks for the coordinating body; (ii) initiate national and international Neuroinformatics activities; and (iii) broaden the awareness in the community and prepare an operational plan for the next Phase.

Phase 2: After approval, the INC needs to be established, as well as an operational plan for the program.

Phase 3: Will include a feasibility assessment on the basis of pilot projects, while

Phase 4: is devoted to the main realization of the program.

APPENDIX

Members of the Neuroinformatics subgroup

Shunichi Amari, Japan.
Jan G. Bjaalie, Norway.
Francesco Beltrame, Italy.
Julio E. Celis, Denmark.
Anne-Marie Coriat, UK.
David Cornwell, EC.
Rodney Douglas, Switzerland.
Erik De Schutter, Belgium.
Peter Dukes, UK.
Alan Evans, Canada.
Alf Game, UK.
Sten Grillner, Sweden.
Michael F. Huerta, USA.
Yasuhiro Itakura, Japan.
Russell E. Jacobs, USA.
Henri Korn, EC.
Stephen H. Koslow, USA. (Chairman)
John Mazziotta, USA.
Line Mathiessen, EC.
Stefan Michalowski, OECD
Geoff Oldham, UK.
Keith van Rijsbergen, UK.
Edmund T. Rolls, UK.
Per Roland, EC.
Arthur W. Toga, USA.
Vincent Torre, EC.
David Van Essen, USA.
Jaap van Pelt, Netherlands.
Paul F.M.J. Verschure, Switzerland.
Andrzej Wrobel, Poland.