Chapter 1

Biodiversity Informatics: A Review
Chapter 1

Biodiversity Informatics: A Review

1.1 Biodiversity and Human Civilization

The most striking feature of Earth is the existence of life, and the most striking feature of life is its diversity, popularly known as – biological diversity or biodiversity (Schnase, 2000). Biodiversity is the biological capital of our planet and it forms the foundation upon which the human civilization is built (Daily, 1997). Biodiversity is fundamental for the Earth’s life support system, as it provides us with clean air, clean water, food, clothing, shelter, medicine, and aesthetic enjoyment. The history of human civilization, material culture and development of economic systems are all directly and indirectly associated with the use and management of biotic and abiotic resources, together called as “natural resources”. Biodiversity and the Earth that support it contribute trillions of dollars to national and global economies and indirectly through biologically mediated services such as plant pollination, seed dispersal, grazing land, carbon dioxide removal, nitrogen fixation, flood control, waste breakdown, and the biocontrol of crop pests (Maier et al., 2001).

However, these natural resources are often taken for granted which provide much essential natural service. Thus, biodiversity – the biological richness of ecosystems is perhaps the single most important factor influencing the stability of our environment, thereby continued existence of human civilization. Hence, it is in the interest of mankind that these resources are used in sustainable manner, cautiously, so as to ensure continued survival of human race on this planet. Therefore, this is one of our most important knowledge domains, vital to a wide range of scientific, educational, commercial and governmental activities. Efficient access to data and information about these natural resources (both biotic and abiotic) and natural processes is essential for their effective conservation and sustainable use (PCAST, 1998). Especially, biodiversity information is critical to a wide range of scientific, educational and governmental uses, and is essential to decision-making in many realms (Canhos et al., 2004a). This growing realization has resulted into emergence of new discipline, “Biodiversity Informatics” that applies information management tools to vast amount of biodiversity data and information.
1.2 **Biodiversity Informatics: the Term and Definitions**

The term "Biodiversity Informatics" was coined to circumscribe the application of information technology tools to biodiversity information, principally at the organismic level. It thus deals with information capture, storage, provision, retrieval, and analysis, focused on individual organisms, populations, taxa and their interaction (Berendsohn, 2001). It covers the information generated by the fields of systematics (including molecular systematics), evolutionary biology, population biology, behavioral sciences and synecological fields ranging from pollination biology to parasitism and phytosociology.

Even before the term “Biodiversity Informatics” was coined, applications of information technology in biodiversity were in practice in various quarters of biological sciences without use of definite term to bring these activities under single umbrella leading to status of a recognized discipline within biological sciences. According to Berendsohn (2001), the term “Biodiversity Informatics” was first used by John Whiting in 1992 while establishing the Canadian Biodiversity Informatics Consortium (CBIC). However, it was used in its broader sense during the discussions of the OCED Megascience Working Group on Biological Informatics in 1996, which later recommended the formation of GBIF, the Global Biodiversity Information Facility.

Soberon and Peterson (2004) defined biodiversity informatics as “the application of information technologies, to the management, algorithmic exploration, analysis and interpretation of primary data regarding the life, particularly at the species level of organization”. Thus, biodiversity informatics appears to be sandwiched between, as well strongly overlapping with environmental informatics and molecular bioinformatics. However, it will provide the skeleton for a generalized scientific information infrastructure in biology.

As stated by Canhos et al. (2004a), the existence of biodiversity data resources from different fields of knowledge available to all interested and the strong demand to integrate, synthesize and visualize this information for different purposes and by different end users has led to development of new field of research, the “Biodiversity Informatics”. Thus, it represents the conjunction of efficient use and management of biodiversity information with new tools for its analysis and understanding.
In fact, Biodiversity informatics distinguishes itself as being most focused on biological knowledge dating back to the earliest dates of recorded history, thus the scope of biodiversity informatics spans the age of the Earth (Sarkar, 2007).

### 1.3 Biodiversity Informatics: Historical Context

At the 1992 Rio de Janeiro “Earth Summit”, two agreements were signed that gave birth to United Nations Framework Convention on Climate Change (UNFCCC), and Convention on Biological Diversity (CBD). Successful implementation of both conventions is highly dependent on combined efforts of countries and international organizations, integration of distributed information systems and deployment of biodiversity informatics. CBD’s Article 17 has addressed the need of “information from all publicly available sources, relevant to conservation and sustainable use of biological diversity” including “results of technical, scientific and socio-economic research”. To achieve this and pursue, Article 18, paragraph 3, CBD is implementing the Clearing House Mechanism (CHM), and internet based network promoting technical and scientific cooperation and exchange of information.

In January 1996, OCED Science Ministers established Megascience Forum Working Group on Biological Informatics. This working group’s Biodiversity Informatics sub-group concluded that (a) the biodiversity information domain is vast and complex, but critically important to society, (b) at present, existing biodiversity and ecosystems information is neither readily accessible nor fully useful, and (c) recent technological and political developments present opportunities for OECD countries to show leadership in the area of biodiversity informatics. It thus, recommended the establishment of “Global Biodiversity Information Facility (GBIF)” as international mechanism to make biodiversity data and information accessible openly worldwide. GBIF was founded in March 2001, and participation is open to any interested country, economy, or recognized international organization that agrees to make scientific biodiversity information available. Since then GBIF has made significant progress and currently its data portal collates and disseminate over 134 million primary species occurrence records.

In May 2003 at a meeting in London on “2010 – The Global Biodiversity Challenge” it was recognized that challenge before Parties to CBD is how to quantify
and measure existing biodiversity and how to quantify its loss or conservation and thus reiterated the need “to make the biodiversity data that exists more readily accessible” and available in timely manner (CBD, 2004). Similarly, in January 2005, during the International Conference on “Biodiversity: Science and Governance”, the then French President Jacques Chirac provided political support for establishing “International Mechanism of Scientific Expertise on Biodiversity (IMoSEB)”, which too has stressed the need for accessibility to biodiversity data and information. This was ratified by over 2000 scientists representing over 100 nations at the same conference and later by over 600 scientists at the first DIVERSITAS Open Science Conference held at Oaxaca, Mexico in November 2005 (Loreau et al., 2006). These activities further emphasized the investment in mechanism such as GBIF.

In March 2007, G8+5 Environment Ministers while supporting the “Postdam Initiative – Biological Diversity 2010” encouraged the development of Global Species Information System (GSIS). This was followed by the Coordination meeting at Brussels in April 2007, where in European Union in collaboration with agencies from US, Australia, Brazil, India and South Africa pledged to contribute to GSIS efforts in the form of SpeciesBase. SpeciesBase would collate and disseminate information on species valuable to scientist, policy-makers, farmers, land-managers, conservationists, students, and even public. It further pledged that SpeciesBase would be complementary initiative to Encyclopedia of Life (EoL) in the United States and Atlas of Living Australia (European Union Press Release, May 9, 2007).

While the above announcement of European Union committing to GSIS was being made on May 9, 2007 at Brussels, at Washington DC, many of world’s leading scientific institutions announced the launch of the Encyclopedia of Life, a decadal effort to document all 1.8 million named species of animals, plants and other forms of life on Earth using the mash-up technology. Based on Edward O Wilson’s concept (Wilson, 2003), over US$50 million has been pledged for this effort which would be housed at http://www.eol.org, and will provide written information and when available, photographs, video, sound, location maps and other multimedia information on each species. Built on the scientific integrity of thousands of experts around the globe, the Encyclopedia will be a moderated wiki-style environment, freely available to all users everywhere (EoL Press Release, May 9, 2007).
This walk through the history till date, leads to the conclusion that both political and technological scenarios provide favorable platform for further progress in the field of biodiversity informatics.

1.4 Biodiversity Informatics: The State-of-the-Art

Early work in the areas of biodiversity informatics could be traced to mid 1970s, when Australian herbaria began digitizing their data cooperatively. Since then significant progress has been made in the area of biodiversity informatics. An attempt has been made here to take stalk of this journey of progress in the area of biodiversity informatics by categorizing it in four categories, viz. (1) Mobilizing Biodiversity Data, (2) Standards, Protocols, and Tools development, (3) Informatics Infrastructure development and (4) Capacity Building, Outreach and Open Access Initiatives.

1.4.1 Mobilizing Biodiversity Data

Beginning with Australian herbaria attempts since mid 1970s, there have been several initiatives undertaken in different regions of the globe to mobilize the biodiversity data, as this form the basic constituents of all biodiversity informatics activities. Most of these initiatives are focusing on species and specimen data as the first necessary information component. However, non-biotic environmental and ecological data are increasingly being used for ecological forecasting purposes.

1.4.1.1 Catalogue of Known Biota

How many life forms Earth’s habitats harbor? is a question often being debated. It is estimated that somewhere 5 to 50 million organisms must be inhabiting this planet (May, 1988), of which nearly 1.8 million organisms have been named and classified so far (Edwards et al., 2000; Wilson, 2003). However, there is no single repository either offline or online that could provide access to baseline data about these 1.8 million known organisms. Several national, regional, and taxon specific efforts have been initiated in last decade or so to collate baseline data about these known organisms. Chavan et al. (2004), enlist some of the significant ones, as well as discussed in Chapter 2, which deal with development of IndFauna, electronic catalogue of known organisms. However, it must be mentioned that GBIF aims at
indexing at least 98% of these 1.8 million known organisms by 2011 (GBIF, 2006). Towards this end, it signed memorandum of cooperation with Species2000 and ITIS Catalogue of Life Partnership to expedite the process of cataloguing all known species (GBIF 2004). On March 29, 2007, Species2000 ITIS Catalogue of Life Partnership achieved major milestone by launching seventh edition of annual checklist containing **1,008,965** species. The present catalogue with over one million species is compiled with sectors provided by 47 taxonomic databases from around the world. Many of these contain taxonomic data and opinions from extensive networks of specialists, so that the complete work contains contributions from more than 3,000 specialists from throughout the taxonomic profession (Bisby et al., 2007).

### 1.4.1.2 Specimen and Observation Data

Specimen and culture collections are the primary archives documenting biological diversity on Earth. It is estimated that 6500 natural history museums spread across the globe together house nearly 3 billion specimens. Associated with these specimens is data on identities, habitats, histories, and spatial distributions of 1.8 million known organisms. Chavan and Krishnan (2003) have reviewed various initiatives aiming towards digitizing these specimens, as they return investments made during 250 years of global biological inventories (Canhos et al., 2004a). Initiatives such as HISPID, ENHSIN, ENBI, BioCASE, Species Analyst, FishNet, MaNIS, HerpNET, ORNIS, REMIB (World Biodiversity Information Network) and Australian Virtual Herbarium (AVH) etc. have contributed immensely to liberate the data associated with the specimens housed in world major natural history museums and make it accessible in public domain.

However, only less than 10% of the worldwide specimens are available in the electronic domain (Krishtalka and Humphrey, 2000). To expedite the process of digitization as well as public domain accessibility, GBIF aims to bring online specimen records in the range of 500 million to 1 billion by 2011 (GBIF, 2006). It has been successful in bringing together over 134 million records digitized and shared by 978 collections through over 200 data providers. To complement this, several observation networks such as ENBI, LifeWatch, and UK’s NBN are providing platform survey and observation data. However, our progress in both ensuring accessibility to both specimen and survey/observation data is far from satisfactory.
However, both technological developments and growing realization about importance of digitization and sharing of such data are quite promising.

1.4.1.3 Environmental and Ecological Data

As stated by Canhos et al., (2004a), efforts to improve understanding of environmental patterns, their variability, their changes over time, and their implications for human welfare and decision-making, depend critically on the quality, accessibility, and usability of diverse environmental and related social science data. Thus, it is necessary to improve access to existing and emerging sources of environmental, biological and socio-economic data, and improve integration of these data in support of disciplinary and interdisciplinary research efforts and applications and related policy-making initiatives (Canhos et al., 2004b). Such data sets are essential for various analysis and modeling studies such as ecological niche modeling. Majority of the times, these datasets are available at global scales and thus could not be used for modeling at local scales as they lack required micro precision. Initiatives such as Earth Observation System (EoS), Long Term Ecological Network (LTER) and recently launched Global Earth Observation System of Systems (GEOSS) need boosting by enabling open access to ecological and environmental data.

This calls for development of rational rules, open source tools for data conversion, visualization and analysis. Institutions and consortia such as Canadian Facility for Ecoinformatics Research, US National Biological Information Infrastructure (US NBII), National Center for Ecological Analysis and Synthesis (NCEAS), Knowledge Network for Biodiversity (KNB), Geosciences Network (GEON), Science Environment for Ecological Knowledge (SEEK), Semantic Prototype in Research Ecoinformatics (SPIRE) are either developing such tools or fostering their development (Jones et al., 2006). However, scientific cooperation and partnership between researchers and institutions working with biological and non-biological environmental and ecological data is just beginning and needs to be encouraged.
1.4.2 Standards, Protocols and Tools development

Biodiversity data being generated by heterogeneous group of researchers for past 250 years of modern biology and remained distributed various institutions and individuals, in multiple forms and formats. One of the challenges in achieving seamless, easy and efficient integration of these datasets is development of tools, standards and infrastructure that can evolve interoperable framework. Towards this end, several initiatives are engaged in development of (1) standards and protocols, (2) collection management tools, (3) geo-referencing and mapping tools, (4) data cleaning tools, (5) modeling tools, as well as (6) web services and computational frameworks.

1.4.2.1 Standards and Protocols

Standards and protocols are essential for integrating data from distributed sources. Biodiversity Information Standards (TDWG), previously known as Taxonomic Database Working Group has developed several standards which include Access to Biological Collections Data (ABCD), Structured Descriptive Data (SDD), Taxonomic Concept Transfer Schema, amongst others. Of these, ABCD aims to define global formats for data exchange and exchange from diverse biological collections. Similar to ABCD, another federated schema called Darwin Core is currently being used by several museums to exchange/share data. GBIF has used DiGIR, an XML based protocol capable of working with configurable federated schemas. Developed by University of Kansas Natural History Museum and Biodiversity Research Center, motivation for DiGIR development was to replace Z39.50 protocol, and also unify diverse networks in a single technology.

Biodiversity Information Standards (TDWG) is currently working on developing additional standards such as Natural Collections Descriptions (NSD), TDWG GeoInteroprobability testbed pilot, Globally Unique Identifiers (GUID), Invasive Species Information System (ISIS), Imaging Standards, Observation and Specimen Records (OSR), and TDWG Access Protocol for Information Retrieval (TAPIR). TAPIR combines and extends features of the BioCASE and DiGIR protocols to create a new and more generic means of communication between client applications and data providers using the Internet.
However, considering the scope and expanse of biodiversity information, its spread, heterogeneity, standards needs to be developed on extracting and integrating multilingual data and information. Further, protocols and standards are required for integration of biodiversity data with the non-biodiversity data.

1.4.2.2 Collection management tools

Chavan and Krishnan (2003), and Berendsohn, et.al. (2003) have listed several natural history collections management software packages. However, I have observed that many of them lack controlled vocabularies; as well fall short in collating region or ecosystem specific information. My experience of developing IndCollections (Chavan et al, 2005a) makes me to believe that web based informatics infrastructures would prove as great asset in expediting speed of collections digitization, as well as ensuring quality control of collated data. Further, we need tools for imaging of specimens to highest possible resolutions so that it could be used in identification exercises, and also act as electronic field guides (EFG).

1.4.2.3 Geo-referencing and mapping tools

Geo-referencing of is crucial for meaningful representation, visualization, and analysis of biodiversity data. In Chapter 4, the need for geo-referencing is argued with compelling reasons. Several methods of converting textual descriptions to spatial coordinates (Williams, 1996, Murphey et al., 2004, Wieczorek et al., 2004, Beaman et al., 2004, Guralnick and Neufield, 2005, Chapman, 2005c) and tools are available. Data can be georeferenced via simple techniques using convenient, automated online gazetteers. Guide to georeferencing as a result of MANIS project (http://elib.cs.berkeley.edu/manis/GeorefGuide.html), establishes a standard methodology to assign geospatial coordinates to historical locality descriptions. Latest result of growing collaboration among biodiversity informaticians is BioGeomancer, a geo-referencing tool specially designed for text-to-coordinate conversion of locality data. It currently encompasses natural language processing (geo-parsing) to interpret descriptive localities, place-name lookup to register localities with known geographic coordinates, and ambiguity analysis of self-document uncertainties in resulting geographic descriptions. However, tools are needed to convert textual locality descriptions with coarser resolution to polygons.
1.4.2.4 Data cleaning tools

Errors in data are common and are to be expected. However, good understanding of errors and error propagation can lead to active quality control and management improvement. Thus, there are several methods for cleaning two primary biodiversity data types’ viz., nomenclature, and occurrence data. Chapman, (2005b, and 2005c) has extensively dealt with principles of data quality and data cleaning methods for both data types. Emerging web based tools for validating geo-references, taxonomic identifications, and collection dates are leading to development of complex automated data validation tools. CRIA SpeciesLink data cleaning tool, BioGeoMancer Workbench, GBIF Data Cleaning Demo Interface, and DIVA-GIS are some of the commonly used tools. However, there is need for tools capable of detecting geographic and ecological outliers, incorrectly geo-referenced localities, misidentified specimens, and nomenclatural discrepancies (Chavan et al., 2005b).

1.4.2.5 Modeling Tools

Existing biodiversity data do not provide enough coverage for direct, detailed environmental decisions. Thus, modeling is required for identifying and filling data gaps, planning future research, assessing conservation priorities, and providing information for environmental decisions. Many modeling tools and techniques are used for Ecological Niche Modeling (ENM) such as BIOCLIM (Nix, 1986), GLM (Austin et al, 1994), GAM (Yee and Mitchell, 1991), CART (Breiman et al., 1984), GARP (Stockwell and Peters, 1999), and ANN (Olden and Jackson, 2002; Peterson et al, 2002), among others. It is hard to state as to which one of them would be best, as underlying algorithms differ in each one of them. BIOMOD (Thuiller, 2003) and OpenModeller (Santana et al., 2006) are based on generic frameworks approach to support development and testing of modeling algorithms. While BIOMOD includes four techniques such as GLM, GAM, CART, and ANN to predict spatial distributions; OpenModeller includes several ENM algorithms (such as BIOCLIM, Climate Space Model, GARP and Euclidean distance techniques) along with SOAP and command line interface, and desktop interface. It is expected that in the near future, generic libraries like OpenModeller will be able to perform tasks in a distributed fashion, including running analysis separately in remote cluster processes via web services or GRID paradigms.
1.4.2.6 Web services and Computational tools

As Internet grows, emerging technological concepts such as web services and grid computing are offering unexplored and unlimited potential to biodiversity informaticians across the globe. Web services have demonstrated its capabilities facilitating single portal access to over 134 million records through GBIF. However, this is just a curtain raiser as web services enable application-to-application interactions, and thus offers enormous scope for biodiversity software production in a cooperative manner (Canhos et al., 2004a). uBio (http://www.ubio.org) is just a prelude to enormous power that could be harnessed through implementation of web services architecture, as it leads to scalable infrastructure that could be made available to both large and small institutions for varied purposes- from developing hassle free maps to answering questions about names, spellings, synonymy, to multiple nomenclatural concepts.

GRID technology is facilitating development of cyber infrastructure for to address complex questions by leveraging unused computational power of processors when they are idle and unused. SEEK (Science Environment for Ecological Knowledge) and BiodiversityWorld are early examples of GRID based problem solving environment for studying biodiversity. However, considering the potentials of both web services and GRID technologies, biodiversity informatics exercises are in its nascent stage, and requires enhancement.

1.4.3 Informatics Infrastructure development

Ever since mid 1970 initiative of Australian herbaria to digitize their data, efforts are on to build informatics infrastructure with exponential technological capacities, computational power, storage capacity, analytical ability. While Environmental Resources Information Network (ERIN) provided geographically related environmental information in 1990’s; HISPID (Herbarium Information Standards and Protocols for Interchange of Data) evolved standard format for interchange of electronic herbarium specimens at the same time in Australia. This encouraged CONABIO and INBio to fully engage themselves in biodiversity informatics activities in Brazil and Costa Rica respectively.
Post 1990, several global, regional, national, and thematic initiatives such as BIN21, US NBII, OBIS, GISIN, CBD CHM, ERMS, ENHSIN, BioCASE, TDWG, ENBI, LifeWatch, ETI, Consortium of European Taxonomic Facilities (CETAF), Fauna Europea, Euro+Med PlantBase, ILDIS, NABIN, IABIN, Regional LOOPS of BioNET International, Species2000, Integrated Taxonomic Information System (ITIS), InfoNatura, CRIA, Canadian Biodiversity Information Facility (CBIF), DiscoverLife, NatureServe, Israels BioGIS, Chinese Biodiversity Infomration Facility, TaiBIF, Australian Biodiversity Information Facility, numerous global thematic and taxonomic database have directly or indirectly have contributed evolving informatics infrastructure. However, this development is not unique across the globe, and thus needs to be rationalizing, as well regionalize into unexplored regions of the world.

1.4.4 Capacity Building, Outreach and Open Access Initiative

Open access to biodiversity data will promote scientific progress, facilitate training of researchers, and maximize value derived from public investments in data collection and archival efforts. However, legal, cultural, and technical restrictions exist, and must be overcome. Gaikwad and Chavan (2006) discussed in detail merits of open access to biodiversity data and tools to analyze such data. According to them open access to biodiversity data would lead to creation of data enriched virtual biodiversity research space. Agencies such as CBD, CODATA, and GBIF are emphasizing the need for liberating biodiversity data, as without access to primary biodiversity data, scientific studies carried out on global, regional and possibly national scales such as impact of climate change on indicator species would not be possible.

However, open access movement can not be infiltrated to bench level researchers and small and medium sized institutions and initiatives unless biodiversity informatics and outreach effort reaches to every nook of globe. The global impact of deployment of the expanded data infrastructure and emerging tools could be felt only when capacity building and outreach activities are innovative and effective in mobilizing and encouraging more individuals and institutions to undertake biodiversity informatics as profession.
1.5 Biodiversity Informatics: An Analysis

Foregoing discussion about status of biodiversity informatics may confuse and misled to interpret that there are sufficient tools, techniques, and information bases are developed. It is believed that there are over 1500 resources developed so far in different regions of the globe. These include information bases, software tools, standards, and protocols etc. However, it is also believed that the pattern of development and use of these resources are isolate, and uneven, creating chaotic situation. This not only results into duplication of efforts and investments, but it also hinders the coherent advancement, and reduces the pace of progress of biodiversity and ecosystem informatics disciplines. One of the major reasons for this growing catastrophe is unavailability of metadata about these resources at a single click of a mouse. There is growing and urgent need for web based repository of biodiversity and ecosystem information resources.

To address this issue, my group has developed the BIR (Biodiversity Information Resources) database to collate metadata about distributed and isolated biodiversity and ecosystem information resources. Annexure I, discuss the rational, and method of developing BIR. Accessible at http://www.ncbi.org.in/BIR/ provides metadata information about 1383 biodiversity informatics resources. In this section, attempt has been to analyse these resources with respect to their taxon and geographic scope, and resource types, to test the hypothesis that development and use of existing biodiversity informatics resources is dispersed, isolated, and even.

These resources could be broadly classified into four categories, viz., information system and networks, database/databank(s), software(s), and standard/protocol(s). Analysis of 1383 resources reveals that over 50% of the resources document data about one or more taxons of animal kingdom (Table 1.1), where in less than 30% of the resources focuses on documenting data about plant kingdom. As depicted in Figure 1.1, 559 resources have global coverage, as against 301 with national, 285 with local and 239 regional coverage. Some of the resources fall into more than one category of geographic scope. Majority of the resources with global coverage are those digitizing biological specimens housed in world’s major natural history collections, holding type specimens. Analysis further reveals that over 90% of the 1383 resources are either databases or databanks (Figure 1.2). Out of the remaining less than 10% resources, 4.84% are information systems and networks,
3.76% software tools, and 0.65% is standards and protocols. Closer look at the databases and databank reveals that over 58% (58.31%) are taxonomic in nature, mostly species catalogues (global and regional checklists), and little over 23% (23.70%) are specimen databases (Figure 1.3). Only 6.51% databases contain images or multimedia artwork. Amongst rest of the databases 2.90% are bibliographic/referral, 2.40% observation/survey, 2% geospatial, 1.76% genomes of specific organisms or groups of organisms, and 1.48% educational databases. Majority of the resources are developed and intended for English speaking users.

Foregoing analysis reveals that progress in biodiversity and ecosystem informatics is uneven and imbalanced similar to biodiversity and distribution as well accessibility of biodiversity data across the globe. Most of the existing resources have been conceived and developed by information centers, laboratories, and research groups located in developed part of our globe, especially Europe, North America, and Australia. This means there is little or at time no involvement of research groups located in developing and under-developed mega-biodiversity regions of the world. Growing involvement of these local research/survey groups from mega-biodiversity world, would not only ensure their participation, but also help in developing tools, and information products that are usable and reusable for effective conservation of local biodiversity. Further, confirming Edwards et. al., 2000 these resources will help in sharing scientific biodiversity data for society, science and a sustainable future. There is a need to encouragement to regional, national and local level biodiversity and ecosystem documentation initiatives. Think Globally, but Act Locally, should be the approach of the biodiversity and ecosystem informatics community, so that micro-level biotic diversity gets documented to its minutest details. Thus, biodiversity documentation should be encouraged at village, block, district/county, state/province, country, and regional level. Initiatives such as PBR, Peoples Biodiversity Register (Gadgil et al, 2000, Gadgil et al, 2006) should be encouraged which have inbuilt process to involve local population in documentation of biodiversity local area.

Majority of databases focused on collating broad species and specimen related data, wherein biodiversity informatics relate to aspects of biodiversity from genes to ecosystems and environment (Costello and Berghe, 2006, Costello et al, 2006), and thus, significant encouragement is needed for resources that can document ecosystem, environmental and genetic diversity data. Thus, while developing and funding
databases emphasize needs to be focused on neglected taxas, specimen digitization from mega-biodiversity developing and under developed nations.

More multimedia databases as well bibliographic and referral databases development should be encouraged. To provide true picture of world’s biodiversity, it is also essential to focus our investment on development of observational and survey databanks, as well geospatial databases. There is scope to develop increasing educational databases.

It was observed that there are very few software packages, tools, standards and protocols that are currently available. There is a need to invest in developing new tools and standards that would expedite the process of biodiversity and ecosystem data documentation, analysis, modeling, prediction, as well dissemination. We also need to encourage development of standards and protocols leading to interactivity, integration, and interoperability between and across the cross-discipline, multidiscipline biodiversity and ecosystem information resources.

There is emergent need to develop biodiversity and ecosystem resources, may it be databases/databanks, information systems/networks, or software/tools that could be used by non-English speaking population of the world. If the goal of our activities in biodiversity and ecosystem informatics is to conserve the biotic resources, then our information products should be in the languages that people understand the best, so that they are sensitized enough to take proactive measures towards their neighborhood and biotic resources that it harbors.

1.6 Biodiversity Informatics in mega-biodiversity World: Why?

As revealed in previous section, mega-biodiversity nations, which harbor rich and diverse biotic resources and needs invest in biodiversity informatics activities for sustainable bioresources they harbor. India, being one of the mega-biodiversity and developing nation, it is imperative that biodiversity informatics activities form cornerstone of our economical, environmental, and social well-being. This immediate need is further justified with three arguments.

1.6.1 Exploding population – A National challenge

It is being predicted by several economic analysts that during 21st century India would lead itself from developing to a developed nation. This is happening
against the backdrop of its exponential population growth. The pressures from increasing population are such that human survival takes precedence to concerns over loss of biodiversity. Faced with such dire prospects, it becomes important to understand the links between biodiversity and benefits to mankind. Unfortunately, all these aspects of biodiversity are very much in the realms of the unknown at present. Thus, while meeting needs and aspirations of exploding human population, and protecting biodiversity and natural ecosystems on other hand, has emerged as a major national challenge. As argued by Chavan and Krishnan (2004), meeting this challenge without efficient and timely access to accurate, sufficient, and authentic information about the status of biotic and abiotic resources of Indian Ocean, and consequences of human centered development on these resources, would be impossible.

1.6.2 Natural Resources based Economy

Economics and biodiversity are closely linked. While, natural resources have immense economic value, economic forces themselves are major reason for biodiversity loss. Once destroyed it is impossible to regenerate and replicate these natural resources, and ecosystems that harbor them. Thus, it is essential that we evolve the mechanism to better manage and use these natural resources. As stated by Emerton, (2000), it is in this very sense that economics is crucial to biodiversity conservation because unless it makes demonstrable economic and financial sense for conservation, it is unlikely that all concerned would take actions to do so. Thus, we need to inculcate the informatics-supported natural resources accounting into our national economy.

1.6.3 Emerging Knowledge Catastrophe

Even though limited geographic coverage of the country has been studied and surveyed so far to study its biodiversity, it has resulted into enormous amount of data. Available datasets are scattered with agencies and individuals within India and outside India. Much of the currently available public domain Indian biodiversity data is from the agencies overseas, especially those from developed world and over 99% of it in English. Much of the non-English vernacular language data and information available or produced by agencies within India is behind strong cultural barriers of exchange and sharing. Thus, most of the time, same data and information is churned and reproduced again and again. This in my opinion would widen the gap of knowledge about true state of Indian biodiversity. It is my fear that thus conservation priorities based on this biased data view, would be the perception and understanding
of overseas experts, and institutions. This to me is emerging knowledge catastrophe (Figure 1.4), which can only be prevented if biodiversity informatics is considered as cornerstone of natural resources conservation, their sustainable use, and social well being in India.

1.7 Biodiversity Informatics in India: Status

Biodiversity and ecosystems information about Indian biodiversity within India and overseas is enormous, but isolated, dispersed, and in heterogeneous formats and without any metadata. During past few years several agencies and institutions have begun working on various aspects of biodiversity and ecosystems informatics (Table 1.2). In addition to this several information centers under “Biotechnology Information System (BTISNet)” (DBT, 2007), “Environmental Information System (ENVIS)” (MoEF, 2007), and “Agricultural Research Information Network (ARISNET)” (Sreenivasulu and Nandwana, 2001) are also working on one or more issues related to biodiversity and ecosystem informatics.

National Biodiversity Strategy and Action Plan (NBSAP) has proposed a comprehensive “Indian Biodiversity Information System (IBIS)”, which would be expansive version of the current Environmental Information System (ENVIS), and would focus on species diversity centered informatics (Javed, 2001, Kothari, 2003). Recently, National Biodiversity Authority (NBA) constituted a working group to explore feasibility of establishing “Indian Biodiversity Information System”. In the meanwhile, Department of Biotechnology has launched “Indian Biodiversity Information Network (IBIN)” (Government of India, 2007) to disseminate data collated as part of its National Bioresources Development Board (NBDB) initiatives. As mandated by the Biological Diversity Act, 2002 (Gadgil, 2003), National Biodiversity Authority (NBA) is also supporting development of Peoples Biodiversity Registers (PBR) to document people’s knowledge of biodiversity. Recently, National Knowledge Commission (NKC) has accepted in principle to support multi-institutional proposal to develop “India Biodiversity and Bioresources Portal” aiming at research, education, bio-prospecting, and conservation of national biodiversity (Bawa, 2007, pers. Comm.).
Chapter 1: Biodiversity Informatics: A Review

However, considering the vast amount of information that is available, and exponential rate at which new data is being generated, these efforts are inadequate. What we lack is a framework or information infrastructure within which these enormous volumes of scattered data sources could be linked together facilitating exchange, and use of data, leading to sustainable use of our biotic resources.

1.8 Recommendations

- In order to overcome uneven spread of biodiversity and biodiversity information engagement of mega-biodiversity developing and under-developed nations is must.
- Mega-biodiversity developing and under-developed nations must treat biodiversity informatics as corner stone of their environmental, economic, and social well-being.
- Biodiversity information resources need to be developed both at mega, and micro scale, and in vernacular languages.
- In India, biodiversity information infrastructure needs to be conceptualized, and developed.

1.9 Summary

Biodiversity informatics is truly a mega-science endeavor. Though institutions and individuals worldwide had undertaken several initiatives, most of them are concentrated in the developed regions. BIR analysis reveals that megadiversity nations need to invest in initiating or strengthening biodiversity informatics activities. Such an investment in biodiversity informatics is essential to cope up with exploding population led stress on natural resources, efficient natural resources accounting, and bridging biodiversity knowledge catastrophe. In India biodiversity informatics activities are in primitive stages and thus need encouragement.
Table 1.1: Taxon analysis (in %) of resources in BIR

<table>
<thead>
<tr>
<th>Kingdom(s)</th>
<th>% of resources available as per kingdom in BIR</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Biota</td>
<td>10.74</td>
</tr>
<tr>
<td>Animalia</td>
<td>51.34</td>
</tr>
<tr>
<td>Plantae</td>
<td>27.38</td>
</tr>
<tr>
<td>Fungi</td>
<td>8.39</td>
</tr>
<tr>
<td>Bacteria</td>
<td>1.14</td>
</tr>
<tr>
<td>Chromista</td>
<td>0.40</td>
</tr>
<tr>
<td>Viruses</td>
<td>0.34</td>
</tr>
<tr>
<td>Protozoa</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Figure 1.1: Geographical scope analysis of resources in BIR reveals that majority of resources are macro-scale in nature.
Figure 1.2: Different resource types (in %) BIR.

[IS/IN-Information System/Information Networks; ST- Software Tools; St/Pr/Sc- Standards/Protocols/Schemas]

Figure 1.3: The ten different types of databases/databanks differentiated in BIR. Taxonomic types of databases are more in numbers than others databases.

[PBRs- Peoples Biodiversity Registers]
Both biodiversity and biodiversity data are unevenly distributed around the world:

Developing World Biodiversity

Developed World Biodiversity Data

Digital Divide  Content Divide  Lingual Divide

Knowledge Divide

Emerging catastrophe

Figure 1.4: Uneven distribution of biodiversity and biodiversity data together with content and lingual divide is leading to “Biodiversity Knowledge Catastrophe”.

<table>
<thead>
<tr>
<th>Biodiversity Information</th>
<th>URL</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agricultural Databases and information on sacred groves</td>
<td><a href="http://www.mssrf.org/">http://www.mssrf.org/</a></td>
<td>MSSRF, 2007</td>
</tr>
<tr>
<td>Agricultural Research Information Network (ARISNET)</td>
<td>--</td>
<td>Sreenivasulu and Nandwana, 2001</td>
</tr>
<tr>
<td>Bibliographic and referral information on Western Ghats</td>
<td><a href="http://ces.iisc.ernet.in/hpg/envis/lkwestern.htm">http://ces.iisc.ernet.in/hpg/envis/lkwestern.htm</a></td>
<td>CES, 2007</td>
</tr>
<tr>
<td>Biodiversity characterization using RS/GIS</td>
<td>--</td>
<td>Roy and Ravan, 1996; Roy and Tomar, 2000; and Roy, et al., 2002</td>
</tr>
<tr>
<td>Biotechnology Information System (BTISNet)</td>
<td><a href="http://www.btisnet.nic.in/">http://www.btisnet.nic.in/</a></td>
<td>DBT, 2007</td>
</tr>
<tr>
<td>Birds of India</td>
<td><a href="http://www.wetlandsofindia.org/">http://www.wetlandsofindia.org/</a></td>
<td>SACON, 2007</td>
</tr>
<tr>
<td>CDROMs on Marine Prawns, Marine Crabs, Mangroves, Lignicolous Fungi and corals of India</td>
<td><a href="http://www.indian-ocean.org/">http://www.indian-ocean.org/</a></td>
<td>NIO, 2005</td>
</tr>
<tr>
<td>Endemic Trees of Western Ghats</td>
<td>--</td>
<td>Datta et al., 1997</td>
</tr>
<tr>
<td>Environmental Information System (ENVIS)</td>
<td><a href="http://www.envis.nic.in/">http://www.envis.nic.in/</a></td>
<td>MoEF, 2007</td>
</tr>
<tr>
<td>Flora of Karnataka</td>
<td>--</td>
<td>Ganeshaiah et al., 2002</td>
</tr>
<tr>
<td>Indian Biodiversity Information Network</td>
<td><a href="http://www.ibin.co.in/">http://www.ibin.co.in/</a></td>
<td>Government of India, 2007</td>
</tr>
<tr>
<td>National and State Forest Vegetation maps and National Basic Forest Inventory (NBFIS)</td>
<td><a href="http://envfor.nic.in/fsi/fsi.html">http://envfor.nic.in/fsi/fsi.html</a></td>
<td>FSI, 2005</td>
</tr>
<tr>
<td>NCL Center for Biodiversity Informatics</td>
<td><a href="http://www.ncbi.org.in/">http://www.ncbi.org.in/</a></td>
<td>NCL, 2005</td>
</tr>
<tr>
<td>Plants of India and Legume Database of South Asia</td>
<td><a href="http://www.nbri-lok.org/bioinformatics1.htm">http://www.nbri-lok.org/bioinformatics1.htm</a></td>
<td>NBRI, 2007</td>
</tr>
<tr>
<td>SAHYADRI</td>
<td><a href="http://wgbis.ces.iisc.ernet.in/biodiversity/">http://wgbis.ces.iisc.ernet.in/biodiversity/</a></td>
<td>Ganeshaiah, 2003</td>
</tr>
<tr>
<td>Sasya Sahyadri</td>
<td>--</td>
<td>Ganeshaiah, 2003</td>
</tr>
</tbody>
</table>

Table 1.2: Major Biodiversity Informatics activities in India