DESIGN AND IMPLEMENTATION OF A BIODIVERSITY INFORMATION MANAGEMENT SYSTEM: ELECTRONIC CATALOGUE OF KNOWN INDIAN FAUNA – A CASE STUDY

A THESIS SUBMITTED TO THE UNIVERSITY OF PUNE
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN BIOINFORMATICS

BY

VISHWAS CHAVAN

UNDER THE GUIDANCE OF

DR. S. KRISHNAN

NATIONAL CHEMICAL LABORATORY
PUNE

SEPTEMBER 2007

DESIGN AND IMPLEMENTATION OF A BIODIVERSITY INFORMATION MANAGEMENT SYSTEM: ELECTRONIC CATALOGUE OF KNOWN INDIAN FAUNA – A CASE STUDY

A THESIS SUBMITTED TO THE UNIVERSITY OF PUNE
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY IN BIOINFORMATICS

BY

VISHWAS CHAVAN

UNDER THE GUIDANCE OF

DR. S. KRISHNAN

NATIONAL CHEMICAL LABORATORY
PUNE

SEPTEMBER 2007

DECLARATION

I hereby declare that the work embodied in this thesis entitled "DESIGN AND IMPLEMENTATION OF A BIODIVERSITY INFORMATION MANAGEMENT SYSTEM: ELECTRONIC CATALOGUE OF KNOWN INDIAN FAUNA – A CASE STUDY" represents original work carried out by me under the supervision of Dr. S. Krishnan, Head, Information Division, National Chemical Laboratory, Pune. It has not been submitted previously for any other research degree of this or any other University.

September 25, 2007

Vishwas Chavan
National Chemical Laboratory
Pune 411008, INDIA

CERTIFICATE

Certified that the work incorporated in the thesis entitled "DESIGN AND IMPLEMENTATION OF A BIODIVERSITY INFORMATION MANAGEMENT SYSTEM: ELECTRONIC CATALOGUE OF KNOWN INDIAN FAUNA – A CASE STUDY" submitted by Mr. Vishwas Chavan, was carried out by the candidate under my supervision. Materials obtained from other sources have been duly acknowledged in the thesis.

September 25, 2007

Dr. S. Krishnan
Head, Information Division
National Chemical Laboratory
Pune 411008, INDIA

Dedicated to Mother Earth who nurtures life,

my parents who brought me in to this world,

my teachers who taught me to understand nature, and

my daughter who I believe would continue to care for life……

Acknowledgements

Biodiversity informatics is a passion for me. During my 15+ years of professional career in this field, I was on look out for friend, philosopher, and mentor who would guide me in my pursuit to contribute to this much essential yet neglected discipline. Dr. S. Krishnan in true sense mentored not only during the course of this work, but also guided me to achieve excellence that is today recognized at national and international level. I thank him for the excellent guidance, encouragement and support that he has extended to transform me from young amateur worker to well recognized mature researcher. I owe him a lot for the freedom of work and expression as well as for the most fruitful association I had with him as my guide and senior colleague.

I am thankful to Dr. S. Sivaram, Director, National Chemical Laboratory (NCL), Pune for encouraging and supporting activities in the field of biodiversity informatics. NCL is known for undertaking work in unconventional disciplines of science and technology beyond boundaries. Dr. Sivaram lived with that reputation of NCL, and supported biodiversity informatics amidst odds, and stiff resistance from individuals and institutions who thought biodiversity informatics is not our area of work.

What I am today is a result of immense contributions of many institutions and individuals. Some of these contributions were noticeable, while many were invisible, yet significant ones. At this juncture, I wish to express my deep sense of gratitude towards each one of them. During my 15+ years of professional career, which I began at the National Institute of Oceanography (NIO), Goa followed by Centre for Cellular and Molecular Biology (CCMB), Hyderabad, and currently, at the National Chemical Laboratory (NCL), Pune several colleagues at all these laboratories ensured that improve intellectually during every possible interaction, and research experience with them. At NIO, Goa it was Dr. D. Chandramohan, Dr. Baban Ingole, Dr. T. G. Jagtap, Dr. C. T. Achuthankutty, Dr. Usha and Dr. Subhash Goswami, Mr. Arvind Ghosh, Mr. Albert Gouveia and Dr. Ehrlich Desa were responsible to bring in much required confidence that I could plan, and implement R&D programs independently. At CCMB, it was my colleague Dr. Yogendra Sharma not only boosted my morale both intellectually, and psychologically, probably at a moment when, I was being almost rejected by the mainstream bioinformatics community for not practicing conventional

bioinformatics. It was Yogendra who encouraged me to write correspondence in "Nature", and also apply for "Fulbright Fellowship" which not only opened up new and unexplored word of biodiversity informatics, but at the same time made me to realize that my conviction to work in the field of "biodiversity informatics" was not only a right decision, but ahead of our time.

During my Fulbright experience at the US Geological Survey, Reston, VA, USA, Dr. Anne Frondorf opened a gateway for me by introducing me to several US colleagues, and each discussion with them led me to believe that I am not only progressing on right track, but also working ahead of others. Ms. Gladys Cotter, Associate Chief Biologists, and her colleagues at US NBII provided an experience of how to plan big, and implement with ease.

While I was cris-crossing continents as expert on various national, and international agencies; Mr. Nandkumar Dahibhate, my colleague at NCL, not only ensured that my group work with same vigor and dedication, but also shoulder lots of my personal planning burden. His support over past 7 years, and critical advice at crucial junctures, is something that I would continue to cherish all the time. He is not only a perfect example of a man with discipline, but also the rare individual with uncommon quality of service to community of friends and close ones without any expectation. I owe him a lot!

Success that is attributed to me are not of mine alone, but its contributions of several colleagues and students who worked with me. At NIO, Goa my colleagues Mr. Devanand Kavlekar and Mr. Vishwanath Kulkarni not only shared work space with me but to certain degree contributed to my ambition to work in marine biodiversity informatics. At NCL, Pune my colleagues Mr. Manegsh Deshapande, Mr. Siddharth Paralikar assisted me to the best of their ability. I would always cherish my intellectual, philosophical, and motivational conversations with my buddy Dr. M. Karthikeyan. Encouragement that I received from Mr. G. Prabhakaran, and Mr. D. B. Pradhan needs special mention. I enjoyed working with most of the members of group. Notable amongst them are Dr. Aparna Watve, Mr. Nilesh Rane, Mr. Jitendra Gaikwad, Dr. Swapna Prabhu, Ms. Manisha Londhe, Ms. Vaishali Jadhav, Ms. Asavari Navlakhe, Dr. Aniruddha Khadkikar, Ms. Abhilasha Sen, Ms. Rashmi Rajhansa, Mr. Chandan Badapanda, Ms. Prajakta Tembe, Ms. Varunjeet Kaur, Mr. Gurushant Upase, Mr. Vitthal Kudal, Mr. Shriram Dawkhar, Mr. Santosh Gaikwad,

Ms. Priya Gaikwad, Ms. Bhagyashree Kumbhalkar, Mr. Aditya Kakodkar. I thank each one of them, and those to who I might have forgotten to mention here.

I wish to place on record appreciation that I received from Prof. Madhav Gadgil for my efforts in biodiversity informatics. Several of my colleagues and friends viz., Dr. Prashant Naik, Dr. Urmila Kulkarni-Kale, Ms. Shubhada Nagarkar, Mr. Sanjay Londhe were always source of encouragement during this whole process.

During my 15+ years of professional career, I came across many young boys and girls, with diverse ambitions, work style and high degree of attitude. Many of them contributed to my success indirectly, and to this dissertation directly. However, Mr. Jitendra Gaikwad provided me immense opportunity to shape a career. He is not only teachable, flexible to changes, with high degree of ambition, but also inclusive in his nature. Apart from assigned duties at my lab, he became very close member of my family, and thus contributed significantly in sharing some of my personal assignments, so that I could have time enough to devote to this work. I wish him all the very bests in future.

My parents came to Pune from remote village in Vidharba region of Maharashtra in 1960's. They did not had an opportunity to learn much, but they imbibed great dream for their kids, to see them progressing. Had it not been my parents who are great dreamers, but also GOD gifted with uncommon strength to stand for right things against all odds, I would not have been shaped as a scientist. They worked to ensure that we progress. I would always cherish their love, care, and passion to see their kids progressing.

My father-in-law Late Principal Maruti Salwe, was teacher by heart, who succeeded amidst adverse situations. Had he been alive, he would have been extremely happy to share his joy on this occasion. In more than one sense he was my philosophical father with whom I got very little time to exchange/share views. He was always very proud of his son-in-law, and today would have been overjoyed with great sense of pride. I must also mention the support and words of encouragement that I received from my mother-in-law all along. I am thankful to numerous relatives who offered unconditional support to me during adverse situations.

My wife and best companion ever, Pradnya is a great dreamer, and has always encouraged me to move forward in life. I cannot explain in word painstaking efforts she has taken to ensure that I work beyond commoners in order to achieve my and her dreams. During last 11 years of our married life, she stood by me at every highs and

lows that came our way. She believed in my abilities to succeed. I would continue to cherish our companionship and friendship all the time.

Both Pradnya and I thank GOD to have blessed us with wonderful and loving daughter Maitri, whose love and caring nature, as well oozing confidence is source of our energy. I am confident that our daughter to carry forward our legacy from the point where we cease to perform, and would achieve unthinkable heights of success.

May GOD give me, and my family the strength and wisdom to continue working the field of biodiversity informatics, and achieve greater heights of success in all spheres of life, benefiting to the society, mankind and all living being!

Vishwas Chavan
September 25, 2007

CONTENTS

ABSTRACT

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. Biodiversity is the biological capital of our planet and it forms the foundation upon which the human civilization is built. It is fundamental for the Earth's life support system. The history of human civilization and material culture and development of economic systems are all indirectly associated with the use and management of biotic and abiotic resources, together called as "natural resources". These natural resources are often taken for granted which provide much essential natural service. 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. Efficient access to knowledge base on these natural resources and process is essential for their effective conservation and sustainable use.

The term "Biodiversity Informatics" was coined to circumscribe the application of information technology tools and technology to biodiversity information, principally at the organismic level. It thus deals with information capture, storage, provision, retrieval, and analysis, focused on individual organisms, populations, and taxa, and their interaction. It covers the information generated by the fields of systematics (including molecular systematics), evolutionary biology, population biology, behavioural sciences, and synecological fields ranging from pollination biology to parasitism and phytosociology. Biodiversity Informatics is considered a part of biological informatics sandwiched between - and strongly overlapping with - environmental informatics and molecular bioinformatics. It will provide the skeleton for a generalized scientific information infrastructure in biology.

However, there is disparity and uneven distribution of biodiversity and biodiversity information across the globe. Similarly, the progress of biodiversity informatics is currently concentrated outside mega-biodiversity regions of the world. Thus, for mega-biodiversity developing nation such as India, it is essential that we realize the biodiversity informatics as cornerstone of our economic, ecological and social well being. Therefore, the aim of this exercise is to design and implement biodiversity information management system for mega-biodiversity nation such as India, with cataloguing Indian fauna as a case study.

This dissertation deals with (i) Review of biodiversity informatics, and its significance for mega-biodiversity developing nation like India, (ii) IndFauna, electronic catalogue of known Indian fauna, (iii) IndFauna, Data Quality, Taxonomic Scrutiny, and Lessons Learnt, (iv) JaivaNaksha: Web mapping of occurrence data and its georefrencing, and (v) National Biodiversity Information Infrastructure: Defining the Roadmap.

Chapter 1: Biodiversity Informatics: A Review

Biodiversity informatics is emerging discipline, which deals with collection, collation, analysis, prediction, and dissemination of data and information related to biotic resources of the earth. This chapter reviews the global progress in the field of biodiversity informatics by grouping it in four categories, viz. (i) mobilizing biodiversity data, (ii) standards, protocols, and tools, (iii) informatics infrastructure building initiatives, and (iv) capacity building, outreach, and open access initiatives.

There are estimated to be 1500 biodiversity information resources. Since information about these resources is distributed, it is difficult to review the progress made in biodiversity informatics. Thus, BIR, Biodiversity Information Resources Database was developed. Analysis of BIR confirms that similar to uneven distribution of biodiversity, biodiversity information and informatics activities are unevenly distributed. While, biodiversity is concentrated within tropical mega-biodiversity regions which is both developing and under-developed; biodiversity information and informatics activities are concentrated in non mega biodiversity, developed world institutions. Majority of the biodiversity information resources are coarse in nature (global, regional, or national), and in English. There is a need for micro-focus databases in vernacular languages. Further, there needs to be an impetus on development of biodiversity information handling protocols, tools, and standards so to achieve interoperability amongst the resources. BIR analysis clearly emphasizes the need of biodiversity informatics activities in these regions, if our goal of sustainable use and conservation of biotic resources is to be achieved.

Biodiversity informatics activities are critically significant for the mega-biodiversity, developing economy such as India. Three strong points of argument for undertaking biodiversity informatics activities in India includes, (a) exploding population – a national challenge, (b) natural resources and economy, and (c) emerging biodiversity knowledge catastrophe. Form the review it appears that biodiversity informatics activities in India are in its nascent stage, and needs support.

Chapter 2: IndFauna, Electronic Catalogue of Known Indian Fauna

India is known to harbor 89,451 faunal species which is 7% of the world's known faunal diversity. While, this estimate is nearly 9 years old, numerous new species have been described in recent past. However, there is no single repository where baseline data and information regarding these species could be accessed. Currently, this data is distributed with several individuals and institutions, and majority of the times it is in non-interoperable forms and format. In order to address this, I have conceived and developed the IndFuana, electronic catalogue of known Indian fauna.

While briefly reviewing the faunal diversity in India, I have attempted to reason the development of IndFuana. IndFauna collates following baseline information viz., (a) valid scientific name with authority, year of publication according to accepted taxonomic opinion, (b) systematics of the species from kingdom to forma level, (c) synonyms with authority and publication year, (d) common, local, or vernacular names with languages, and regions, (e) occurrence along with details of source data, (f) bibliographic details, and (e) multimedia artwork, etc.

Orcale 9i has been implemented to develop the database structure which is described in detail in this chapter. One of the unique features of IndFauna, is its entirely transparent process for which web based set of data management, and data curation modules. Another feature of IndFauna is LinkOut with other datasets such as sequence databases. LinkOut has also been provided with GoogleImages. Similar LinkOut could be established with other databases such as Google Scholar, Google Books, Barcode of Life Database (BOLD), Species2000 ITIS Catalogue of Life, uBIO, PubMed, Scirus, and ZooBank, etc.

Another unique feature of IndFauna is its data cleaning process, which is based on approach of both prevention and correction. Over 100+ self volunteered "Taxon Experts" contributed towards enhancing data quality, taxonomic authenticity and validity of collated data. Process of data cleaning and taxonomic scrutiny has been described in detail.

IndFauna currently collates baseline data about 94500 known Indian faunal species. It not only facilitates easy access to Indian faunal diversity data, but also would provide sound base for resolving conflicts in taxonomies, planning future

research and analysis. IndFauna, thus have strong potential for act as “central registry of names of organisms”, which could form the backbone of national, and regional biodiversity information system collating and disseminating data on host of other parameters and factors that are responsible for dynamics, and health of our natural resources and their habitats.

Chapter 3: IndFauna: Data Cleaning, Taxonomic Scrutiny, and Lessons Learnt

Key purpose of electronic catalogue of known organisms such as IndFauna is to collate and disseminate species occurrence data. The uses of species and occurrence data are wide and varied and encompass virtually every aspect of human endeavor – food, shelter and recreation; art and history, society, science and politics. However, efficient and effective applicability and use of this data depends on quality of data. Thus, data cleaning enhances the “fitness for use” of data. Since, IndFauna has collated data from secondary sources; an approach of “prevention and correction” was adopted for enhancing the quality of data. These approaches and development of data curation modules has been described in this chapter. Taxonomic scrutiny which forms part of correction approach of data cleaning was carried out using offline and online data curation modules.

IndFauna has also been subjected to quantitative and qualitative analysis. While quantitative analysis was aimed as identifying gaps in collated and accessible data, qualitative analysis was attempted to identify taxonomic discrepancies, and help resolve them. During IndFauna development, it was learnt that one of the major reason for these discrepancies is disparity in availability of nomenclature change literature to the taxonomists of the developing world and availability of taxonomic papers published by developing world scientists to their counterparts in developed part of the globe. However, development of electronic catalogues of names of known organisms would help in pointing out these issues. I have attempted to highlight a few of such discrepancies found while developing IndFauna, an electronic catalogue of known Indian fauna and comparing it with existing global and regional databases. These discrepancies can be grouped into three categories, viz., (a) hierarchical differences, (b) spelling differences, and (c) homonymies.

Resolving these discrepancies is a matter of taxonomic discussion. They need to be resolved using nomenclatural rules. However, through the examples quoted in the chapter, I am attempting to demonstrate the role of electronic catalogues in bringing issues or discrepancies to the knowledge of taxonomic community, starting a

dialogue between taxonomists across the globe and identifying issues of common concern. In order to notice such discrepancies and resolve them quickly, it is essential that a wrapper be developed which traverses through various electronic catalogues searching for taxonomic anomalies. This calls for increasing collaboration among the various electronic catalogues of names of known organisms, which is far from happening in developing world. Thus, it justifies the investment in development of national electronic catalogues as core of the biodiversity informatics activities in mega-biodiversity regions of the world.

Chapter 4: JaivaNaksha: Web mapping of Occurrence data and its Geo-referencing

One of the important goals of biodiversity databases is to provide location information of species in order to empower decision-making in context of planning, developmental projects, conservation and invasive species management. Most biodiversity data usually has been collected from older collections when accurate maps and global positioning systems were unavailable. Consequently, species in museum collection or older literature seldom are associated with geographic coordinates. Locality references in most cases are in form of textual descriptions. In this chapter, while detailing the development of JaivaNaksha, web mapping application for indFauna; I review the challenges of assigning geo-coordinates to such descriptive locations, and standards to be adopted while doing so, as majority of the locality records documented in current exercise falls in this category.

Thus, one of the important challenges while designing JaivaNaksha is to accommodate the various types of locality records. These ranged from general references like ‘Throughout India’ to ‘Maharashtra’ (state names) to ‘Sindhdurg district’ (district name) to point locations (villages, towns). Apart from such textual descriptions, there also exist references to both arbitrary and precisely defined regions, examples of which include ‘Southern India’ for the former and ‘Thar Desert’ or ‘Rajaji National Park’ for the latter. The third type of location data that was encountered was in the form of river/water body names including lakes, mangroves, lagoons and estuaries. Our philosophy has been to provide an accurate representation of the described localities without taking decisions on their appropriateness. Hence, although ‘Throughout India’ would be discounted today for most species, we have sought to plot the same on the map.

JaivaNaksha, a user friendly web-based geographic information system (GIS) has been developed considering all these complexities and based on all these issues, JaivaNaksha and its backend database schema was designed using a combination of open source and proprietary technologies. A primary challenge in its development was to dynamically create maps on the fly, for which PHP Mapscript and session variables were used. It further generate the report which provides detailed information for the species with respect to occurrence data, location type and the source in which that location has been described. The maps are expected to be further refined as more occurrence data is georefrenced. The process of georefrencing of IndFauna occurrence records is also described together with the data curation module developed for this purpose.

During development of JaivaNaksha, it was realized that there is resistance to share both spatial data as well spatial data products such as shape files. While the need for open access to such data and products has been emphasized several times in the past, there is no open access repository where such data and products could be contributed. Thus, Open Access Geospatial Data Repository (OAGDR) was developed to exchange / share shapefiles of commonly used geographic features.

Chapter 5: National Biodiversity Information Infrastructure: Challenges, Potentials and Roadmap

During development of IndFauna, and allied products described in earlier sections of the thesis, urgent need was felt to evolve planned mechanism to collect, collate, and disseminate data and information about Indian biodiversity. As noted earlier such data and information is currently distributed, isolated, in heterogeneous forms and format, and most seriously locked up in institutional and individual cupboards under the misconceptions of national security, intellectual property related sensitivity. While, there are few sporadic, isolated efforts being made in recent past, there is a need to coordinate these activities under one single over-arching umbrella. Thus, there is a need to conceive and establish "**National Biodiversity Information Infrastructure (NBII)**". Current technological and political scenario presents ample scope to undertake establishment of such a facility that is capable of collation, analysis, and dissemination of biodiversity and ecosystem related information form / to distributed sources.

NBII should be an interoperable network of biodiversity databases, information networks and systems, traditional knowledge, peoples biodiversity

registers, and information technology tools that will enable users to navigate and put to use the nation's vast quantities of biodiversity and ecosystem information to produce national economic, environmental, and social benefits. Thus, it would be an overarching information facility, which would leverage on progress made so far by the various information systems, networks and databases spearheaded by various individuals, institutions and groups within and outside India.

In this chapter, I have attempted to elaborate on the vision and operational objectives, potential work programs, major milestones, as well performance indicators of such a system. While highlighting the challenges, I have further discussed some of the technical implementation related issues such as use of web services architecture for evolving such an infrastructure, with its merits and demerits, as learned from similar implementations in other regions of the globe. Further, I have discussed the governance structure of NBII.

It is my belief that such a facility will contribute towards economic growth, ecological sustainability, and social outcomes through increasing the utility, availability and completeness of new and existing biodiversity and ecosystem information resources.

Annexure I: BIR, Biodiversity Information Resources Database

Our progress in biodiversity informatics is similar to that of the uneven distribution of biodiversity and biodiversity data. While, most of the development of information bases, standards and tools is happening in developed part of our globe, mega-biodiversity developing nations are lagging behind in collation and dissemination of data about their biodiversity. Currently, metadata of the biodiversity information resources themselves is not accessible at a single click of a mouse. To overcome this development of BIR, Biodiversity Information Resources database was undertaken. Annexure I, describe the development of BIR which has collated metadata for over 1300+ biodiversity information resources. Thus, BIR facilitate up-to-date and current documentation of existing and new biodiversity and ecosystem information resources. It was felt that metadata repository such as BIR needs to be constantly updated, if our goal is to bridge the imbalance between the biodiversity and ecosystem informatics products and distribution of biodiversity and its data.

Annexure II: Open Access Geospatial Data Repository (OAGDR)

In recent times, the need of having an easily accessible spatial data infrastructure has been emphasized by several communities such as scientists,

technologists, academicians, planners and even commoners. However, simple geospatial products such as shapefiles are not available when they are needed the most. To overcome this impediment, and foster a community-driven effort towards building a geospatial data infrastructure, an Open Access Geospatial Data Repository (OAGDR) has been developed. This annexure describe the purpose and development of OAGDR in detail.

Annexure III: Connecting Diversity: Pilot project for development of an interoperable framework for connecting distributed and heterogeneous bioresources databases

Chapter 5 of this dissertation deals with development of National Biodiversity Information Infrastructure (NBII). However, during my 15+ years of work in the area of biodiversity informatics, I was always queried if such a grand vision can ever be implemented? About two years back with support from Government of India's Department of Biotechnology, I was awarded a pilot project to implement web services architecture to interconnect biodiversity and bioresources databases. Annexure III, deals with progress of this pilot project which has been able to harvest together over 500,000 records from 7 distributed and heterogeneous databases through 2 data providers. This data is accessible through IBIF*prototype* portal (http://www.ibif.net.in/). Experience of this pilot project once again reaffirm that technology is not an impediment in bridging and interconnecting the data, but its mindset of individuals who hold this data.

ABBREVIATIONS

- ABCD Access to Biological Collections Data
- ANN Artificial Neural Network
- ARISNET Agricultural Research Information Network, India
- ATREE Ashoka Trust for Research in Ecology and Environment, India
- AVH Australian Virtual Herbarium
- BIN21 Biodiversity Information Network
- BioCASE Biological Collection Access Services
- BIOCLIM Bioclimatic prediction system
- BioGIS Isreal Biodiversity Information System
- BIR Biodiversity Information Resources Database
- BLOB Binary Large Object
- BMC Biodiversity Management Committee
- BNHS Bombay Natural History Society, India
- BoC Board of Consortium (as conceptualized in this dissertation)
- BOLD Barcode of Life Database
- BSI Botanical Survey of India, India
- BTISNet Biotechnology Information System, India
- CART Classification and Regression Trees
- CBD Convention on Biological Diversity
- CBIC Canadian Biodiversity Informatics Consortium
- CBIF Canadian Biodiversity Information Facility
- CDAC Centre for Development of Advanced Computing, India
- CDFD Centre for DNA Fingerprinting, India
- CDROM Compact Disc Read Only Memory
- CES Centre for Ecological Sciences, Indian Institute of Sciences, India
- CETAF Consortium of European Taxonomic Facilities
- CGI Common Gateway Interface
- CHM Clearing-House Mechanism, CBD
- CODATA The Committee on Data for Science and Technology
- CoL Catalogue of Life
- CoML Census of Marine Life
- CONABIO Comisión nacional para el conocimiento y uso de la biodiversidad
- CPCB Central Pollution Control Board, India
- CRIA The Centro de Referência em Informação Ambiental
- CSIR Council of Scientific & Industrial Research, India
- CSS Cascading Style Sheets
- CZA Central Zoo Authority, India
- DADI Data Access and Data Interoperability
- DATA&DSS Data Use, Applications and Decision Support System
- DBT Department of Biotechnology, India
- DEM Digital Elevation Model
- DHTML Dynamic Hypertext Markup Language
- DiGIR Distributed Generic Information Retrieval
- DM Data Manager
- DNA Deoxyribonucleic acid
- DoD Department of Ocean Development, India
- DRDO Defense Research and Development Organization, India
- DSIR Department of Scientific & Industrial Research, India
- DSN Data Source Number

- DSS — Decision Support System
- DST — Department of Science and Technology, India
- ECAT — Electronic Catalogue of Known Organisms
- EEZ — Exclusive Economic Zone
- EMBL — European Molecular Biology Laboratory
- ENBI — European Network for Biodiversity Information
- ENHSIN — European Natural History Specimen Information Network
- ENM — Ecological Niche Modeling
- ENVIS — Environmental Information System, India
- EoL — Encyclopedia of Life
- EoS — Earth Observation System
- ERIN — Environmental Resources Information Network
- ERMS — European Register of Marine Species
- ETI — Expert Center for Taxonomic Identification
- FaunaEuropea — The Fauna Europaea project
- FBI — Fauna of British India
- FGDC — Federal Geographic Data Committee
- FRI — Forest Research Institute, India
- FRLHT — Foundation for Revitalisation of Local Health Traditions, India
- FSI — Fisheries Survey of India, India
- FSI — Forest Survey of India, India
- G8+5 — Group of G8 nations plus outreach countries
- GAM — Generalized Additive Models
- GARP — Genetic Algorithm for Rule-set Prediction
- GBIF MAPA — GBIF Mapping and Analysis
- GBIF — Global Biodiversity Information Facility
- GBPIHED — GB Pant Institute of Himalayan Environment & Development
- GEON — Geosciences Network
- GEOSS — Global Earth Observation System of Systems
- GIS — Geographic Information System
- GISIN — Global Invasive Species Information Network
- GLM — Generalized Linear Models
- GPS — Global Positioning System
- GRID — phrase in “Distributed Computing”
- GSD — Global Species Database
- GSI — Geological Survey of India, India
- GSIS — Global Species Information System
- GTI — Global Taxonomy Initiative
- GUID — Globally Unique Identifiers
- HISPID — Herbarium Information Standards and Protocols for Interchange of Data
- HTML — Hyper Text Markup Language
- IABIN — Inter American Biodiversity Information Network
- IBIF — Indian Biodiversity Information Facility (Connecting Diversity prototype as described in Annexure III)
- IBIN — Indian Biodiversity Information Network
- IBIS — Indian Biodiversity Information System
- IBSD — Institute of Bioresources and Sustainable Development, India
- ICAR — Indian Council of Agricultural Research, India
- ICBN — International Code of Botanical Nomenclature
- ICFRE — Indian Council of Forestry Research and Education
- ICMR — Indian Council of Medical Research

- ICTI — Information and Communication Technology Implementation
- ICTVDB — The Universal Virus Database of the International Committee on Taxonomy of Viruses
- ICZN — International Code of Zoological Nomenclature
- ICZN — International Union of Zoological Nomenclature
- IGNFA — Indira Gandhi National Forest Academy, India
- IIFM — Indian Institute of Forest Management, India
- IISc — Indian Institute of Sciences, India
- IISER — Indian Institute of Science Education Research, India
- IIT — Indian Institute of Technology, India
- IITM — Indian Institute of Tropical Meteorology, India
- ILDIS — International Legume Database and Information Service
- IMD — India Meteorological Department, India
- IMoSEB — International Mechanism of Scientific Expertise on Biodiversity
- IN/IS — Information Systems/Information Networks
- InBIO — Instituto Nacional de Biodiversidad, Costa Rica
- INCOIS — Indian National Centre for Ocean Information Services, India
- IndFauna — Electronic Catalogue of Known Indian Fauna
- InfoNatura — Conservation and educational resource on the birds, mammals and amphibians of Latin America and the Caribbean
- ION — Index to Organism Names
- IPIN — International Plant Names Index
- IPR — Intellectual Property Rights
- ISIS — TDWG Invasive Species Information System interest group
- ITIS — Integrated Taxonomic Information System
- IUCN — The World Conservation Union
- JNCAR — Jawaharlal Nehru Center for Advanced Research, India
- JSP — Java Server Pages
- KEW — Kew Botanical Garden, London
- KNB — Knowledge Network for Biodiversity
- LepIndex — The Global Lepidoptera Names Index
- LIFE WATCH — European plan to link ecological monitoring data
- LSID — Life Science Identifiers
- LTER — Long Term Ecological Network
- MaNIS — Mammal Networked Information System
- MaPSTeDI — Mountains and Plains Spatio-Temporal Database Informatics Initiative
- MHRD — Ministry of Human Resources Development, India
- MoC — Memorandum of Cooperation
- MoEF — Ministry of Environment and Forests, India
- MoU — Memorandum of Understanding
- MSSRF — MS Swaminathan Research Foundation, India
- NABIN — North American Biodiversity Information Network
- NAS — Non-indigenous Aquatic Species
- NASA — National Aeronautics and Space Administration, USA
- NatureServe — A Network Connecting Science with Conservation
- NBA — National Biodiversity Authority, India
- NBDB — National Bioresources Development Board, India
- NBII — National Biodiversity Information Infrastructure (as conceptualized in this dissertation, ***see*** Chapter 5)
- NBN — National Biodiversity Network, UK
- NBRI — National Botanical Research Institute, India

- NBSAP National Biodiversity Strategy and Action Plan
- NCAOR National Centre for Antarctic and Ocean Research, India
- NCBI National Centre for Biotechnology Information, USA
- NCCS National Centre for Cell Sciences, India
- NCEAS National Center for Ecological Analysis and Synthesis
- NCL National Chemical Laboratory, India
- NCMRWF National Centre for Medium Range Weather Forecasting, India
- NGO Non Governmental Organization
- NHM, London Natural History Museum, London
- NHM, Paris Natural History Museum, Paris
- NIC National Informatics Centre, India
- NII National Institute of Immunology, India
- NIO National Institute of Oceanography, India
- NIOT National Institute of Ocean Technology, India
- NKC National Knowledge Commission, India
- NLWRA National Land and Water Resources Audit
- NMNH National Museum of Natural History, India
- NODC National Oceanographic Data Centre, USA
- NSD Natural Collections Descriptions
- NSDI National Spatial Data Infrastructure
- OAGDR Open Access Geospatial Data Repository
- OBIS Ocean Biogeographic Information System
- OCB Outreach and Capacity Building and IPR
- OECD Organization for Economic Co-operation and Development
- OMII Open Middleware Infrastructure Institute
- ORNIS Ornithological Information System
- OSR TDWG Observation and Specimen Records interest group
- PBR Peoples Biodiversity Register
- PCAST President's Council of Advisors on Science and Technology, USA
- PHP PHP Hypertext Processor
- PIR Protein Information Resource
- QC Quality Controller
- RDBMS Relational Database Management System
- REMIB World Biodiversity Information Network
- RFP Request for Proposals
- SACON Salim Ali Center for Ornithology and Natural History, India
- SBB State Biodiversity Board
- SDD Structured Descriptive Data
- SDI Spatial Data Infrastructure
- SDSS Spatial Descision Support System
- SEC US Securities and Exchange Commission
- SEEK Science Environment for Ecological Knowledge
- SNMNH Smithsonian National Museum of Natural History
- SOAP Simple Object Access Protocol
- SPiRE Semantic Prototype in Research Ecoinformatics
- SRTM Shuttle Radar Topographic Mission
- SSC Science Sub Committee
- ST Software Tools
- St/Pr/Sc Standards/Protocols/Schemas
- SWAT Strength, Weakness Advantages and Threats
- Swiss-Prot Manually curated biological database of protein sequences

- TaiBIF — Taiwan Biodiversity Information Facility
- TAPIR — TDWG Access Protocol for Information Retrieval
- TBGRI — Tropical Botanical Garden Research Institute, India
- TDWG — Taxonomic Database Working Group (*now* Biodiversity Information Standards)
- TE — Taxon Editor
- TK — Traditional Knowledge
- TKDL — Traditional Knowledge Digital Library
- TKR — Traditional Knowledge Repository
- TrEMBL — Translated EMBL, very large protein database in Swiss-Prot format
- TSN — Taxonomic Serial Number
- uBio — Universal Biological Indexer and Organizer
- UDDI — Universal Discovery Description Integration
- UGC — University Grants Commission
- UNFCCC — United Nations Framework Convention on Climate Change
- URL — Universal Resource Locator
- US NBII — US National Biological Information Infrastructure
- UTM — Universal Tranverse Mercator
- WII — Wildlife Institute of India
- WWW — World Wide Web
- XML — Extensible Markup Language
- Z39.50 — International Standard, ISO 23950
- ZSI — Zoological Survey of India, India

LIST OF TABLES

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 Sciecne 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 atleast 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 consortiums 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 exchage/share data. GBIF has used DiGIR, an XML based protocol capable of working with configurable federated schemas. Developed by University of Kansas Natural Histroy 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 GeoInteroprability 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.berkely.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 Infromation 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 maga-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 maga-biodiversity and developing nation, it is imperative that biodiversity informatics activities form corner stone 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.).

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 recourses 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.

Kingdom(s)	% of resources available as per kingdom in BIR
All Biota	10.74
Animalia	51.34
Plantae	27.38
Fungi	8.39
Bacteria	1.14
Chromista	0.40
Viruses	0.34
Protozoa	0.27

Table 1.1: Taxon analysis (in %) of resources in BIR

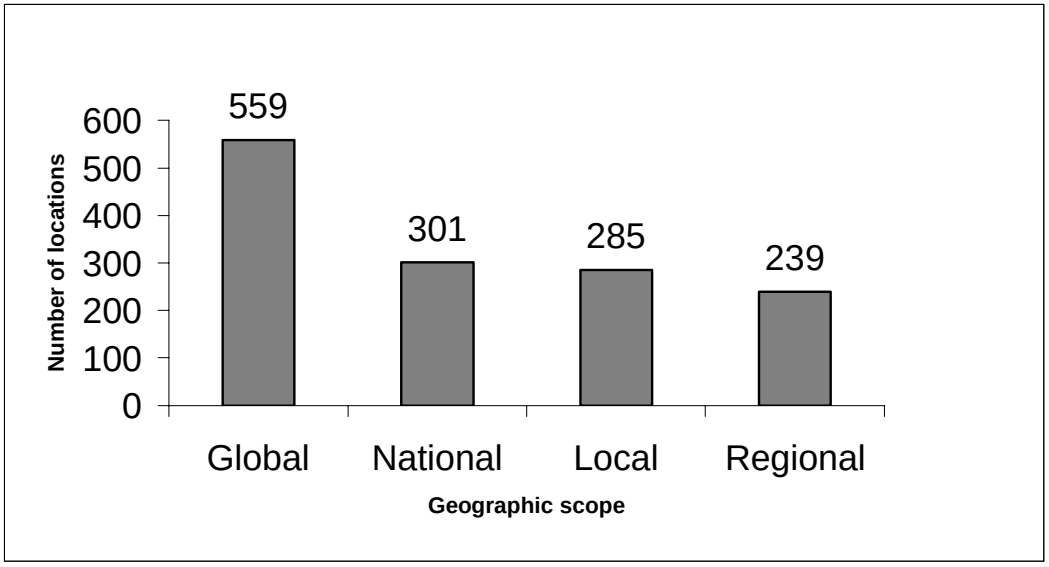


Figure 1.1: Geographical scope analysis of resources in BIR reveals that majority of resources are macro-scale in nature.

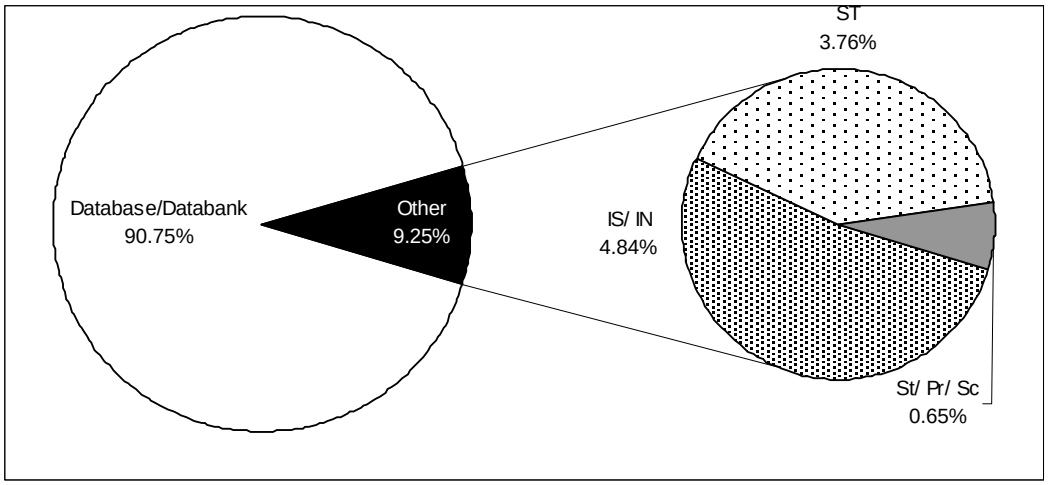


Figure 1.2: Different resource types (in %) BIR.

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

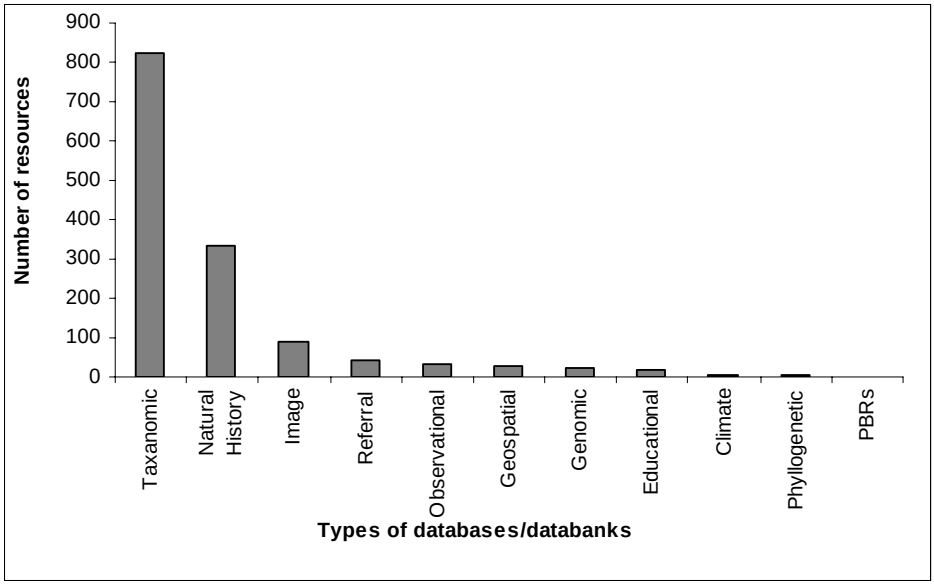


Figure 1.3: The ten different types of databases/databanks differentiated in BIR. Taxonomic types of databases are more in numbers than others databases.

[PBRs- Peoples Biodiversity Registers]

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

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

Table 1.2: Major Biodiversity Informatics activities in India

Chapter 2

IndFauna, Electronic Catalogue of Known Indian Fauna

Chapter 2

IndFauna, Electronic Catalogue of Known Indian Fauna

2.1 Electronic Catalogue of Known Organisms

The most practical and widely applicable measure of this biodiversity is "Species". They are the common currency for biodiversity research and management, and the only measure of biodiversity with a well-established standardized code of nomenclature (Costello, 2000). The presence of a species can indicate the habitats present, environmental quality and state of knowledge of biodiversity such as rates of discovery and extinctions. The relative richness of species in comparable samples can be a good indicator of environmental health. The most important aspect of biodiversity is species composition. From checklists of species taken over time, the rates of emigration, extinction, and turnover of species in a community can be measured and modeled. These dynamics measure the stability of biodiversity in ecosystems. Thus, species names or scientific names are thus at the foundation of quality control in biological studies. Further, scientific names are fundamental to biodiversity research as they are means of communicating information across the globe.

About 1.8 million species are "known" to the world so far, in the sense that they have been described and named by taxonomists (Edwards et al., 2000), however, it is estimated that anywhere from 3 million to more than 100 million species exist in the world today (May, 1999). Spatial and temporal patterns in biodiversity distribution can be analyzed by linking these names with information on nomenclature, taxonomy, ecology, distribution and abundance. Developing a single repository for such information is vital for future studies in biodiversity. Electronic cataloguing (ECAT) of known organisms provides an effective tool for collation, analysis and dissemination of information about biological diversity. Such national, regional, and global ECATs can be used for effective biodiversity management and policymaking (Peterson et al., 2002).

2.1.1 ECAT: Global Status

During last decade or so number of ECAT development activities have been started in different parts of the globe (Edwards et al., 2000). ECAT falls into two categories (a) Global Species Catalogues, and (b) Regional or National Species Catalogues. Global species catalogues are usually dedicated to specific taxa such as

AmphibiaWeb (http://www.amphibiaweb.org/), Fishbase (http://www.fishbase.org/), Cephbase (http://www.cephbase.utmb.edu/), Coleoperta (http://www.coleoptera.org/), ScaleNet (http://www.sel.barc.usda.gov/scalenet/scalenet.htm), AlgaeBase (http://www.algaebase.org/), ILDIS (http://www.ildis.org), ICTVDB (http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm), and Hexacorls (http://hercules.kgs.ku.edu/hexacoral/anemone2/index.cfm) etc. Regional or National species catalogues are aimed at inventorying species known to occur in specific region or country, such as European Register of Marine Species (Costello, 2000), Eur+Med PlantBase (http://ww2.bgbm.org/EuroPlusMed/query.asp), Flora of China (http://flora.huh.harvard.edu/china/), New Zealand Plant Names etc.

Three major cataloguing initiatives that need special mention are Integrated Taxonomic Information System (ITIS), Species2000 and OBIS. The goal of ITIS (http://www.itis.gov/) is to create an easily accessible database with reliable information on species names and their hierarchical classification. The database is reviewed periodically to ensure high quality with valid classifications, revisions and additions of newly described species. ITIS includes documented taxonomic information of flora and fauna from both aquatic and terrestrial habitats. However, it places greater priority on North American species. Species2000 (http://www.sp2000.org/) is global in its coverage and initiated to compile a "Catalogue of Life" (CoL), using distributed networking on the Internet (Bisby, 2000). It aims at creating validated and uniform index of world's known organisms (animals, plants, fungi, microbes and viruses). In early 1999, worlds leading marine scientists initiated Census of Marine Life (CoML) with ultimate goal of developing detailed series of online atlases that will facilitate researchers visualize where marine organisms once lived, where they are now, and where they might be found in future. Ocean Biogeographic Information System (OBIS), the data and information component of CoML accessible at http://www.iobis.org/ catalogue 80000 marine species integrated through 237 databases.

However, these as well as many national and regional databases are incompatible to each other in more than one ways. Realizing this, Global Biodiversity Information Facility (GBIF) was formed with major goal of providing a mechanism to promote and enhance the development of standards required for interoperability (Edwards et al., 2000). GBIF aims at indexing at least 95% of the 1.8 million known species names and its associated baseline data by 2011 (GBIF, 2006). In its first

phase, GBIF is interoperably linking species- and specimen- level databases, helping to complete the electronic catalogue of names of known organisms and devising a plan of outreach and capacity building, so that individuals in all countries will have access to scientific biodiversity data. On December 15, 2003, GBIF signed a memorandum of cooperation with Spceies2000 and ITIS Catalogue of Life Partnership to expedite the process of cataloguing all known species (GBIF 2004). Catalogue of Life is partnership between Species2000 and ITIS organization, which decided to work together in 2001 with a target to complete cataloguing of all 1.8 million known species by 2011. On March 29, 2007, it 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).

Preceding discussion reveals that majority of these initiatives are from the agencies in the developed regions of the globe. However, highest concentration of biodiversity is in the tropical region, especially in developing and under-developed nations. Similar to distribution of biological specimens in the museums of developed nations, biodiversity databases too are being developed by institutions from developed world. It is therefore essential to encourage developing nations such as India, to undertake development of electronic catalogue of its known biotic resources.

2.1.2 ECAT: National Status

Indian scenario is in many ways representative of the difficulties faced by developing countries in biodiversity cataloguing. The rich diversity of Indian biota has posed considerable challenge to generations of taxonomists in India and across the world. In addition to the two hotspots of biodiversity, Western Ghats and Eastern Himalaya, specialized ecosystems such as islands, oceans, deserts and mountains across India are rich in flora and fauna. The available information about flora and fauna is distributed in various sources and is not available from a single source. What makes India a more interesting nation are its multiple religions, ethnic communities with diverse lifestyles, habits, languages and cultures. This provides another dimension to documentation of biodiversity, as single species is known in different regions and languages with variety of local or vernacular names. There is an urgent

need to develop widely accessible, up-to-date repository collating information on scientific names and its common or local names in various regions, and languages.

In recent past, sporadic efforts have been made to electronic documentation of known Indian biota. Ashoka Trust for Research in Environment and Ecology (ATREE), Bangalore has released CDROM titled “Sasya Sahyadri” (Ganeshaiah, 2003), collating baseline information on flora of Western Ghats. Legume database of South Asia developed by the National Botanical Research Institute, Lucknow has been integrated in Global Legumes Database (NBRI, 2007). National Institute of Oceanography, Goa has developed taxon specific CDROM titles on Marine Prawns of India, Marine Crabs of India, Mangroves of India, Lignicolous Fungi of India (NIO, 2007). Jawaharlal Nehru Centre for Advanced Research (JNCAR), Bangalore has developed database on flora of Karnataka (Ganeshaiah et al., 2002) indexing 4758 floral species that occur in Karanataka. Salim Ali Centre for Ornithology and Natural History (SACON), Coimbatore is developing database of Birds of India (SACON, 2007). FRLHT, Bangalore is developing medicinal plants database (FRLHT, 2007), and French Institute of Pondicherry has released database on Endemic Plants of Western Ghats (Datta et al., 1997).

This indicates that electronic cataloguing of known life in India is happening in various distributed and isolated pockets. These datasets are restricted to some geographical regions or to certain taxonomic groups. Since, most of these are offline in nature, access to information requires special efforts. There is no interaction between individual developers and due to lack of uniform standards; most of them are incompatible to each other, raising serious interoperability issues.

This calls for coordinated and integrated approach to collect, maintain and provide baseline information on Indian biota so as to develop web accessible electronic catalogues of known Indian species. For a mega-biodiversity nation such as ours, it is critical to have anytime, anywhere access to baseline information about our biotic resource for their efficient sustainable use and management.

2.2 Indian Fauna

2.2.1 Faunal diversity in India

According to recent estimates, there are about 89,451 known faunal species in India, which are about 7% of the total animal species in the world (Alfred, 1998). However, only less than 50% of the geographic region of the country has been surveyed so far (Pushpangadan and Nair, 2001). The earliest studies on fauna date

back to 1800s from which Fauna of British India (FBI) was put together by various researchers until 1940. However, since then, several new species have been described and taxonomic status of the species has undergone many revisions (Das, 2003). New species are discovered every year from various parts of India and there is no centralized system to disseminate secondary information regarding these descriptions.

The exploratory phase in Indian taxonomy can continue for a long period as several areas such as Eastern Himalayas and Andaman and Nicobar Islands have not been totally surveyed so far. Several invertebrate phyla, viz., Nemertinea, Nematomorpha, Pogonophora, Priapulida and Pentastomatida are yet to be explored thoroughly from India (Alfred, 1998). The lower groups of organisms especially insects are still to be documented in detail. A large number of invertebrate taxa are mainly known from collections in museums across the world to which Indian taxonomists have limited access. Although FBI is available as baseline literature, many invertebrate taxa were not covered in FBI and as a result their information is only available in monographs, collection records and catalogues published outside India. The information from various taxonomic studies carried out so far in India is distributed with several organizations and individuals making it difficult to access adequate and accurate information on variety of aspects of faunal diversity.

2.2.2 Faunal diversity studies in India

Zoological Survey of India (ZSI) is the central institution dealing with documentation of Indian faunal diversity. Established in 1916, ZSI has published monographs and taxonomic revisions of many taxa together with around 3929 new records during 1916 –1991 (Das, 2003). In addition, several research groups in universities, colleges and research institutions across the country are also working on faunal taxonomy. Information related to the ecology, population biology, biogeography, phylogeny and traditional knowledge regarding fauna is available through various research projects and surveys by institutions such as Bombay Natural History Society (BNHS), Wildlife Institute of India (WII), Forest Research Institute (FRI), Salim Ali Centre for Ornithology and Natural History (SACON), ZOO Outreach etc. This information is mostly available in the published text format as survey reports or research papers and is not easily accessible to all.

2.3 IndFauna, Electronic Catalogue of Known Indian Fauna

2.3.1 IndFauna: Why?

At present there is no single repository to provide information such as scientific names, common names, occurrence of organisms, their spatial and temporal distribution and bibliography. Users such as conservationists, policy makers, environmental managers and para-taxonomists feel the need for this basic information about Indian fauna. Taxonomists themselves often feel the need of a single information source on Indian fauna and quick access to references. The diversity of languages and cultures across the India should also be taken into consideration while disseminating information. It is therefore necessary at this stage to create an information system to collate existing information, create tools for receiving new information with facilities to integrate, exchange, and disseminate it in multiple ways. ECAT provides best approach to compile and to integrate or exchange information. Therefore, IndFauna, Electronic Catalogue of Known Indian Fauna!

2.3.2 IndFauna: Features

IndFauna, was conceived as web accessible catalogue that would encompass all taxa of known Indian fauna including protozoans covering current political geographic coverage of India. Thus, IndFauna would encompass both terrestrial as well aquatic (freshwater and marine) fauna recorded so far from it terrestrial jurisdiction, as well 7000 km coastline and 2.1 million sq. km Exclusive Economic Zone (EEZ). Being web accessible catalogue, IndFauna was conceived as dynamic data collection, collation, management, and dissemination information system, which should be able to provide collaborative environment to data contributors and data validators, the “Taxon Editors”. Unlike other electronic catalogues, data collation through IndFauna is live and transparent in the sense that there is no residency period for data acquisition and its dissemination.

After review of major global and regional electronic catalogues, it was decided that IndFauna would collate following baseline data/information for each species.

1. Valid scientific name with authority, year of publication according to accepted taxonomic opinion.
2. Systematics of the species from Kingdom to forma level
3. Synonyms with authority and publication year
4. Common / local / vernacular names with language, region and references.
5. Details on references (DSN, Data Source Number).

6. Latest taxonomic scrutiny (Name of "taxon editor" and date of latest taxonomic scrutiny).
7. Biogeography, the occurrence within India with location maps and references.
8. Multimedia artwork such as Images, sketches, photographs, and audio-video clippings.

Thus, IndFauna adapts a species-centric approach, where scientific name is the nucleus of the database (Fig. 2.1) to which taxonomic, synonym, common name, biogeography and other information is linked. Similar to Species2000 and ITIS (Bisby, 2000), IndFauna brings together taxonomic treatments from authors and institutions to provide a centrally collected system for Indian fauna.

2.3.3 IndFauna: Development and Processes

From conception in early 2001 to production on http://www.ncbi.org.in/, process of development of IndFauna could be divided into five distinct steps listed below.

(a) Conceptualization, Implementation plans and monitoring process,

(b) Design of database architecture

(c) Data Management

(d) Data Curation and Taxonomic Scrutiny

(e) Data Dissemination and Feedback

However, the whole process of IndFauna development was dynamic in nature that each of these processes evolved and progressed concurrently. For instance, initially it was conceived to collate only valid scientific names and their synonyms, together with occurrence details. However, as the development and implementation began, it was realized that collating common names and vernacular names of organisms practiced in various languages and regions of the country together with sources of data and artwork would add the value to IndFauna, and thus enhance its usability. Initially, visualization of occurrence data into web based geo-referenced environment was not the objective, and the database was not catering to the process of geo-referencing, which was added up at later stage of the project. Several features were developed because of needs, experiments, and experiences as project moved forward. Thus, development of biodiversity information system is ever evolving process and always work in progress, and IndFauna justify this in more than one way.

2.3.4 IndFauna: Architecture

The complex and interlinked biodiversity data and its dynamic nature posed many challenges for data management, its interlinking, integration as well dissemination. While there are tools available for creating offline inventory and descriptive information systems such as Linneaus II of ETI (Schalk, 1998), Platypus (Platypus, 2003), ITIS workbench, I could not come across tools or programs, which can be used for developing web-interfaced electronic catalogue. Further I realize that many of them were not able to encode biodiversity data collected from (a) disparate sources from legacy literature to electronic data files, worksheets, and databases, and (b) data types from referral, and bibliographic, taxonomic, geographic, environmental, and even multimedia. It was therefore necessary to set up a unique cost effective and easy to use information system for providing faunal database to assist searching of locations, taxonomy and other information of the fauna. In addition, precision and ease in data entry was required to deal with rigorous task of entering data records of about 90,000 species.

The taxonomy data explains the relation between the species based on certain characteristics. These characteristics on which the species are defined may vary in time due to discovery of a new class of characteristics, or corrections to previously recorded characteristics, etc. The system had to be flexible enough to accommodate frequent changes in taxonomic hierarchy, which is a common feature of all biodiversity data. Systematic data for a species is not uniform in different taxonomic systems and causes problems in building standard data sets. These faunal data types have deeply nested relationships within and between themselves. This difficulty was overcome by matching to the hierarchy with standards of ITIS, Species2000, and also referring with International Code of Zoological Nomenclature (ICZN).

Hence, the design goal was to create a database which can accommodate separate ownership of biodiversity data by different departments and deal with disparity in data management standards which is the main difficulty in biodiversity information exchange and sharing, and comprehensive processing. Therefore, IndFauna database structure were created keeping in view that it should facilitate data acquisition, storage, query support, unique species digitization, restrictive data access and easy to use with bare minimum infrastructure.

2.3.4.1 Database structure

IndFauna consists of 22 tables. Data content of each table is summarized in Table 2.1. Out of the 21 tables, 14 are primary which collate scientific data. Structures of these primary tables are detailed in Tables 2.2 – 2.15. There are 8 supporting or secondary tables that hold secondary data and are designed to facilitate easy data management. In order to capture of data management, such as (a) who entered, (b) when entered, (c) who edited, (d) when edited, (e) who deleted, and (f) when deleted, appropriate fields were incorporated into all primary and relevant secondary tables. Between the tables one-to-many, or many-to-many relationships were established (Fig. 2.1). These were required to link data component such as scientific name, synonym, common name, and locality with the DSN (Data Source Number) which is the reference from which data has been culled out and entered into the system. In order to ease the data cleaning and taxonomic scrutiny process, soft copies of these literature sources were also uploaded into the database as BLOB (Table 2.15). In case of multimedia artwork of species, it was uploaded into pre-assigned folder, and file path was stored into the "image" table. Thus, depending on the type of data to be collated NUMBER, VARCHAR, DATE, and BLOB data types were assigned to each of the fields in the tables.

The entity relationship diagram (Fig. 2.1) depicts the relationship between all these 21 tables. The "sciname" table (Table 2.2) is central to the entire database. The DSN table (Table 2.14) has one to many- relationship with the "sciname" (Table 2.2), "synonym" (Table 2.3), common name (Table 2.4) and "sci_loc" (Table 2.12). Table "sci_loc" collate occurrence or biogeography details of the species. It is linked with six supporting or secondary tables namely country, state, district, locality, and waterbody. The taxonomic hierarchy is managed through seven tables namely, "kingdom" (Table 2.5), "division" (Table 2.6), "class" (Table 2.7), "orders" (Table 2.8), "family" (Table 2.9), "genus" (Table 2.10), and "species" (Table 2.11). These are linked with each other using primary key and foreign key features. Table "species" is linked with "sciname" table using similar feature. Table "image" (Table 2.13) collates the metadata of the multimedia artwork of the species. It is linked with "contributor" table collating details of contributor of the image. Table "users" stores the login and password and privilege levels of the "data managers" (DM), and "taxon editors" (TE) for administration and taxonomic scrutiny purposes respectively.

2.3.4.2 Data Flow

IndFauna development could be grouped into three types of functionalities, viz., (a) Data Management and (b) Data Curation or Taxonomic Scrutiny, and (c) Data Dissemination and Feedback. Each of these functionalities has set of modules whose features are described in detailed in subsequent sections of this and next chapter (Chapter 3).

The dataflow in different data management modules (add, edit, delete and search) are bimodal as depicted by the Fig. 2.2, 2.3, and 2.4 respectively. The valid username and password entry leads a "Data Manager (DM)" to various Data Management modules (Fig. 2.2). The processes begin with the search of a particular record, which if exist in turn will lead to the edit or delete modules. If the record is not present in the database the data manager is directed to the Add module (Fig. 2.3).

2.3.4.3 Security and Privileges

For the purpose of security and integrity of data, different "Data Managers (DM)" and "Taxon Editors (TE)" has varied data access, add, edit, and delete privileges For instance, "Super User" has complete access and all privileges to add, edit, and delete the data. However, DM can add data, and edit only data that he or she has added. Further, DM can only flag data entered by him or her for deletion, however can not delete the data. Super User is only authorized to delete the data. Further, "taxon editors" do not delete the records directly, but they simply mark them for deletion. Super user in consultation with panel of experts takes final decision on permanent removal of the record. System also generates history (date, time and details of data contributed) for each of the records contributed.

2.3.4.4 Tools

The database creation was one of the major tasks in implementing the system to store the data. Considering the quantum of data to be handled, integration of variety of data types, secured collaborations, scalability, and hassle-free storage and archival requirements, Oracle 9i is used as a database server (Corey et al., 2002). JSP (Java Server Pages) is used to create user-friendly web interfaces. Apache is used as a web server, and Tomcat 4.1 is used as a container for JSP applications.

2.3.5 IndFauna: System Modules

As mentioned in previous section, IndFauna modules could be grouped into three types of functionalities, viz., (a) Data Management, and (b) Data Curation and Taxonomic Scrutiny, and (c) Data Dissemination and Feedback (Fig. 2.5). IndFuana

being web based, all these modules are accessible on the web and thus could be used from distributed locations over the World Wide Web. Of these, (a) Data Management and (c) Data Dissemination and Feedback modules are described in following sections. Modules dealing with (b) Data Curation and Taxonomic Scrutiny are described in Chapter 3.

2.3.5.1 Data Management

Data Management or Administration modules have been developed to perform add, edit, and delete functions for management of (A) Data Source, (B) Scientific Name and Taxonomic hierarchy, (C) Synonyms, (D) Common Name (E) Occurrence, (F) ArtWork, (G) Terminology, (H) Data Managers Performance Report, and (I) LinkOut with other databases.

(A) Data Source

In order to maintain the quality, authenticity of data, for every byte of data collated in IndFauna, there is corresponding reference in the form of "Data Source Number (DSN). This could be peer reviewed research paper, monographs, thesis, survey reports, and global, regional or national species checklists. More about the selection of these reference literatures is described in Chapter 3.

In order to avoid duplicate entries for same reference, addition precedes search of DSN record. Search is conducted on any of seven fields, viz., DSN number, title, author, year, publisher, availability, and user_id (Data Manager). Once it is ascertained that reference is not present in the database, DM can proceed with add feature. Upon adding the details of the reference, unique DSN number is assigned to it, which is then used for representing any data culled out and integrated into the database. As depicted in Fig. 2.6, using edit feature, details of data source could be updated at any time. For each valid scientific name in the database and its relevant information is supported by one or more DSNs. However, data source record entered by one DM can not be updated by other DM. When, DMs wish to delete the specific data source records, they can only flag it for deletion and Super User would take final decision to delete the record.

(B) Scientific Name and Taxonomic Hierarchy

This is the core data management module of IndFauna. In order to avoid multiple entries of scientific name and its valid synonyms, search is carried out in both 'sciname' and 'synonym' tables. If a name is not present in the database, it facilitates collating data on its authority, year when described, systematic (valid,

accepted, or provisionally accepted), threat (extinct, threatened, and vulnerable, etc.), and invasive (invasive or non-invasive) status. Data Manager is also required to add the DSN number and DSN page number from where the data is collated. Without DSN number, module would not accept the record for further processing (Fig. 2.7).
Once this data is submitted, taxonomic hierarchy can be added for which "bottom to top" approach of data management has been adopted. Using "bottom-top" approach taxonomic hierarchy is entered in un-conventional manner i.e. from species to kingdom level. Since, scientific name consists of genus + species and its variants, both "species" and "genus" forms draw it directly from scientific name field. While, entering the higher taxonomy, there is drop down options to select the appropriate level name (Fig. 2.7). This prevents any spelling mistakes and also enhances the speed of data management. Further, at any stage, if higher taxonomic level hierarchy is present, then it provides option to select it, so that data management could be expedited, as well common errors could be avoided.

In case, either scientific name or its valid synonym is present, modules leads to edit option, where scientific name details could be edited, and synonyms, common names, and occurrence details could be added or edited.

(C) Synonyms

This module has been designed to collate the synonyms of valid scientific names. However, taxonomy being dynamic in nature valid scientific name according to one taxonomic opinion could be synonym according to other. Thus, before collating synonym data, valid scientific name for which synonym is to be added is searched in both "sciname" and "synonym" table, so as to ascertain that it is not entered in synonym table and vice versa. Once, this is assured new synonym(s) could be added into the database. In case, more than one synonym is to be added, multiple synonym management form could be used (Fig. 2.8).

(D) Common Name

Common Name module has been devised to collate common names of a species used in various regions of India, as also the English common names. Single species can have multiple common names in different languages, and regions. Sex of an organism as well life stage (adult, juvenile, etc.) could also be reasons for common names. Single common name or multiple common names of same species could be collated using "single common name form", or "multiple common name form", respectively (Fig. 2.9). In case, same common name is being practiced in two regions,

languages, or for more than one organism, record could be edited appropriately using edit option.

(E) Occurrence

This module collates all the known occurrence details of species in time and space. In order to avoid the duplication of same observation record being entered more than once, it is mandatory to perform search into "sci_loc" table, which stores the occurrence records. Once it is ascertained that observation record is new for a given species, or the locality is new for entire database, its details could be added using this module. This module provides add-on feature to link every locality with single or multiple districts, or states incase of terrestrial organisms, and rivers and other water bodies including seas or oceans in case of aquatic organisms (Fig. 2.10). In case of precise localities, module facilitates to incorporate accurate geo-coordinates. If the locality record is broad area such as 'western ghats', its geo-coordinate range could be added. Period of occurrence of species can be recorded as precise date of observation, or period of observation. Associated with this is the data on altitude, which would prove important in analysis of spatial and temporal distribution of each species. Similar to other modules, it is also mandatory to provide appropriate DSN number for each collated locality details.

(F) ArtWork

This module facilitates acquiring multimedia artwork for each species. These could be photographs, illustrations, as well as audio, and video clippings. For each contributed artwork data on contributor and detailed description of artwork need to be submitted, before the actual artwork is uploaded to a database (Fig. 2.11) by a contributor or taxon editor. Such an artwork could later be used in visual identification of the species based on characteristics as captured in the graphics.

(G) Terminology

This module was designed to build in dictionary of various terms used in describing fauna, as well its biology, habitat, etc. For instance, term "Protozoa" is also referred as "Protists", where as place "Pune" has been referred in literature as "Poona", "Punyanagari" etc. As depicted in Fig. 2.12 for a given term, its aliases along with description or meaning of term could be added, edited.

(H) Data Managers Performance Report

As discussed earlier IndFauna is web based information system, which means several of its contributor and DMs were distributed all over the Internet. This module

was developed to review the performance or contribution of each of the DM, more importantly it was also developed keeping in mind to ascertain which byte of data has been added or edited by which DM. As depited in Fig. 2.13, it facilitates periodic report for all the DMs or a particular DM.

(I) LinkOut with other databases

IndFauna being web-based information system, it is possible to establish LinkOut with other datasets developed by outside agencies. Linkages with the databases such as sequence or molecular data, geospatial and climate data, and ecological and ecosystems data, will enable "data mining" never before possible (Edwards et al 2000). I always feel this simple looking LinkOut exercise will facilitate the exploration of questions that, at present, can not readily be answered. During development of IndFauna, an attempt has been made to LinkOut each species record with public domain sequence databases such as nucleotide sequences (NCBI, and EMBL), protein sequences (NCBI, EMBL, Swiss-Prot, PIR, and TrEMBL), and protein structures (PDBSum). Out-link has also been provided with GoogleImages. Similar out-links can be attempted with complementary resources such as Google Scholar, Google Books, Barcode of Life Database (BOLD), Species2000 ITIS Catalogue of Life, uBIO, PubMed, Scirus, and ZooBank, etc.

2.3.5.2 Data Curation and Taxonomic Scrutiny

This forms the core of IndFauna functionalities which ensures the quality and taxonomic authenticity of the data being collated and disseminated by it. IndFauna modules developed for data quality control, assurance and taxonomic scrutiny are described in Chapter 3, as it deals with data quality, data assurance, and taxonomic scrutiny.

2.3.5.3 Data Dissemination

User interface is the main feature of the IndFauna, because it is the only way through which user will get connected to the system. Hence, user interface should have good visualization interface for the results produced by the system. The establishment of user needs was a very important aspect, because most of the end-users are biologists. Thus the result could be presented in a systematic, easy to understand and in a simple text format.

The search module facilitates the web users to query on scientific name, common name, synonym, and occurrence details. Wildcard searches can be made using options such as "contains", "is", "begins with", and "ends with" for each of

these categories (Fig. 2.14). Hyperlink feature of the web has been used to facilitate retrieval of data on other parameters irrespective of search category. For instance one can search for specific locality using occurrance search module, and also retrieve data on taxonomic hierarchy, synonyms, common names, and ArtWork of each species recorded from a given locality.

In order to represent the occurrence data over the web so that it can make sense to users, exclusive geospatial data dissemination module called "JaivaNaksha' was developed. Rational for developing JaivaNaksha, its features and development has been described in Chapter 4.

2.3.5.4 Suggest a New Species

As discussed in section 2.3.2, IndFauna has been conceived as dynamic data collection, collation, management and dissemination information system which can provide collaborative environment to data providers and experts to validate it. As first step in this direction, and to encourage potential data providers to contribute data, "Suggest a new Species" module has developed. If while searching an expert notice that species of his or her interest is not present in the database, he or she can communicate the same to IndFauna developers. In addition to contributing data about species that is currently not present in IndFauna, contributor is also prompted to provide his contact details along with the peer reviewed published data source where more details of the species could be sought (Fig. 2.15).

2.3.6 IndFuana: Testing

IndFauna was tested using VV testing approach which has four types of testing viz., (a) Unit test, (b) Integration test, (c) System test, and (d) Acceptance test (Patton, 2006). Unit test or also called as "White Box test", involves checking syntax, typographical errors, as well testing execution of all programs and codes, where in tester has the knowledge of system. Integration test or "Black Box test" verifies inter-linking between sub-systems, and the tester does not have knowledge of system. Compatibility test was carried out to verify the compatibility of IndFauna application with various web browsers, hardware, and operating systems. For instance, initially the web server was on windows, but it was not able to give sufficient security from viruses. Thus the site was shifted to Linux.

Functional test was carried out to validate behavior of each feature, using normal and erroneous data input. At times, multi user data entry was going on correctly, but data miss-linking happened because of the lack of higher level of

transaction. The higher level of transaction was given to the connections of oracle and java. Another serious problem was that Oracle did not accept string containing an apostrophe (e.g. Common name: Marshall's Iora). Replacing the single apostrophe by double solved this problem.

Performance test was carried out to identify bottlenecks with high use applications during normal, peak and exceptional load conditions. For instance, during continuous data entry, the oracle was not able to store the huge data in to memory. An error ("Number of cursers open exceeded") used to occur when multiple searches were done on oracle data. It occurred when open connection exceeded the limit. Increasing the open cursor limit of the oracle solved this problem.

Beta users were invited for acceptance test, and their suggestions were taken into consideration to improve the system. Acceptance test also resulted into set of suggestions and directions to DMs for efficient data management practices. Initial, data addition process was lengthy and thus time consuming. Beta testers suggested deployment of various pull-down and pop-up lists showing pre-defined attributes, and attributes generated from database at run time that eased out the data entry sequence, and ensured consistency in data entry process.

2.4 IndFauna: Data collection and management

The data incorporated in the database is collected from multiple sources. As on date, data for IndFauna has been collated from over 12500 data sources. Main focus was to collate data from published literature including research papers, faunas, and monographs as sources of authentic and reviewed information. The collections of specimens in various museums are also important as they provide primary information regarding identification and occurrence of certain species. This is especially important in case of invertebrate taxa for which published literature is not available. Recently many individuals and institutions have created web-based databases and checklists that are useful for getting information regarding Indian fauna. However, the information is carefully scrutinized for validity and accepted only if it is from reputed taxonomic institutions or experts. Non-taxonomic research papers on ecology, physiology, animal behavior, distribution etc. are a secondary source of information especially for information supplementary to the scientific name.

In case of invertebrate taxa such as Lucanidae, Embioptera, Anthicidae there is no published checklist for India. In this case, personal communication with taxonomists across the world was made to acquire data on Indian taxonomy. Thus, for

collecting information, highest importance is given to “faunas” and “monographs” followed by “published research papers”, “online and offline databases”, “region and taxon specific web sites” followed by “personal communications with experts”, and at the end to “non taxonomic publications”. Detailed analysis of these data sources has been discussed in Chapter 3.

2.5 Data Curation and Taxonomic Scrutiny

Cleaning and validation of data; taxonomic scrutiny of taxonomic data; and geo-referencing of occurrence data is the key to provide accurate, authentic, and appropriate baseline information on various Indian faunal species. Thus, in order to achieve the acceptable quality of data various processes were incorporated at various levels of IndFuana development. These include (a) Orientation, and Hand-on-Training to DMs, (b) Error detection mechanism in Data Management and Curation modules, (c) Taxonomic Scrutiny, and (d) Geo-referencing of occurrence data. Of these (a), (b), and (c) are discussed in Chapter 3, while (d) has been described in Chapter 4. Thus, owing to the data curation, taxonomic scrutiny, and geo-referencing as in built processes in IndFauna development, it has potential to be used as peer-reviewed online publication accessible to all.

2.6 IndFauna: Significance and Future

In case of biodiversity research, scientific names identify entities, determine the relationship between entities and facilitate location, function/role. Current exercise of creating IndFauna has demonstrated that electronic catalogue is powerful tool to collect, analyse and disseminate biodiversity information, anywhere, anytime. This unique single source for information on Indian fauna would provide sound base for resolving conflicts in taxonomies, planning future research and analysis. IndFauna collates and disseminate baseline data about nearly 94500 known faunal species in India, thus surpassing the ZSI estimation of 89451 (Alfred, 1998). Together with valid scientific names, and their synonyms, IndFauna documents 147,937 names, 15102 common names, and 176,447 occurrence records representing 6577 unique localities spreading over 250 years of modern day biology (as on June 16, 2007).

As described earlier, the lack of central registry for names of organisms is a major impediment in tracking the number of species in India. IndFauna offers an effective method of creating a unique register. Owing to the rules of acceptance of scientific names, the names cannot be registered as valid before the publication of description in a journal. To solve this, precedent can be set that in case of each new

description; along with the type specimen a provisional registration number in the national ECAT, such as IndFauna should be quoted. This provisional number will be later changed to permanent registration number after furnishing the proof of valid publication of the species. Provisional registrations, which fail to be converted to permanent registration, will automatically be considered invalid after a certain period to be agreed by the taxonomic fraternity in the country. This method will lead to standardization of procedure for new name publication, will be useful for searching new descriptions and will electronically track the species publications in future.

ECAT can also be effectively used for linking information on species within diverse databases, as has been demonstrated in current exercise with LinkOut to sequence and image databases. In the same way LinkOut can be established with ecological, ecosystem and climate database, or with host-parasites, pray-predator, and food plant databases. Similarly, LinkOut with databases of natural history collections will be of much help in taxonomic research.

ECATs such as IndFauna are only the starting point for biodiversity management and research. These lists provide a single nomenclature for species, which will generate further research to clarify anomalies (Costello, 2000). ECATs not only form basis for more elaborate and specific databases on groups of organisms or species, but also would benefit cooperation amongst scientists, leading to increased communication and interest in the management and use of taxonomic data. ECATs will provide standard working list of names for non-specialists to use. Analysis of ECATs will identify where identification guides are most needed, in what taxa most species remain to be discovered, and where the expertise is weakest. It is anticipated that ECATs will become a standard reference and technological tool for biodiversity training, research and management. It can be used (a) to check the spelling or find the correct name of a species and the authority, (b) to find information on the distribution of species, (c) correct taxonomic information of species or group of taxa, and (d) indicate the level of knowledge of a group of species by analyzing the rate of discovery of species. However, for this to happen taxonomic scrutiny of collated data is essential, and IndFauna provides the platform to undertake such scrutiny exercise. Some of the experiences of scrutiny events for three Orders, viz. Coleoptera, Lepidoptera, and Hymenoptera are discussed in Chapter 4.

For many organisms information is available only in the text form. Use of images, audio and video clippings along with scientific names will be beneficial for

future identification. ECATs can serve as basis for electronic field guides. Use of GIS and mapping tools to display and develop dynamic maps of species distribution would enhance the quality of end results, and create much required awareness amongst common people about state of species distribution.

Since, multiple cultures, with diverse lifestyles, habitats, languages and dialects co-exists in India, it is essential that databases be made available in multiple languages which will make it friendly to users from all parts of India (Chavan et al., 2003, 2004). This will help in dissemination and acquisition of data from distributed sources. Further, it would help in overcoming the geographic and language barriers in biodiversity information.

The experience of developing IndFauna raises one question. Why an exercise of inventorying nations known biota was not initiated before? It is often argued that we do not have resources, man-power, and standardized procedure for undertaking such an exercise. It is my feeling that more than these essentials, we need determination to begin and lead it to successful completion, with determination to collaborate and coordinate between cross-discipline, and diverse expertise.

2.7 Recommendations

- Electronic catalogues of known biota are crucial in order to document and disseminate biodiversity data and information, and thus should be encouraged.
- Development of ECATs is a lengthy process that requires collaboration between domain experts, data managers and IT experts.
- ECATs should be used to overcome taxonomic impediments, and thus needs to be adopted as basic tool in taxonomic studies, and research.

2.8 Summary

One of the major mega-biodiversity nations, India is known to harbor rich and diverse biotic resources within the length and breadth of its territory. Data and information regarding these resources remains distributed with several organizations and individuals, making it difficult to access adequate and accurate information about them easily and efficiently. This calls for development of well-constructed electronic catalogues (ECAT) of known biotic resources of India. IndFauna, electronic catalogue of known Indian fauna, accessible at http://www.ncbi.org.in/, provides baseline data about over 94500 known faunal species in India. Experience of developing this web-interfaced catalogue of known Indian fauna, IndFauna, demonstrates that information

technology plays vital role in documenting, and disseminating biodiversity data and information. IndFauna, not only demonstrated that ECATs is powerful tool to collect, collate, analyze, and disseminate biodiversity information anywhere, anytime, to anyone; it put forth the model of collaboration between domain experts and IT managers. Further, IndFauna evidences the feasibility of biodiversity documentation efforts turning successful within restricted resources through determination, collaboration, and cooperation within institutes, organizations, and experts. Thus, it advances the agenda of overcoming taxonomic impediments, and better sustainable use and conservation of our biotic resources.

Table name	Data content
Primary Tables	
sciname	SCID (auto-generated), valid scientific name of a species, authority, year described, taxonomic status, threat status, invasive status, Organism type, Data Source Number and Data Source Page Nos., Scrutinized by, Scrutiny date
synonym	SCID (auto-linked), SNID (auto generated), synonym, synonym authority, year described, synonym status, cause of synonym, remarks, Data Source Number, and Data Source Page Nos.
commonname	SCID (auto-linked), CNID (auto-generated), common name, language, region, sex, life stage, remarks, Data Source Number, and Data Source Page Nos.
kingdom	KID (auto-generated), Kingdom, sub-kingdom
Division	KID (auto-linked), DID (auto-generated), division or phylum, sub-division or sub-phylum
Class	DID (auto-linked), CID (auto-generated), superclass, class, sub-class, infra-class
Orders	CID (auto-linked), OID (auto-generated), super-order, order, sub-order, infra-order
Family	OID (auto-linked), FID (auto-generated), super-family, family, sub-family, infra-family
Genus	FID (auto-linked), GID (auto-generated), tribe, sub-tribe, genus, section, sub-section, series, sub-series
Species	GID (auto-linked), SID (auto-generated) species, sub-species, forma, sub-forma, clone, cultivar, breed
Sci_loc	SID (auto-linked), LOCALITY_ID (auto-generated), Locality Name, geo-coordinates, observation date, observation period, altitude/depth, remarks,
Images	SCID (auto-linked), IMGID (auto-generated), Image Name, Image path, media type, media nature, contributor ID, date of contribution, copyright status, keywords, description
DSN	DSNNO, source title, authors, author affiliations, year of publication, journal name, volume no., pages, publisher name, ISSN/ISBN No., keywords, abstract, status of availability, accession no in information center/library, obtained from where, completion status
Secondary tables	
Users	user id, name of data manager / taxon editor, password, and user privileges
Country	Country_ID (auto-generated), country name, geo-coordinates
State	Country_ID (auto-linked), Sate_ID (auto-generated), state or province name, geo-coordinates
Distrisct	State_ID (auto-linked), district_ID (auto-generated), district or county name, geo-coordinates
Locality	District_ID (auto-linked), Locality_ID (auto-generated), name of the locality or tehsil, geo-coordinates
Waterbody	Waterbody_id (auto-generated), waterbody name, geo-coordinates
Contributor	CONTRI_ID (auto-generated), name of contributor, affiliation, address, telephone, fax, email, url
DSN_softcopy	DSNNO, File Name, softcopy

Table 2.1: Tables in IndFuana database.

Column Name	Data Type	Nullable	Primary Key
SCID	NUMBER(15,0)	No	
SPID	NUMBER(15,0)	Yes	
SCINAME	VARCHAR2(100 Bytes)	Yes	
AUTHOR	VARCHAR2(100 Bytes)	Yes	
YEAR	VARCHAR2(15 Bytes)	Yes	
STATUS	VARCHAR2(100 Bytes)	Yes	
DSNNO	VARCHAR2(1000 Bytes)	Yes	
ID	NUMBER(15,0)	Yes	
DELETED	VARCHAR2(1 Bytes)	Yes	
EDITEDBY	VARCHAR2(100 Bytes)	Yes	
EDITEDDT	VARCHAR2(25 Bytes)	Yes	
SCRUTINIZEDBY	VARCHAR2(100 Bytes)	Yes	
SCRUTINIZEDDT	VARCHAR2(50 Bytes)	Yes	
THREAT	VARCHAR2(100 Bytes)	Yes	
ENTEREDBY	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT	VARCHAR2(50 Bytes)	Yes	
ENTEREDDT2	DATE	Yes	
EDITEDDT2	DATE	Yes	
INVASIVE	VARCHAR2(100 Bytes)	Yes	
DSNPAGES	VARCHAR2(1000 Bytes)	Yes	
ORGANISM_TYPE	VARCHAR2(100 Bytes)	Yes	

Table 2.2: Structure of "sciname" table.

Column Name	Data Type	Nullable	Primary Key
SCID	NUMBER(15,0)	Yes	
SYNONYMS	VARCHAR2(100 Bytes)	Yes	
AUTHORITY	VARCHAR2(100 Bytes)	Yes	
YEAR	VARCHAR2(15 Bytes)	Yes	
CAUSE	VARCHAR2(20 Bytes)	Yes	
REMARKS	VARCHAR2(1000 Bytes)	Yes	
DSNNO	VARCHAR2(100 Bytes)	Yes	
ID	NUMBER(15,0)	Yes	
SNID	NUMBER(15,0)	Yes	
DELETED	VARCHAR2(1 Bytes)	Yes	
EDITEDBY	VARCHAR2(100 Bytes)	Yes	
EDITEDDT	VARCHAR2(25 Bytes)	Yes	
ENTEREDBY	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT	VARCHAR2(50 Bytes)	Yes	
ENTEREDDT2	DATE	Yes	
EDITEDDT2	DATE	Yes	
DSNPAGES	VARCHAR2(1000 Bytes)	Yes	

Table 2.3: Structure of "synonym" table.

Column Name	Data Type	Nullable	Primary Key
SCID	NUMBER(15,0)	No	1
COMMONNAME	VARCHAR2(100 Bytes)	Yes	
LANGUAGE	VARCHAR2(50 Bytes)	Yes	
REGION	VARCHAR2(50 Bytes)	Yes	
SEX	VARCHAR2(6 Bytes)	Yes	
LIFESTAGE	VARCHAR2(10 Bytes)	Yes	
REMARKS	VARCHAR2(1000 Bytes)	Yes	
DSNNO	VARCHAR2(100 Bytes)	Yes	
ID	NUMBER(15,0)	Yes	
CNID	NUMBER(15,0)	Yes	
DELETED	VARCHAR2(1 Bytes)	Yes	
EDITEDBY	VARCHAR2(100 Bytes)	Yes	
EDITEDDT	VARCHAR2(25 Bytes)	Yes	
ENTEREDBY	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT	VARCHAR2(50 Bytes)	Yes	
ENTEREDDT2	DATE	Yes	
EDITEDDT2	DATE	Yes	
DSNPAGES	VARCHAR2(1000 Bytes)	Yes	

Table 2.4: Structure of "commonname" table

Column Name	Data Type	Nullable	Primary Key
KID	NUMBER(2,0)	No	1
KINGDOM	VARCHAR2(100 Bytes)	Yes	
SUBKINGDOM	VARCHAR2(100 Bytes)	Yes	

Table 2.5: Structure of "kingdom" table

Column Name	Data Type	Nullable	Primary Key
DID	NUMBER(15,0)	No	1
KID	NUMBER(2,0)	Yes	
DIVISION	VARCHAR2(100 Bytes)	Yes	
SUBDIVISION	VARCHAR2(100 Bytes)	Yes	

Table 2.6: Structure of 'division" table

Column Name	Data Type	Nullable	Primary Key
CID	NUMBER(15,0)	No	1
DID	NUMBER(15,0)	Yes	
SUPERCLASS	VARCHAR2(100 Bytes)	Yes	
CLASS	VARCHAR2(100 Bytes)	Yes	
SUBCLASS	VARCHAR2(100 Bytes)	Yes	
INFRACLASS	VARCHAR2(100 Bytes)	Yes	

Table 2.7: Structure of "class" table

Column Name	Data Type	Nullable	Primary Key
OID	NUMBER(15,0)	No	1
CID	NUMBER(15,0)	Yes	
SUPERORDER	VARCHAR2(100 Bytes)	Yes	
ORDERS	VARCHAR2(100 Bytes)	Yes	
SUBORDER	VARCHAR2(100 Bytes)	Yes	
INFRAORDER	VARCHAR2(100 Bytes)	Yes	

Table 2.8: Structure of "orders" table

Column Name	Data Type	Nullable	Primary Key
FID	NUMBER(15,0)	No	1
OID	NUMBER(15,0)	Yes	
SUPERFAMILY	VARCHAR2(100 Bytes)	Yes	
FAMILY	VARCHAR2(100 Bytes)	No	
SUBFAMILY	VARCHAR2(100 Bytes)	Yes	
INFRAFAMILY	VARCHAR2(100 Bytes)	Yes	

Table 2.9: Structure of "family" table

Column Name	Data Type	Nullable	Primary Key
GID	NUMBER(15,0)	No	1
FID	NUMBER(15,0)	Yes	
TRIBE	VARCHAR2(100 Bytes)	Yes	
SUBTRIBE	VARCHAR2(100 Bytes)	Yes	
GENUS	VARCHAR2(100 Bytes)	No	
SUBGENUS	VARCHAR2(100 Bytes)	Yes	
SECTION	VARCHAR2(100 Bytes)	Yes	
SUBSECTION	VARCHAR2(100 Bytes)	Yes	
SERIES	VARCHAR2(100 Bytes)	Yes	
SUBSERIES	VARCHAR2(100 Bytes)	Yes	

Table 2.10: Structure of "genus" table.

Column Name	Data Type	Nullable	Primary Key
SID	NUMBER(15,0)	No	1
GID	NUMBER(15,0)	Yes	
SPECIES	VARCHAR2(100 Bytes)	No	
SUBSPECIES	VARCHAR2(100 Bytes)	Yes	
VARIETY	VARCHAR2(100 Bytes)	Yes	
SUBVARIETY	VARCHAR2(100 Bytes)	Yes	
FORMA	VARCHAR2(100 Bytes)	Yes	
SUBFORMA	VARCHAR2(100 Bytes)	Yes	
CLONE	VARCHAR2(100 Bytes)	Yes	
CULTIVOR	VARCHAR2(100 Bytes)	Yes	
BREED	VARCHAR2(100 Bytes)	Yes	

Table 2.11: Structure of "species" table

Column Name	Data Type	Nullable	Primary Key
ID	NUMBER(15,0)	No	
SCID	NUMBER(15,0)	Yes	
LOCALITY_ID	NUMBER(15,0)	Yes	
DSNNO	VARCHAR2(100 Bytes)	Yes	
USER_ID	VARCHAR2(15 Bytes)	Yes	
DELETED	VARCHAR2(1 Bytes)	Yes	
EDITEDBY	VARCHAR2(100 Bytes)	Yes	
EDITEDDT	VARCHAR2(25 Bytes)	Yes	
OBDT	VARCHAR2(25 Bytes)	Yes	
OBPERIOD	VARCHAR2(100 Bytes)	Yes	
ENTEREDBY	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT	VARCHAR2(50 Bytes)	Yes	
MSL	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT2	DATE	Yes	
EDITEDDT2	DATE	Yes	
DSNPAGES	VARCHAR2(1000 Bytes)	Yes	

Table 2.12: Structure of "sci_loc" table

Column Name	Data Type	Nullable	Primary Key
IMGID	NUMBER(15,0)	No	
SCID	NUMBER(15,0)	No	
IMGNAME	VARCHAR2(1000 Bytes)	No	
PATH	VARCHAR2(1000 Bytes)	Yes	
MEDIA_TYPE	VARCHAR2(1000 Bytes)	Yes	
MEDIA_NATURE	VARCHAR2(1000 Bytes)	Yes	
AID	VARCHAR2(1000 Bytes)	Yes	
CONTRI_ID	VARCHAR2(1000 Bytes)	Yes	
DATE_OF_CONTRIBUTION	VARCHAR2(1000 Bytes)	Yes	
COPYRIGHT_STATUS	VARCHAR2(1000 Bytes)	Yes	
KEYWORDS	VARCHAR2(1000 Bytes)	Yes	
DESCRIPTION	VARCHAR2(1000 Bytes)	Yes	
SCINAME	VARCHAR2(1000 Bytes)	Yes	
DELETED	VARCHAR2(1 Bytes)	Yes	
ENTEREDBY	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT	VARCHAR2(50 Bytes)	Yes	
ID	NUMBER(15,0)	Yes	

Table 2.13: Structure of "images" table

Column Name	Data Type	Nullable	Primary Key
DSNNO	NUMBER(12,0)	No	1
TITLE	VARCHAR2(1000 Bytes)	Yes	
AUTHOR	VARCHAR2(1000 Bytes)	Yes	
AFFILATION	VARCHAR2(1000 Bytes)	Yes	
YEAR	VARCHAR2(1000 Bytes)	Yes	
VOLUME	VARCHAR2(1000 Bytes)	Yes	
PAGES	VARCHAR2(1000 Bytes)	Yes	
PUBLISHER	VARCHAR2(1000 Bytes)	Yes	
ISSNNO	VARCHAR2(1000 Bytes)	Yes	
KEYWORDS	VARCHAR2(1000 Bytes)	Yes	
ABSTRACT	VARCHAR2(1000 Bytes)	Yes	
AVAILIBILITY	VARCHAR2(1000 Bytes)	Yes	
JOURNAL	VARCHAR2(1000 Bytes)	Yes	
ID	NUMBER(15,0)	Yes	
DELETED	VARCHAR2(1 Bytes)	Yes	
EDITEDBY	VARCHAR2(100 Bytes)	Yes	
EDITEDDT	VARCHAR2(25 Bytes)	Yes	
ENTEREDBY	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT	VARCHAR2(50 Bytes)	Yes	
ENTEREDDT2	DATE	Yes	
EDITEDDT2	DATE	Yes	
OBTAINED	VARCHAR2(1000 Bytes)	Yes	
ACCESSION	VARCHAR2(1000 Bytes)	Yes	
COMPLETED	VARCHAR2(100 Bytes)	Yes	

Table 2.14: Structure of 'DSN" table.

Column Name	Data Type	Nullable	Primary Key
REC_NO	NUMBER(15,0)	No	
DSNNO	NUMBER(15,0)	Yes	
FILENAME	VARCHAR2(500 Bytes)	Yes	
ENTEREDBY	VARCHAR2(100 Bytes)	Yes	
ENTEREDDT	VARCHAR2(100 Bytes)	Yes	
EDITEDBY	VARCHAR2(100 Bytes)	Yes	
EDITEDDT	DATE	Yes	
DELETED	VARCHAR2(10 Bytes)	Yes	
DELETEDDT	DATE	Yes	
SOFTCOPY	BLOB	Yes	

Table 2.15: Table structure for "dsn_softcopy"

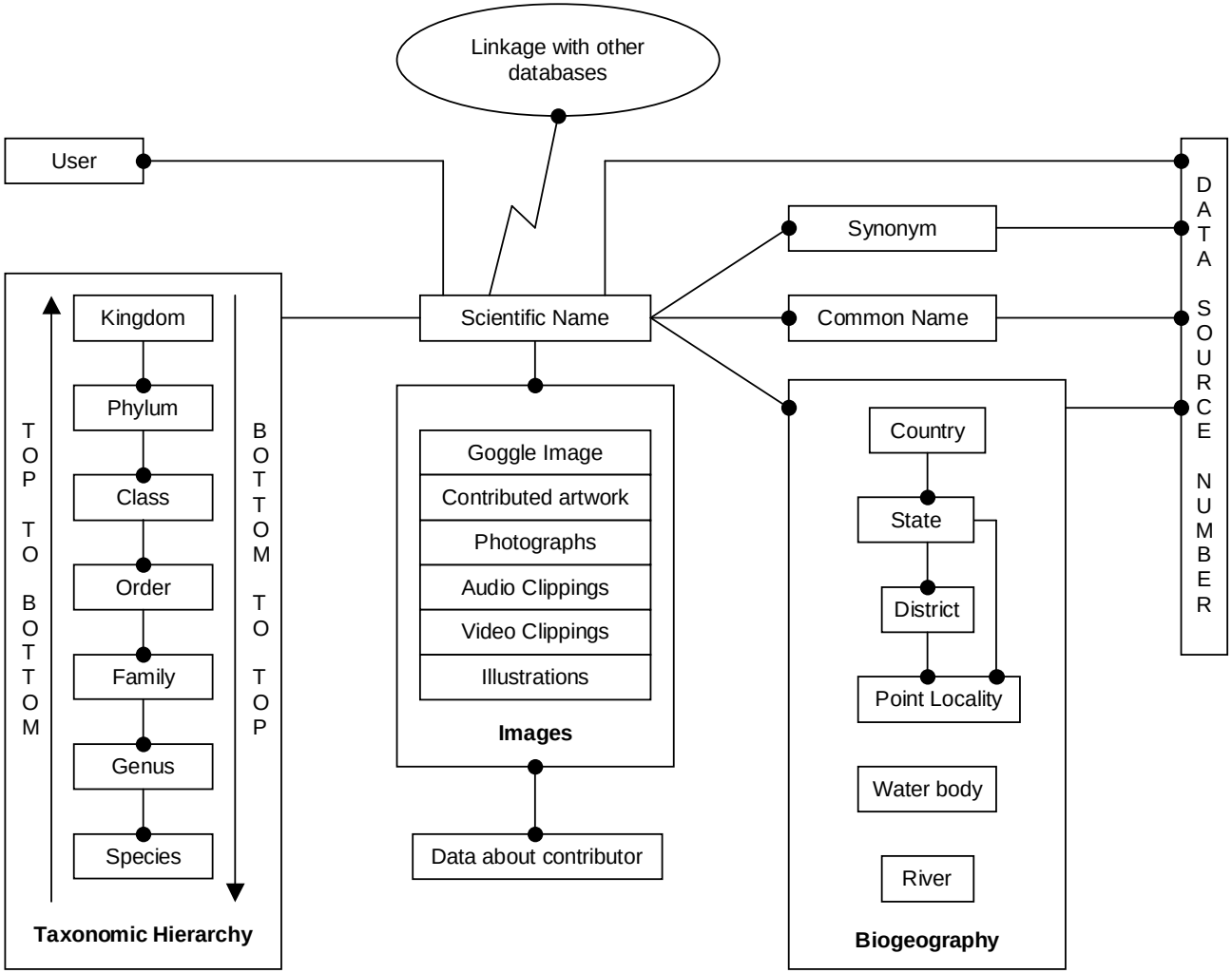


Figure 2.1: Architecture of IndFauna information system along with relationships between entities.

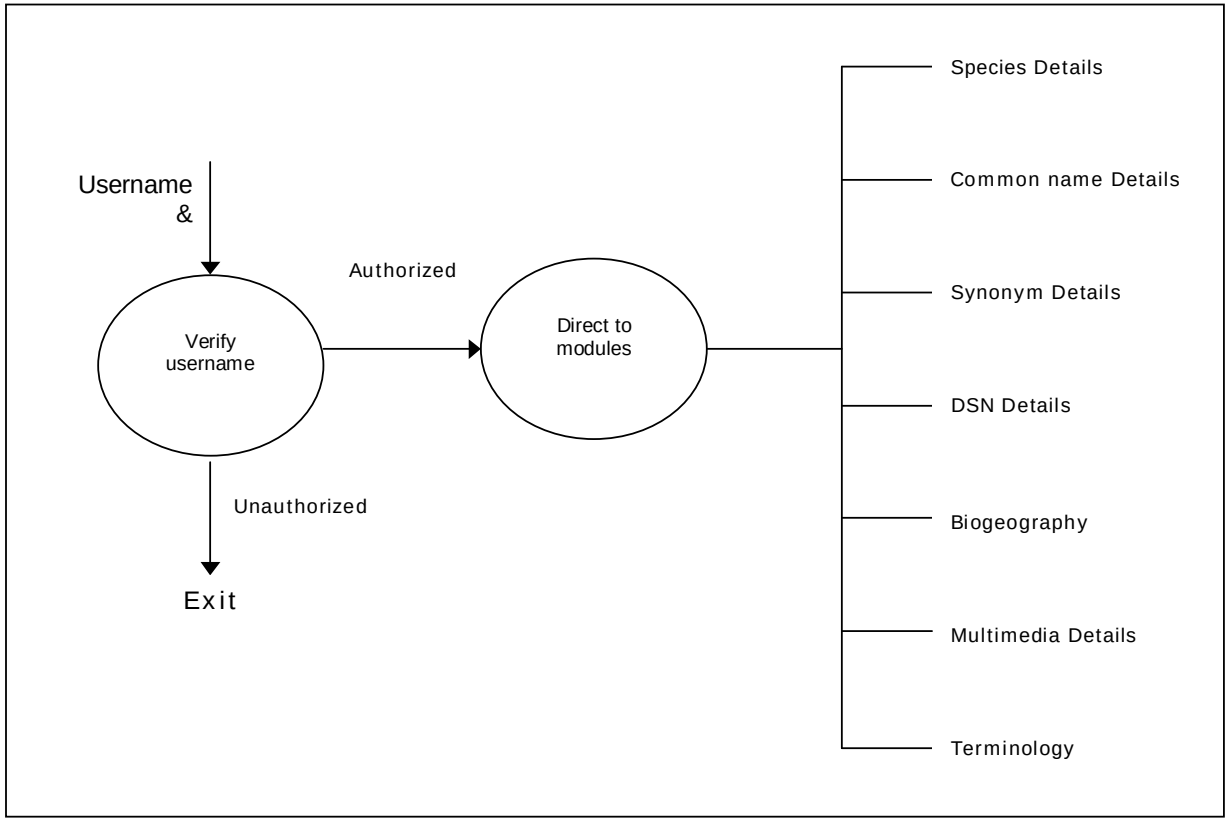


Figure 2.2 Data Flow Diagram (DFD) for IndFauna Data Management

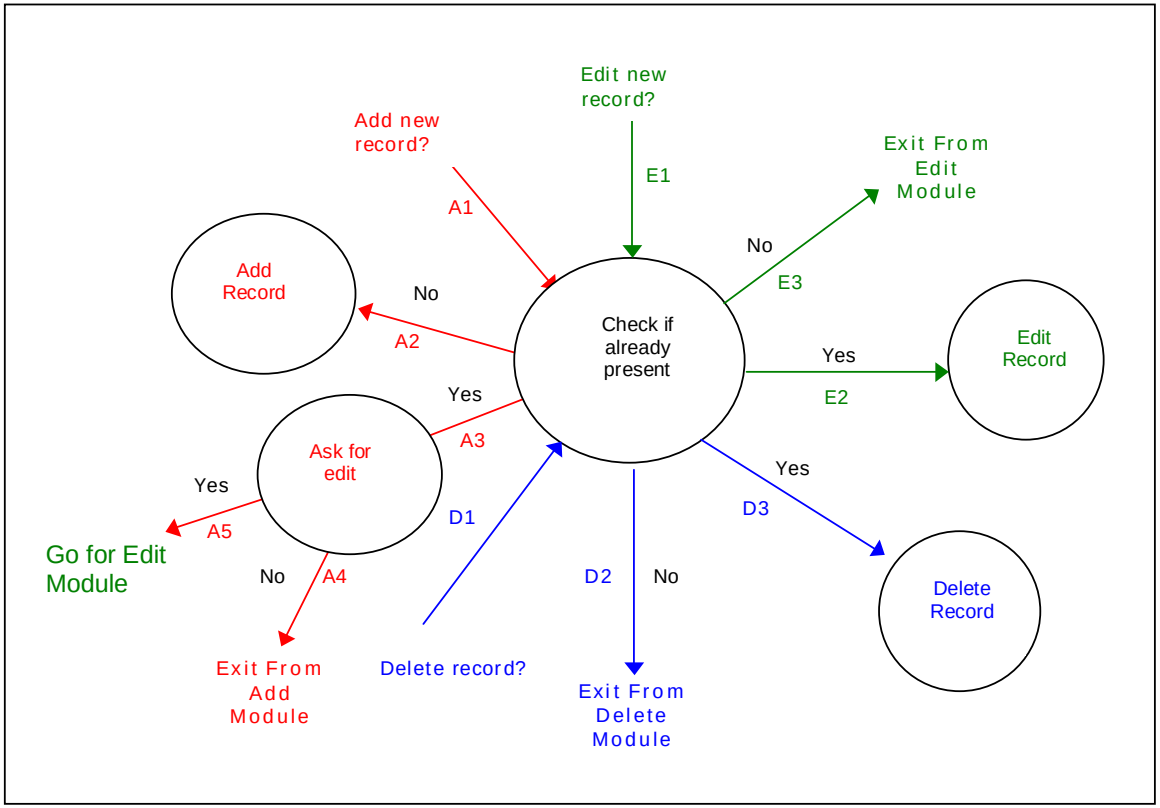


Figure 2.3: Data Flow Diagram for "ADD", "EDIT", and "DELETE" functions of Administrative modules.

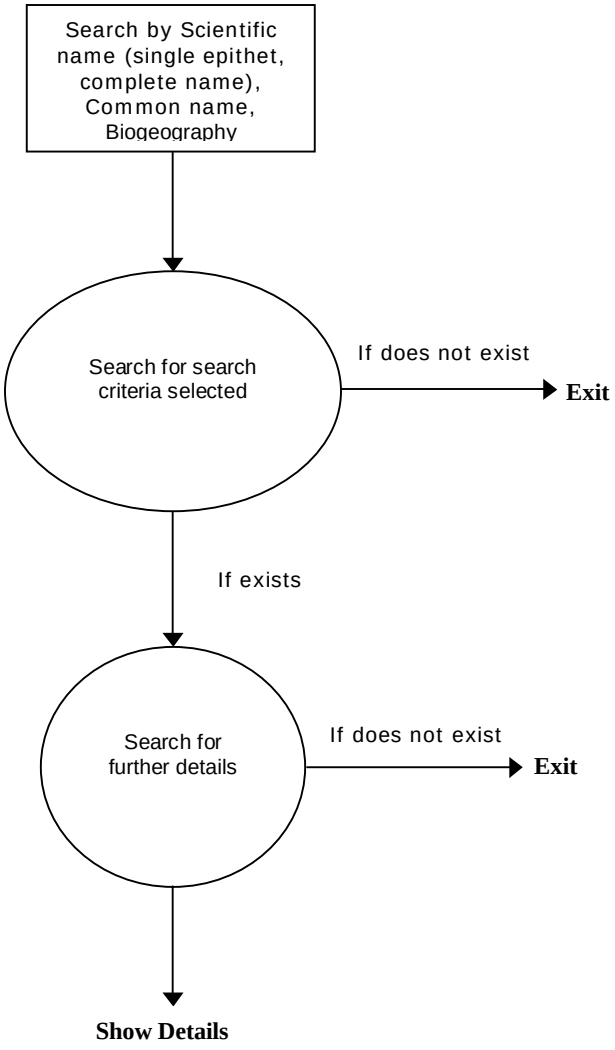


Figure 2.4: Data Flow Diagram for Search function of the Administrative modules.

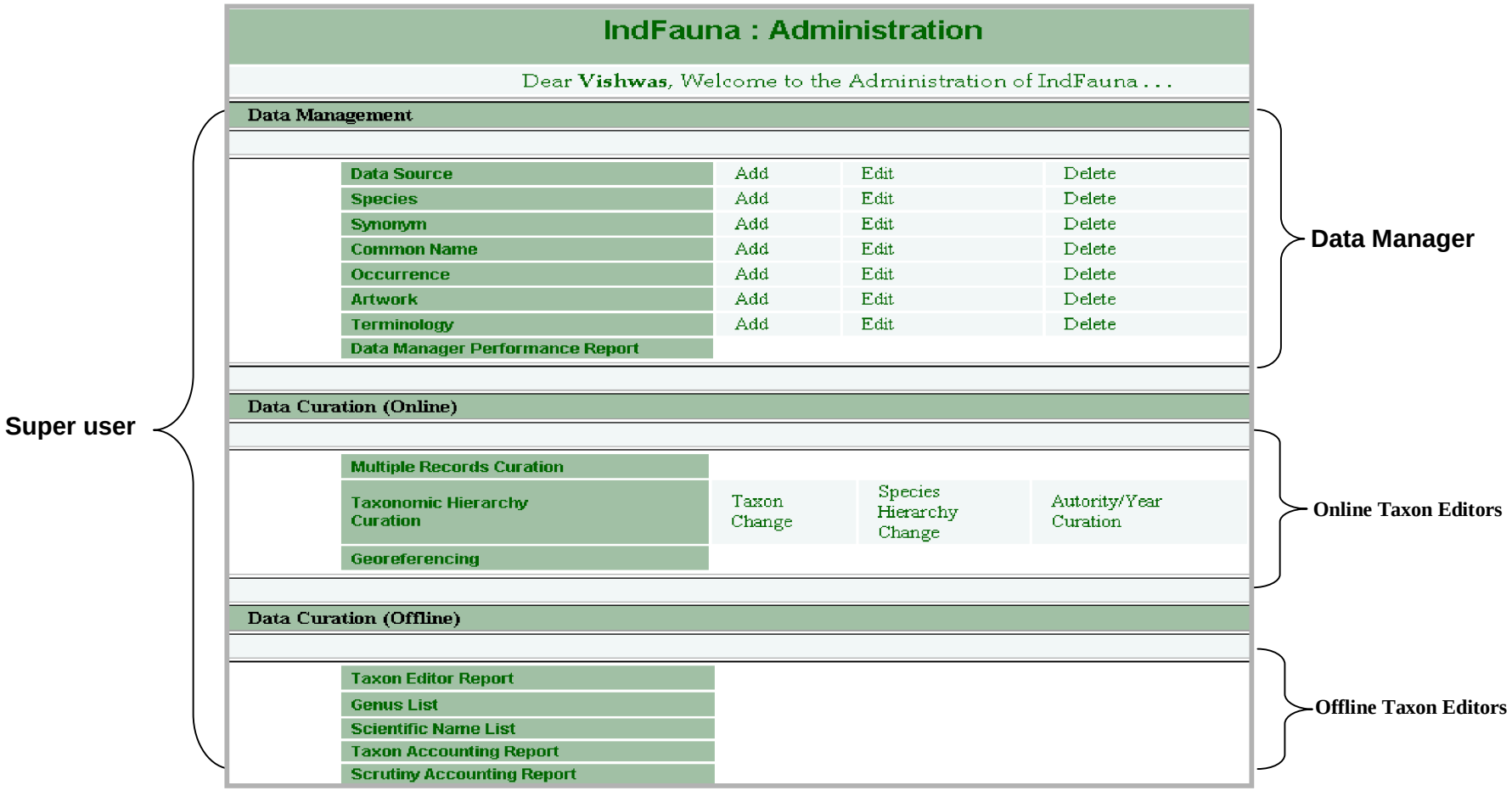


Figure 2.5: IndFuana Administration Module has four levels of privileges, (a) Data Managers can add, edit, and mark records for deletion, (b) Online Taxon Editors can curate records over web, (c) Offline Taxon Editors can generate reports for editing purposes, and (d) Super User has all of the above privileges including deleting of records.

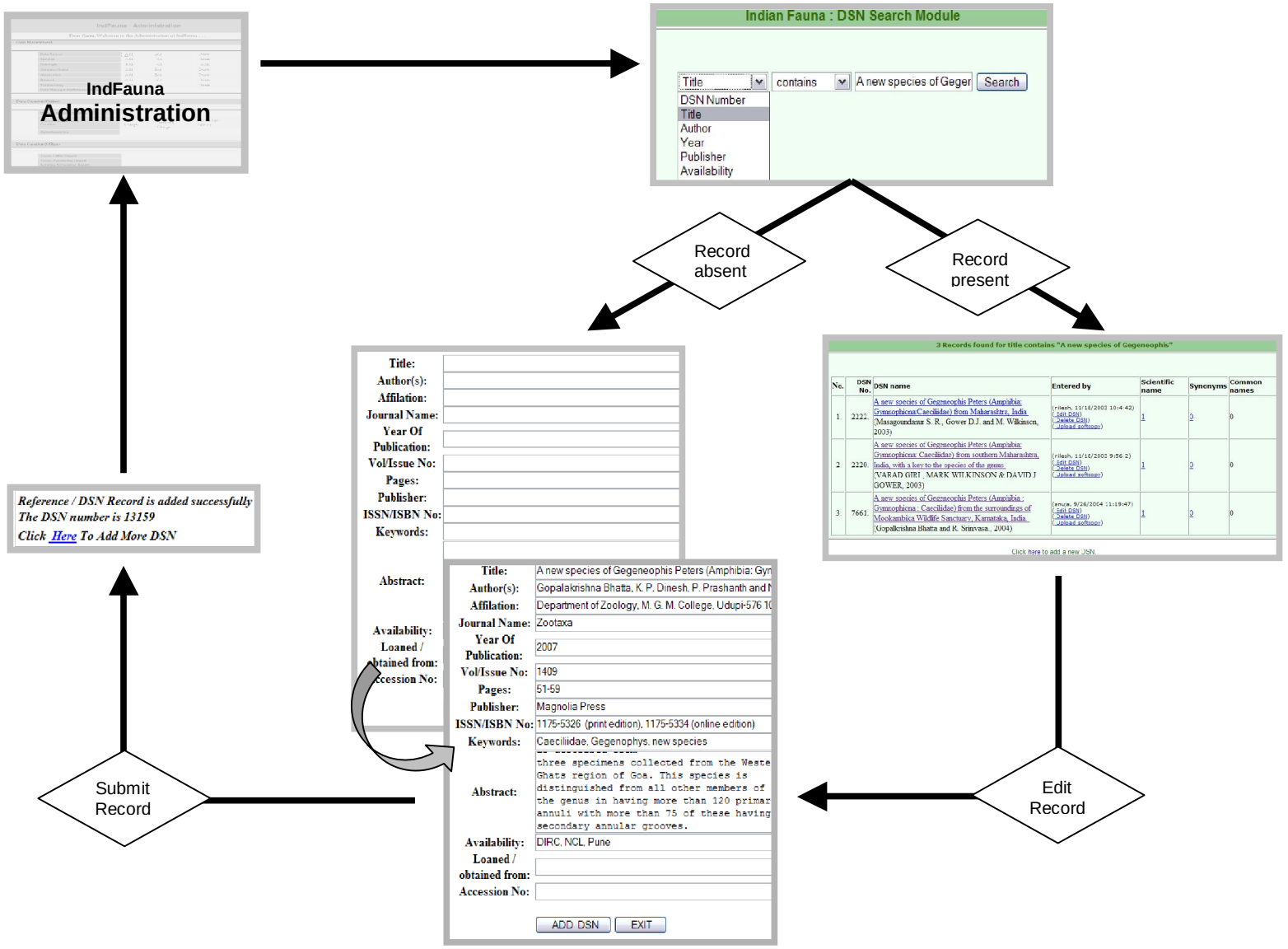


Figure 2.6: IndFuana Data Source module.

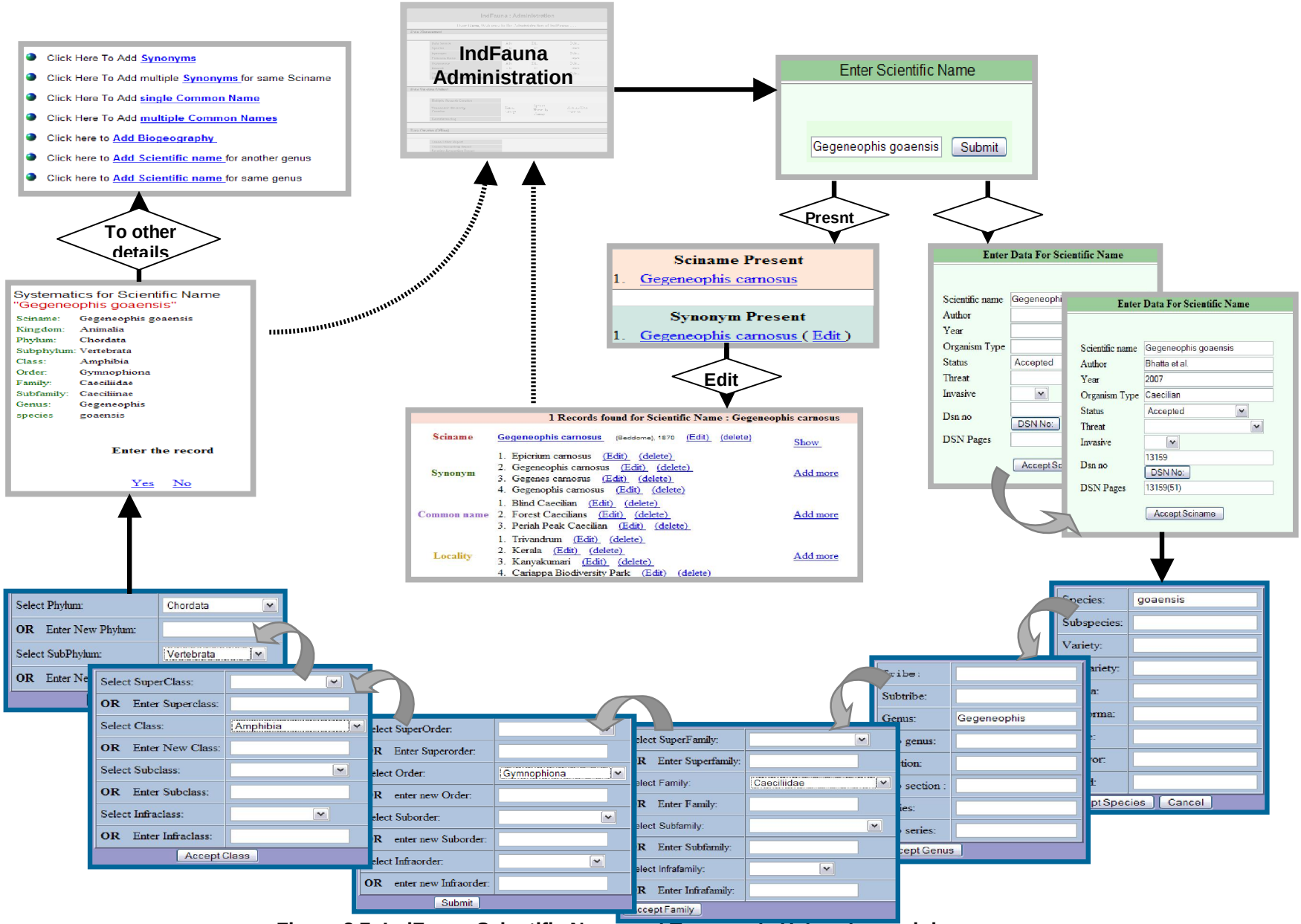


Figure 2.7: IndFauna Scientific Name and Taxonomic Heirarchy module.

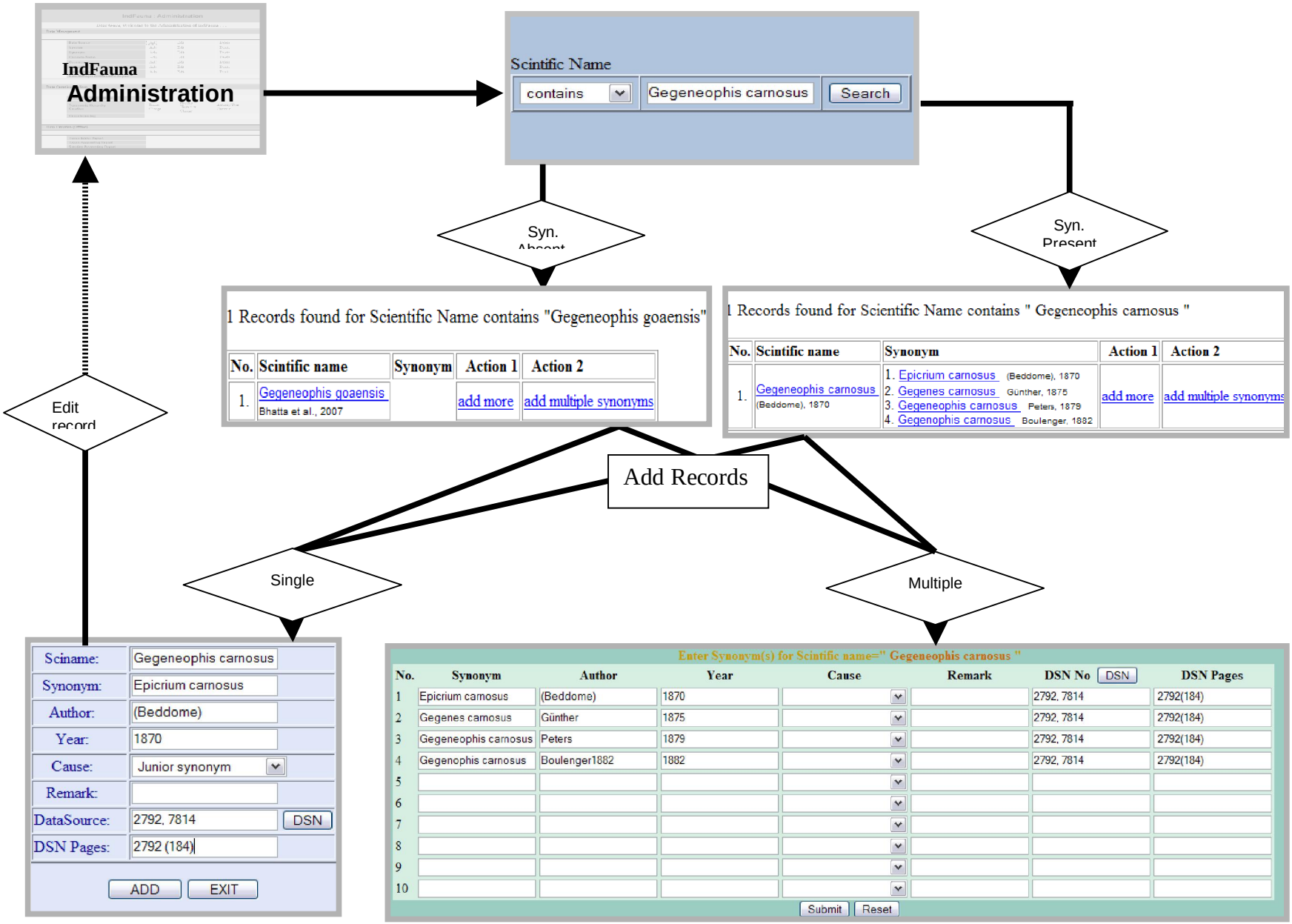


Figure 2.8: IndFauna Synonym module.

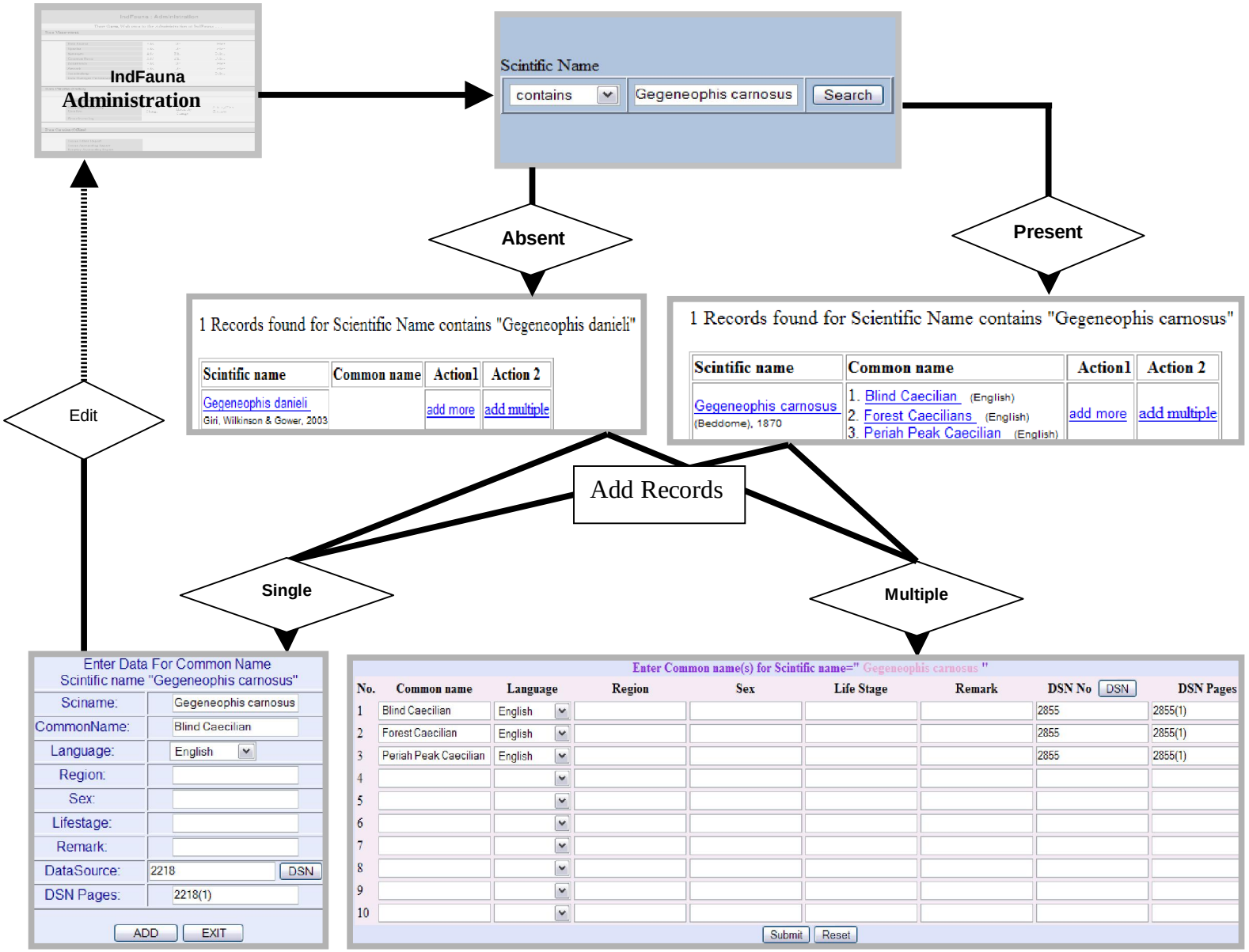


Figure 2.9: IndFauna Common Name module.

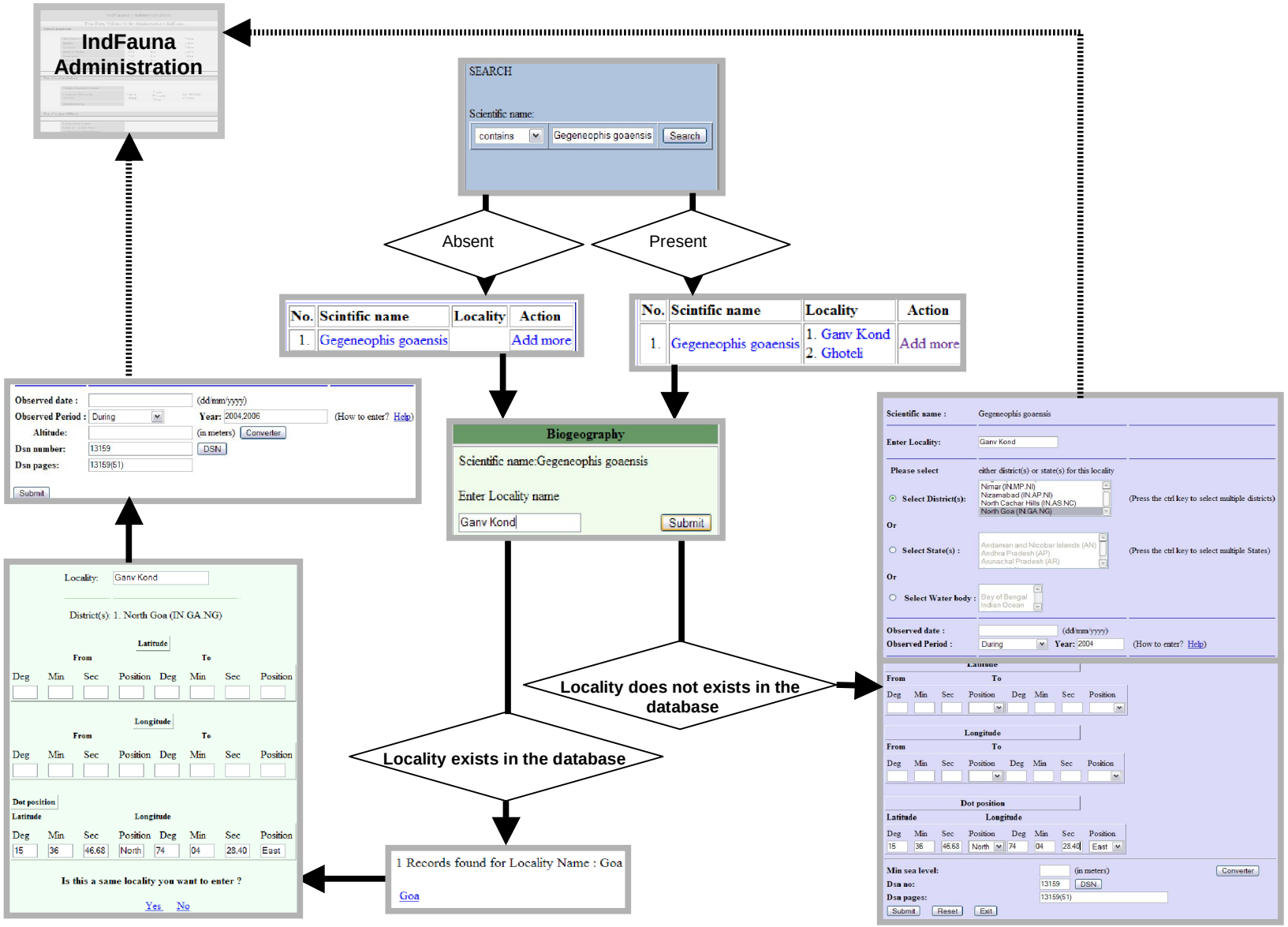


Figure 2.10: IndFauna Occurrence module.

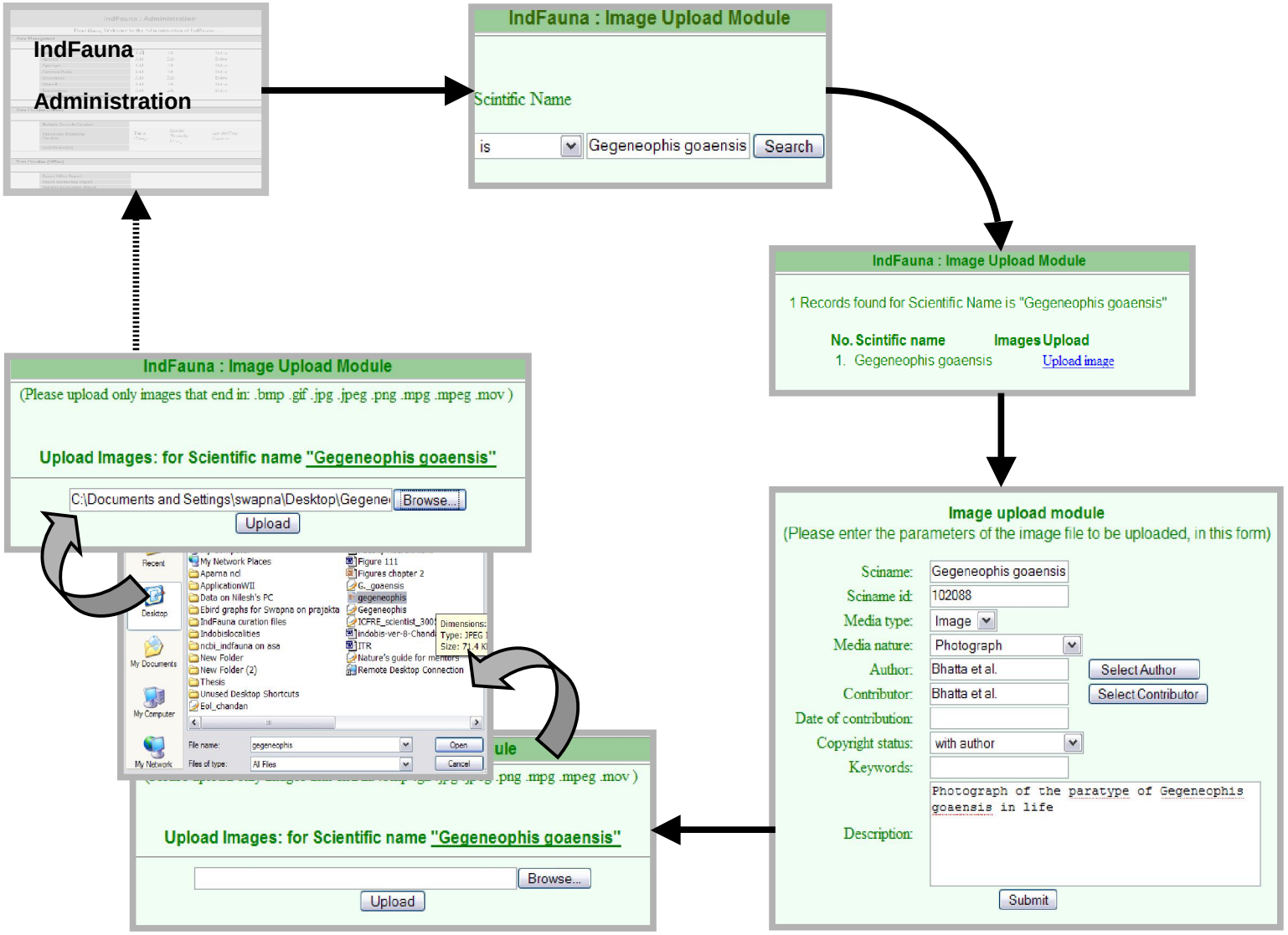


Figure 2.11 IndFuana ArtWork module.

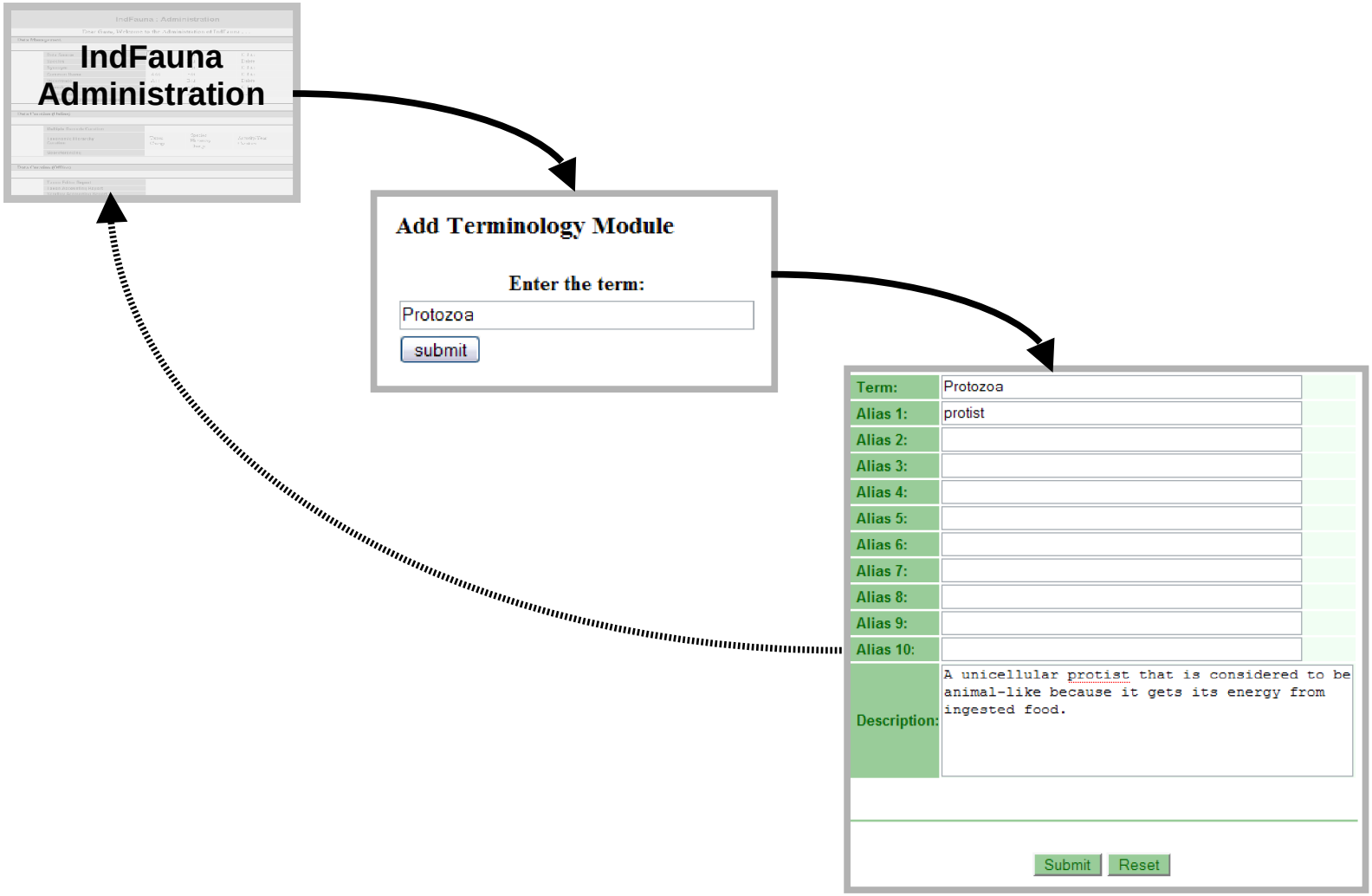


Figure 2.12 IndFauna Terminology module.

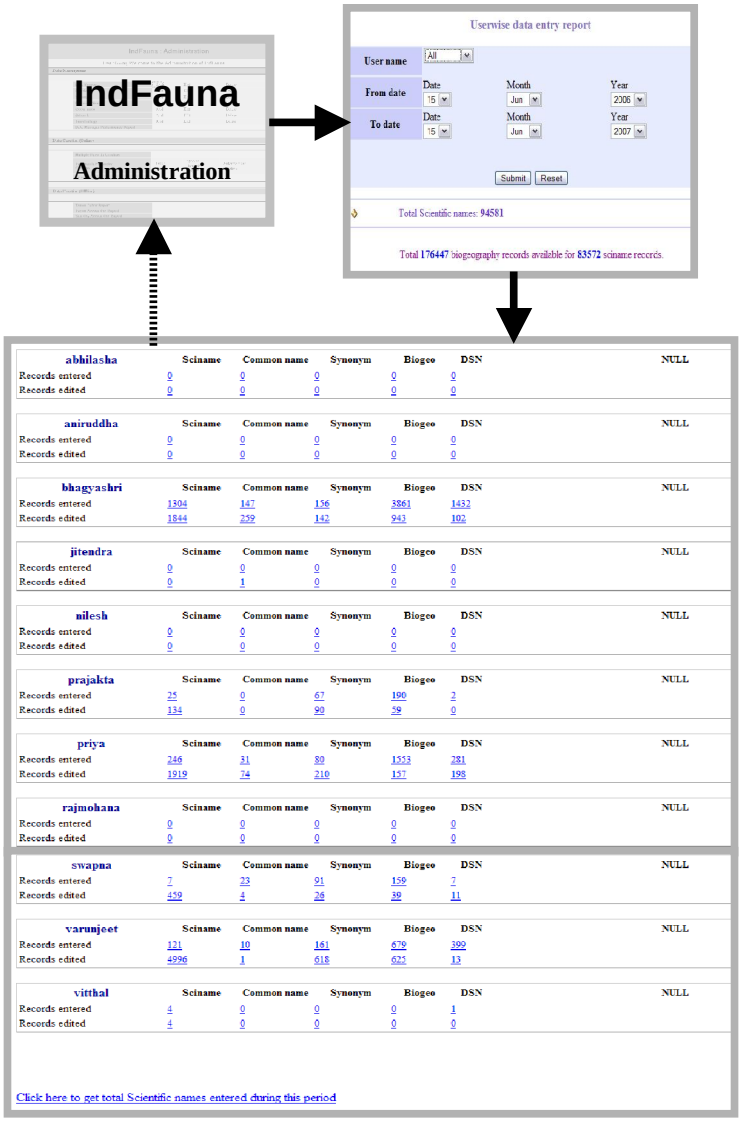


Figure 2.13: Data Managers Performance module.

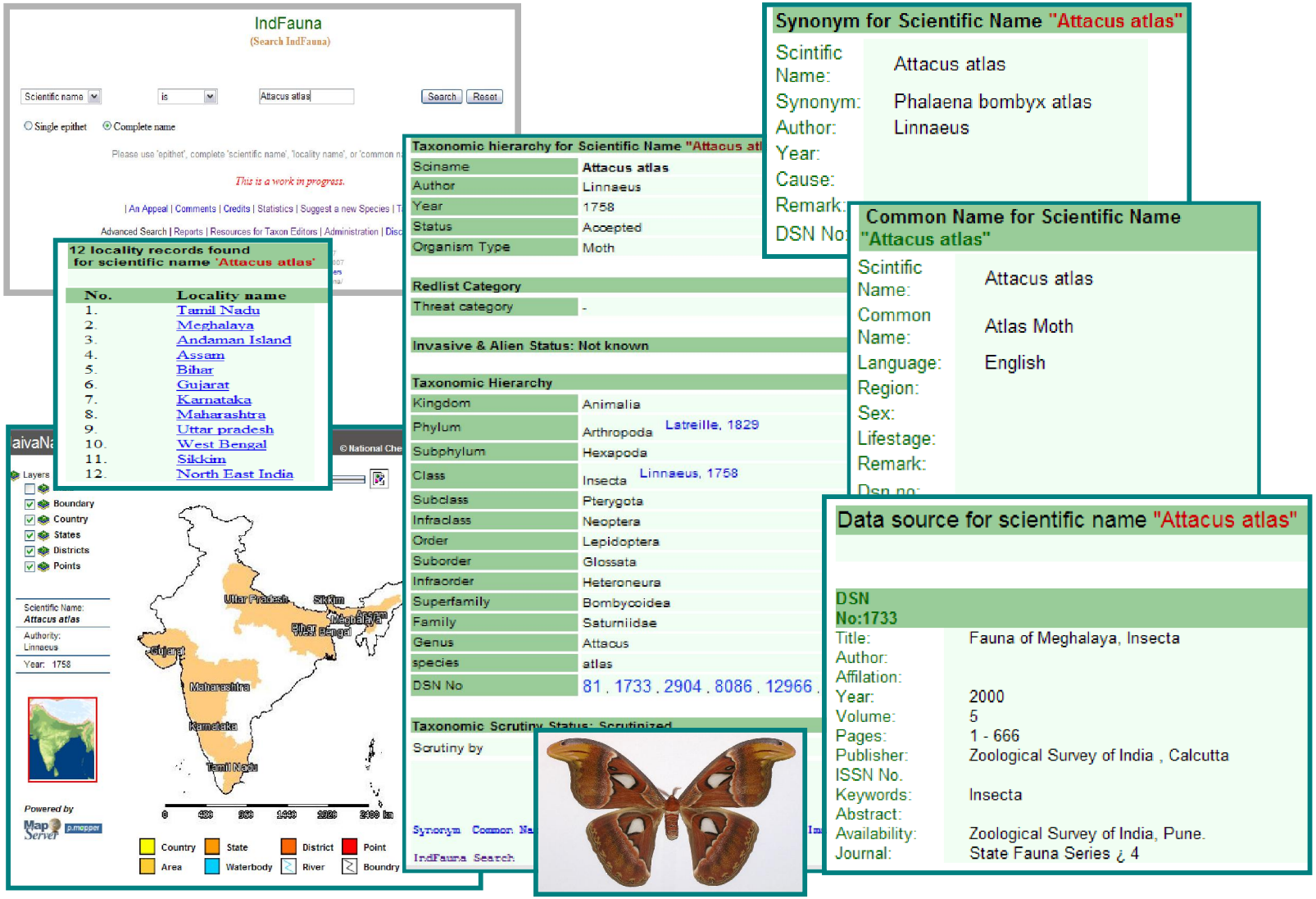
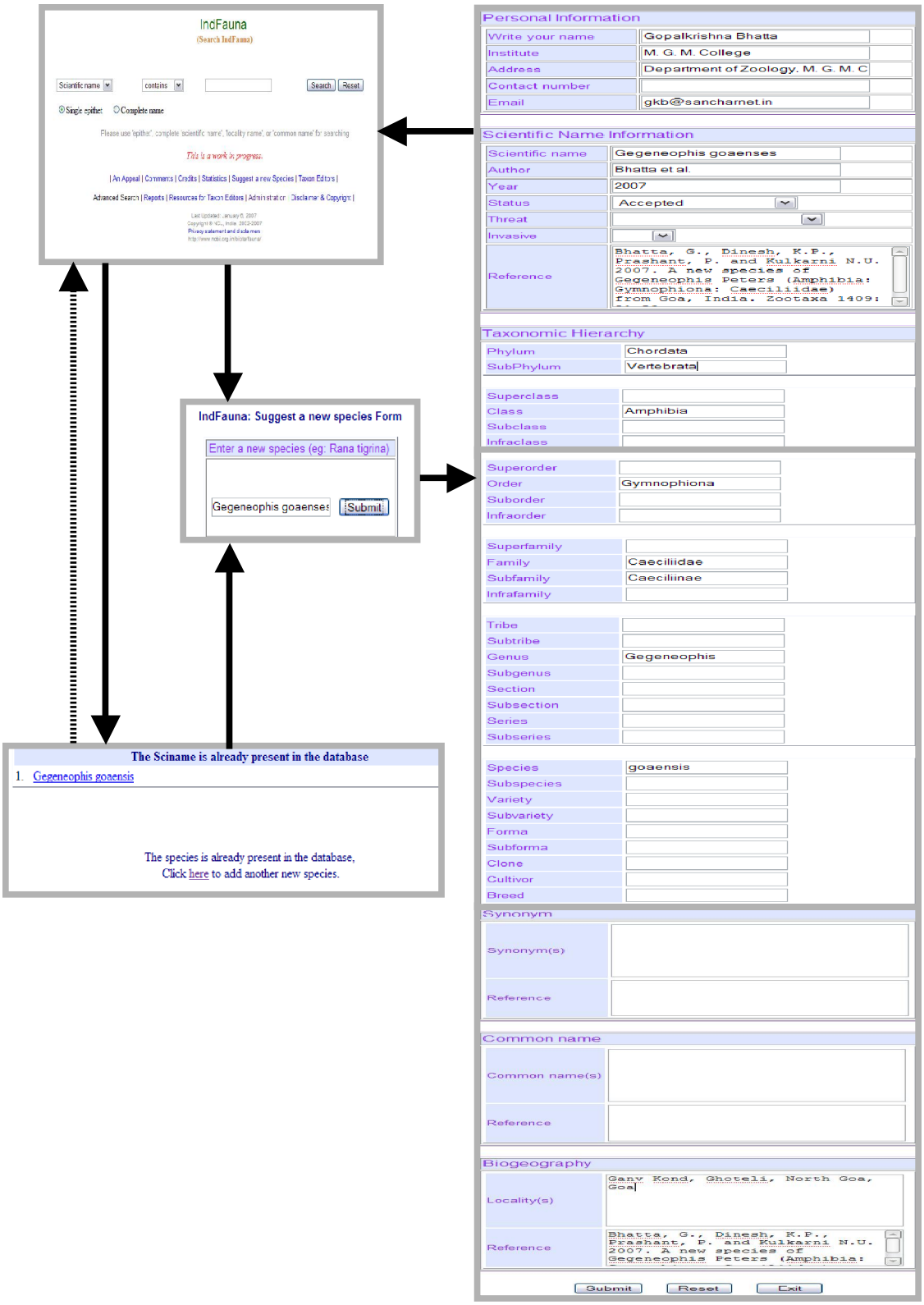


Figure 2.14: Output through IndFauna search module.



2.15: IndFauna: Suggest a New Species module

Chapter 3

IndFauna: Data Cleaning, Taxonomic Scrutiny and Lessons Learned

Chapter 3

IndFauna: Data Cleaning, Taxonomic Scrutiny and Lessons Learnt

3.1 Uses of Species and Occurrence Data

A key purpose of electronic catalogue of known life, such as IndFauna, is to provide data with cost effective method of querying and analyzing that data. This data provide not only present day information on the locations of these entities, but also historic information going back several hundred years (Chapman and Busby, 1994). The uses of species and occurrence data are wide and varied and encompass virtually every aspect of human endeavor – food, shelter and recreation; art and history, society, science and politics (Chapman, 2005a). The increased of data on species is opening up new and improved methods of dealing with various issues. Availability of species and occurrence data in on-line databases is improving science, reducing costs by providing for more efficient and effective biological survey, freeing up scientists to spend more time on research, and leading to a more rapid build-up of knowledge of our environments leading to its improved conservation and sustainable use. The ability to search databases all around the world for spatially referenced species occurrence data has opened up the information to a range of uses, many of which have previously not been possible (Chapman, 2005a).

From Fig. 3.1, it is evident that the uses of species occurrence data are endless. However, one of the key limiting factors is the quality, and authenticity of the data, which determine the confidence level of results of analysis. Thus, it is imperative that every bit of species and occurrence data is processed to ascertain highest degree of quality. Therefore, data quality principles have become a core business practice in fields such as business (SEC, 2002), medicine (Gad and Taulbee, 1996), GIS (Zhang and Goodchild, 2002), remote sensing (Lunetta and Lyon, 2004), and many others over recent times, but are only now becoming universally accepted by biodiversity community, especially taxonomic community (Chapman, 2005b).

3.2 Data Cleaning: Needs and Principles

As discussed in previous section, data quality principles are not universally accepted within the biodiversity community, especially with the taxonomic community. However, with rapid increase in the availability and exchange of taxonomic and species-occurrence data has now made the consideration of data

quality principles an important agenda items, as the users of the data (Fig. 3.1) begin to require more and more detail on its quality.

There are many data quality principles that apply when dealing with species data and especially with the spatial aspects of those data. These principles are involved at all stages of the data management process (Fig. 3.2). A loss of data quality at any one of those stages reduces the applicability and uses to which the data can be adequately put. These stages includes, (a) Data capture and recording at the time of gathering, (b) Data manipulation prior to digitization, (c) Digitization of data, (d) Documentation of data (metadata creation), (e) Data storage and archiving, (f) Data presentation and dissemination, and (g) Data use.

All these have an input into the final quality or "fitness for use" of the data and all apply to all aspects of the data – the taxonomic or nomenclatural portion of the data – the "what", the spatial portion – the "where" and other data such as the "who" and the "when"(Berendsohn, 1997).

Experience has shown that treating data as a long-term asset and managing it within a coordinated framework produces considerable savings and ongoing value (NLWRA, 2003). Thus, principles of data quality need to be applied at all stages of the data management process. There are two keys to improvement of data quality – prevention and correction. Although considerable effort can be and should be given to the prevention of error, the fact remains that errors in large datasets will continue to exist (Maletic and Marcus, 2000) and data validation and correction can not be ignored. However, it is advisable to undertake data cleaning early in the information management chain as cost of error correction increases as one progress along the chain.

As increased data becoming available from many sources, users want to know which source they rely on, and which follow documented quality control procedures. Reputation can alone be the deciding factor on where a user may source their data. Dalcin (2004) suggest that there is a need to develop a certification and accreditation process that informs users of organizations that confirm to minimum data quality documentation standards and procedures. A quality certification of taxonomic data could involve there aspects: primary data sources (the raw material), the information chain (the process), and the database (the product).

3.3 IndFauna: Data Cleaning

As discussed earlier major objective of IndFauna, electronic catalogue of known Indian fauna is to collate and disseminate species and occurrence data about known Indian fauna. Thus, it primarily focused on two types of data collation and dissemination, viz. (a) Taxonomic and Nomenclature Data, and (b) Spatial Data. There are many methods and techniques that can aid in the cleaning of errors in species-occurrence databases (Chapman, 2005c). In this chapter, I focus on data cleaning of taxonomic and nomenclature data with specific reference to taxonomic scrutiny process followed during development of IndFauna, wherein Chapter 4, deals with cleaning of spatial data.

3.3.1. IndFauna Data Cleaning: Need and Approach

There are range of methods used for species (taxonomic or nomenclature) data cleaning. These include methods that have been operating in traditional taxonomic and occurrence data delivery to automated methods that are still largely untested. However, one of the major differences between IndFauna and many other species-occurrence databases is the source of data (raw material). While majority of the species-occurrence databases use collection repositories (specimens), and field based observation/survey data as raw material, IndFauna has used secondary sources of data (peer reviewed papers, monographs, and taxonomic databases) as raw material. Thus, the approach for ensuring data quality of IndFauna, was combination of traditional methods and modern automated methods (Fig. 3.3).

As depicted in Fig. 3.3, secondary data sources such as taxonomic publications of all types and taxonomic database, and species checklists formed the source of data. Thus some degree of confidence has to be posed into these sources, as they are published following some rigorous traditional scrutiny and review process. However, not all publications follow the same level and types of scrutiny and review, and thus IndFauna was posed with this uneven quality of data as its source. Further, the entire process of IndFauna was transparent and dynamic in the sense that users could actually track the development beginning with entry of first byte of data. Thus, it was necessary that errors are irradiated in shortest possible timeframe.

To overcome this, data cleaning were mandated at two different steps in database development (Fig. 3.3).

- **Level I** - First level of data cleaning begins when the DM collates the data. It is subjected to quality checks by Quality Controllers (QC). These are often individuals with experience in handling species-occurrence data and sound taxonomic expertise. Their suggestions are incorporated back into the database by DM. This is being referred as “First Checks” in subsequent sections of this thesis.
- **Level II** - Second level of data cleaning is carried out buy group of “Taxon Editors” either online or offline. Taxon Editors are domain experts with long standing taxonomic expertise on specific group of organisms. Their suggestions are communicated back to the QC, who with the help of DMs incorporates the changes. Subsequent sections elaborate these processes with details and suitable examples. This is being referred as “Taxonomic Scrutiny” in the subsequent sections of the thesis.

3.3.2 IndFauna Data Cleaning: First Checks

As noted in section 3.3.1(A), First checks were performed by the Quality Controllers (QC), who are individuals with experience in handling species-occurrence data and sound taxonomic expertise. First checks were intended to ensure accuracy, completeness, and consistency of the data entered by the DMs. QCs performed checks to detect amongst others typographical mistakes, such as spelling, case sensitivity, spaces within scientific names, absence or presence of special characters, and abbreviated status flags, compliance to naming convention as prescribed by ICZN (International Union of Zoological Nomenclature) Code, compliance with taxon level suffix convention. In addition they also performed checks to determine if DMs have collated all possible data from a given data source or part of it has been omitted or left to collate inadvertently. First checks were given also to ascertain if data has been entered by DMs unaltered with respect to the data source.

First checks were also performed to ascertain that taxonomic hierarchy of the species is entered correctly and that any nomenclatural rank (taxon level) is not missing. QCs also checked if same name is not entered as accepted or valid scientific name, and invalid scientific names or synonyms. They also performed quick scientific names checks against authoritative databases or authority files to ensure that nomenclaturally invalid names are not entered in the database. Another motive of such cross check was to ensure that all names associated with valid or accepted names

are entered. Another measure to determine consistency of names data was to verify that for each known organism there is only one accepted names, and may be multiple non-accepted names.

Since, every data byte has been collated from one or the other published or public domain data source, critical checks were performed to ensure that every byte is linked with appropriate DSN (Data Source Number). Similarly, QCs ensured that Data Source information is properly entered as per accepted style and format. Similarly, for each contributed ArtWork appropriate metadata has been submitted, and copyright expressions are clearly stated.

Quality Controller (QC) bought these errors to the notice of DMs, so that DMs attempt not to commit same errors in subsequent data management activities. The errors were corrected using suites of "Data Curation Modules". Each of these modules and their functions are described in subsequent sections.

3.3.3 IndFauna Data Cleaning: Taxonomic Scrutiny

This process of data cleaning was devised to ensure that IndFauna, disseminate up-to-date, current, authentic, and validated data. Since, secondary data (peer-reviewed publications, and public domain taxonomic literature, online and offline databases) constitute data source for IndFauna, to a great degree it could be considered as peer-reviewed database. However, owing to dynamic and changing nature of taxonomic opinions, it is necessary to conduct taxonomic scrutiny of collated data. Taxonomic Scrutiny was conducted by pool of "Taxon Editors". Taxon Editors are individuals with long standing and proven experience and expertise on specific group of organisms and who have volunteered to correct, authenticate, and validate the data collated for IndFauna. As on date IndFuana is being curated by over 100+ Taxon Editor, over 60% of which are overseas experts. While few "taxon editors" prefer to carry out online taxonomic scrutiny, majority of them preferred to conduct it offline, signifying the reluctance to use cyberspace to strengthen the taxonomic knowledge. In order to facilitate the process of taxonomic scrutiny, set of "Data Curation Modules" are developed. These data curation modules are described in next sections.

3.3.4 IndFauna: Data Curation Modules

In total there are 10 data curation modules, five each for online and offline "taxon editors" respectively (Fig. 3.4). Of the five data curation modules for online

taxon editors, one if for geo-referencing, this is discussed in Chapter 4. Rests of the nine modules are described in subsequent sections.

(A) Multiple Records Curation

This module is developed to implement correction in same field in multiple records. This could be carried out in four tables' viz., Sciname, Synonym, Commonname, and DSN. Once the desired table and field is selected, choose the records in which changes are needed by selecting the check box. Write the correct "syntax" to be replaced, and click the UPDATE. This would curate chosen field of desired records with the correct syntax (Fig. 3.5).

(B) Taxonomic Heirarchy Curation: Taxon Change

This module has been designed to incorporate changes in taxon level of particular species record. These changes could be implemented in major six taxon tables'viz., phylum, class, order, family, genus, and species. Select the table, followed by a specific species record for in which change has to be induced. Make appropriate changes in major taxon level or its sub-levels and UPDATE the change and it would get implemented in the taxonomic hierarchy of particular species record (Fig. 3.6).

(C) Taxonomic Hierarchy Change: Species Hierarchy Change

This module differ from previous module as it facilitate the taxonomic hierarchy change of a particular species record at one go. Upon choosing the scientific name for whose taxonomic hierarchy needs to be changed, content of all concerned tables'viz., sciname, kingdom, phylum, class, orders, family, genus, and species could be appropriately edited. When UPDATE is pressed; the changes gets incorporated in taxonomic hierarchy of that particular species (Fig. 3.7).

(D) Taxonomic Hierarchy Change - Author/Year Curation

This module has been developed to add, edit the authority and year of specific taxon level. As shown in Fig. 3.8, upon selecting the taxon level record (e.g Family Caeciliidae), its authority and year could be added or edited. This change gets effected to all the species records whose Family is Caeciliidae.

(E) Taxon Editor Report

This is one of the five modules to facilitate offline taxonomic scrutiny by "taxon editors" who are comfortable to scrutinize taxomonic or nomenclature data in traditional way similar to checklist scrutiny. This module facilitates the report for specific taxon level similar to checklist. As shown in Fig. 3.9 upon choosing specific taxon level and its details ranging from Phylum to Sub-Genus (e.g. Genus Liocichla),

all the species belonging to this taxon level along with detailed taxonomic hierarchy, accepted scientific names, synonyms, common names, occurrence details, along with DSN are presented in report format similar to traditional taxonomic checklists. For each DSN detailed data source is listed below. As depicted in Fig. 3.10, "taxon editor" can then scrutinize this report offline, which is then incorporated in the database by Quality Controller.

(F) Genus wise Taxon Editors Report

This module was developed to create genuswise taxon editors report for specific taxon level. This was developed to overcome hanging or no-response problem with "Taxon Editors Report"module due to large number of species belonging to specific taxon levels (e.g. Class – Insecta). Once you choose the taxon level (phylum to tribe); all genus belonging to it would be listed. For each of the genus "taxon editors report"could be generated (Fig. 3.11).

(G) Scientific Name List

This module was developed for both "taxon editors" and "Quality Controller"for rapid taxonomic scrutiny and first checks, respectively. This facilitates creation of scientific name list building for specific taxon level. Such a list is further sorted by scientific names in alphabetical order (A-Z), or by date of entry or entered by specific DM. As shown in Fig. 3.12, once specific taxon level with its detail is selected, scientific name list is generated along with its Authority, Year, DSN, DM no. This could be subjected to rapid scrutiny and first level quality checks, such as typographical errors.

(H) Taxon Accounting Report

This report format was devised to provide quick account of nos. of child taxons of specific parent taxon up to species level. As shown in Fig. 3.13, when specific taxon level (Phylum to Genus) and its details are selected (e.g. Order Gensiotrocha), its all child taxon levels upto species level are listed in traditional checklist pattern providing the total number of species belonging to each child taxon level. Nos. of species under each of the child taxon levels are hyperlinked, whose details could be viewed for generating genus wise species report.

(I) Scrutiny Accounting Report

This module was developed to provide at a glance account of scrutinized and unscrutinized species records of a particular taxon level (Order, Family and Genus). As depicted in Fig. 3.14, once specific taxon level and its details are slected (e.g.

Family Nymphalidae) report provides an account of total nos of species belonging to this taxon level, nos of scrutinized species record and unscrutinized species record, with hyperlink to enlist the scientific names of these species.

3.3.5 Taxonomic Scrutiny: A Process

As noted in previous chapter (Chapter 2), IndFauna so far collate baseline data about 94500 known Indian faunal species, collated from over 12500 data sources spreading over 250 years of modern biology. Over 65 man years have been invested to collate this data from various distributed and culturally heterogeneous centers (institutions, and individuals). Data cleaning and taxonomic scrutiny of such vast data can not be achieved at one go, owing to the fact that both data cleaning and taxonomic scrutiny are intelligent processes, and thus needs involvement of skilled individuals and taxonomic expertise within Indian and outside, which itself is rare to locate, and engaged for this purpose. Thus, it is going to be time consuming process, which calls for consistent and persistent persuasion with this pool of expertise.

However, with support financial support from the Global Biodiversity Information Facility (GBIF), data cleaning and taxonomic scrutiny of three Orders viz., Coleoptera, Hymenoptera, and Lepidoptera has been completed, accounting nearly 50% of the species collated in IndFauna. While the experiences and lessons learned from this scrutiny process are summarized subsequent sections of this chapter, it is evident that data cleaning and taxonomic scrutiny would be always work in progress, due to changing and dynamic nature of science of systematics or taxonomy.

3.4 IndFauna: Qualitative and Quantitative Analysis

As on date, baseline data regarding 94624 known faunal organisms has been collated along with 53361 known synonyms, 15104 common names, linked to 6574 localities distributed across India. This data has been sourced from 12954 data sources. Qualitative and quantitative of this exercise from various perspective put fourth exciting facts.

3.4.1 Taxonomic coverage

While, IndFauna collation of higher taxa such as vertebrates are complete, that of some of the lower taxa, such as insects, bryozoa (ectoprocta), mesozoa, protozoa are yet to be completed, mainly due to time required to gather information on them from dispersed sources or scarcity of information. In order to understand the taxonomic coverage of phyla collated into IndFauna compared to that with estimated nos. of known species in Indian phyla vis-à-vis global coverage, a bar chart was

produced (Fig. 3.15). According to Alfred (1998) less than 30 % of world's Phylum: Phoronida species coverage is recorded from India. However, IndFauna coverage of this phylum suggest that Indian coverage to be 90% of world's total coverage (Fig. 3.15). Among other phyla's exceeding coverage compared to national and global coverage (Alfred, 1998) are Mesozoa, Cnidaria, Nemertinea, Sipuncula, Priapulida, Chaetognatha and Chordata. On the other hand Phyla's such as Protozoa, Porifera, Gastrotricha, Acanthocephala, Echiura, Entoprocta, Tardigrada, Echinodermata and Hemichordata are yet to be completed as compared with the known number of estimated species from India. The first observation could be attributed to the fact that the national estimation (Alfred, 1998), is nearly nine years old, and during which period new species have been discovered. In case of second observation it could be reasoned out that these being less studied taxa's it is hard to place hand on data sources covering them.

3.4.2 Number of Species v/s. Number of Publications

In order to understand the relationship between numbers of publications and discovery of coverage of species per phylum, another bar chart was produced (Fig. 3.16). This bar chart reveals that Phlya such as Arthropoda, Chordata, which are well studied, produced maximum number of peer-reviewed or public domain literature, and thus nos. of species coverage was more i.e. in proportional to number of publications. However, in case of Phlya's such as Annelida, Brachiopoda, Bryozoa, Chaetognatha, Ciliophora, Echinodermata to Mollusca, to Hemichordata, though the nos. of publications is less, the number of species documented are very high, signifying that rate of new species discovery amongst these phylas are extremely high. In case of Phyla's Nemeta and Platyhelminthes, where number of publications are high in comparison to species count, indicating that rate of species discovery amongst these phyla's is slow in comparison with others.

3.4.3 Taxonomic literature during 1750 to 2007

When cumulative numbers of taxonomic publications per decade were plotted from 1750 till 2007, it reveals that publications on Indian faunal taxonomy increased during 1920's because of explorations and consequent studies by the Zoological Survey of India in 1916 (Fig. 3.17). The taxonomic work was further boosted up exponentially during 1940s and shows a striking ascend till 2000. The number of species newly recorded or renamed per year corresponds to the number of publications with a steady increase in records from 1950 till 1982 (Fig. 3.17 or 3.18),

which represents the post-independent boost to survey, exploration and inventorying activities.

3.4.5 Taxonomic studies in Indian states

A state wise distribution of species records and related number of publications reveal that the states such as Andaman and Nicobar, Sikkim, Tamil Nadu, Assam, west Bengal, Uttar Pradesh, Meghalaya and Kerala are highly studied and represent high number of species accordingly (Fig. 3.19). However, if compared to the geographic coverage of each state the large states like Rajasthan, Madhya Pradesh, Andhra Pradesh and Gujarat are very poorly represented by both, the number of publications and the number of species records.

3.5 Taxonomic discrepancies

During the exercise of data quality control, when IndFauna was cross-checked with some of the global species directories or authority files such as Species 2000, ITIS, and ION, it revealed several discrepancies which seem to be coexisting for several years and decades together, without being attempted to resolve. In this section, I am attempting to enlist some of these discrepancies. These discrepancies could be grouped into four categories, viz. (a) hierarchical differences with other known global databases and authority files, (b) diffrences in taxonomic hierarchy, (c) spelling differences, and (d) homonymies.

3.5.1 Hierarchical differences with other known global databases

Most of the global data bases like Species 2000 follow eight-kingdom classification in which Kingdom Animalia is distributed in 38 phyla and 93 classes (including 11 uncategorized classes), whereas Kingdom protozoa is distributed into 17 phyla. IndFauna includes both the Kingdoms Animalia as well as Protozoa with 36 phyla further divided into 141 classes including 20 uncategorized ones. Thus the phylum Apicomplexa, Phylum Ciliophora, Phylum Protozoa and Phylum Sarcomastigophora that belong to Kingdom Protozoa are additional.

Nematomorpha and Kynorhyncha are two distinct phyla in IndFauna, which are treated as classes under Phylum Cephalorhyncha in Species 2000. Phylum Placozoa of Species 2000 represented by single genus and two species (Trichoplax adhaerens and T. reptans) is absent in IndFauna.

This differences needs to be resolved, but, not before evolving national consensus amongst the taxonomic community in India. However, brining in such a

consensus seems to be long term process, and calls for strong advocacy amongst the distributed and disparate group of taxonomists.

3.5.2 Difference in taxonomic hierarchies

Several examples were found where the taxonomic hierarchy of organisms' followed in India did not match that used by ITIS. This is especially true in case of some fishes, nematodes and insects. This is the result of differences in taxonomic opinions or the provisional nature of certain data in ITIS and a consensus is often difficult. In this case, the information managers can display the placement of the taxon according to alternative schemes. In spite of this option, it is necessary to confirm with the international taxonomic opinion, to make the datasets interoperable with those developed in other parts of the world. This is an issue that needs to be discussed and resolved by taxonomists working in India. Although making changes in taxonomic hierarchy is technically possible in case of the electronic datasets each change needs to be validated by taxonomic community as some taxa may or may not be confirming to that change. Some of the examples where taxonomic hierarchy is different are given in Table 3.1. As per ITIS and other taxonomic resources Sub Class Elasmobranchii is placed under Class Chondricthyes, while Systema Naturae 2000 (Brands, 1989-2005) still recognizes it as Class Elasmobranchii.

3.5.3 Differences in spellings

The most common problem faced while digitizing the data was different spellings of organisms' names. Some examples are given in Table 3.2. Order Cheilostomata as per ITIS is named differently as Order Cheilostomida by ERMS (Costello, et al, 2006). Even the hierarchy under this order is not same for many species given in these two databases.

In some cases these were misspellings especially typographical errors. However, to follow the taxonomic norms, each misspelling needs to be reported alongwith the valid scientific name for avoiding future problems. Usually such wrongly spelled scientific names are reported as synonyms with a prefix "sic" as per the International Code of Zoological Nomenclature. In some case the difference in spelling was due to different taxonomic opinions, and choosing which to use requires discussion with taxonomists. For example the Blue Whale Shark *Rhincodon* which is spelled *Rhiniodon* (Talwar and Kacker, 1984) in Indian literature. ITIS shows *Rhiniodon* as a synonym of *Rhincodon* that was suppressed by a ruling.

3.5.4 Homonyms

A few homonyms were identified in the cataloguing process. *Chaunoproctus* & *Microcosmus* were observed in common use, which is a direct contradiction to nomenclature rules, and hence taxonomic opinion was sought to resolve the problem. *Chaunoproctus* is widely in use and *Chaunoproctus* Pearse, 1906 (Insecta, Acari) is a group of mites as per the Indian Faunas (ION, 2005; Sanyal and Bhaduri, 1986; Sanyal et al, 2003), while *Chaunoproctus* Bonaparte, 1850 (Aves, Passeriformes) is a bird as per ITIS (ITIS, 2005), which is now extinct (IUCN, 2005). A detailed search of literature revealed that Family Chaunoproctidae Balogh 1961 was created for which *Chaunoproctus* Bonaparte, 1850 is the type genus (Hallan, J. pers.comm.). According to Hallan (Hallan, 2003) it holds one more genus *Chaunoproctellus* S. Mahunka 1992. Genus *Chaunoproctus* Pearse 1906 is given as equal to *Zetorchella* Berlese 1916 [*Zetorchella pedestris* Berlese 1916: type] and also equal to *Caloppia* Balogh 1958 [*Caloppia basilewskyi* Balogh 1958: type] with one species *Chaunoproctus cancellatus* Pearse 1906. However based upon the publication dates we conclude that the generic name *Chaunoproctus* Pearse, 1906 is a junior homonym of *Chaunoproctus* Bonaparte, 1850. The Indian literature thus has to be corrected now following the rules of nomenclature.

Similarly the generic name *Microcosmus* Heller, 1878 (ITIS, 2005) [1877 (ION, 2005)] (Chordata, Ascidiacea) is a homonym of *Microcosmus* Chaudoir 1878 (Insecta, Coleoptera, Carabidae) (Saha et al, 1992). Both names are in use and refer, respectively, to Ascidiacea as per the ITIS (ITIS, 2005) and ION (ION, 2005), and for a beetle as per Indian literature (Saha et al, 1992). Opinion was sought from carabidologists and finally Wolfgang Schiller resolved this issue. *Microcosmus* genus was described by Chaudoir 1878 in Ann. Soc. ent. Belgique XXI p.85 and P.139 for the tribe Panageini within family Carabidae (Order Coleoptera: Class Insecta: Phylum Arthropoda) for the species *M. flavopilosus*. Because the same genus *Microcosmus* was already designated by Heller in 1877 under family Pyuridae, (Class Ascidiacea, Phylum Chordata), Emberik Strand proposed in 1936 in Folia Zool. Hydrobiol. IX p. 169 the new name *Microcosmodes*. So until today the carabid species is cited as *Microcosmodes flavopilosus*. But still Zoological Survey of India records show presence of *Microcosmus* instead of *Microcosmodes* until recent publications.

Polypodium hydriforme, which is under phylum Cnidaria, is still considered as a valid scientific name under Phylum Pisces alongwith Phylum Cnidaria by ION

(ION, 2005). Genus *Doto* described by Oken, 1815 is considered valid for phylum Mollusca while; Smithsonian NMNH database (2004) displays it under phylum Arthropoda as well as phylum Mollusca. In these cases it is really difficult to place the organism in a specific hierarchy. Another example is of the Genus *Cyaniris* which is present under the Family Lycaenidae, Order Lepidoptera of Class Insecta (Bingham, 1907). The same genus name *Cyaniris* is also given to insects belonging to family Chrysomelidae, Order Coleoptera of Class Insecta as per the database of the holotype collections from India present in the Belgium Museum (Institute Royal des Sciences naturelles de Belgique, 2002). The genus is not included in the ITIS database while the other web sites and literature sources shows presence of genus *Cyaniris* in Order Lepidoptera. While checking hierarchy for a species namely, *Idia pristis* which is placed under Class: Hydrozoa (Ritchie, 1910), surprisingly we came across genus *Idia* which is also present under Class Insecta, Order Lepidoptera, Family Noctuidae as per the checklist of Noctuoidea of Ontario present on the website hosted by Canadian Biodiversity Information Facility (CBIF, 2003). There is also no ready reference available for this genus designation anywhere.

3.6 Taxonomic and Nomenclatural Issues

Main issues, which evolved during database development, were more of taxonomic nature and need to be dealt with collaboratively by biodiversity scientists and IT managers. First of these issues is the availability of authentic information on Indian fauna. A thorough search of zoological literature is in progress to collect information on species. It has revealed many deficiencies in the data. Even after 200 years of research work, information is available mainly on vertebrates, while invertebrates in general are grossly understudied. Even though the scope of this cataloguing exercise was limited to Indian fauna, it was extremely difficult to locate data sources about known species of invertebrates. This is especially true in case of Class Insecta, Order Coleoptera, which is the largest order in India. Even FBI, the primary source of information on Indian fauna, does not cover some families of Coleoptera. Recent faunas and monographs being published by ZSI are available only for certain groups such as Aphidoidea (Ghosh 1982, David and Ghosh 1982), Scorpions (Tikader and Bastawade 1983), Spiders (Tikader 1982), Dermaptera (Srivastatva 2003) etc. Thus, information for other taxa not covered in these two major works, is dispersed in research papers in various national and international journals, spread over a period of 100 years and accessing this is a major challenge.

ZSI has made a major contribution by publishing a sourcebook of all taxa published so far by ZSI scientists (Das 2003). However, those published by other workers in other research institutions cannot be sourced through a single repository. There exists no centralized system within the country for registering the new taxa, name changes or new combinations. Our effort in this case is to systematically go through the international Zoological records to note the new taxa published so far from India. The question still remains about the names, which have appeared in journals not abstracted in zoological records. This leads to serious problems in estimating biodiversity richness esp. the number of species. No single source gives the entire list of 89451 species as per the recent estimates. Some of the old and also recent fauna volumes include species from adjacent areas such as Nepal, Burma (=Myanmar), Bangladesh and Ceylon (=Sri Lanka), and a mere addition of all the numbers in Faunas will lead to wrong estimates of species found within India. No reliable estimate can be provided for the number of subspecies and varieties present within India. An actual count of subspecies, varieties can help in analysis of origin and zoogeographic studies of Indian fauna. The problem has been intensified due to inclusion in recent literature of species, which are only identified to the genus level. It is not advisable to include these in ECAT, however it is not clear whether they have been counted in the total species estimates for India.

International Code of Zoological Nomenclature set up in 1895, is the international authority that rules on scientific names. It publishes the rules universally accepted as governing the application of scientific names to all organisms, which are treated as animals. It also gives rulings on individual nomenclatural problems brought to its attention, so as to achieve internationally acceptable solutions. Several million species of animals are recognized, and more than 2000 new generic names and 15000 new specific names are added to the zoological literature every year. With such a multiplicity of names problems are bound to occur, hence it is necessary for individual researcher to adhere to and inform the ICZN regarding new species and nomenclatural changes. But in many cases in India, the new names or combinations published in journals are not brought to the attention of the Commission. ICZN also has a quarterly journal, the Bulletin of Zoological Nomenclature, in which problems needing a formal decision by the Commission are published for discussion by the zoological community. However, very few taxonomic institutions have access to this journal. Hence zoologists hardly ever use the international system for name

registering and keeping up with the multiple descriptions, nomenclatural changes etc. are the burden of individual taxonomists. This is an extremely hard work considering the poorly equipped taxonomic research laboratories with meager funding to libraries. It is not surprising that scientists do not get updated information about internationally accepted taxonomic changes.

This has given rise to major problem, which is the great divide between taxonomic systems followed within India and elsewhere in the world. We found several instances where taxonomic system and nomenclature followed in India does not match the ITIS. This is especially true in case of fish and Nematodes. Even within India, there are many opinions about the correct taxonomic hierarchy and placement of groups. The bifurcation in taxonomy is in fact bifurcation in taxonomic opinion and a consensus on the issue is most often impossible. In this case, the only alternative left for information managers is to display the placement of the taxon according to different alternative schemes. In spite of this option, it is necessary to confirm with the international taxonomic opinion, to make the datasets interoperable with those developed in other parts of the world. This is an issue that needs to be discussed and resolved by taxonomists working in India. Although making changes in taxonomic hierarchy is technically possible in case of the electronic datasets each change needs to be validated by taxonomic community as some taxa may or may not be confirming to that change. FBI describes Order Rhynchota (Distant 1903, 1904, 1910, 1916). Now, several alternative classifications are available for the group, and recent trend is to avoid term Hemiptera and to treat Heteroptera and Homoptera as Orders (Nearctica, 1998). However, ITIS shows all three as distinct valid orders. In this case the challenge to our information management was to accordingly update the classification of species in former order Rhynchota of FBI. Although ITIS was followed mainly for this purpose, confirmed decision could not be made about placement of some species.

Hence it is not merely a technical task but essentially a taxonomic revision task, which is needed for confirming to recent taxonomic hierarchies. It is extremely difficult as it requires access to literature, type specimens, keys and protologues, but can be made possible by scientific collaborations across the world. Electronic catalogue in this case can form baseline dataset for revision work.

Another challenge faced while cataloguing is the change in the political boundary of India since British period. This has led to many of the localities noted in

early literature being now actually in the neighboring countries. Even recently published faunas include organisms from Nepal and Sri Lanka. Although biodiversity transcends political boundaries, the distinction between species strictly in India and those in neighboring countries is necessary for policy makers and managers, especially in case of endemic and rare species. Species which are only known from localities outside India such as Tenasserim , Sylhet, Allahabad, which were in Pre-Independence India are excluded from IndFauna as they are now in Myanmar, Bangladesh and Pakistan respectively. The distributions provided by the literature often vary from point localities such as villages to regional distribution such as certain districts or states, or ecosystem coverage such as Western Ghats, Himalaya, oceans, rivers etc. ECAT provides alternatives to include all of these data and links it to the states within India. Old literature often refers to regions such as United Provinces, Bombay presidency etc. which no longer exist. The post independence India has also undergone major rearrangements, most recent being creation of Uttaranchal, Chhattisgarh etc. To compare the historical distribution data with today's geopolitical maps can be achieved by coupling multi-layered mapping facility with ECAT. In this we can overlay various maps with distribution at one historical place and time and compare it with present reports.

The analysis of data so far indicates that some ecosystems like freshwater river systems, wetlands, and oceans are yet to be surveyed in detail. The data is also deficient in case of local names in multiple languages. It is possible that in case of invertebrates the local people have only broad category names for certain taxa e.g. Koli for spider. For vertebrates, esp. birds, mammals, names in only a few languages have been documented.

At present the data about diversity is mainly textual. Images and artwork is available in case of some common taxa especially of vertebrates. An illustration is most often provided for new species, however, it does not aid in visual identification. Audio and video data regarding animal behaviour can also help in identification and documentation of species. However, it is at present restricted to more 'popular' species such as Mammals, Snakes and Birds.

Identifying and documenting synonymy is a major challenge in taxonomy, which created certain unique problems in cataloguing. In some cases, a single species had as many as 100 synonyms according to recent revision. In addition to that some literature also quotes reports of names, which are essentially not synonyms, but it is

difficult to differentiate between a report and synonym in these faunas. Several of the names used in FBI and other old literature have now changed. But most of the recent works do not include a citation of FBI or old literature, making it difficult to check up synonymisation. Indian floras as a rule include citation of Flora of British India, and this rule should be applied to faunal literature as well. In most cases, the published literature only included few of the more important synonyms owing to the limitation of printing space. Although electronic media has no such limitations, some rules are necessary to define the number of synonyms and reports essential to provide complete history of a particular scientific name. The electronic catalogue can provide a means to identify potential synonyms. For ex. *Lucilia indica* Robineau-Desvoidy 1830 is reported as a synonym of two species first of *Orthellia indica* (Robineau-Desvoidy) (van Emden 1965, Mitra et al 2002) and second of *Orthelia lauta* (ZSI, 1997). The documentation does not allow us to guess whether the two new names are of different organisms, or are in fact synonyms of each other. Several such instance of taxonomic ambiguity can be pointed out by use of the electronic catalogue. The most common among these problems is of spellings. Documentation is available in cases such as change of name in tortoise shell beetle, Aspidomorpha to Aspidimorpha (Borowiek and Swietojanska 2002), but in many cases, the research papers do not provide support to the specific spelling of scientific name that is followed.

The role of electronic catalogue in this case has been to raise these issues and put forth them in public domain for further discussion. It is hoped that the collaborative efforts in future can contribute positively in dealing with the above-mentioned ambiguities in the biodiversity information.

Identification of organisms is fundamental to biodiversity studies. Owing to this, the discipline of taxonomy, especially scientific nomenclature has gained immense importance. Taxonomy provides a vocabulary to discuss the world (Knapp et al., 2002). Each name is unique and its representative organism is precisely described. It is estimated that about 1.8 million species of organisms' have been formally named from the world (May, 1999) and each is recognized by a unique binomial. More than 2000 new generic names and 15000 new specific names are added to the zoological literature every year and with such a multiplicity of names, problems are bound to occur. International mechanisms such as the International Code of Zoological Nomenclature (ICZN) (ICZN, 2003) and the International Code of Botanical Nomenclature (ICBN) are rulebooks that govern how organisms' are named

and they provide clear instructions on how to go about the process (Knapp et al., 2002).

International codes of nomenclatures require taxonomic actions to be published and the data thus made available (Agosti and Johnson, 2002). However, nomenclatural additions or changes have to be conveyed to the ICZN or ICBN by the authors and is usually done when ratification is needed from the international authority. Similarly, discrepancies in the nomenclature are brought to the notice of the ICZN and ICBN by scientists, which are later reviewed. This process requires a long time and the availability of a large amount of literature to the scientists discussing nomenclature. In several cases, especially for taxonomists in developing countries, recent taxonomic literature including the code itself is unavailable. Very few libraries around the world have the financial capacity to carry the full range of literature in which systematic results are published (Agosti and Johnson, 2002). Hence, nomenclature changes are in many cases unavailable or become available much later to the developing country scientists than to their counterparts in developed world. This leads to use of old or outdated nomenclature.

On the other hand taxonomic papers by developing country scientists published in journals with regional scope, which are not scientifically abstracted, remain isolated and unnoticed by the wider scientific audience and taxonomic changes proposed or used in such papers are often neglected. This obviously leads to many discrepancies in the information available, especially about the current or correct taxonomic hierarchy of organisms'. It is thus necessary to create a system, which will lead to rapid identification of taxonomic discrepancies, and their resolution. In addition, a permanent mechanism for registering and validating scientific names of organisms' needs to be created at a national as well as a global level. Web-based electronic catalogues can be effective in creating such a central repository of taxonomic information. In this paper, we demonstrate the use of web-based electronic catalogues (ECATs) in identifying taxonomic discrepancies in Indian context.

However, the information is carefully scrutinized for validity and accepted only if it is from reputed taxonomic institutions or experts. For each species, the taxonomic hierarchy used by Indian faunas is crosschecked diligently with that used in global taxonomic inventories such as Integrated Taxonomic Information System (ITIS), Species2000, Catalogue of Life (Bisby et al., 2007), Index to Organism Names

(ION, 2005), European Register for Marine Species (Costello et al., 2006), Systema Naturae 2000 (Brands, 1989-2005) etc. Incase of any problems regarding taxonomic placement of the species concerned taxonomy experts are contacted and as per their suggestions the species are being entered in the database.

3.7 Recommendations

A consensus on these issues is a matter of taxonomic discussion. They need to be resolved by using nomenclatural rules, which requires further detailed research. However these examples effectively demonstrate the potential of electronic catalogues (ECATs) in bringing issues or discrepancies to the knowledge of taxonomic community, starting a dialogue between taxonomists across the globe and identifying issues of common concern. In order to notice such discrepancies and resolve them quickly, it is essential that a wrapper be developed which traverses through various electronic catalogues searching for taxonomic anomalies. This calls for increasing collaboration among the various ECATs.

The information available so far on the Internet is limited to the names and citation alone. But with the growing use of information and communication technologies in biodiversity research it should be possible to make the taxonomic literature itself available on the Internet, which can be used for checking inconsistency in taxonomy, used worldwide. Although taxonomists from around the world have been effectively dealing with these discrepancies, it is a time consuming and tedious process to identify, check, and correct them using the traditional media such as published literature. Modern information and communication tools can be of immense help for identifying taxonomic discrepancies quickly and resolving them in a collaborative manner leading to globally acceptable standardized inventories. With the use of Internet, there can be a true two-way exchange of information between taxonomists from developed and developing countries. Active collaboration and commitment of taxonomists and information managers are required to work towards the goal of developing information systems to bring uniformity and precision to taxonomic inventories across the world. Many of the discrepancies arise because taxonomists are unable or find it difficult to check up on taxon names especially for taxa outside their field of expertise.

Hence to build up easy communication pathways and reduce the time input, it would be extremely helpful to have a web-based central registry system for taxonomic names. Checking of names being used in publication with the central registry would

definitely eliminate many of the commonly encountered discrepancies described above. Thorne (Thorne, 2003) also proposed the need for registration of new taxa names in a central registry of names. The journal *Nature* has already taken a step towards registering of names (*Nature*, 2002) by requiring the authors of papers featuring new taxonomy to file the information with a recognized institute such as Linnaean Society of London. Central registry will be a repository of scientific name information or an index for scientific names in use, along with their history. It will be a dynamic register for proposed scientific names (which will be provisionally accepted, noted), which can later be added to the repository after annotation. These will serve as a reference for scientists describing new names to check if same has been used before, and in which context. This will eliminate generation of homonyms. It can also provide a point of "normalization" for data.

ECATs offer an effective method of creating unique electronic registers. Owing to the rules of acceptance of scientific names, the names cannot be registered as valid before the publication of taxon description in a journal. To solve this, a precedent can be set that in case of each new description, together with the type specimen deposition number, a provisional registration number in the global, regional or national web based ECAT should be quoted. There will be two way information exchanges with other ECATs. The registry will compare between ECATs information, find out any points of mismatches/ conflicting data, and also pick up new information automatically from the ECATs. Using this, a single number reference system for each scientific name can be developed. The central registry can provide a minimum standard and starting point for use in other databases.

Global Biodiversity Information Facility (GBIF) together with the Taxonomic Database Working Group (TDWG) is currently seeking requirements for globally unique identifiers (GUIDs) for biodiversity informatics and to establish infrastructure to support their use. GUIDs once developed can overcome most of the current problems, such as (a) identification of same data records served from multiple locations, (b) referring to data from outside network, irrespective of frequent change of URLs, and (c) referring to taxon concepts in reliable and consistent way. Page (2005) suggests a system of Life Science Identifiers (LSIDs) as unique numerical identifiers for scientific names and ITIS currently employs a system of unique Taxonomic Serial Numbers (TSN"s). Databases could be mapped to ITIS Taxonomic Serial Numbers (TSN's).

Many organizations are working towards building up registers of published scientific names of taxa such as for beetles (Coleoptera, 2005), mites (Hallan, 2003) etc. Plant names can be checked using International Plant Names Index (IPNI, 2005). Index to Organism Names (ION, 2005) database can be used to check up zoological names. The most holistic effort are of the Species 2000 and ITIS Catalogue of Life (Leslie, 2005) and GBIF (GBIF, 2006), which aim to create an index of all 1.8 million known species by 2011. This is a major step towards developing a central register of names, and increasing collaborations between similar efforts worldwide should shorten the time required for this. Links to other databases like image/ DNA/ protein sequences/locality maps etc. eliminating homonymies, a date attached to each scientific name linking names with collection catalogue numbers will be major step. It will help scientists quickly track all collections and know where they are deposited. Applications could be built such as those on the ITIS Canada website or Ubio that display multiple classifications. In addition, a taxonomist's time is saved by having a tool that can readily compare the taxonomist's data with that in the central file (such as ITIS' Taxcompare tool). The ability to quickly check for homonymies will also save the taxonomist time. In addition, development and use of national ECATs should be encouraged to collate information at the national levels and make it available to the global users. This is especially important, as these national registers will be able to easily access locally available primary taxonomic information.

It is necessary to track the changes in species concepts over time. ITIS has developed a capability for change tracking, but has not yet implemented it. Availability of specimens, images, protologues, classifying characters which are in use in different countries, comparing between specimens of a species with wide distribution- transcending political boundaries and building biogeographic distribution maps, language barrier - translating of Latin diagnosis, picture data are some of the capabilities required to ensure accurate results in biodiversity research projects.

These advances suggest that in future, the taxonomic discipline would make broad use of the web-based information and benefit from it. Therefore, it is crucial at the moment to build up or improve the collaborative activities among domain experts, information managers and users of taxonomic information. This would ultimately help in strengthening the biodiversity research necessary for conservation and management of global natural resources.

3.8 Summary

The uses of species and occurracne data are wide and varied, and encompass virtually every aspect of human endeavour. However, increasing use of species-occurrence data demands high degree of "fit-to-use" state of data. However, the implementations of data cleaning principles that determine the "fit-to-use" of data are in its nascent stage in biodiversity and especially taxonomic data management. IndFauna, which collate species and occurrence data from secondary sources literature, adopted two levels of data curation, viz., Fast cheks and Taxonomic Scrutiny. For this purpose it has deployed data curation modules, which are used by Quality Controllers, and online "taxon editors", where in taxon editors report modules were implemented for taxonomists who prefer to help curate the data offline.

Qulaitative and quantitative analysis of IndFauna data, and its data sources reveals some interesting facts, emphasizing the need for more detailed studies of less known taxas, and survey and exploration activities in less studies regions of the country. The discrepancies found while developing IndFauna and comparing it with existing databases can help to solve several issues like taxonomic ambiguities, inadequate documentation and incorrect placements of species. Development of electronic catalogues of names of known organisms' (ECATs) will help in pointing out these issues. International organizations like GBIF are trying to make all biodiversity data accessible to the largest possible section of the human population. Recently GBIF, Species 2000, ITIS and uBio (GBIF, 2005) have decided to cooperate on compiling and utilizing taxonomic information resources. National and regional resources such as IndFauna, after solving the types of discrepancies described here, can make valuable contributions to preparing of global taxonomic databases and standards.

Table 3.1: Difference in hierarchies used in various sources

Sr. No.	Taxon or Scientific name	Sources Referred in IndFauna	Other Sources	Remarks
1	**Species: *Appendicularia histnae***	Kingdom: Animalia Phylum: Chordata Subphylum: Urochordata Class: Larvacea Order: Oikopleurida Family: Appendicularidae Genus: *Appendicularia* Species: *histnae* (Das 2003, Dhandapani 1977)	Kingdom: Animalia Phylum: Chordata Subphylum: Tunicata Class: Appendicularia Order: Copelata (ITIS)	ITIS does not include genus *Appendicularia*
2	**Geus: *Pillaia***	Kingdom: Animalia Phylum: Chordata Class: Actinopterygii Order: Perciformes Family: Chaudhuriidae Genus: *Pillaia* (Rao et al., 2000)	Kingdom: Animalia Phylum: Chordata Class: Actinopterygii Order: Synnbranchiformes Family: Chaudhuriidae Genus *Pillaia* (ITIS)	In ITIS Genus ***Pillaia*** is placed under different Order of Class Actinopterygii
3	**Genus: *Zenarchopterus***	Kingdom: Animalia Phylum: Chordata Class: Actinopterygii Order: Atheriniformes Sub Order: Exocoetoidei	Kingdom: Animalia Phylum: Chordata Class: Actinopterygii Order: Beloniformes Sub Order: Belonoidei	In ITIS Genus ***Zenarchopterus*** is placed under different order

		Family: Hemiramphidae Genus *Zenarchopterus* (Rao et. al, 2000)	Family: Hemiramphidae Genus: *Zenarchopterus* (ITIS)	
4	**Species: *Chazara heydenreichi***	Kingdom: Animalia Phylum: Arthropoda Subphylum: Hexapoda Class: Insecta Order: Lepidoptera Suborder: Macrolepidoptera Superfamily: Papilionoidea Family: Nymphalidae Subfamily: Satyrininae Genus: *Chazara* Species: *heydenreichi* (Evans, 1932; Wynter-Blyth, 1957)	Kingdom: Animalia Phylum: Arthropoda Subphylum: Hexapoda Class: Insecta Order: Lepidoptera Suborder: Macrolepidoptera Superfamily: Papilionoidea Family: Nymphalidae Subfamily: Satyrininae Genus: *Satyrus* Subgenus: *Chazara* Species: *heidenreichii* (LepIndex)	Natural History Museum (NHM), London's LepIndex shows both the species S. *heydenreichi and S. heidenreichii as provisionally accepted names.*
5	**Genus: *Discophora***	Kingdom: Animalia Phylum: Arthropoda Subphylum: Hexapoda Class: Insecta Order: Lepidoptera Superfamily: Papilionoidea Family: Amathusiidae	Kingdom: Animalia Phylum: Arthropoda Subphylum: Hexapoda Class: Insecta Order: Lepidoptera Superfamily: Papilionoidea Family: Nymphalidae	Most of the species under Family Amathisiidae are placed under Tribe Amathusiini of Family Nymphalidae in LepIndex.

		Genus: *Discophora* (Soubadra Devi and Davidar 2001)	Subfamily: Morphinae Tribe: Amathusiini Genus: *Discophora* (LepIndex)	
6	**Genus: *Neocentrophyes***	Kingdom: Animalia Plylum: Kinorhyncha Class: - Order: Homalorhagida Family: Neocentrophyidae Genus: *Neocentrophyes* (ZSI 1991, SNMNH, 2007, ITIS)	Kingdom: Animalia Plylum: Cephalorhyncha Class: Kinorhyncha Order: Homalorhagida Family: Neocentrophyidae Genus: *Neocentrophyes* (Sp2000)	Kinorhyncha is treated as Phylum in ITIS and as Class in Sp2000.
7	**Genus: *Meiopriapulus***	Kingdom: Animalia Plylum: Priapulida Class: - Order: - Family: Tubulichidae Genus: *Meiopriapulus* (SMNNH, 2007 and ITIS)	Kingdom: Animalia Plylum: Cephalorhyncha Class: Priapulida Order: - Family: Tubilichidae Genus: *Meiopriapulus* (Sp2000)	Priapulida is treated as Phylum in ITIS and as Class in Sp2000.
8	**Families: Gordiidea and Chordodidae**	Kingdom: Animalia Plylum: Nematomorpha Class: - Order: Gordioidea Family: Gordiidae	Kingdom: Animalia Plylum: Cephalorhyncha Class: Nematomorpha	Nematomorpha is treated as Phylum in ITIS whereas in Sp2000, it is Class containing only one order; and Order

		Family: Chordodidae (Camerano, 1912; Schmidt-Rhaesa, A. and A. K. Yadav, 2004)	(Sp2000)	Gordioidea is absent.

Table 3.1: Difference in hierarchies used in various sources

Table 3.2: Misspellings and difference in hierarchies in various sources

Sr. No.	Indian sources	Other sources	Remarks
1	Genus: *Corynosoma* Species: *streemosum* (Bhattacharya 1998)	Genus: *Corynosoma* Species: *strumosum*	ITIS
2	Genus: *Psudorca* Species: *crassidens* (Agarwal V.C. 1998)	Genus: *Pseudorca* Species: *crassidens*	ITIS
3	Genus: *Ambylopharyngodon* (Aditya & Raut 2001)	Genus: *Amblypharyngodon*	Genus previously present in ITIS but not found as on date.
4	Kingdom: Animalia Phylum: Chordata Class: Actinopterygii Order : Cyprinodontiformes Family: Cyprinodontidae Genus: *Oryzias* Species: *melanostigma* (Nandi 1993)	Kingdom: Animalia Phylum: Chordata Class: Actinopterygii Order: Beloniformes Family: Adrianichthyidae Genus: *Oryzias* Species: *melastigma*	The hierarchy of Genus *Oryzias* given by Indian sources does not match with ITIS and Species 2000 after Class level.
5	Genus: *Hestina* Species: *nama* (Gupta and Shukla 1988)	Genus: *Hestinalis* Species: *nama*	LepIndex

6	Genus: Mycalesis Species: *aemata* (Gupta and Shukla 1988; ZSI, 2000)	Genus: Mycalesis Species: aemate	LepIndex
7	Family: Nymphalidae Genus: Catapoecilma Species: delicatum (Evans 1910)	Family: Lycaenidae Subfamily: Theclinae Genus: Catapoecilma Species: dolicatum	LepIndex
8	Family: Nymphalidae Genus: Doleschallia Species: andamana (Zoo Outreach, 1995; ION 2005)	Family: Nymphalidae Subfamily: Nymphalinae Tribe: Kallimini Genus: Doleschallia Species 1: andamanensis Speices 2: andamanica	LepIndex shows two valid species D. andamanensis and D. andamanica, however D. andamana is not present in it.
9	Family: Nymphalidae Genus: Mesoacidalia Species: aglaja (Kunte et al, 1999; ION 2005)	Family: Nymphalidae Genus: Mesoacidalia Species: aglaja	LepIndex

10	Order: Lepidoptera Superfamily: Papilionoidea Family: Amathusiidae Genus: Discophora	Order: Lepidoptera Superfamily: Papilionoidea Family: Nymphalidae Subfamily: Morphinae Tribe: Amathusiidae Genus: Discophora	LepIndex

Table 3.2: Misspellings and difference in hierarchies in various sources

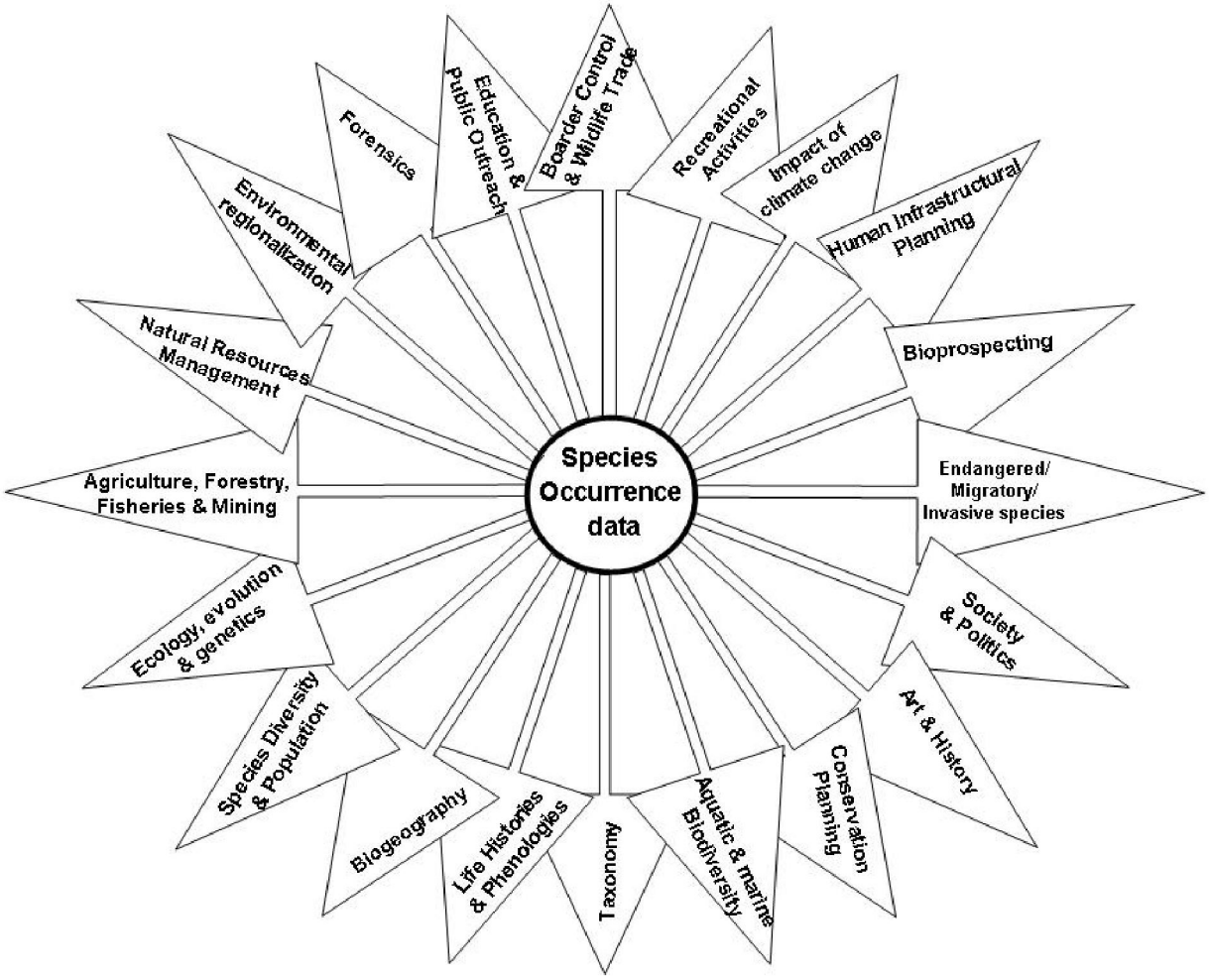


Figure 3.1: Uses of Species Occurrence data

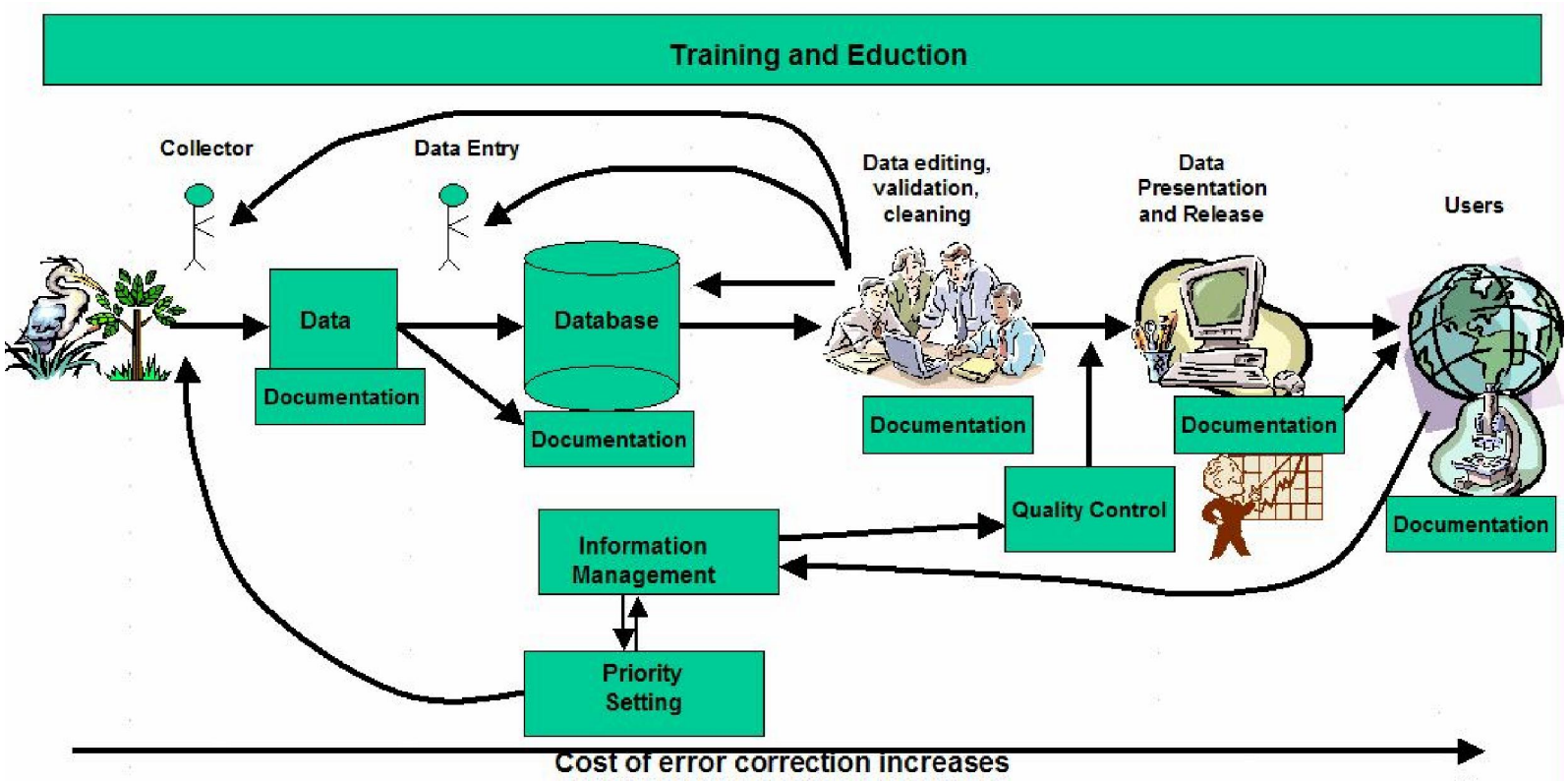


Figure 3.2: Information Management Chain depicting that the cost of error correction increases as one move along the chain. (Source: Chapman, 2005a)

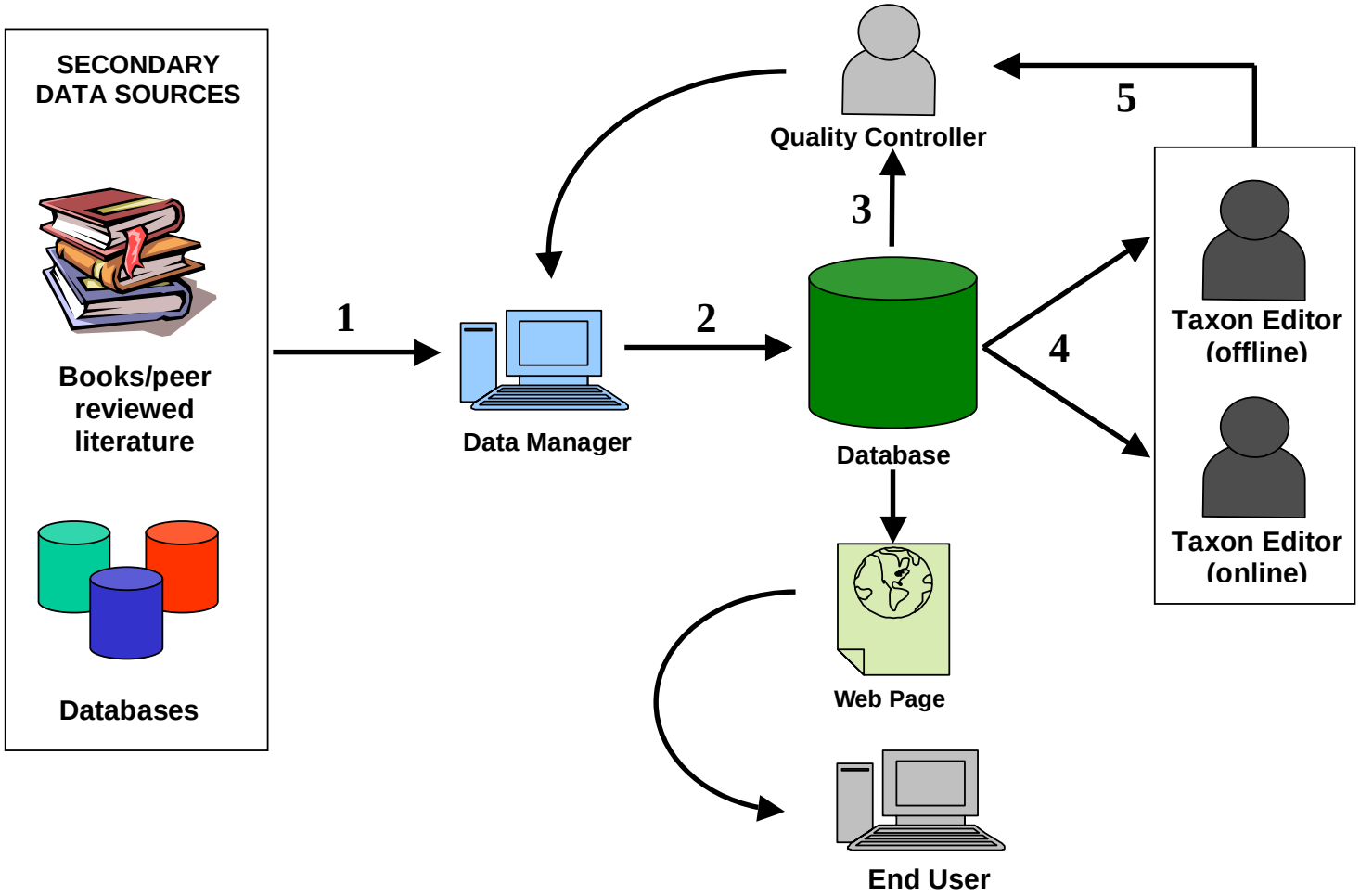


Figure 3.3: IndFauna approach for Taxonomic Scrutiny

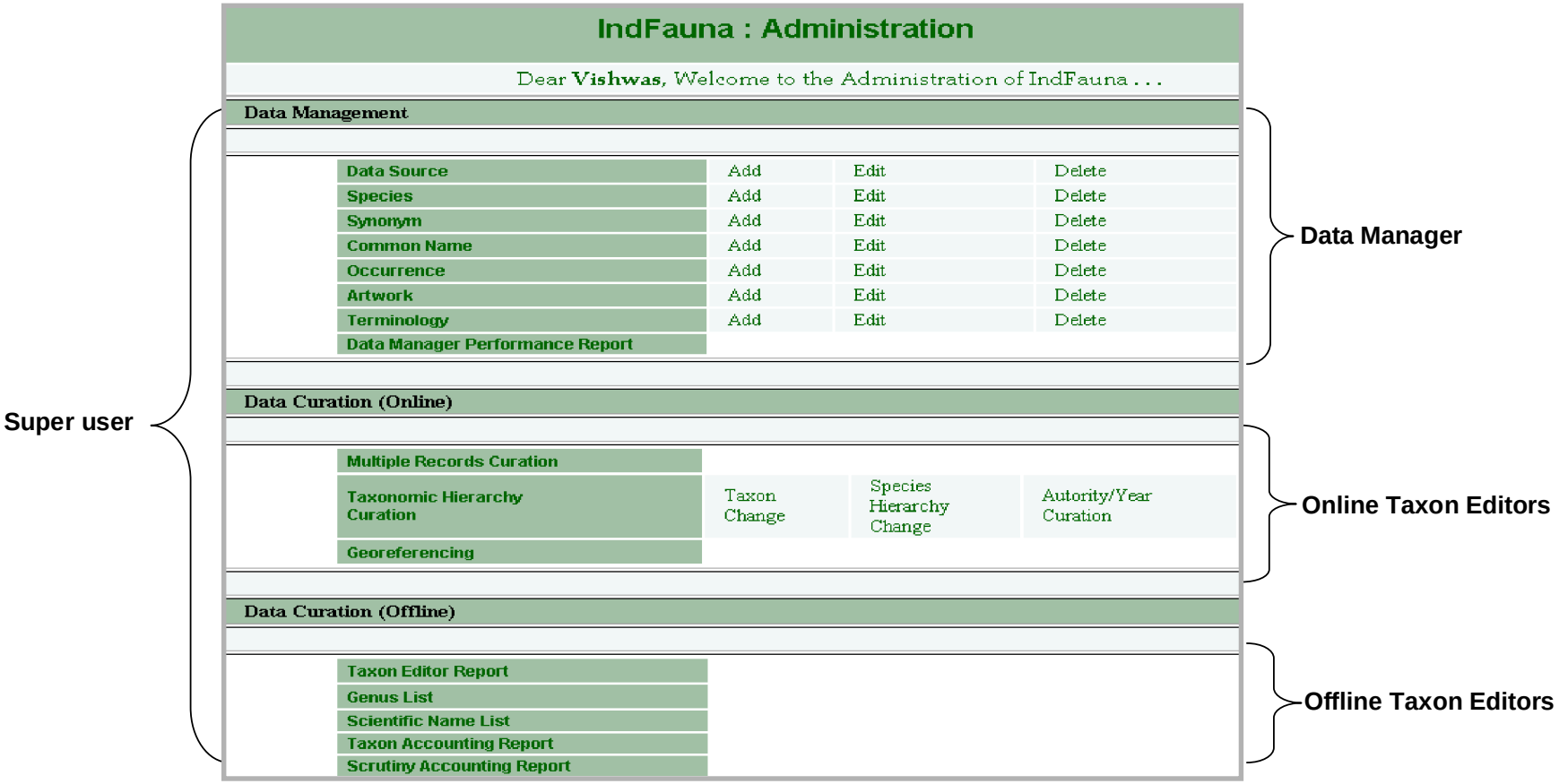


Figure 3.4: IndFauna Data Curation Modules are used by (a) Quality Controllers, (b) Online Taxon Editors, and (c) Offline Taxon Editors

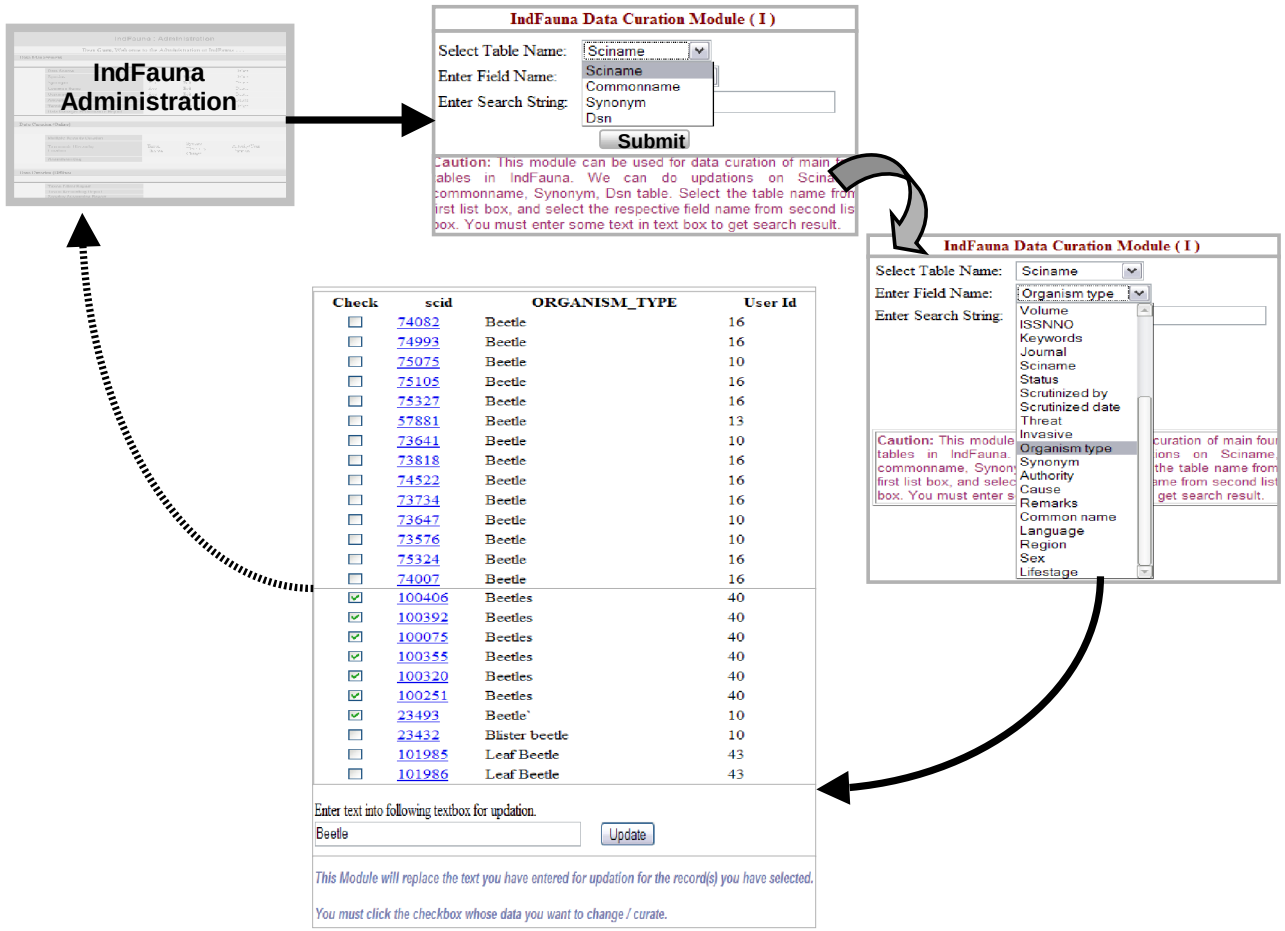


Figure 3.5: Multiple Records Curation Module

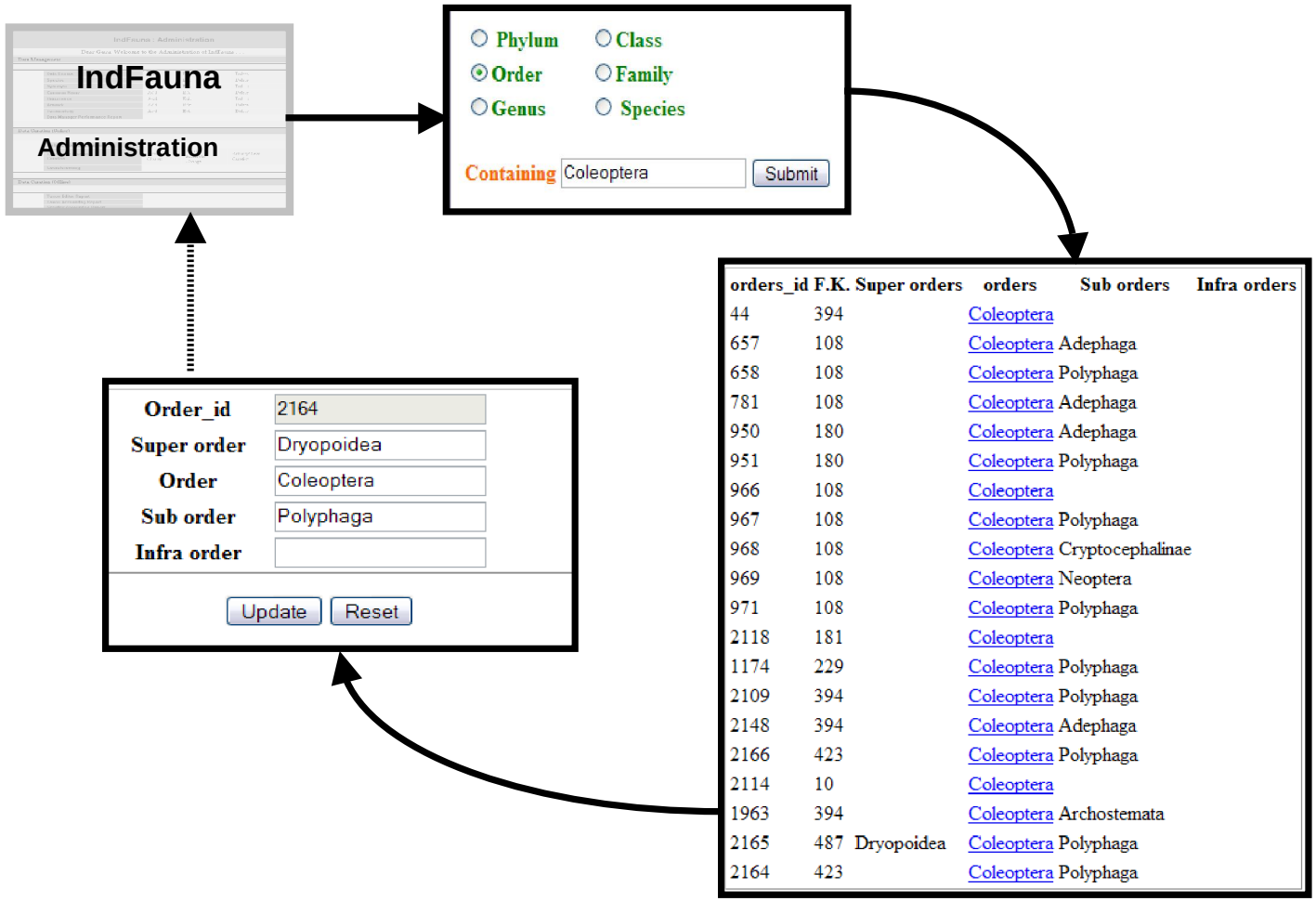


Figure 3.6: Taxon Change Module

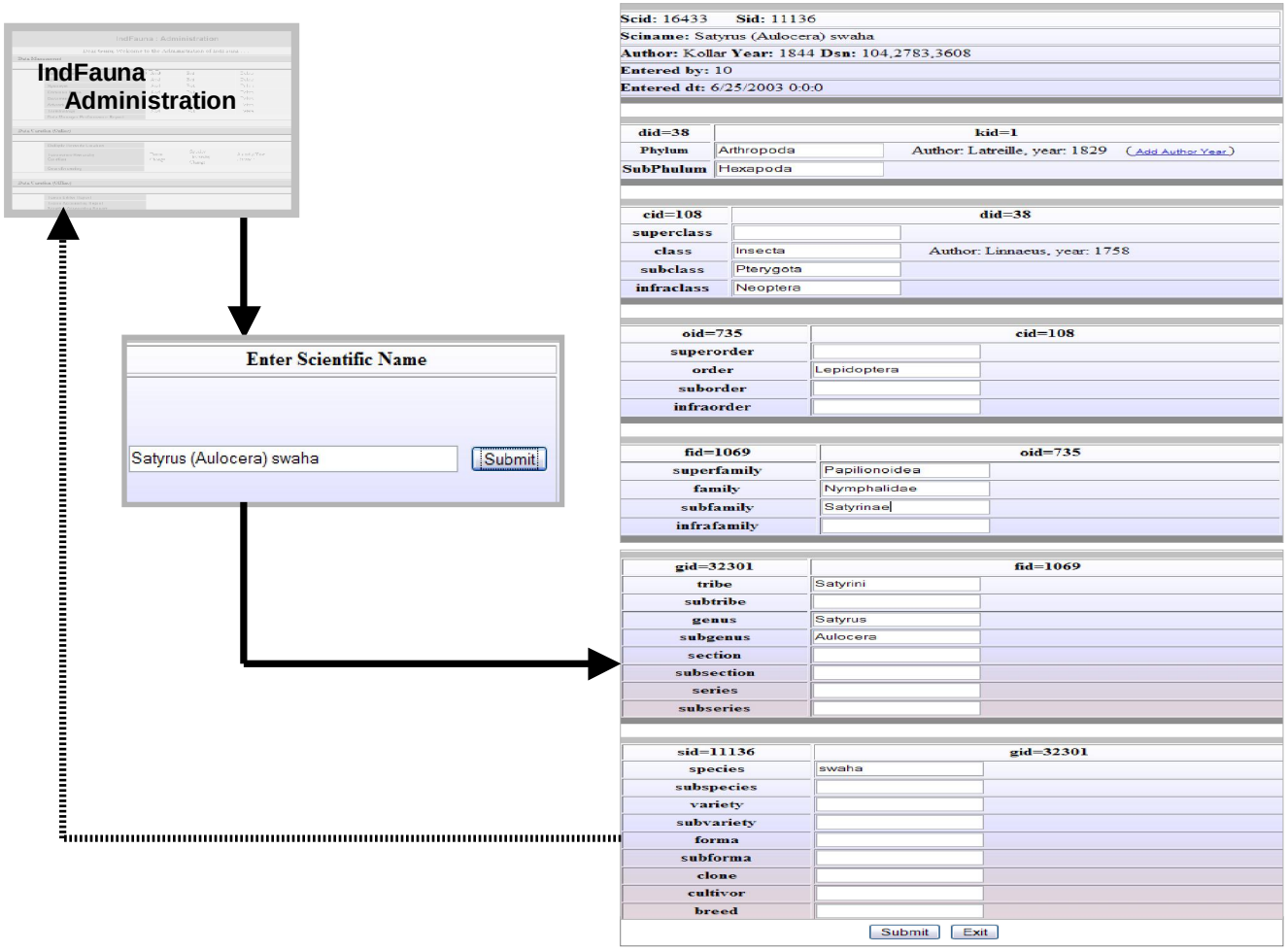


Figure 3.7: Species Hierarchy Change Module

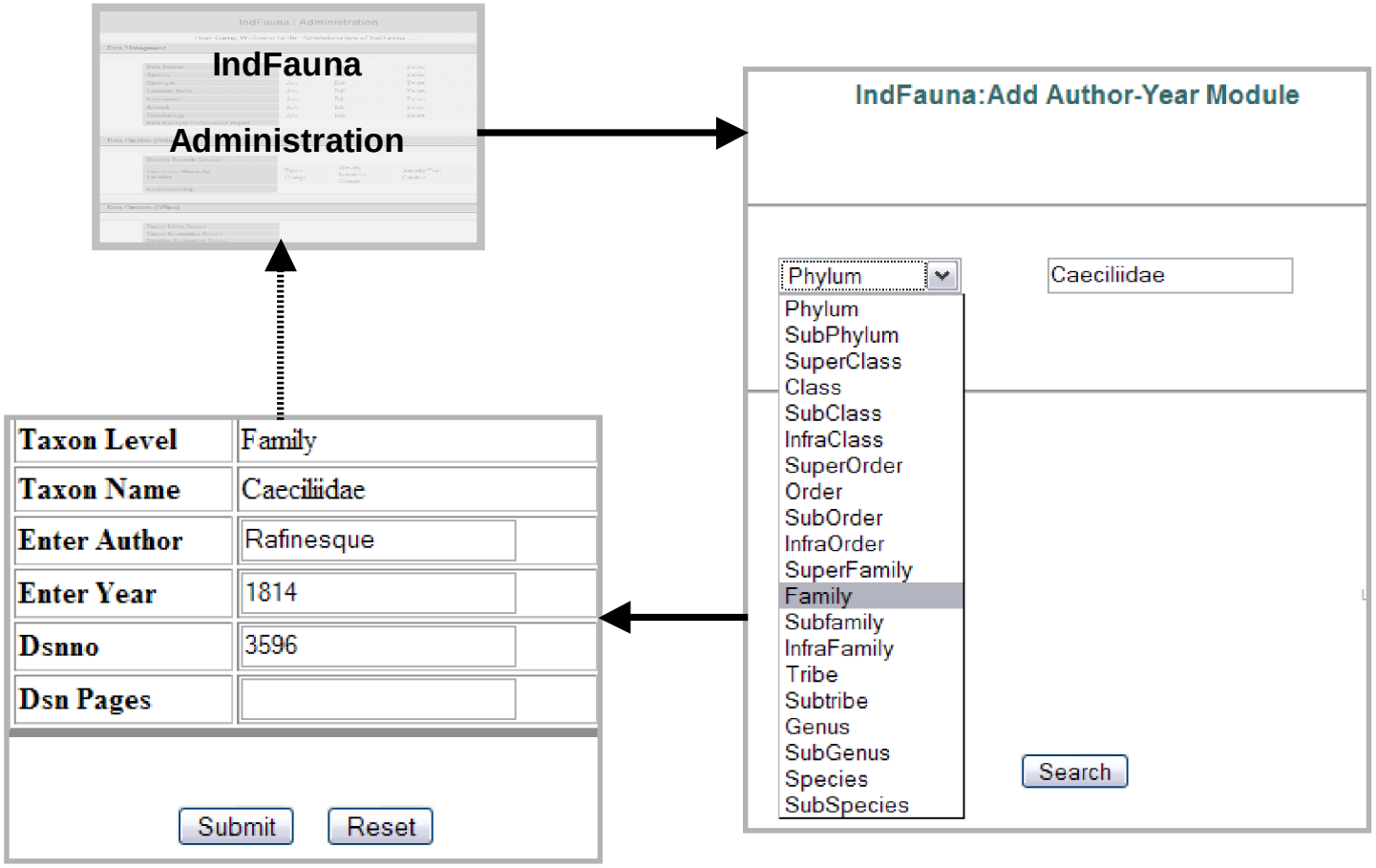


Figure 3.8: Taxon Author/Year Curation Module

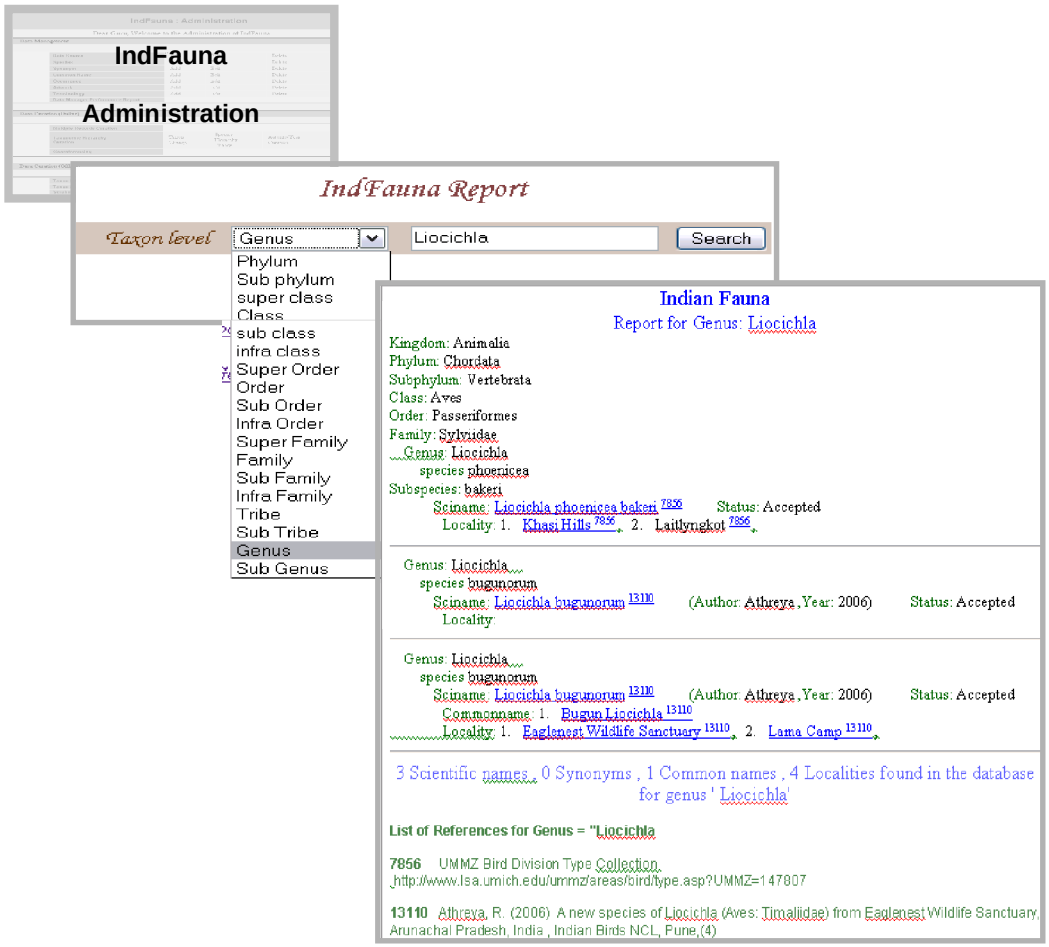


Figure 3.9: Taxon Editors report

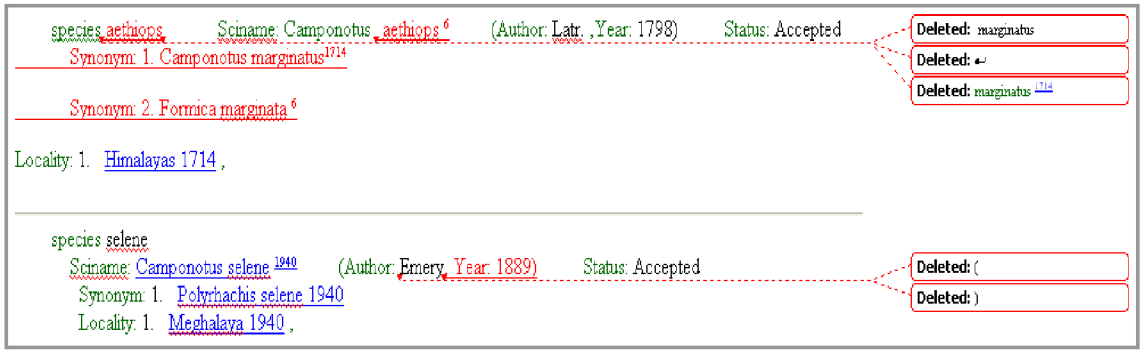


Figure 3.10: Curated Taxon Editors Report

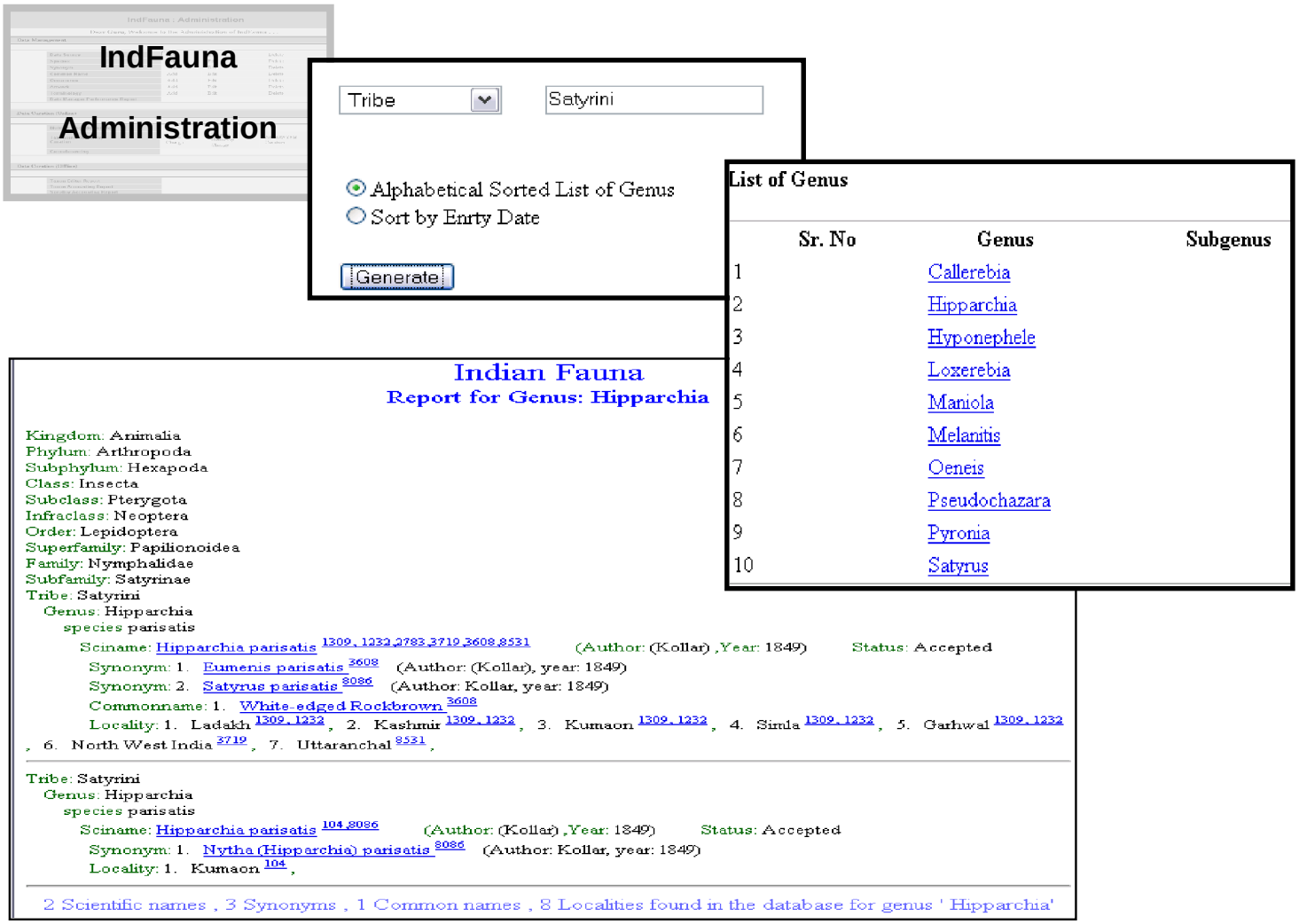
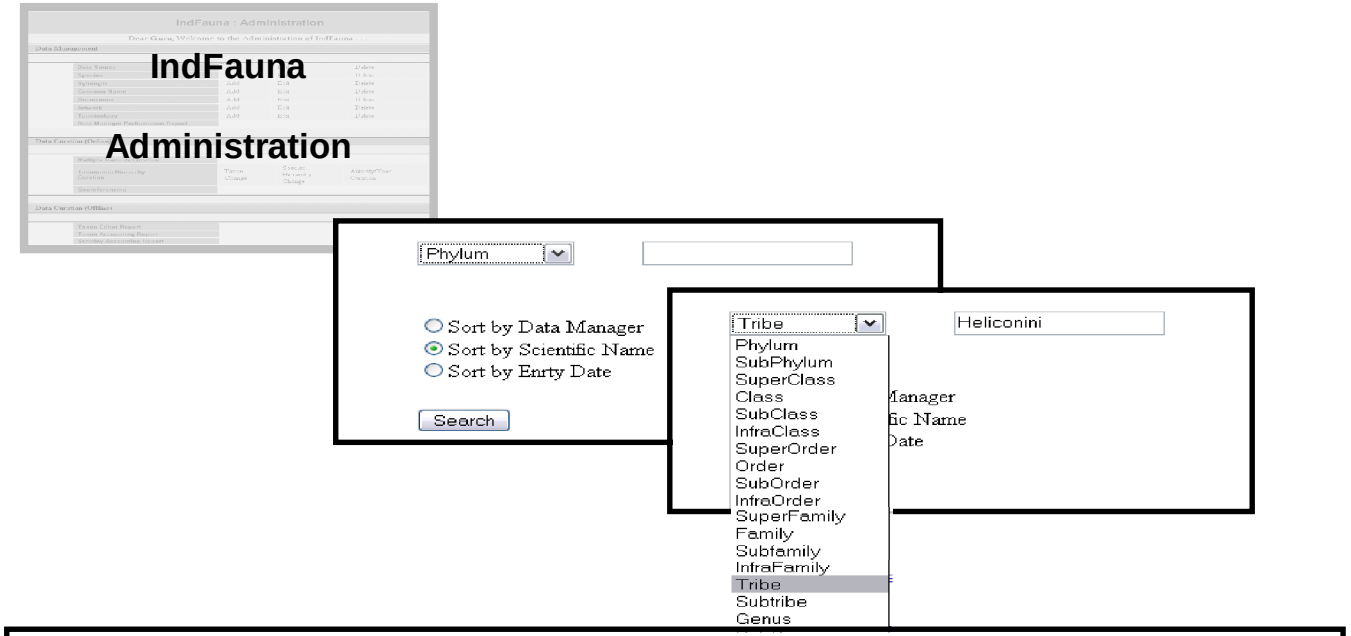


Figure 3.11: Genus specific Taxon Editor Report



List of sciname

Sr. No	Sciname	Author	Year	DSN No	User Id
1	Phalanta alcippoides	(Moore)	1900	3608	10
2	Vagrans macromalayana	(Fruhstorfer)	1912	4264	10
3	Vindula asela	Moore	1872	116	9
4	Vindula erota	(Fabricius)	1793	94,13147	9
5	Vindula saloma	(Swinhoe)	1889	100,3608	11

Figure 3.12: Scientific name list report

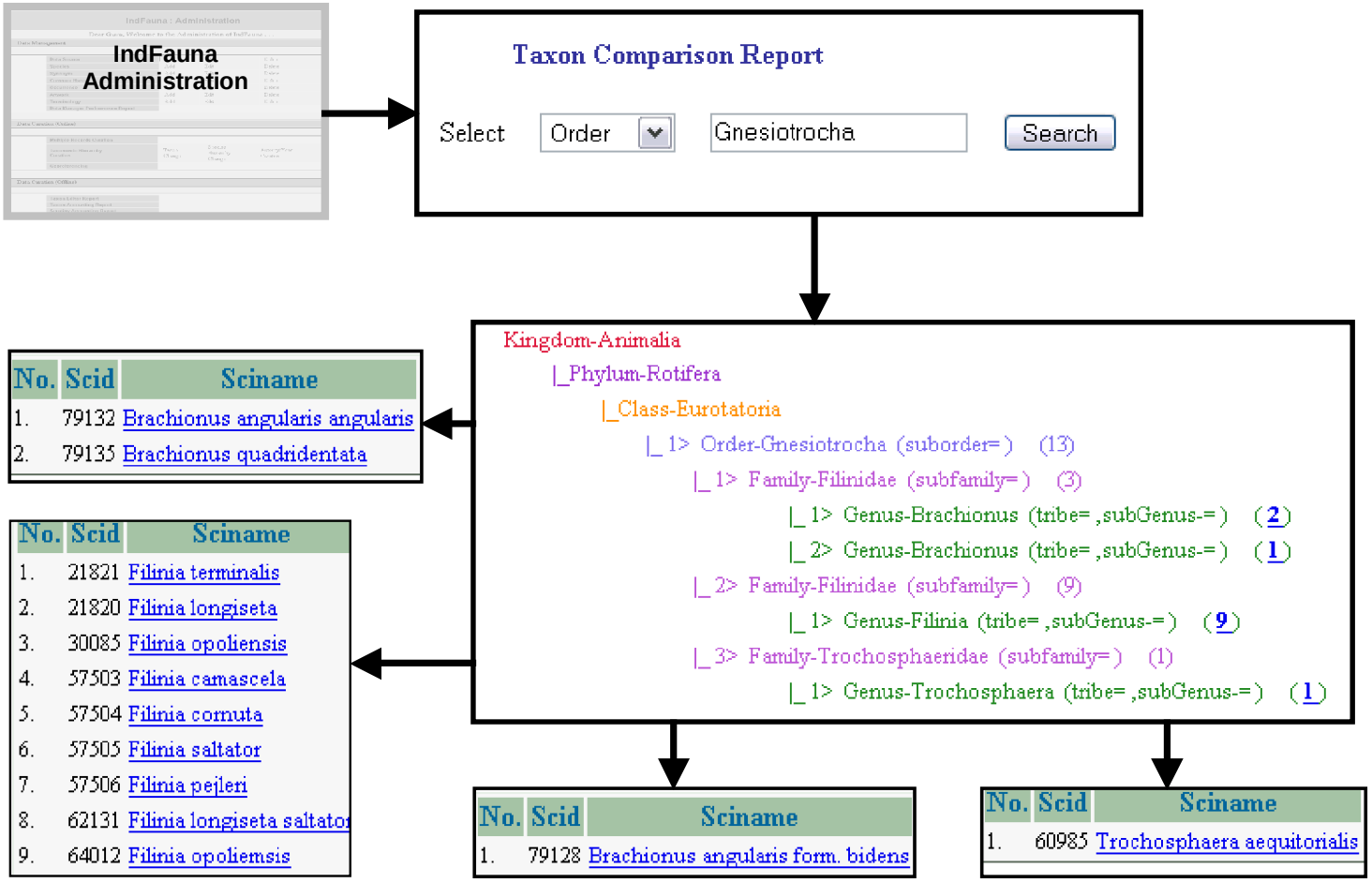
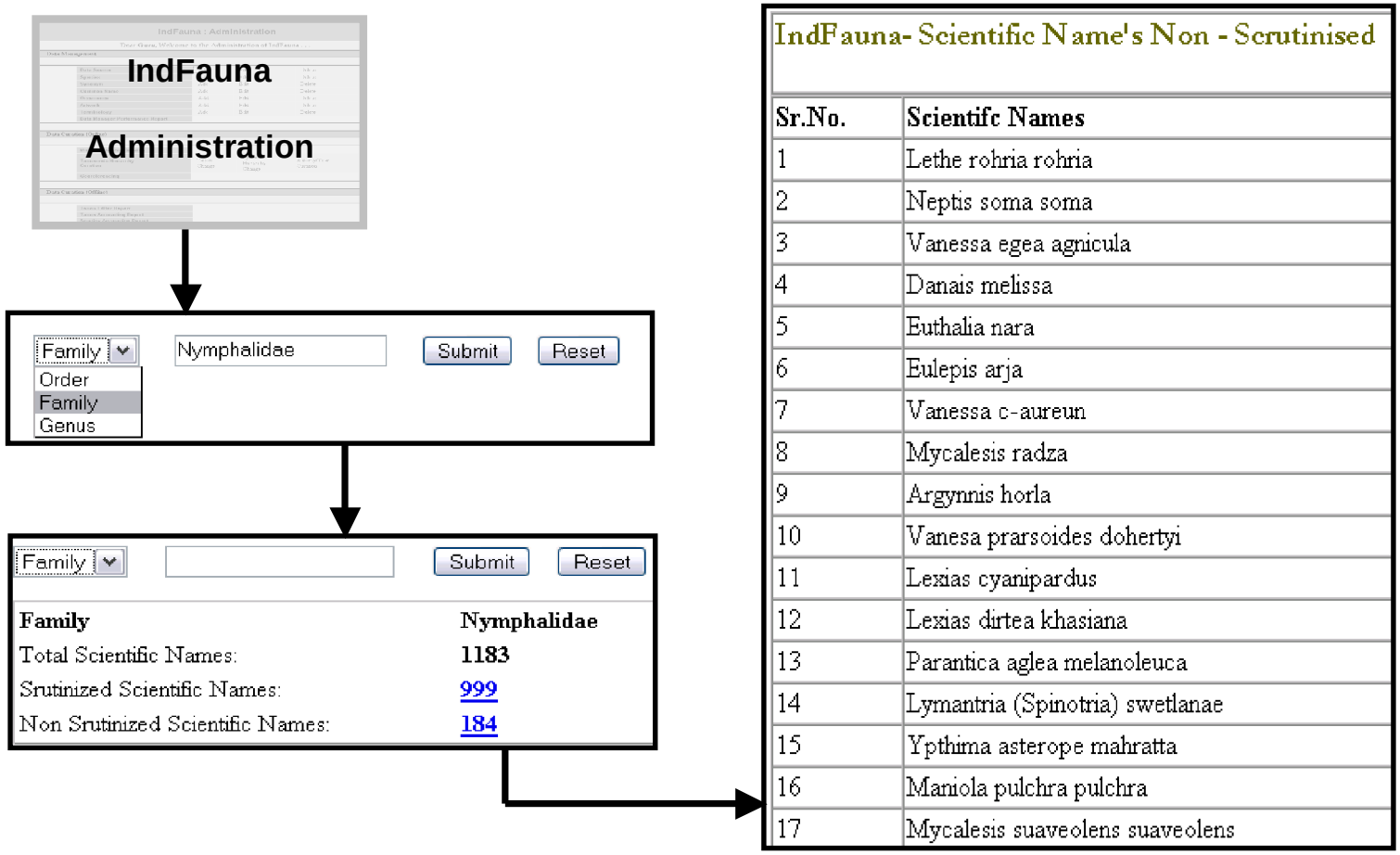


Figure 3.13: Taxon Accounting Report



IndFauna- Scientific Name's Non - Scrutinised

Sr.No.	Scientifc Names
1	Lethe rohria rohria
2	Neptis soma soma
3	Vanessa egea agnicula
4	Danais melissa
5	Euthalia nara
6	Eulepis arja
7	Vanessa c-aureun
8	Mycalesis radza
9	Argynnis horla
10	Vanesa prarsoides dohertyi
11	Lexias cyanipardus
12	Lexias dirtea khasiana
13	Parantica aglea melanoleuca
14	Lymantria (Spinotria) swetlanae
15	Ypthima asterope mahratta
16	Maniola pulchra pulchra
17	Mycalesis suaveolens suaveolens

Figure 3.14: Scrutiny Accounting Report

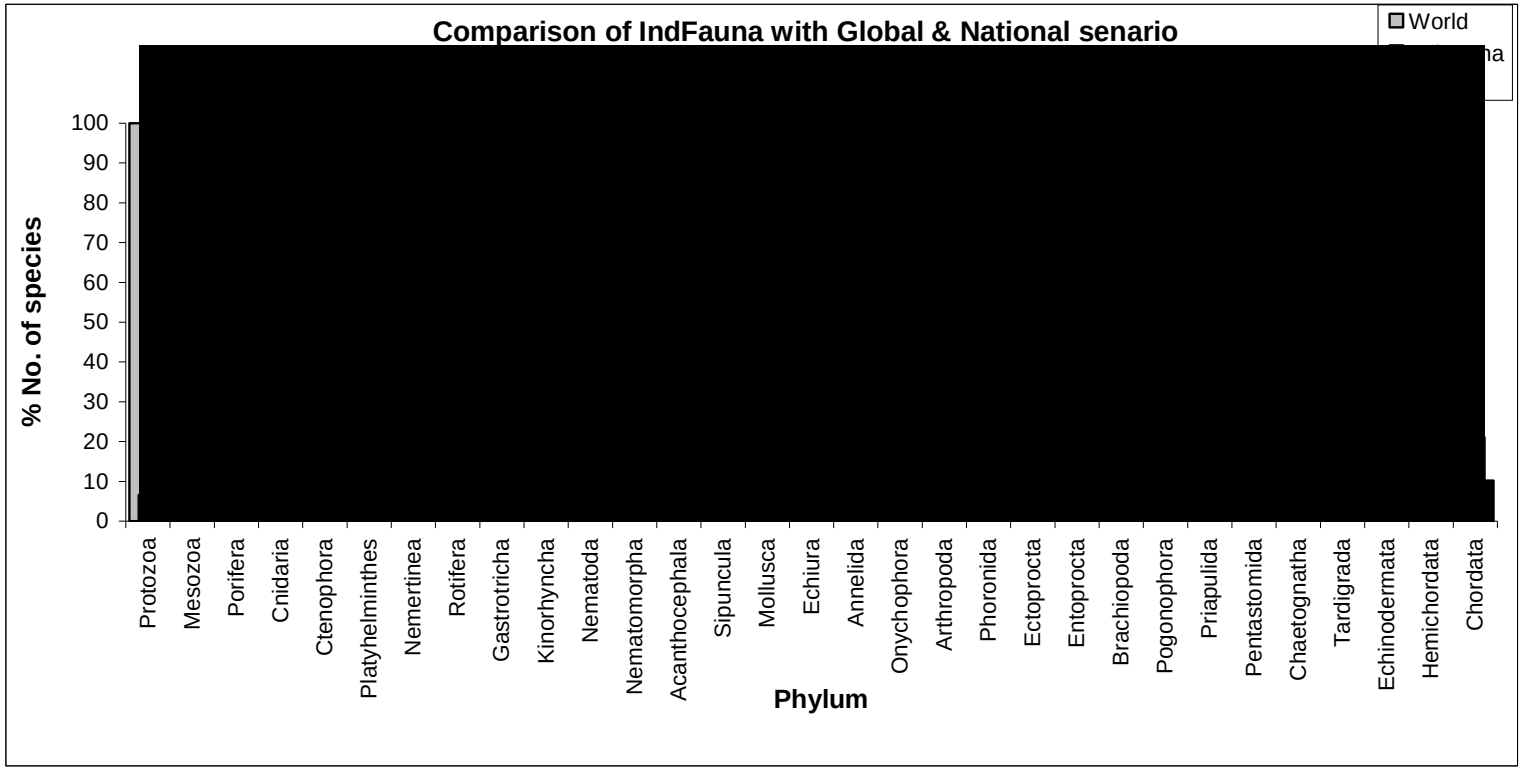


Figure 3.15: Comparison between total number of species per phylum in world, India and IndFauna. (Data on world and India: Alfred 1998)

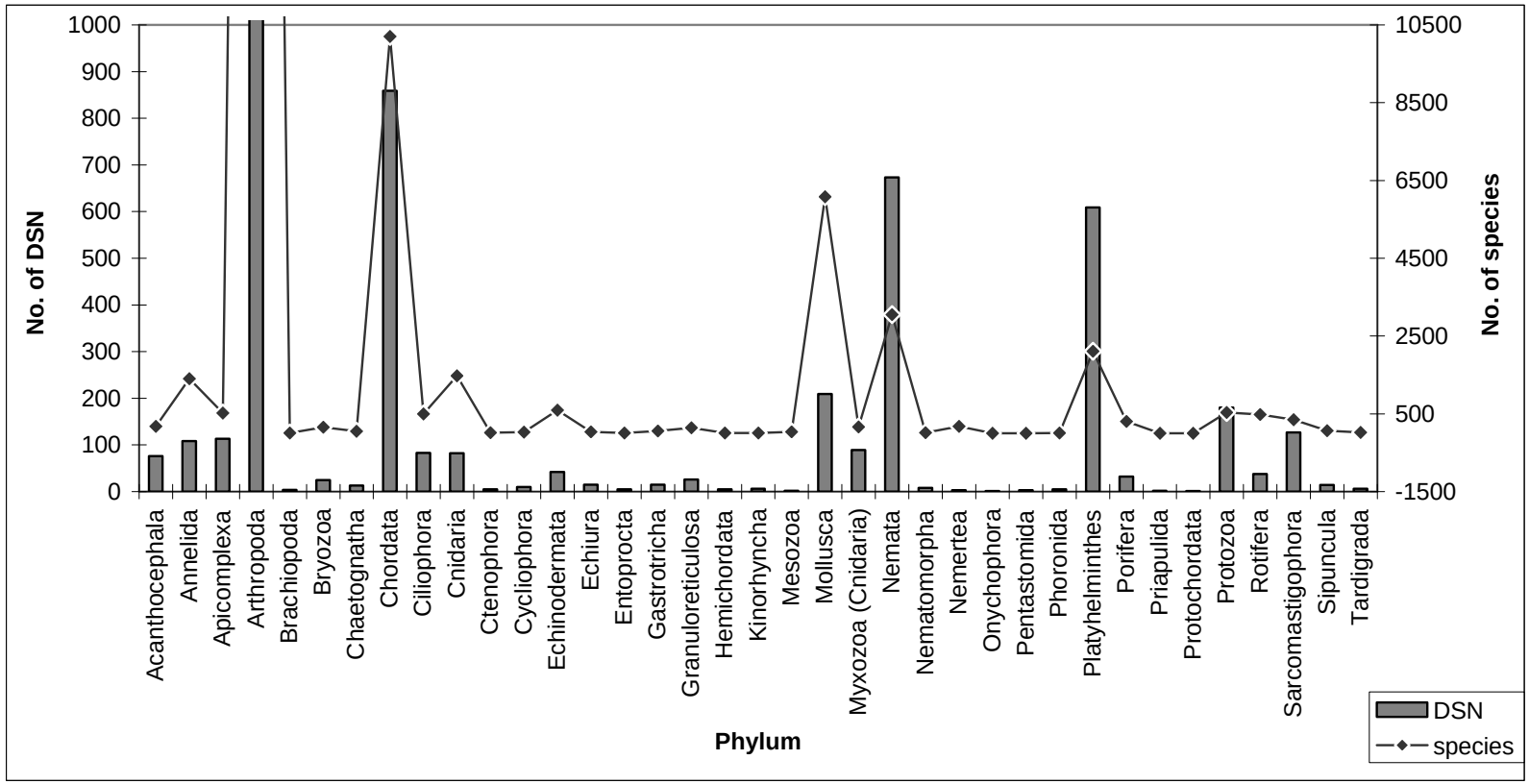


Figure 3.16: Species coverage in comparison with nos. of publications per Phylum.

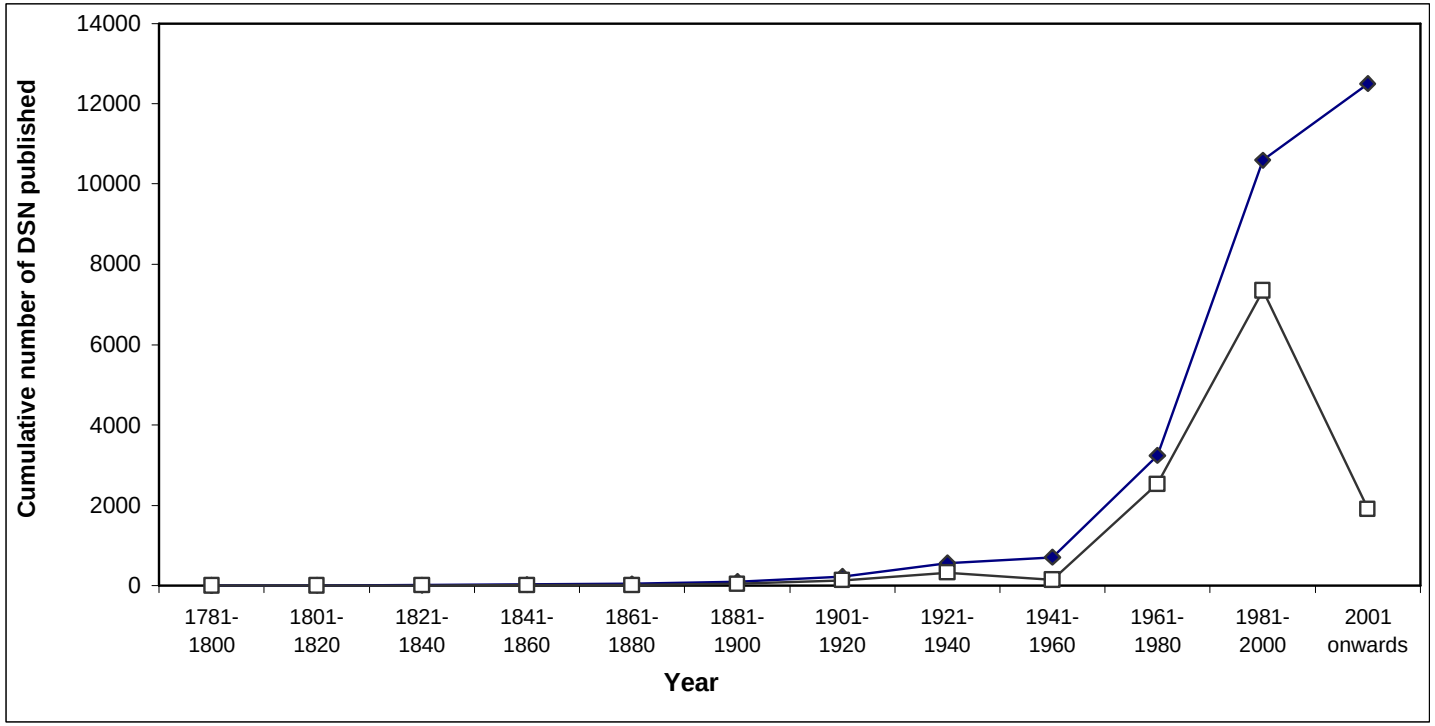
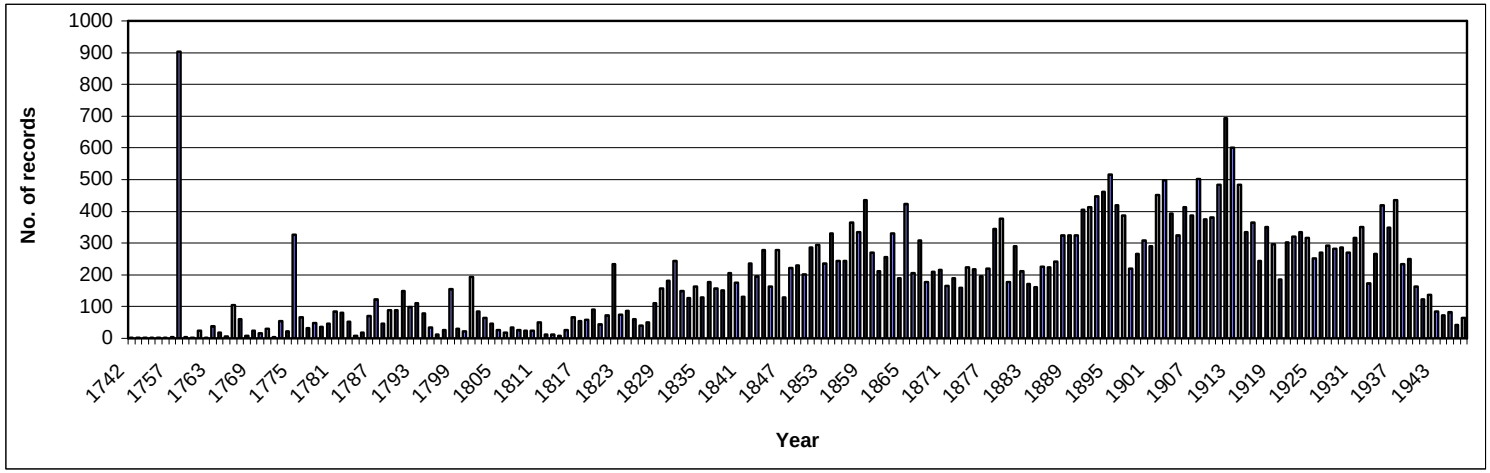
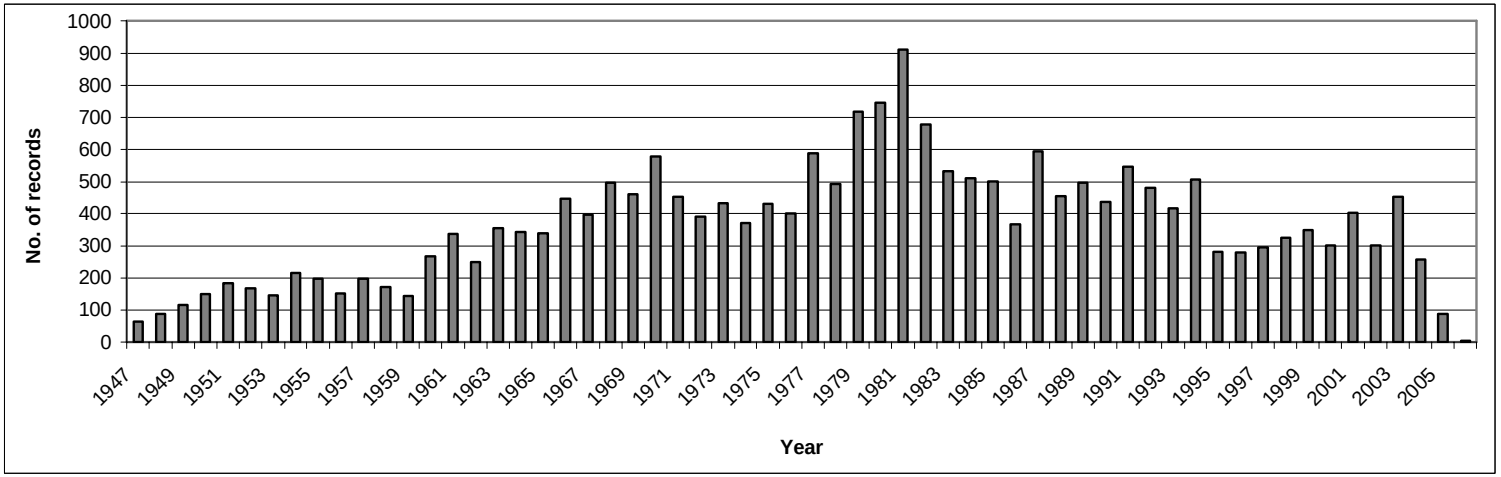


Figure 3.17: Nos. of Taxonomic publications during 1750 to 2007 (n = 12499)



a)



b)

Figure 3.18: The number of species records published per year during a) years 1706 to 1947 and b) years 1947 to 2006.

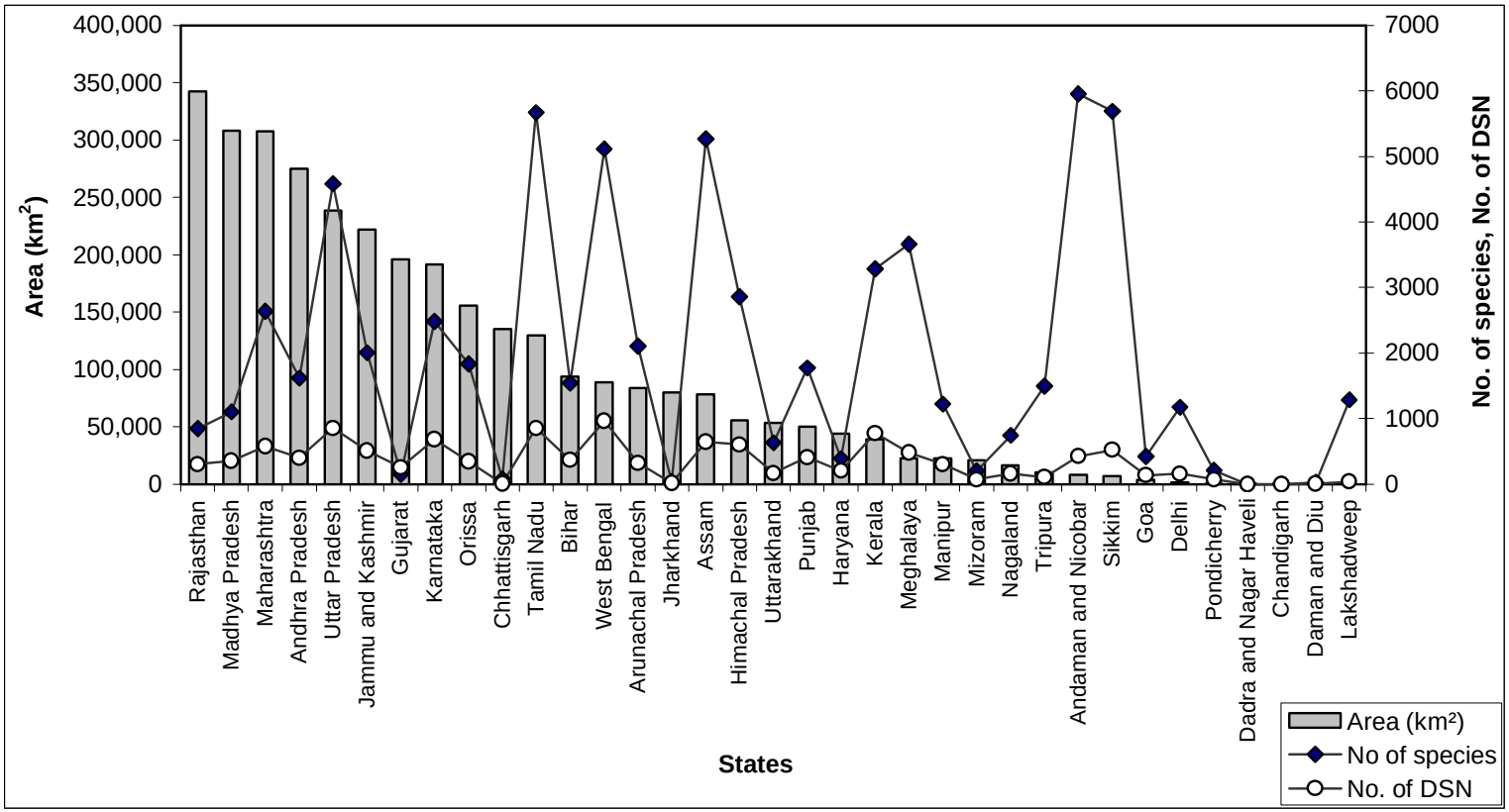


Fig. 3.19 State wise distribution of geographical area, number of species recorded and publications available (Twenty six records for which states are not assigned and 329 records for which the locality is a water bodies from Indian Ocean system are excluded from this analysis).

Chapter 4

JaivaNaksha: Web mapping of Occurrence data and Geo-referencing

Chapter 4

JaivaNaksha: Web mapping of Occurrence Data and Geo-referencing

4.1 Significance of Occurrence data

One of the basic goals of biodiversity databases is to provide occurrence data on species (Salem, 2003), in order to empower decision-making in context of planning, development, and conservation, etc. In chapter 3, I have discussed the various uses of species occurrence data. As described by Chapman (2005a), these uses ranges from biogeographic studies (Longmore, 1986; Peterson et al, 1998), conservation planning (Faith et.al., 2001), selection of protected area (Margules and Pressey, 2000), development of environmental regionalization (Thackway and Cresswell, 1995), climate change studies (Chapman and Milne, 1998; Pouliquen and Newman, 1999; and Peterson et.al., 2002), agriculture, forestry, and fishery production (Booth, 1996; Nicholls, 1997, Peterson and Veiglas, 2001), etc.

In this context, recent development of GPS and GIS technologies appear ideally suited to conservation efforts because they empower ecologists to expeditiously acquire, store, analyze, and display spatial data on organisms and their environment (Johnston, 1998; Wadsworth & Treweek, 1999). However, current spatial data on species-occurrence alone is not sufficient for biogeography and other studies listed above. Infact, it is the legacy data which forms major constituent of such studies. As discussed earlier this legacy species occurrence data is collated from three sources (a) museums specimens, (b) field surveys/observations, and (c) literature. It is in this sense, that database like IndFauna, which collates species occurrence details spreading over 250 years of modern biology literature attains immense significance, at the same time poses challenges to use the data effectively and efficiently.

4.2 Mapping Occurrence data over the Web

Online mapping is the key of biodiversity information system as it allows users to explore the spatial context of biodiversity information visually and assemble quickly the datasets needed to ask and answer biodiversity research and management questions (Guralnick and Neufeld, 2005). Further, online mapping simplifies the process of data exploration and ultimately lowers the cost barrier to analyses that have not been attempted, leading to potential novel research findings and wider use of data by researchers and planners.

There exist several examples on the web related to biodiversity mapping. The MaPSTeDI Geomuse (http://mapstedi.colorado.edu/geomuse.html) project, the Non-indigenous aquatic species (NAS:http://nas.er.usgs.gov/) project of the USGS, the Canadian Biodiversity Information Facility mapper (http://www.cbif.gc.ca), a biodiversity mapping facility developed by the Taiwan GBIF node (http://gis2.sinica.edu.tw), and GBIF-MAPA, the Global Biodiversity Information Facility Mapping and Analysis Portal Application (http://gbifmapa.austmus.gov.au/mapa/) to cite a few (Flemons et al, 2007). A common feature in all these exercises is the treatment of locality data as points. This could be due to the higher quality of biodiversity data (i.e. Availability of point location data) or could also be due to a conversion from polygons to points through mapping of the centroid.

4.3 Geo-referencing Occurrence Data: Why?

There are three challenges in using species-occurrence data for analysis and modeling leading to decision making, (a) capture the information, (b) making distributed information available on the net, and (c) increasing value of the species-occurrence records by converting the textual descriptions of places into their corresponding geographic coordinate – to geo-reference them. Much of the information on species diversity and distributions has neither been digitized, nor georeferenced, exists in different data formats, or cannot be aggregated or integrated with existing computer applications. The number of species-occurrence records worldwide that are not currently georeferenced is astounding; more than 99% of the specimen records (Beaman and Conn, 2003; Beaman et al, 2004 and Guralnick et al., 2006). Currently catalogues which provide valuable information about the species distribution pattern, do not indicate specific locations (Murthy et al., 2003).

This severely limits the degree to which past and current distributions of species can be mapped and analyzed in combination with spatial data from other disciplines (e.g. climatology, geology, geography, social sciences) (Graham et al, 2004). In fact, our ability to use increasingly available occurrence data effectively for ecological research management and conservation depends on how quickly and how accurately we are able to geo-reference these records.

If this textual occurrence data is to be used for analysis, leading to decision making, it needs to be converted into spatial coordinates (Murphey et al., 2004). Geo-referencing of biodiversity data is absolutely necessary for biogeography (Canhos et

al., 2004a) as well other studies. The current situation for geo-referencing is characterized by the following major shortcomings: (a) The process is slow – an average of several minutes per locality, (b) With a few exceptions, the accuracy and precision of assigned coordinates is unknown, (c) A large fraction of available coordinates are demonstrably inconsistent with the rest of the locality information, (d) The materials and methods used are poorly documented, and (e) Many localities are geo-referenced many times over – but not likely with the same results (Graham et al., 2004).

4.4 Geo-referencing Occurrence Data: How?

Use of biodiversity data in biogeographic studies has imposed an extra focus on issue of data quality. Errors are common and are to be expected, but cannot be ignored. Good understanding of errors and error propagation can lead to active quality control and management improvement (Chapman, 2004). With legacy data covering the past and knowing that it can not be replaced with by new surveys even if necessary funds were available, owing to loss of biodiversity and habitat changes and the unique nature of each record. Geocoding historical data can also be very complex, and is an additional potential source of errors (Soberon and Peterson, 2004).

Much has been written about 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) and several techniques are available. Collaborations among biodiversity informaticians are leading to geo-referencing protocols with standardized methods for determining both the spatial coordinates for a location and the error and uncertainty regarding the assigned points. Guide to geo-referencing as a result of MANIS project (http://elib.cs.berkely.edu/manis/GeorefGuide.html), establishes a standard methodology to assign geospatial coordinates to historical locality descriptions. It also establishes a standard means to assign an uncertainty value or maximum error distance associated with geospatial coordinates. 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, not all species distribution data are amenable for conversion to such precision. The challenge then faced is in the visual representation of imprecise data. Imprecise data would thus include references to administrative boundaries, parks, national forests or ecological regions, water bodies etc. Depending on the scale of the application some of these aerial distributions may be represented as points. However in most cases large regions cannot be converted by any means to precisely defined point location data.

4.5 Occurrence data in IndFauna: Characteristics

IndFauna has so collated over 176444 occurrence records for 6577 unique localities. Most of these records have been collected from secondary literature sources (Chavan et al., 2005b), and this majority of them are in the form of textual descriptions. These ranged from general descriptions like ‘Throughout India’ to ‘Maharashtra’ (state names) to ‘Sindhdurg district’ (district name) to point locations (villages, towns). Apart from such textual descriptions, there also exist descriptions to both arbitrary and precisely defined regions, examples of which include ‘Southern India’ for the former and ‘Thar Desert’ or ‘Rajaji National Park’ for the latter. The third type of location data that was encountered was in the form of river/water body names including lakes, mangroves, lagoons and estuaries. Thus, these 6577 unique localities could be categorized into seven types, viz. (a) points, (b) rivers or streams, (c) water bodies, (d) areas, (e) districts, (f) states, and (g) country. Table 4.1 lists representative examples of these occurrences. These occurrence types could be geometrically represented as points, lines, and polygons. However, considering the coarse resolution of textual occurrence records, most of them would needed to be represented as polygons. Subsequent sections describe the approach adopted for mapping of this imprecise data, along with approach for dynamic rendering of maps for each species.

4.6 JaivaNaksha

4.6.1 JaivaNaksha Schema: Considerations

Considering the unique characteristics of IndFauna occurrence data (Section 4.5), and to facilitate mapping of imprecise locations, JaivaNaksha schema was planned to draw from set of distinct locations and their corresponding polygons. This is because these locations can not be depicted as precise point locations (e.g. Country, State, District, Water bodies, ecological and geomorphic regions such as Western Ghats and deserts, etc.) because of the vast geographic coverage that they represent.

Uncertainty in describing locations of was another consideration that influenced the design of JaivaNaksha schema. For instance, when it is reported that species occurred at particular location, e.g. Western Ghats, it could have been one of the following scenario. (a) Entire Western Ghats was surveyed systematically following survey conventions and standardized methods, and species was found to be occurring every where within the Western Ghats, (b) Only part of the Western Ghats were surveyed, however, author considering that similar habitat exists in other regions of Western Ghats, concluded that it is present in entire Western Ghats, and (c) Inaccurate description of location or generalization by author could also lead to locality being mentioned as Western Ghats. However, my principal was to represent localities on "as is" basis, without tampering or altering what they denote. Thus, it was imperative to represent these occurrence records as polygons, and not to denote them with points and ranges. To resolve and represent the complexity two tables were designed, (i) one containing a distinct set of administrative boundaries for each administrative level with unique identification number, administrative boundary name was devised, and (ii) another containing distinct set of geomorphic, ecological, and arbitrarily defined regions having a unique identifier, description (name) and a geometry column.

In case of precise occurrences, rather than having set of distinct localities, point location data was populated for each species. As adopted by MAPSTEDI (Murphy et al., 2004), such an approach has two reasons, (a) it is extremely difficult to maintain a set of over million localities, and (b) there are many localities with same name, but geographically distinct.

4.6.2 JaivaNaksha Schema: Structure

Closer analysis of IndFauna's database structure revealed that concerns and considerations explained in previous section could not be addressed by using existing structure. Thus, it was felt appropriate to design schema specific to fit the purpose of JaivaNaksha. JaivaNaksha schema consists of 14 tables (Table 4.1) of which 6 are intermediary tables for country, state, district, river, water body, and point localities respectively (Table 4.2 – 4.14). In addition to this there are two tables one for shapefiles (Table 4.15) and contributor of the shapefiles (Table 4.16). Relationships between entities are depicted in Fig. 4.1, where in data flow within and outside of the JaivaNaksha schema is shown in Fig. 4.2.

4.6.3 JaivaNaksha - Spatial data management

In order to curate the textual locality descriptions and to maintain accuracy and precision of geospatial data, "Geo-referencing module" was developed as part of IndFauna data cleaning process. Thus the objective of this module is to georeference textual descriptions in IndFauna, and stores them as spatial objects in Oracle9i, which has been used as RDBMS for IndFauna.

As depicted in Fig. 4.4, it necessitated searching textual locality descriptions for a given scientific name which are stored in sci_loc table. Georefrencing module facilitates proceeding for further geo-referencing using a decision tree process. First and foremost decision that needs to be taken is weather a textual description is a precise locality (i.e. point), river, water body, area covering entire country, state, district or geomorphic or ecological features (e.g. Thar desert, or Western Ghats), or historic geopolitical/administrative area (Bombay Presidency, which now consists of several western states). In order to rectify the errors that might happen in such a decision making process, an option was revoking the decision and revising it was incorporated in the module. In order to prevent duplication of efforts by freshly creating shapefiles of polygons representing these occurrences, it was felt to use web as collaborating platform, by creating facility which could be used to develop repository of such commonly used shapfiles. Thus, OAGDR (Open Access Geospatial Data Repository) was devised. Details of OAGDR are described in Appendix II.

Thus, descriptions which are smaller in area compared to average districts are considered as point localities and their geo-coordinates were determined as those of centorid of the polygon that could best representing such a textual description. For instance, if the textual description is "Matheran", then geo-coordinates of the centroid of Matheran have been adopted as geo-coordiantes representing Matheran. Such geo-coordinates are either searched through GEONET gazetteer linked with this module, or separately determined. Since, such a decision making process

For rapid geo-referencing of localities such as states and country, feature of simultaneous geo-referencing was developed. This facilitated curation of over 50,000 out of over 178,000 occurrence records within a day, and therefore easing the process of transition from text to map.

4.6.4 JaivaNaksha: Tools and Development

IndFauna is developed using Oracle 9i as back end which led to a use of both proprietary and open source tools for developing JaivaNaksha. As detailed in Fig.

4.3, three tier approaches was adopted to evolve infrastructure of JaivaNaksha, viz. client or user interface, application server, and database server. Oracle 9i formed back end for database development. Java Server Pages were used to develop data cleaning module “geo-referencing” (Fig. 4.4).

(A) Application Server

Apache (http://www.apache.org/) along with PHP was used as application server. Another important component used for application server is MapServer. Developed by the University of Minnesota, MapServer, is a CGI (Common Gateway Interface) program and performs job of rendering maps based on vector and/or raster data. The vector data may be stored as native shape files (for vector data) or in a database that supports storing of spatial objects. Since, IndFauna has been developed using Oracle9i, Oracle Spatial was used to store shapefiles. Mapserver acts as an intermediate that receives parameters from the web server and processes data accessed from the database.

The parameters that are required to define the map image output are stored in a 'map' file. The map file is a text file that contains numerous configuration parameters such as the map extent, layer definitions, symbol and label definitions and classes along with details on legend and scale. Although these parameters may be stored as a static map file, JaivaNaksha demanded that the mapfiles be created dynamically, as species distribution map of any one of the 94000+ species needs to be served dynamically.

In order to enable this PHP Mapscript was employed, which allows access to Mapserver's MapScript functions and classes available in a PHP environment. Along with PHP Mapscript, by using the 'Scientific ID' (unique species identifier of IndFauna) as a session variable, the desired dynamism for rendering maps for any number of species was achieved. Session can be defined as a period of time in which a user can communicate with a server. Hence session variables may take different values, one at a time during a session. Also, every client may have different values for the same server. This allows the server to service multiple sessions, each session uniquely identified by the session identifier.

(B) User interface

User interface for web mapping was developed using “pmapper”, an open source client available at http://sourceforge.net/projects/pmapper. Choice of “pmapper” among other considerations was influenced because of its compatibility

with range of bandwidths, ranging from dial-up to broadband connection. In order to develop this user interface Javascript, DHTML and PHP was used, which provided a broad framework for developing applications using Mapserver and PHP / Mapscript.

The design of the web mapping client was based on several issues. First and foremost was its usability and availability of features essential for viewing the species distribution maps. Key features that were deemed mandatory include (a) ability to zoom in and out and zoom to the full extent, (b) map panning and panning using the index map, (c) layer management (ability to turn on/off layers), and (d) Biogeography report generation. These features were thought mandatory considering that IndFuana users fall in two categories viz., general users and taxonomic or research community.

4.7 Lessons Learnt

Apart from providing data to policy makers, biodiversity data is also sought by biologists and the common man alike. In order to meet the common requirements of this diverse user-base, JaivaNaksha was developed as a value addition to IndFauna (Fig. 4.5), and well to the data collated and disseminated by www.ncbi.org.in. While achieving this, various aspects of database design, quality of user-interface and spatial data representation were addressed, some of which has been discussed in preceding sections. Here, I would focus on some of the lessons learned and fututeristics of web mapping on species occurrence data.

4.7.1 Do polygons represent right status of species occurrence?

Experience of JaivaNaksha development was unique as well differ from similar exercise. Out of 170000+ occurrence records that have been collated by IndFauna, over 95% of them are textual descriptions of localities. Closer analysis of these records reveled that it would be difficult to georeference them to a degree of precision so that precise geo-coordinates could be assigned. Thus, it was decided to maintain “as is” status of these occurrence records and map them as polygons in JaivaNaksha. Arguments were made that representing them as polygons would not serve any purpose as well add value to IndFauna. However, I believe that an approach of not converting polygons to points by using a centroid and range circle approach for large administrative regions, and geomorphological, ecological geopolitical, as well historic administrative areas is the best way of representing such localities. Such a data may not be of direct use in analysis, however, it help directing the future survey and conservation activities as it certainly sensitize about potential areas from where a specific organism could be recorded. Further, if these coarser occurrence records such

as “throughout India” are subjected to geo-referencing to assign geo-coordinates to them following best practice guidelines, they would certainly lend different interpretation to the plotted data.

4.7.2 Sharing coarse resolution occurrence records

It was also observed that “as is” philosophy of representing occurrence records, also leads to complications in data sharing, as current XML schemas require that occurrence records should be geo-referenced with each assigned with geo-coordinates. This excludes billions and probably many trillions of such coarser observation records being in web mapping exercise, thus depriving potential users to sensitize themselves with potential areas of presence of these organisms. Therefore, it is imperative that future XML schemas should be able to exchange/share coarser occurrence records.

4.7.3 Occurrence records base for web based spatial decision support systems

The World Wide Web provides an apt infrastructure for representing and sharing spatial information – anytime, anywhere (Dragicevic, 2004, Kraak, 2004). This has been a prime reason for the spurt in the development of decision support systems (DSS) and/or spatial decision support systems (SDSS), the latter having a strong geographic component. Such web based SDSS have been developed for health applications, potato crop management (Sante et al., 2004) and wetland management (Mathiyalagan et al., 2005), to name a few.

In current its form JaivaNaksha is quite away from being treated as SDSS, as the focus of current exercise is to provide visualization support for occurrence data. However, I foresee possibilities of extending the functionality of JaivaNaksha to include features that would aid in the development of a SDSS. Such features would necessarily include (a) integration of species distribution data with environmental layers, (ii) Creation of buffer zones, (iii) species diversity related statistics at District, State, and well geomorphic, ecological areas. Together with this also modeling tools such as those used for ecological forecasting could also be coupled.

4.7.4 Impediments in sharing geospatial occurrence data and products

One of the reason that IndFauna used secondary literature for collating species occurrence data as there are unfortunate, and un-necessary stigmas attached with sharing species occurrence data, as well geospatial data products. This was evident from several requests made to various research groups who use shapefiles of some of the common areas such as “Western Ghats”, and many water bodies. Thus OAGDR

was developed to provide platform for sharing commonly used shapefiles. However, irrespective of well recognized need for sharing geospatial data (Ramachandran, 2000), research groups, especially those working in the area of biodiversity and ecology needs to come forward to exchange / share species occurrence data and geospatial products of such data sets.

4.8 Recommendations

- Geo-referencing of legacy data is essential and needs encouragement.
- Easy to use tools for geo-referencing are need of the hour, which can also expedite the speed of geo-referencing.
- Geo-referencing needs to carried out on “as is” basis than altering and manipulating the scale of locations.
- Exchange / sharing of geo-spatial products needs encouragement, and thus Open Access Geospatial data repository is essential.

4.9 Summary

Biodiversity databases primarily collate and disseminate species occurrence data, which has varied uses and empower sustainable development and conservation related decision making processes. Geospatial technologies significantly contribute such initiatives. However, vast amount of species occurrence data is not geo-referenced. Geo-referencing of legacy occurrence data poses challenges, and is time consuming process. However, initiates such as BioGeoMancer facilitate conversion of textual occurrence records into geo-coordinates. In recent past, several national, international and regional species occurrence web mapping such as MAPSTEDI, and GBIF MAPA has been initiated. However, owing to difficulties in geo-referencing of legacy occurrence data, such web mapping of species occurrence data initiatives are not initiated in many maga-biodiversity countries such as India.

JaivaNaksha, a web mapping application was developed for to represent IndFauna species occurrence records over web. Since, majority of IndFauna collated occurrence records are coarse in their resolution, adopting “as is” philosophy they were represented as polygons. JaivaNaksha was developed using proprietary and open source tools. In order to facilitate exchange / sharing of geospatial products such as shapefiles, OAGDR was devised.

Table Name	Description	Relationship with other tables
Sciname_Loc	Intermediary table for fetching scientific names and occurrences from IndFauna schema	Many-to-Many
Country	List of Countries	
Country_Sci	Intermediary table between scinames and countries	Many-to-Many
State	List of states	
State_Sci	Intermediary table between states and sciname	Many-to-Many
District	List of districts	
Disctrict_Sci	Intermediary table between districts and sciname	Many-to-Many
River	List of rivers, streams, estuaries	
River_Sci	Intermediary table between rivers and sciname	Many-to-Many
Waterbody	List of water bodies (Sea, Oceans, Lagoons, Lakes, dams, etc.)	
Waterbody_Sci	Intermediary table between waterbody and sciname	Many-to-Many
Gazetter	List of geo-referenced point localities	
Pnt_Loc	Intermediary table between point and locality	Many to Many
Shapefile	Shapefiles	
Contrfibutor	List of Shapefile Contributors	

Table 4.1: Tables in JaivaNaksha Schema along with their purpose and relationships

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SCI_ID	NUMBER(10,0)	Yes	1	
STATUS	VARCHAR2(10 BYTE)	Yes	2	
DATE_MOD	DATE	Yes	3	
CURATED_BY	VARCHAR2(30 BYTE)	Yes	4	
LOCALITY_ID	NUMBER(15,0)	Yes	5	
DSN_NO	NUMBER(6,0)	Yes	6	
LOCALITY_TYPE	VARCHAR2(15 BYTE)	Yes	7	

Table 4.2: Table for Sciname_Loc (intermediary table for fetching scientific names and occurrences from IndFauna schema, M-M)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
COUNTRY_ID	NUMBER(3,0)	No	1	1
COUNTRY_NAME	VARCHAR2(50 BYTE)	Yes	2	
COUNTRY_GEOM	SDO_GEOMETRY	Yes	3	

Table 4.3 Table for Countries (List of countries)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SCI_ID	NUMBER(10,0)	Yes	1	
COUNTRY_ID	NUMBER(3,0)	Yes	2	
DSN_NO	NUMBER(6,0)	Yes	3	
LOCALITY	VARCHAR2(50 BYTE)	Yes	4	

Table 4.4 Table Country_Sci (intermediary table between scinames and countries, M-M)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
STATE_ID	NUMBER(3,0)	No	1	1
STATE_NAME	VARCHAR2(50 BYTE)	Yes	2	
STATE_GEOM	SDO_GEOMETRY	Yes	3	
COUNTRY_NAME	VARCHAR2(50 BYTE)	Yes	4	

Table 4.5 Table States (List of States)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SCI_ID	NUMBER(10,0)	Yes	1	
STATE_ID	NUMBER(3,0)	Yes	2	
DSN_NO	NUMBER(6,0)	Yes	3	
LOCALITY	VARCHAR2(50 BYTE)	Yes	4	

Table 4.6 Table States_Sci (intermediary table between states and sciname, M-M)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
DIST_ID	NUMBER(3,0)	No	1	1
DIST_NAME	VARCHAR2(50 BYTE)	Yes	2	
DIST_GEOM	SDO_GEOMETRY	Yes	3	
STATE_NAME	VARCHAR2(50 BYTE)	Yes	4	

Table 4.7 Table Districts (List of Districts)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SCI_ID	NUMBER(10,0)	Yes	1	
DIST_ID	NUMBER(3,0)	Yes	2	
DSN_NO	NUMBER(6,0)	Yes	3	
LOCALITY	VARCHAR2(50 BYTE)	Yes	4	

Table 4.8 Table District_Sci (intermediary table between district and sciname, M-M)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
RIVER_ID	NUMBER(3,0)	No	1	1
RIVER_NAME	VARCHAR2(50 BYTE)	Yes	2	
RIVER_GEOM	SDO_GEOMETRY	Yes	3	

Table 4.9 Table Rivers (List of Rivers)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SCI_ID	NUMBER(10,0)	Yes	1	
RIVER_ID	NUMBER(3,0)	Yes	2	
DSN_NO	NUMBER(6,0)	Yes	3	
LOCALITY	VARCHAR2(50 BYTE)	Yes	4	

Table 4.10 Table Rivers_Sci (intermediary table between Rivers and Sciname, M-M)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
WB_ID	NUMBER(3,0)	No	1	1
WB_NAME	VARCHAR2(50 BYTE)	Yes	2	
WB_GEOM	SDO_GEOMETRY	Yes	3	

Table 4.11 Table Waterbody (List of water bodies)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SCI_ID	NUMBER(10,0)	Yes	1	
WB_ID	NUMBER(3,0)	Yes	2	
DSN_NO	NUMBER(6,0)	Yes	3	
LOCALITY	VARCHAR2(50 BYTE)	Yes	4	

Table 4.12 Table Waterbody_Sci (intermediary table between waterbody and Sciname, M-M)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
LOCN_NAME	VARCHAR2(50 BYTE)	Yes	1	
LATITUDE	NUMBER	Yes	2	
LONGITUDE	NUMBER	Yes	3	

Table 4.13 Table Gazetteer (List of gazetters)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SCI_ID	NUMBER(10,0)	Yes	1	
PNT_NAME	VARCHAR2(50 BYTE)	Yes	2	
PNT_GEOM	SDO_GEOMETRY	Yes	3	
DIST_NAME	VARCHAR2(50 BYTE)	Yes	4	
DSN_NO	NUMBER(6,0)	Yes	5	

Table 4.14 Table Pnt_Loc (intermediary table between point and Locality, M-M)

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
SHAPEFILE_ID	NUMBER(10,0)	No	1	1
NAME	VARCHAR2(50 BYTE)	No	2	
FILENAME	VARCHAR2(50 BYTE)	No	3	
LATITUDE	VARCHAR2(10 BYTE)	No	4	
LONGITUDE	VARCHAR2(10 BYTE)	No	5	
DESCRIPTION	VARCHAR2(200 BYTE)	No	6	
PRIMARY_SOURCE	VARCHAR2(50 BYTE)	No	7	
PRIMARY_URL	VARCHAR2(30 BYTE)	No	8	
PRIMARY_CONTACT	VARCHAR2(50 BYTE)	No	9	
ACCURACY_NOTES	VARCHAR2(300 BYTE)	No	10	
METHODOLOGY	VARCHAR2(300 BYTE)	No	11	
CONTRIBUTION_DATE	DATE	No	12	
CONTRIBUTOR_ID	NUMBER(10,0)	No	13	

Table 4.15 Table Shapefiles

Column Name	Data Type	Nullable	COLUMN ID	Primary Key
CONTRIBUTOR_ID	NUMBER(10,0)	No	1	1
NAME	VARCHAR2(20 BYTE)	No	2	
AFFILIATION	VARCHAR2(50 BYTE)	No	3	
POSITION	VARCHAR2(50 BYTE)	No	4	
EMAIL	VARCHAR2(30 BYTE)	No	5	
TELPHONE	VARCHAR2(20 BYTE)	No	6	

4.16 Table Contributor

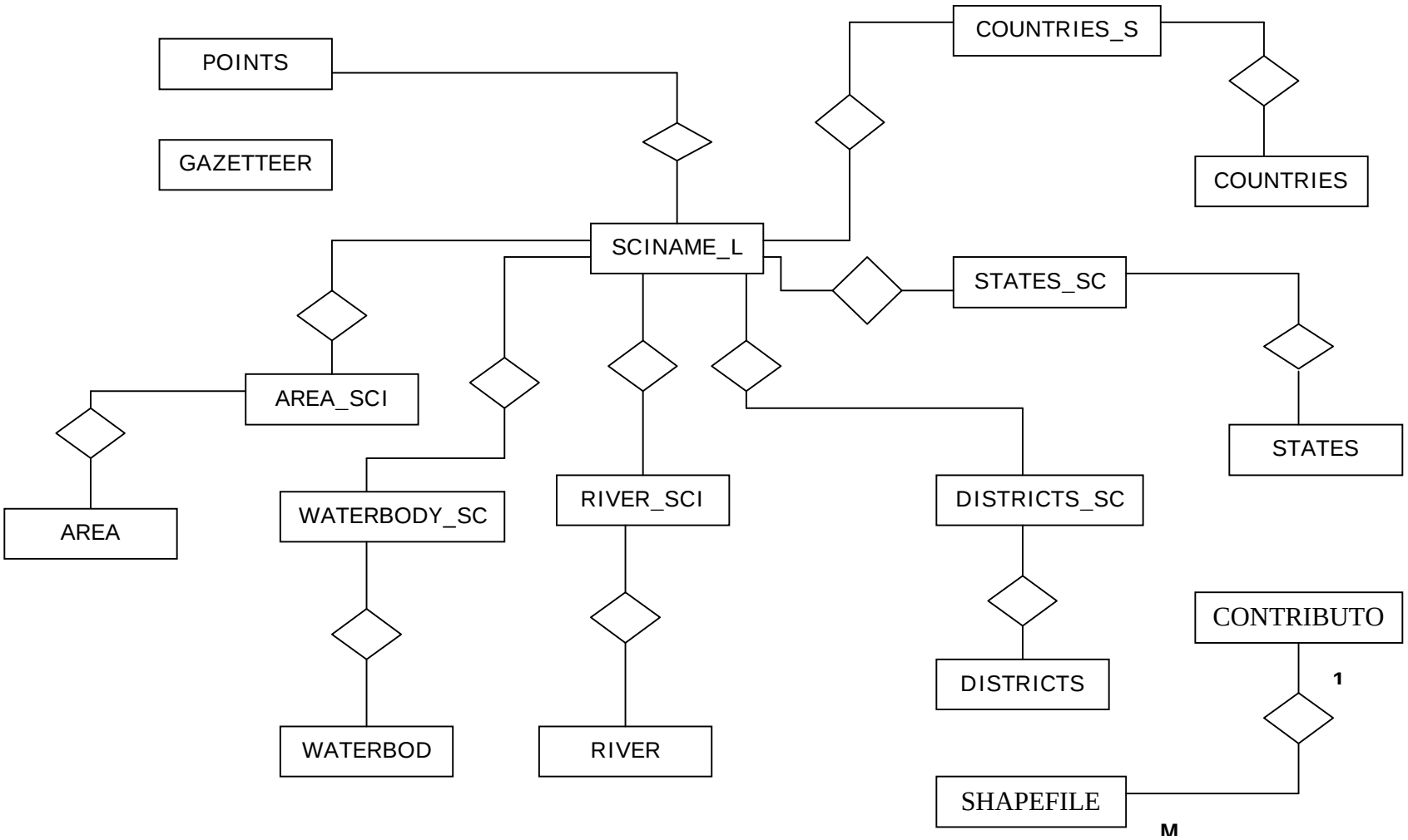


Figure 4.1: Entity Relationship Diagram for JaivaNaksha

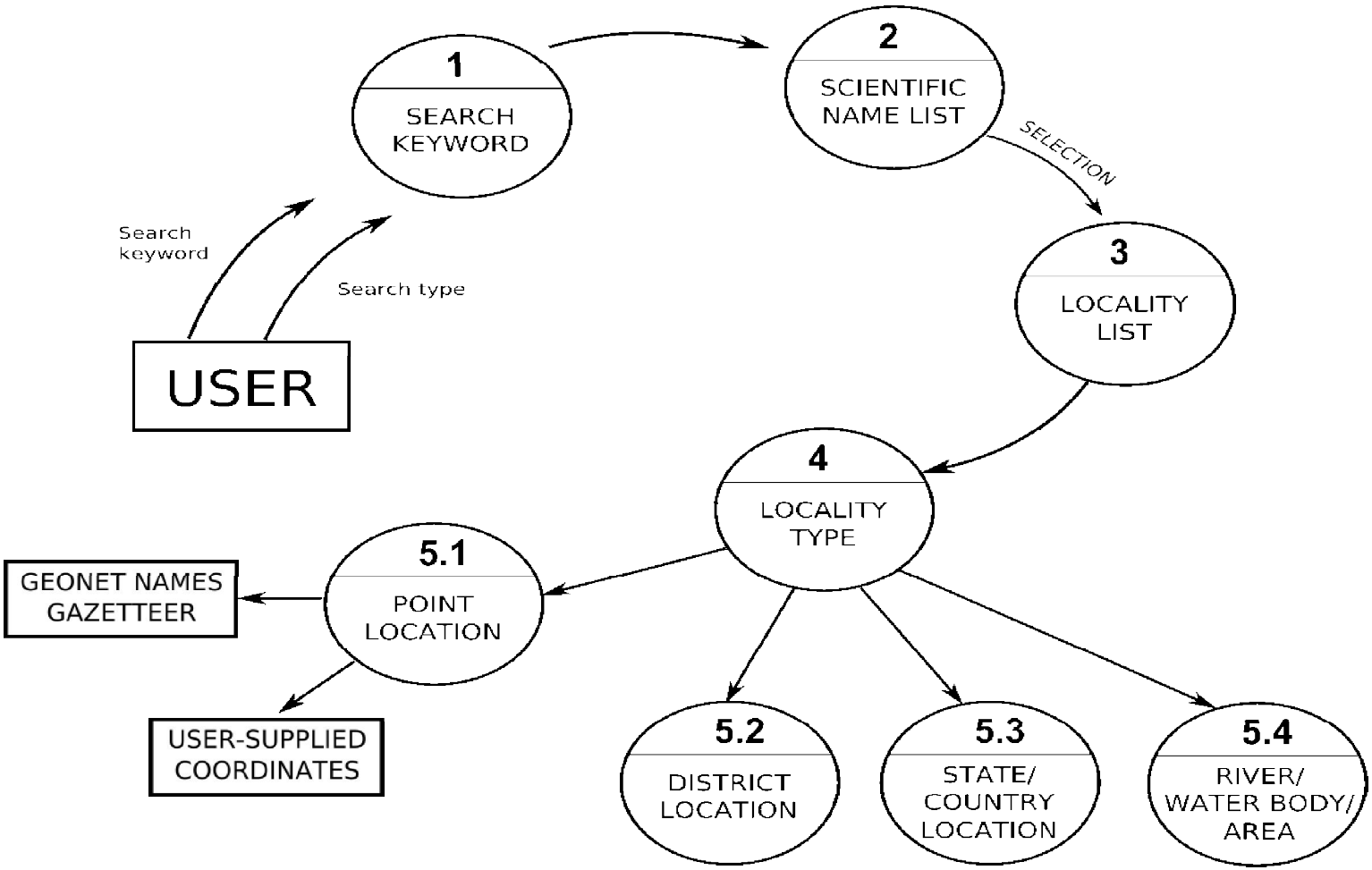


Figure 4.2: JaivaNaksha Data Flow Diagram representing data flow when search is performed.

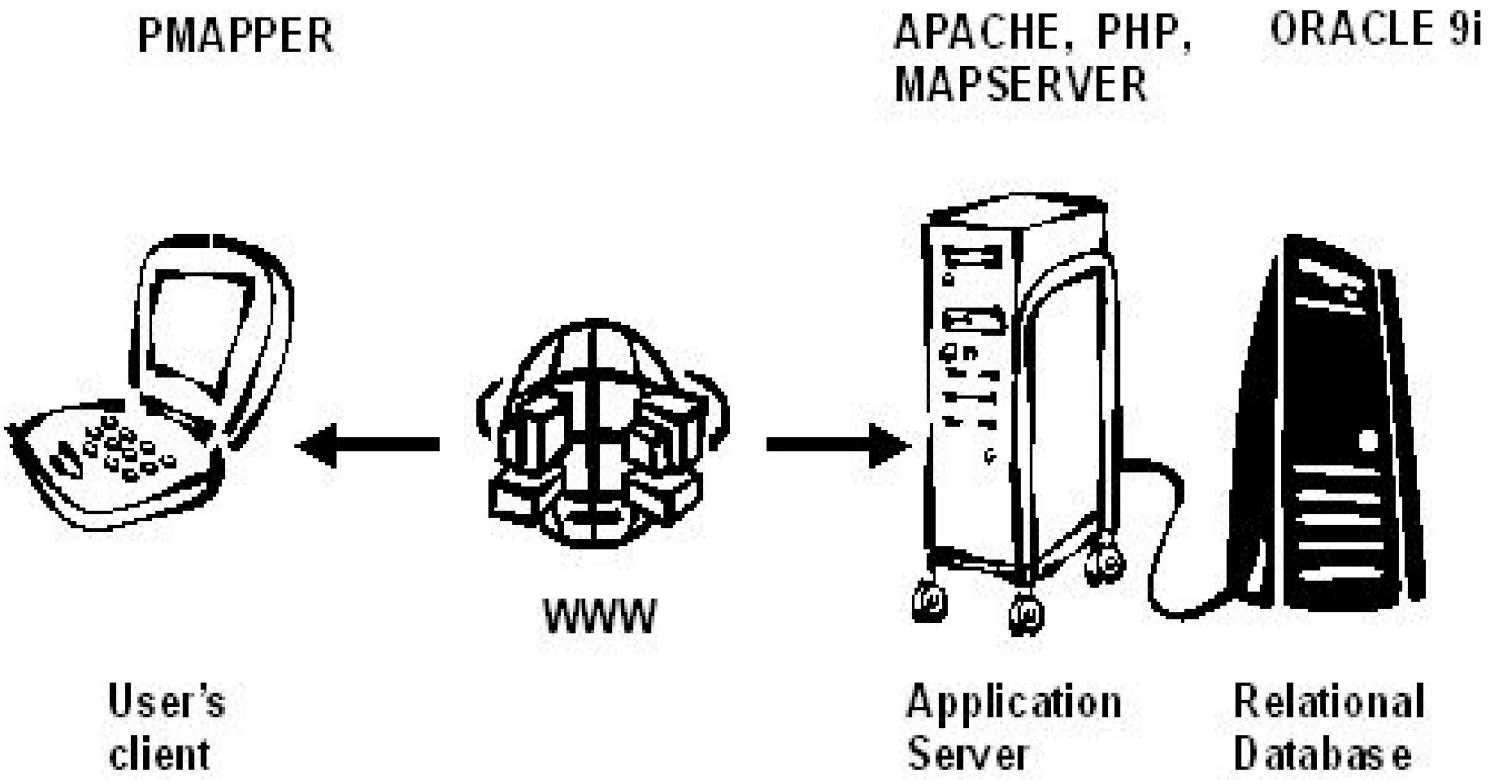


Figure 4.3: Informatics architecture and technologies used for JaivaNaksha

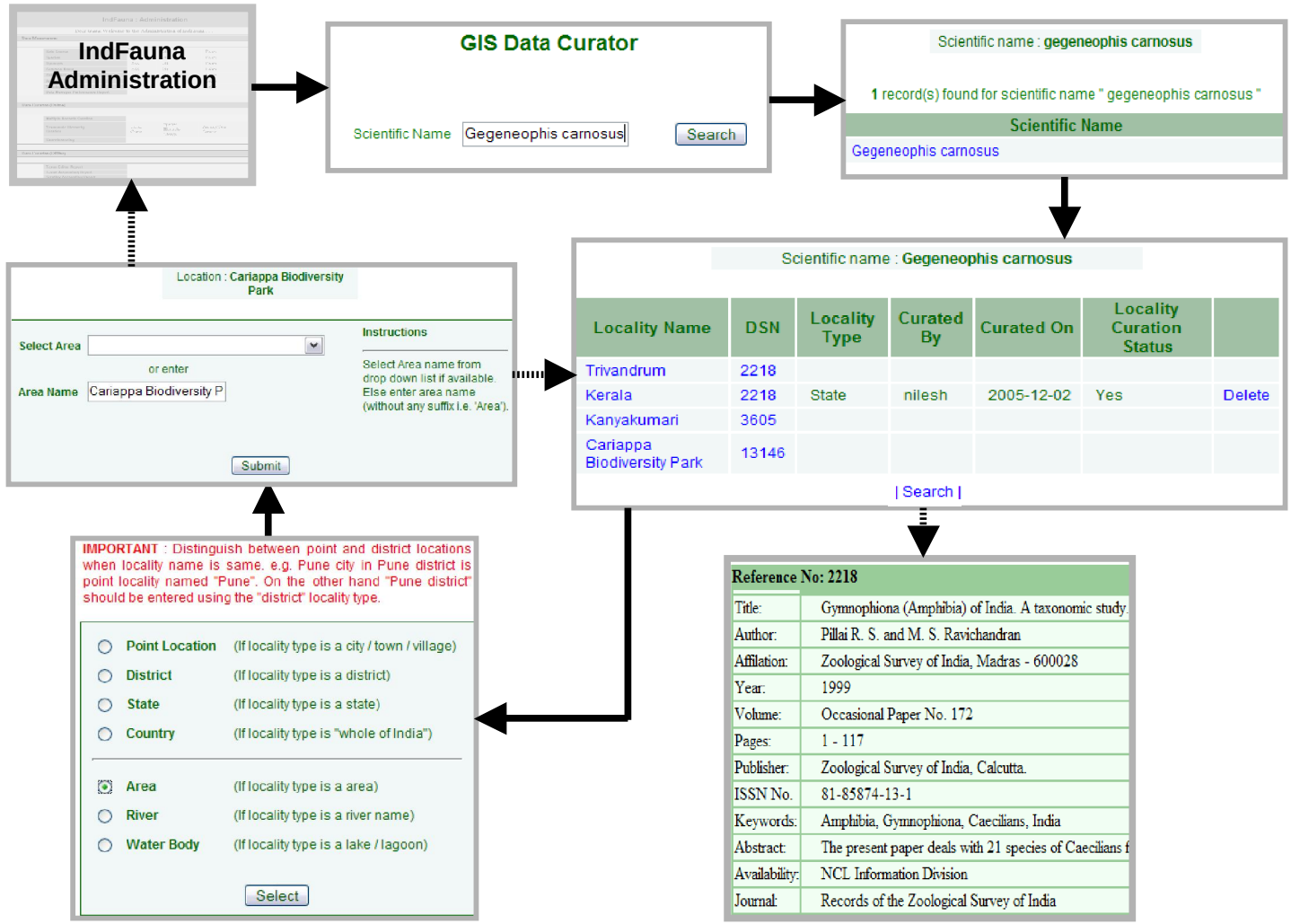


Figure 4.4: Spatial Data Curation Module which curate locality descriptions collated in IndFauna.

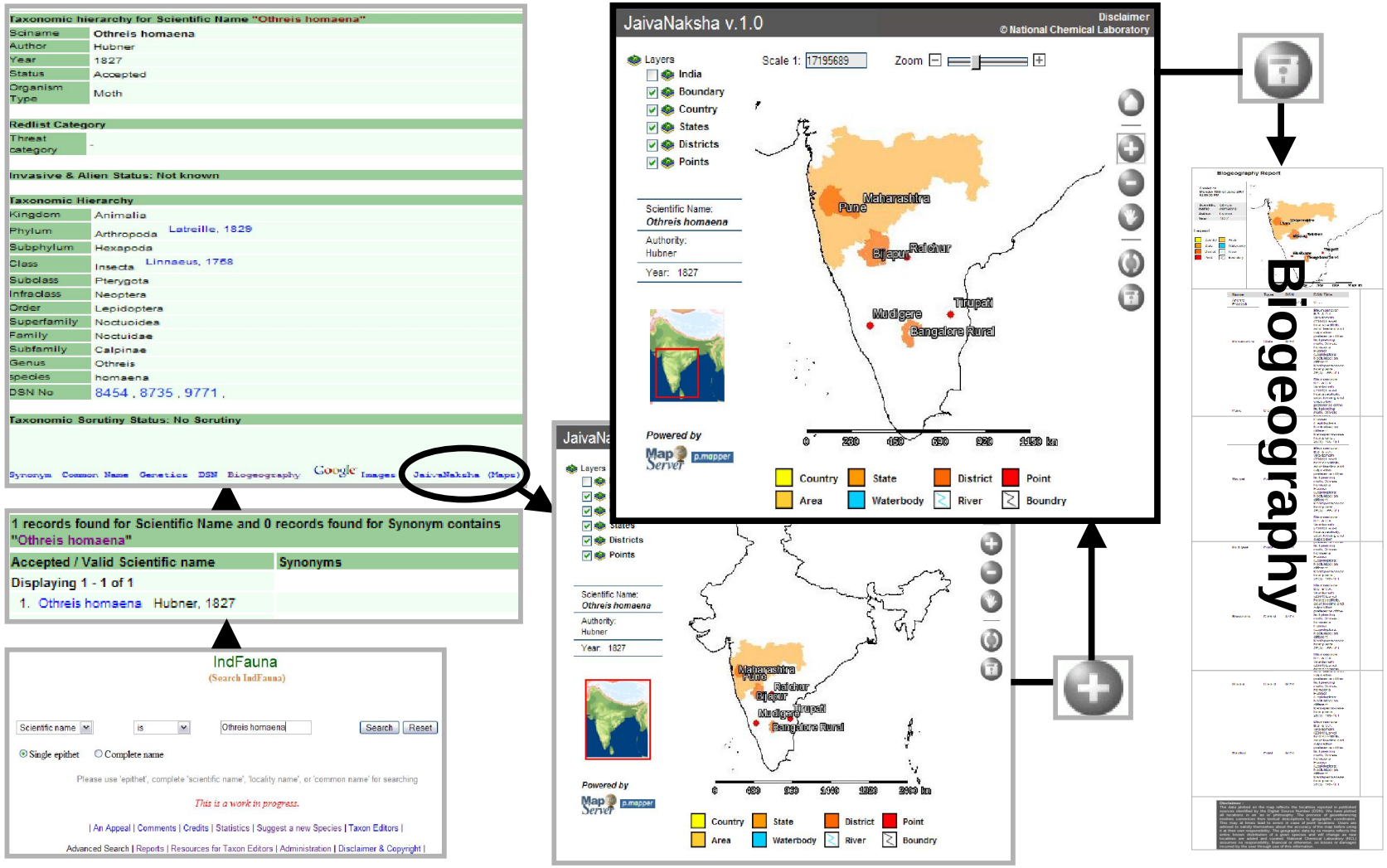


Figure 4.5: Web Mapping of IndFauna Occurrence data.

Chapter 5

National Biodiversity Information Infrastructure: Challenges, Potentials, and Roadmap

Chapter 5

National Biodiversity Information Infrastructure: Challenges, Potentials and Roadmap

5.1 National Biodiversity Information Infrastructure: Why?

During development of IndFauna, and allied products described in earlier sections of the dissertation, urgent need was felt to evolve planned mechanism to collect, collate, and disseminate data and information about Indian biodiversity. As noted earlier such data and information is currently distributed, isolated, in heterogeneous forms and format, and most seriously locked up in institutional and individual cupboards under the misconceptions of national security, intellectual property related sensitivity. Thus, there is a need to conceive and establish National Biodiversity Information Infrastructure. In Chapter 1, I have argued the need for such a national information infrastructure, especially for mega-diversity, developing country such as ours, which is aiming to move forward into the group of elite developed nations. Further, if Biological Diversity Act 2002 is to be implemented, there is an urgent need to set up biodiversity information system of unparalleled size and complexity (Gadgil, 2003).

Indian Biodiversity information domain is vast and complex but critically important to society. At present existing biodiversity and ecosystem information is neither readily accessible nor fully useful. There are few sporadic, isolated efforts being made in recent past. There is an urgent need for development of a web based interoperable and collaborative framework for interconnecting these distributed and heterogeneous databases by establishing resource discovery and access control mechanism. Further current technological and political scenario presents ample scope to undertake establishment of such a facility that is capable of collation, analysis, and dissemination of biodiversity and ecosystem related information from / to distributed sources.

In this chapter, an attempt has been made to define the needs, aims and objectives, scope, modalities as well implementation mechanism together with performance indicators of such a national information infrastructure, dedicated to collation and dissemination of data and information so that informed decisions could be taken up for sustainable use of biotic resources and their conservation. This vision and plan is an outcome of my experience of over 15 years with biodiversity

informatics initiatives at national, regional, and global level. This call for national biodiversity information infrastructure is also based on review of variety of information infrastructures or networks in various countries (e.g. Canada, United Kingdom, United States of America, Mexico, Brazil, Costa Rica, Australia, many EU countries), regions (e.g. North American Biodiversity Information System, Inter American Biodiversity Information System, Amazon Basin Biodiversity Information System, etc.), themes (e.g. Global Invasive Species Information System, Global Invasive Species Program, Ocean Biogeographic Information System, etc.) and those with global coverage (e.g. Global Biodiversity Information Facility, Global Earth Observation System of Systems, Encyclopedia of Life, etc).

5.2 National Biodiversity Information Infrastructure: What is there in name?

While conceptualizing national biodiversity information infrastructure described in subsequent sections of this chapter, I am aware that there would be debate and discussions as to how should it be titled, what acronym to be used. During my interactions with experts and potential contributors several names have been tossed up, such as "Indian Biodiversity Information System (IBIS)", "Indian Biodiversity Information Facility (IBIF)", "Indian Biodiversity Information Network (IBIN)", etc. I conceive this system of systems which would encompass many databases, and several information systems/networks, and hence, to me it is infrastructure. Thus, I have opted to term it as "National Biodiversity Information Infrastructure (NBII)".

5.3 NBII: Defining the purpose

As argued in Chapter 1 (Section 1.6), such an information infrastructure is needed for economic, ecological and social well being. This purpose of information infrastructure of national scale dedicated to biodiversity could be best summarized through vision statement and set of objectives which are able to signify it aptly.

5.3.1 Vision

> ***"National Biodiversity Informatics Infrastructure (NBII) will contribute towards economic growth, ecological sustainability, and social outcomes through increasing the utility, availability and completeness of new and existing biodiversity and ecosystem information resources"***

5.3.2 Objectives

If the vision of the NBII is increasing the utility, availability and completeness of new and existing biodiversity data, it can not be achieved by single institution or

group of individual, since data itself is isolated, dispersed, distributed, and in heterogeneous forms and formats. Considering this, operational objectives of NBII would be;

- to work closely with all providers of biodiversity data and information. It will have the characteristics of a large, distributed public domain databases with a number of interlinked and interoperable modules (databases, software, and networking tools, search engines, analytical algorithms, etc.) that will enable users to navigate and put to use the nation's vast quantities of biodiversity and ecosystem information,
- to seek interoperability amongst biodiversity and ecosystem database, and other associated datasets (such as sequence, weather, climate, economic, traditional knowledge),
- to be established as free-standing national organization as federated consortium of potential data providers, users, and stake-holders to fulfill the obligations spelt in Biodiversity Act 2002.

5.4 NBII: Features

The NBII will be an interoperable network of biodiversity databases, information networks and systems, traditional knowledge, peoples biodiversity registers, and information technology tools that will enable users to navigate and put to use the nation's vast quantities of biodiversity and ecosystem information to produce national economic, environmental, and social benefits. Thus, NBII would be an overarching information infrastructure, which would leverage on progress made so far by the various information systems, networks and databases spearheaded by various individuals, institutions and groups within and outside India. In true sense it would be a registry of shared biodiversity and ecosystems databases developed by various agencies and individuals (Fig. 5.1)

The purpose of establishing NBII is to design, implement, coordinate, and promote the compilation, linking, standardization, digitization, and nation wide dissemination of the nations biodiversity and ecosystem data within an appropriate framework for property rights and due attribution. While doing so, it is essential to maintain the autonomy of the various databases, information systems. If NBII has to prosper and sustain, it is critical that scope for expression of creativity and originality of the designers and managers of various information systems and databases is

ensured. Further it should protect all legitimate intellectual property rights, benefit and access sharing of tribal and indigenous communities, or country as a whole.

The NBII should be designed to work in close co-operation with established national, regional, and global programs and organizations that compile, maintain, and use biological resources. These collaborative and coordinated efforts will established and support a distributed information system that will enable users to access and utilize vast quantities of new and existing biodiversity and ecosystem information to generate new knowledge, wealth and ecological sustainability.

NBII will:

- be a distributed facility, while encouraging co-operation and coherence;
- be national in scale, though implemented by various national, state, district, village level, and thematic players (agencies, institutions, individuals, communities);
- be open to participation by individuals, communities, agencies, institutions, from all walks of life, and offering potential benefits to entire nation and its constituents;
- help bridge human language barriers by promoting standards and software tools designed to facilitate their adoption into multiple languages, character sets, and computer encoding;
- serve to disseminate technological capacity by drawing on and making available scientific, technical, and social information; and
- while aiming to make biodiversity information openly, and freely available, will facilitate credit to its contributors, respect and fulfill national aspirations such as intellectual property rights, economic wealth, bio-security, and social wellbeing.

5.5 NBII: Work Programs, Milestones and Performance Indicators

5.5.1 NBII Work Programs

In order to achieve ambitious looking vision and objectives as detailed in preceding sections, entire operations of needs to be grouped into functions as well content specific work programs. These work program fall into three themes viz., Content, Informatics and Participation (Table 5.1). Amongst them, these work programs are bound to have both synergy and overlap between them, and are likely to run concurrent to each other.

The purposes of these "work programs" (Table 5.1) are a) to facilitate the full use of biodiversity and other databases by establishing an information architecture that enables interoperability and facilitate data-mining; b) to facilitate the expansion of biodiversity knowledge by having legacy and newly acquired data digitized and dynamically accessible; c) to make integrated and multi-lingual searching possible, as well as to facilitate the exploration and rapid expansion of biodiversity knowledge; d) to bridge biodiversity information technology "digital divide", "content divide" and "knowledge divide" through training and capacity building, targeted seed funding opportunities ensuring that people in every "nook n corner" of the country have access to, and can easily contribute to this knowledge building endeavor; e) to provide, in real time, a complete compendium of knowledge about Indian biodiversity, its biotic resources, their environs, its uses, and sustainable management, etc., drawn from distributed information sources.

These work programs and their anticipated achievements during a defined period of 5 and 10 years is detailed in Table 5.1. At the NBII Secretariat, each of the above work program would led by the Program Officer, except the last one i.e. Information and Communication Technology Implementation (ICTI), which would be articulated and implemented by the Deputy Director (Informatics). ICTI needs to be conceptualized, evolved and implemented by an information technology expert considering needs as well technologies available, through seamless interactions with all Program Officers, members of BoC, Data Providers and Nodes.

5.5.2 NBII Milestones and Performance indicators

For successful implementation of initiative of this nature, it is crucial to have well planned time bound milestones and performance indicators. Table 5.1, gives an outline of such milestones and performance indicators could be laid down. Many would debate if such ambitious and time bound milestones would really be feasible to achieve? However, it is better to keep performance bar high and work towards it, as it is likely to draw synergy and energy into whole program.

While implementing these work programs, NBII's strategy should be 1) focus on mission and specific goals, 2) inclusiveness in the manner in which it would advice, 3) openness in data sharing and software development, 4) cost-effective partnerships, and 5) adherence to basic principles of interoperability.

If, adopted and meticulously implemented, NBII has the potential to advance by orders of magnitude our national ability to exploit the Web's power, giving society

true, nationwide, usable biodiversity information-at-our-fingertips that will contribute to scientific innovation and progress and towards a sustainable society. NBII will do for biodiversity and ecosystem information what the printing press did for the sharing of recorded information during the Renaissance – it will make recorded knowledge the common property (within the framework of national aspirations and concerns) of everyone, in national context. Even better, NBII will do it electronically, so the resources will be dynamic, interactive, and ever evolving. Within five years, we dream NBII portal will be the most-used gateway to Indian biodiversity and other national biological data on the Internet.

5.6 NBII: Implementation

Successful implementation of such a National Biodiversity Information Infrastructure (NBII) would require four key components:

(a) The institutional framework that defines the policy and administrative arrangements, dealing with issues such as organizational roles and responsibilities, custodianship, pricing, licensing conditions, confidentiality, intellectual property, education and training;

(b) The technical standards to ensure consistency in the characteristics of the datasets;

(c) The fundamental datasets that will be made accessible; and

(d) The clearinghouse network, which is the means by which the fundamental datasets are made accessible to the community.

In following sections author discuss NBII implementation in two parallel and coordinating, yet interconnected procedures, viz. governance and technical implementation.

5.6.1 NBII Technical Implementation

Review of various options, and lessons learned from ongoing international exercises, suggests that “web services architecture” is the best suitable technology option for achieving the mission. It would facilitate efficient, cost effective, flexible networking of distributed, heterogeneous, cross-discipline databases and information systems, in an interoperable and collaborative manner. This means that there are distributed data providers, a central registry of them, and a central portal to access the data, thematic, regional portal, and multilingual portals to access theme and region specific data, in various languages, all communicating using standardized XML messages (Fig. 5.2).

5.6.1.1 Scope and Objectives

NBII information infrastructure can be considered consisting of three interrelated but separate elements; (a) Data Standards and Protocols, (b) Infrastructure services, and (c) the NBII network.

(a) Data standards and protocols is the most fundamental element in the information infrastructure. In the interoperability framework document (Chavan and Krishnan, 2006), author has discussed some of the existing standards in operation by similar information systems and networks. However, considering the scope and uniqueness of our datasets, NBII would need to invest its energy in developing few new standards. This needs to done together by Secretariat staff (ICTI), DADI Program Officer, DADI Science Sub Committee and other relevant national and international agencies such as TDWG, GIS Community, Multilingual Community, and bioinformatics community, etc.

(b) Infrastructure Services is an integrated system of hardware, software, standards and protocols that facilitates integration and discovery of data resources. ICTI, together with DADI Program Officer and other relevant national and international agencies needs to develop an implementation of these infrastructure services.

(c) The NBII Network is the collaboration of NBII Participant Nodes and independent data providers ("Data Nodes") to share their data using the infrastructure services developed by NBII. Modalities and processes involved in developing NBII Network facilitating Data Nodes to share data has been described in detail in the interoperability framework document (Chavan and Krishnan, 2006).

Expected end result of these three interrelated elements is the NBII data portal. Fig. 5.2, depicts the various stages and processes that lead to dissemination of data through NBII Data Portal. Chavan and Krishnan (2006) discussed in detail the steps and processes to develop and implement NBII services. The NBII data portal encompasses all the central infrastructure services that would be operated by NBII for sharing biodiversity data.

5.6.1.2 Mirroring and replication of NBII Data Services

As the NBII reach in operational phase within its first 5 years of operations, millions of records would be served by hundreds of data nodes (participant and independent data providers). Data providers would announce availability of increasing quantum of data in NBII registry. NBII Data portal would index the records on the

providers and offers a user interface for searching, browsing, displaying, and downloading the data. Operating such a complex yet dynamic service has high availability requirements. There are several reasons why NBII should consciously work on establishing at least two mirrors within fist 5 years of its operations. Some of the reasons include redundancy, load balancing, speed of access, indexing, helpdesk operations, localization, ownership and buy in, etc.

While it is not necessary to mirror the UDDI registry service, we need to mirror or distribute the 'master data chache', 'data index' at the mirror site at scheduled intervals through auto scheduler. On top of this the portal application facilitating searching, browsing, displaying and downloading applications would operate. Through the dynamic mirroring it is feasible to re-route the traffic to nearest mirror site, rather that awaiting response of the NBII data portal hosted at its Secretariat. Ideally, it is recommended that two mirror sites be established along in two geo-climatically different zones, for e.g. if NBII Secretariat is established in Nagpur with its "Data Portal", then mirror sites could be up in north and down in south of India. Institutions hosting mirror site should have basic required infrastructure and Internet bandwidth of 2 mpbs. Currently several disaster recovery initiatives and investments in some of the major institutions across country provides scope and wider array to select few of them as potential mirror sites.

5.6.1.3 Thematic, Regional, and Lingual Portals

These would have several benefits such as (a) showing the data relevant to theme/region, (b) customizable for thematic or regional needs such as language, (c) easy to use and most effective way of marketing NBII in the region or thematic community, and (d) technically it serves as test bed for user interface component for the national data portal.

State Biodiversity Boards (SBB) being Participant Nodes, if desired could become strong candidates to host region specific portals and interfacing languages being used in the specific regions. However, if SBB do not have infrastructure and technical expertise to operate such regional language portal, it can work together with any institution in the regions, which is independent data provider (Data Node). For instance, SBB in Goa may tie-up with the National Institute of Oceanography (NIO) to establish and operate the Goa specific regional portal in Konkani language.

R&D and educational institutions or NGOs with proven domain expertise are natural choice for establishing the thematic portals. Majority of the occasions, a single

institution may be equipped, and eligible to establish and operate such mirror, however, in cases such as disciplines practiced by many, there may be conflict of interest amongst multiple organizations. For instance, Botanical Survey of India (BSI) as well National Botanical Research Institute (NBRI), Tropical Botanical Garden Research Institute (TBGRI) might be interested in establishing and operating thematic node on floral diversity. In such a scenario either these agencies share the responsibility and work load for such activity and establish single thematic portal jointly, or one of the institute develop, commission, and operate a portal and others mirror the same to provide better access to the portal in their geographic region.

5.6.2 NBII: Governance

There are several governance models that have been adopted by the existing information infrastructures, networks and facilities. After reviewing them, and considering the advantages and potential challenges in Indian scenario, following governance model is proposed. Author believes that networks, and information systems which have adopted similar models at global and national levels in various regions of the world, (e.g. GBIF, OBIS, IABIN, US NBII, etc.) have been able to achieve milestones as they were able to generate sense of ownership amongst its stakeholders, and thus could create immense urge, and inertia amongst them to make such initiatives successful.

NBII will be established as free-standing national facility supported directly by the Planning Commission to fulfill the obligations spelt in Biodiversity Act 2002, and as crafted by the National Biodiversity Authority, it's Board of Consortium (BoC), Science Council and its sub-committees. It will be by a small secretariat that will work nationally to coordinate national, regional, local efforts and bring focus to its organization and its activities (Fig. 5.3).

5.6.2.1 Validity, Authority and Jurisdiction

- NBII should come into existence through ordinance or executive order of the government of India to fulfill the expectations of Biological Diversity Act 2002. This would not only provide much needed legality, and authority over several potential contributors and constituents.
- Nations finance planning body (Planning Commission) should ensure financial support to this new initiative similar to its support and encouragement to missions of national significance, urgency, and national pride.

- NBII is a mega-science facility that will (a) enable scientific research that has never been possible, (b) facilitate the use of scientific, community, traditional, and socio-economic data in biodiversity policy- and decision-making, and (c) make whole national biodiversity and ecosystem information—data that are currently exceedingly difficult to access – freely, and openly accessible via Internet while attributing credits to contributors, and respecting national aspirations such as intellectual property, economic wealth, bio-security, and social well-being.
- It is expected that after initial proof of concept or development phase of 5 years, NBII would move forward into implementation phase (5 years) where in data / information collated can be used in decision making, as well sustainable management process by cross-sectional users.
- If, NBII has to achieve its vision and objectives, majority of its work programs and action plan implementation requires energy and coordination and coherence amongst its stakeholders, which would constitute its Board of Consortium (BoC). Hence, NBII should be established as a national facility under a Memorandum of Understanding amongst its promoters (Govt. of India through NBA, and Planning Commission) and agencies, institutions, organizations, communities (both governmental and non-governmental) which holds potential to contribute existing and new information regarding nations biodiversity and ecosystems.

5.6.2.2 Board of Consortium (BoC)

- Board of Consortium will consist of one representative from each participant. Board of Consortium together with its promoters (NBA and Planning Commission) will oversee the performance of the Secretariat, headed by Executive Director.
- Board of Consortium in consultation with NBA, Planning Commission, and Science Council (and its science sub-committees) will develop detailed long-term action plans, milestones, and oversee completion of these work-programs.
- However, operational budget to ensure achievement of the vision and mission of NBII is provided by the Planning Commission on recommendation of the Board of Consortium, Science Council, following its consultative process and

modalities as applied to various scientific departments, and agencies of national significance.

- Potential members of the Board of Consortium can fall into two categories of nodes viz. participating nodes and independent data providers (Data Nodes) (Fig. 5.4). These two categories are described section 5.6.2.5.
- BoC would meet once a year, and if required more than once if required, which might be the case in the formative years of NBII. Annual BoC would discuss the long-term work plan, annual work program and also rules of operations; as well evolve the 5-year plan and annual budget. It would also oversee the performance of the Secretariat and would offer guidance to the Executive Director.
- If feasible, BoC can have virtual inter-session meetings, interactions using one of the electronic communication media such as litserve, wiki portal on issues of common interests.
- The Chairperson would chair BoC. Chairperson and two Vice-Chairpersons would be elected amongst the candidates nominated by the official delegates to BoC.

5.6.2.3 Science Council and its Science Sub-Committees

- In order to provide guidance, develop strategic action plan, and budget for each work program a small Science Sub-Committee (SSC) would be formed. Program Officer concerned is the Convener of the SSC. The BoC would elect the chairpersons of the SSC from the candidates nominated by the delegates to BoC. Each delegate can also nominate individuals with domain expertise that could be useful for specific work program SSC. Chairperson in consultation with Program Officers, with approval of the Science Council would select not more than 10-12 individual as members of SSC. Tenure of each SSC would be for a period of 2 years.
- SSC would advice program officer in developing work program, its strategic plan and evolving annual budget for work program and would endorse the same. It would then forward the annual or plan period work programs to Science Council for further ratification.
- Science council would be comprising of chairs of each SSC, Executive Director of the Secretariat, nominees of NBA, and Planning Commission and

few (3-4) co-opted men and women of repute whose expertise, and experience can provide guidance in evolving long term and short term work plan, strategic plan, as well developing budget for seamless operations.

- Science Council would be headed by a Chairperson and backed up by two Vice Chairpersons, which would be elected by BoC amongst the candidates nominated by delegates to BoC.

5.6.2.4 NBII Secretariat

- Secretariat is essential to ensure nation-wide, collaborative, coherent, cost-effective operations that are sufficiently effective and visible to maintain focus, leadership and momentum once NBII is operational.
- The role of the Secretariat is to build coalitions amongst ongoing efforts, encourage new developments, and provide mechanism for coordinating separate national, regional and local investments and forging international agreements.
- The Secretariat shall be small and cost-efficient as possible but this should be balanced against the risk of under-funding it. Initially, it will comprise of an Executive Director, Dy. Director (Informatics), 9 Program Officers, and appropriate support, technical and administrative staff. Staffs recruitment would be phased in as necessary as operation expands and continue to achieve the milestones. Operational workflow of NBII is depicted in Fig. 5.5.
- In the initial phases secretariat can be hosted/housed in one of the participating agency/institution, preferably with ability to provide technical and administrative support. The Secretariat should preferably be accommodated in a scientific environment that would allow rich professional interactions and relationship with ongoing biodiversity and ecosystem informatics, biological data management, and analysis research efforts. It should also have access to library, meeting facilities and enough space and associated facilities to build its own robust computing infrastructure with high-performance computing capabilities.
- Later on, Secretariat can have its independent and self-sustaining operational campus. In long term sustenance and operational point of view Secretariat should be in the vicinity of agencies and institutions where robust computing

infrastructure with high performance capabilities, technical and domain expertise can be leveraged on.

5.6.2.5 Data Nodes and Participant Nodes

- IBIF would have two categories of data providers. First is the "Participating Node", and other independent data providers or "Data Nodes"
- **Participant nodes** are the ones who would generate new data as a result of implementation of Biodiversity Act, 2002. This means most of the SBBs, BMCs, and agencies providing "Peoples Biodiversity Register" data would fall in this category, as the PBR is one of the activity mandated by the Biodiversity Act, 2002.
- **Data Nodes** or independent data providers are agencies/institutions, which are engaged into biodiversity, ecology, environment, bioinformatics and other associated data management activity even before the Biodiversity Act, 2002 came into force. Nearly, 75-80% data providers would fall into this category, such as various government departments and their autonomous bodies, universities and colleges, Non Governmental Organizations (NGOs), regional and as well international agencies such as GBIF, GISIN who hold data about Indian bioresources, and museums and data centers overseas.
- Data nodes and participant nodes are the key to NBII's success; the Secretariat would assist the Nodes to carryout the challenging tasks ahead (Fig. 5.6).
- While "Data Nodes" might have technical capabilities to establish "data provider service", "Participant Node" would require support (financial and technical) and mentoring to establish the data provider service.
- These nodes would be enrolled into NBII Board of Consortium through Memorandum of Understanding (MoU), and would be represented by delegate designated by Head of the institution. Such a delegate or his/her alternate would participate in Annual Board of Consortium (Annual BoC) meeting, which would discuss the long-term work plan, annual work program and also rules of operation, as well evolve the 5-year plan and annual budget.
- With a quick review of current state of biodiversity informatics within Indian and globally, it is expected that there would be approximately 1000+ Nodes (Participant and Data) would be able to contribute, exchange/share data through NBII information infrastructure during first 5 years of its implementation.

5.7 Financial requirements

It is beyond doubt that initiatives of such a magnitude with potential of long ranging impacts, investment would be in multiples of 100's of crores. The straw man's guess is that NBII would require investment of nearly Rupees 800 - 1000 crores for a period of first 5 years (US$ 20 – 25 million). An investment to the tune of Rupees 800-1000 crores would make NBII as expensive proposal to implement. However, several national and international information management initiatives are attracting significant investment.

For instance in mainstream bioinformatics initiative such as National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) with approximately 46,400,000 sequence records in its kitty has an recurring budget of US$ 50 million per annum. Protein Data Bank (PDB) with over 31000 annual records has recurring budget of US$ 5 million (GBIF, 2006). In the field of biodiversity informatics, Species2000-ITIS Catalogue of Life (CoL) which has recently documented over 1 million of the known 1.8 million species needed an investment of UK £ 11 million during 1994 – 2001 for central operation alone. This is in addition to UK £ 77 million invested during the same period for 47 GSDs that constitute Species2000-ITIS CoL (Bisby, 2007, pers. Comm...) At regional scale, IndOBIS, Indian Ocean Node of OBIS alone needed US$150,000 to kick start its 2 year moderate operations, and if its operations have to deliver the goal of documenting known life from Indian Ocean, it would need an investment of over US$500,000 per annum for a period of next 10 years.

GBIF in its first 5 years of operations invested approximately US$3.5 million per year in order only integrate over 80.5 million data records from distributed data providers. This is in addition to on an average US$ 20 million per annum spent by each of the 13 member countries, and organizations surveyed (GBIF, 2005). During its next phase GBIF would require on an average US$ 5 million per annum only to run its current operations, and additional US$5 million per year to add up new modules to its work programs as envisioned in its 2007-2011 strategic plan (GBIF, 2006). Thus, considering the scope and magnitude of involved, investment anticipated in NBII is not too much. Recently launched Encyclopedia of Life (EoL) is seeking an investment of US$50 million over next 10 years (EoL, 2007).

Thus, an investment of Rupees 800-1000 crores to realize the dream of NBII, is justified. However, such a requirement needs to be carefully planned, and evolved.

It is best that subsequent to acceptance of the concept and in principal approval for its establishment an experienced consultative group may be commissioned. Alternatively, high-level task group (3-4 members) may be constituted. Such a consultant or task group may be mandated to develop detailed implementation plan, together with budget for first 5 to 10 years with annual split. After the SWAT analysis, and risk assessment, financial proposal could be approved and sanctioned.

5.8 NBII: Can dream be a reality?

Though, the financial requirement of NBII is significant and is enough a reason for national funding agencies to raise an eye brows with regard to confidence of achieving expected outcomes. However, it is my firm belief that a dream of NBII could certainly be a reality, given the experience of pilot project that my group implemented in 2006-2007 with funding from the Government of India's Department of Biotechnology (DBT). During the experiment using the open source UDDI registry, could successfully integrate over 500,000 species and specimen data records from 7 distributed databases through 2 providers (Fig. 5.7). Annexure III provides further details of pilot phase experience of "Connecting Diversity".

Lessons of this experience and considering that increasing number of standards, protocols, software(s) and techniques are being developed to integrate distributed biodiversity data and its interoperable integration with other cross discipline data, technology would not be hurdles in realizing the dream of NBII, at least for nation such as India. However, what is really required is working beyond individual, institutional, regional, national and international boundaries. If this could be guaranteed, NBII would be a reality for sure in next 5-10 years.

5.9 Future of NBII

However, transforming a 500 year tradition of (slow) information transfer by lines of type on paper into a digital (rapid) interchange among thousands of distributed, heterogeneous, and multilingual databases, while at the same time dealing with complexities of the information itself as well as the means of handling it, is no simple task, nor can this task possibly be accomplished in only 5-10 years. This is why NBII would certainly prove itself as a mega-science activity.

While, nation would start enjoying benefits of accessing and using data information pulled together by NBII in next 4-5 years, the demands of data use cases, data use patterns, data synthesis and analysis tools, decision support system tools are bound explode, as peoples imagination would suddenly be boosted. My vision is that

within 5-10 years after its inception, NBII will have become premier biodiversity information source for the wildest possible array of users.

In 2012, NBII portal would be able to address one or many of requests such as *"a lawyer for a group of indigenous peoples who needs to establish the exact identity of a plant on which they claim rights of intellectual property"*; or *"a robotics researcher who needs inspiration from nature about how to solve a particular engineering problem"*, and so on. The applications and utility of biodiversity data are endless, and of inestimable value, and thus NBII too!

5.10 Recommendations

- National Biodiversity Information Infrastructure (NBII) is must for economic, ecological, and social well-being.
- NBII should be a distributed facility, national in scale, open to participation, and bridging human language barriers.
- NBII should be implemented by adopting web services architecture as its technological basis.
- Core funding of Rupees 800-1000 crores for a period of 5 years should be made available by national planning agency, i.e. Planning Commission to implement NBII, in its first phase with assurance to continued support in subsequent phases.

5.11 Summary

For mega-biodiversity developing nation such as India, which is aspiring to be major developed super-power, it is essential to have data enriched "National Biodiversity Information Infrastructure", for its economic, ecological and social well-being. NBII needs to be developed as distributed facility, with open participation by potential data providers, in order to leverage on existing investment. For the magnitude and scale of biodiversity and associated data that exists and is being generated, web services architecture is best suited as technological framework for implementing NBII. Though, exchange of data is one of the major challenges for realizing dream of NBII, current political, social, and economic scenario provide excellent opportunity to pursue this dream. With guaranteed core funding support and mandate from the appropriate functionaries of the Government of India, NBII is not impossible to implement, thus liberating biodiversity data, making it available to

anyone, anytime, anywhere respecting the intellectual property as well national security, and ecological sensitivity.

Table 5.1 NBII would achieve its vision and objectives through implementation of functions and content specific work programs

Program	Objective	Goals / Tasks	Milestones/Performance Indicators
CONTENT			
Registry of Known Organisms and SpeciesBank (ECAT)	To facilitate development of national registry of all known organisms (ECAT), and complete compendium of knowledge about all known species	• Review, adopt existing standards and implement information infrastructure that can integrate existing species checklists, databases. • Develop new standards for data and metadata for new data types such as common names, occurrence etc. • Data about 95% of the known organisms is documented by 5th year. • Review, adopt and develop standards for data and metadata to implement information infrastructure to collate Species(Information)Bank which would facilitate in real time, complete compendium knowledge about each species • By 5th year populate SpeciesBank with complete compendium knowledge about 20% of known organisms mostly economical and ecologically significant ones • By 10th year complete SpeciesBank development with complete compendium knowledge about all known organisms.	• By 2nd year 40% of the accepted names of known organisms (including synonyms) electronically available. • By 5th year 95% of the accepted scientific names of known organisms (including synonyms) electronically available. • By 5th year common and vernacular names in various Indian languages, for various life stages, and sex be electronically available for 40% known organisms. • By 5th year taxonomists / systematicians starts using such a roster of names of organisms, as valid "Registry of Names of Organisms" and also register new names even before submitting research communications to journals. • By 5th year Species (Information) Bank is populated with integrated data/information about minimum of 20% of the known organisms.
Digitization – Specimens, Literature Bank, Archival and Rescue of orphaned data sets (DIGIT)	To facilitate digitization of all biological specimens and associated data housed in Indian museums and specimens of Indian origin in overseas collections To facilitate digitization of all legacy literature and ensure access to it	• Estimate the universe of biological specimens in Indian museums, and specimens of Indian origin in museums abroad. • Digitize all specimens and associated data housed in Indian museums by 10th	• By 3rd year minimum 15% of the specimens housed in various national, regional, and local collections within India are digitized. • By 5th year minimum 25% of the specimens housed in various national,

	together with new literatures through Internet To facilitate rescuer of orphaned and threatened data sets, and their long term archival and preservation	year • Ensure access to data on 100% specimens of Indian origin in overseas museum in next 5 years. • All biodiversity literature (legacy and new) is accessible through National Digital Biodiversity Literature Bank in next 10 years. • All threatened and orphaned data sets rescued and archived by 5th year.	regional, and local collections within India are digitized. • By 5th year data on minimum 40% of the specimens of Indian origin housed in various overseas museums, is accessible through appropriate mechanism. • By 3rd year all new (electronic form) is accessible through Digital Biodiversity Literature Bank. o By 5th year 20% of the legacy literature (print and other non-electronic form) is digitized and accessible through Digital Biodiversity Literature Bank o By 2nd year universe of threatened, and orphaned data sets (electronic and non-electronic form) is determined. o By 5th year 25% of the threatened and orphaned data sets are archived, rescued, and accessible through IBIS. •
Peoples Biodiversity Register (PBR)	To facilitate documentation of knowledge of occurrence, practices of propogation, sustainable harvest, conservation as well economic uses of biodiversity resources that resides with India's local communities	• Develop tools, standards and protocol to acquire People's Knowledge about biodiversity around them, and interlinking of this data with other bioresources and associated datasets. • Develop standards, protocols and tools for data, metadata creation, QA/QC, validation, integration at various levels across networks. • Develop informatics infrastructure for collation and dissemination of PBR data. • By 5th year PBR data from all states (300 out of 500 districts) collated and	• By 5th year PBR data from all states (covering at least 50% of its geographic area) is digitized and accessible. • By 5th year 300 out 500 districts are surveyed in totality thus raising the nation geographic coverage surveyed to nearly 60% • PBR information is digitized and documented in minimum 15 languages and 10 scripts.

		disseminated. • By 10th year entire nations PBR data collated and disseminated • Facilitate data collation and dissemination in 15 languages and 10 scripts by 5th year. • Facilitate data collation and dissemination in all recognized languages by the 10th year.	
Traditional Knowledge Repository (TKR)	To facilitate collation of traditional knowledge about use and bioprospecting of potential bioresources	• Develop to integrate existing digitized data, and for collation of data involving community based organization such as NGOs, CBOs, Schools, and Panchayats, etc. • Develop tools, standards, protocols for new data acquisition, data and metadata, as well for QA/QC, authentication, and validation. • By 7th year 100% of the traditional knowledge data associated with flora is digitized and accessible. • By 10th year 100% of the traditional knowledge data associated with fauna is digitized and accessible. • By 15th year 100% of the traditional knowledge data associated with biotic resources other than flora and fauna is digitized and accessible. • TK information is digitized and documented in minimum 15 languages and 10 scripts by the end of 5 years. • TK information is acquired and disseminated in all Indian languages in next 10 years.	• By 1st year framework is developed to integrate existing digitized data, and for collation of data involving community based organization such as NGOs, CBOs, Schools, and Panchayats, etc. • By 3rd year 30% of the traditional knowledge data associated with flora is digitized and accessible. • By 5th year 60% of the traditional knowledge data associated with flora is digitized and accessible. • By 5th year 25% of the traditional knowledge data associated with fauna is digitized and accessible. • By 5th year 10% of the traditional knowledge data associated with biotic resources other than flora and fauna is digitized and accessible. • TK information is digitized and documented in minimum 15 languages and 10 scripts.
Multilingual Data	To facilitate data collation and	• Develop tools, standards, protocols and	• Data collation and dissemination is

Acquisition, Synthesis, and Dissemination (LINGUAL)	dissemination, also integration of multilingual data with English and other international languages.	best practice guides that can facilitate seamless data collation and dissemination in multiple languages.	feasible in 15 languages and 10 scripts by 5th year. • Data collation, analysis, synthesis, and dissemination in all languages and scripts achieved by 10th year.
INFORMATICS			
Data Access and Data Interoperability (DADI)	To facilitate the full use of biodiversity, ecosystem and other databases by establishing an information architecture that enables interoperability and facilitates data mining	• Review and adopt existing standards for data and metadata being used by other similar initiatives • Develop new standards for data and metadata for new data types such as socio-economic, ecological, traditional knowledge, People biodiversity register, and multilingual data etc. • Develop algorithms to search multiple index and original databases simultaneously • Evolve mechanism to establish deep links between varied data types, facilitating cross-walk across domain and discipline information resources • Continue to develop and evolve, test and adopt, and/customize new technologies	o By 1st year software and hardware infrastructure to enable – The NBII portal, participant node portal, collaborative work environment with BoC members, and other groups. o By 1st year Nodes Development, and Data providers release toolkit released. o By 1st year data and software standards are reviewed and adopted for taxonomic, specimen, literature, traditional knowledge and PBR data sets. o By 2nd year NBII Data Portal with use cases launched. o By 3rd year workshops and working group constituted to adopt or develop data interoperability standards for data sets other than taxonomic, specimen,

Data Use, Data Applications and DSS Development (DATA & DSS)	To facilitate multipurpose use of biodiversity information by cross sectoral users through suites of data application packages, ultimately using data for variety of decision making processes.	• Review, adopt and customize existing tools and packages. • Develop new use and application cases, and develop tools and applications for these new types of uses. • Biodiversity data is usable in various decision-making processes, by variety of user groups, under varied situations. • Atleast 3 demo projects completed by the end of 5th year. • Minimum 10-15 demo projects completed by the end of 10th year.	literature, traditional knowledge and PBR data sets. o By 3rd year first set of tools, and protocols for data QA/QC, Data cleaning, error correction, validation, duplication avoidance are released. o By 5th year Data Portal is accessible 40% constitutionally recognized languages. o By 5th year deep link established with ecological, molecular, and other non-biological component of data/information. o By 5th year minimum 3 (three) demo project rolls out with significant data use, data application and decision support system (DSS) by using data integrated through NBII. o By 5th year minimum 2 mirrors established across the country, and Data Archival and Rescue polity and plan is in place.
Information and Communication Technology Implementation (ICTI)	Plan and implement information infrastructure (hardware, software, standards, services and interfaces) that enables achieving vision, mission and goals of NBII in its totality.	• Review, adopt, customize existing information infrastructure models to suite our needs, and if needed develop new features to accommodate all types of data, its acquisition, sharing, use, leading to data intensive decision supports • By 5th year NBII is capable of providing consultancy and mentorship of nations/economies in the world. • By 10th year NBII information infrastructure is rated as the best and adopted as benchmark by other initiatives.	
PARTICIPATION			

Nodes (Data and Participants) Development	To facilitate establishment of Nodes, and Data Provider mechanism with all potential data custodians within India and overseas agencies holding data of our interest	• Nodes Development kit, Best Practice Guidelines, Curriculum for Capacity Building activities, kit for outreach, and estimation of universe of nodes and quantum of data that they can share is achieved in 2 years. • 300 Nodes operational and exchange data by end of 5th year. • 1000+ nodes operational and exchanging data by 10th year. • Data is shared in all Indian languages and scripts in 10 years. • Mentoring activities aim at strengthening NBII as a truly distributed, but national and collaborative, partnership.	o By 1st year MoU, Data Share, Data Use agreements are finalized and circulated to major stakeholders and potential members of BoC. o By 2nd year minimum 50 data providers starts sharing the data (other than PBR) through provider services. o By 2nd year minimum 15 SBB nodes are commissioned, and ready to share PBR data through their provider service. o By 2nd year minimum 5 data provider workshops are held. o By 2nd year model curriculum for biodiversity and ecosystem informatics is developed. o By 5th year 200 data providers starts

Outreach, Capacity Building and IPR (OCB)	To bridge biodiversity information technology "digital divide", "content divide", and "knowledge divide" through training and capacity building, attracting new data providers, to ensure that every one have open access to and can efficiently use biodiversity information, while crediting contributors and respecting national aspirations such as intellectual property rights, economic wealth, bio-security, and social wellbeing	• Outreach to potential data custodians, and data generators so that increase IBIF participation by 20% each year • Work with organization within and outside India to overcome "divide" that exists • Increase awareness of NBII among potential partner organizations and potential data users • Develop "biodiversity informatics" training courses • Encourage and promote synergies amongst various key players within and outside India • Attract new donor agencies/organizations both inkind and funds	sharing the data (other than PBR) through provider service. o By 5th year 25 SBB nodes are commissioned, and ready to share PBR data through their data provider service. o By 5th year minimum 15 data provider workshops are held. o By 5th year minimum 20 data providers and nodes becomes operational through mentor-mentee mode of capacity building. They are funded through separate RFP. o By 5th year minimum 2 short term Schools in Biodiversity Informatics are organized. o By 5th year minimum 5 universities commission graduate and postgraduate level courses in "Biodiversity and Ecosystem Informatics". o By 5th year minimum 2 workshops on "Ecological Forecasting and Natural Resource Accounting" are organized. o By 5th year Memorandum of Cooperation (MoC) are signed with major neighboring, and global initiatives (GBIF, OBIS, GISIN, Species2000, GTI, BioNET International, etc.) and museums (NHM London, NMH, Paris, Kew Botanical Garden, Smithsonian, etc.) and data centers (NASA, NODC, NBII, ETI, etc.) for exchange/share of data, capacity building, etc. o By 5th year donor conference to attract investment and collaborations with national, regional, global commercial players in the area of biodiversity conservation, natural resources bioprospecting, pharmaceutical, and information technology is organized. o By 5th year NBII is included in environmental and nature conservation

Table 5.1 NBII would achieve its vision and objectives through implementation of functions and content specific work programs

Figure 5.1: NBII should be perceived as registry of shared biodiversity databases.

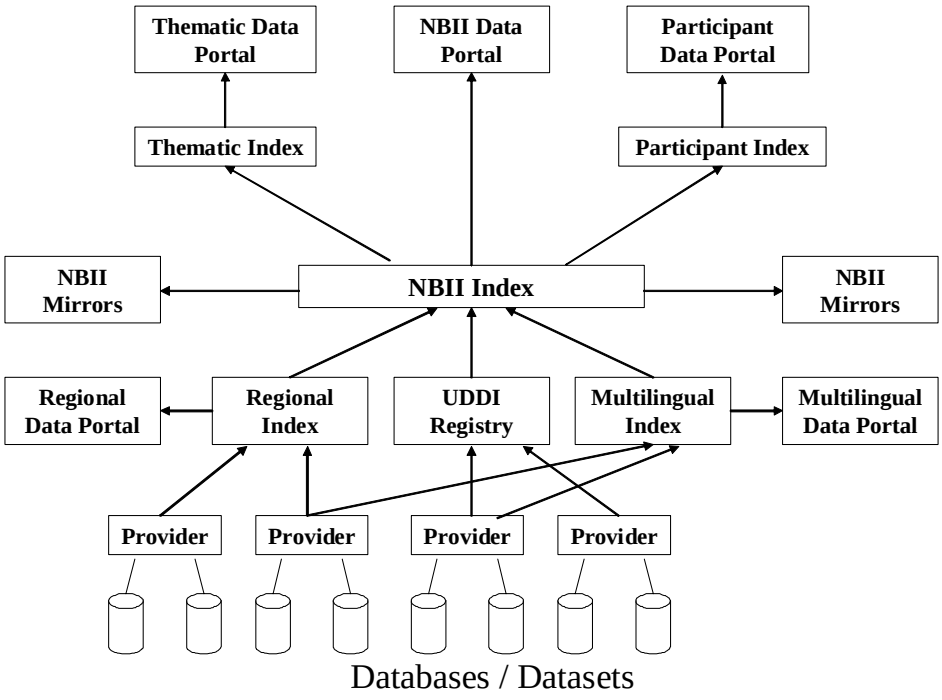


Figure 5.2 High level Architectural Overview and flow of processes in NBII Implementation.

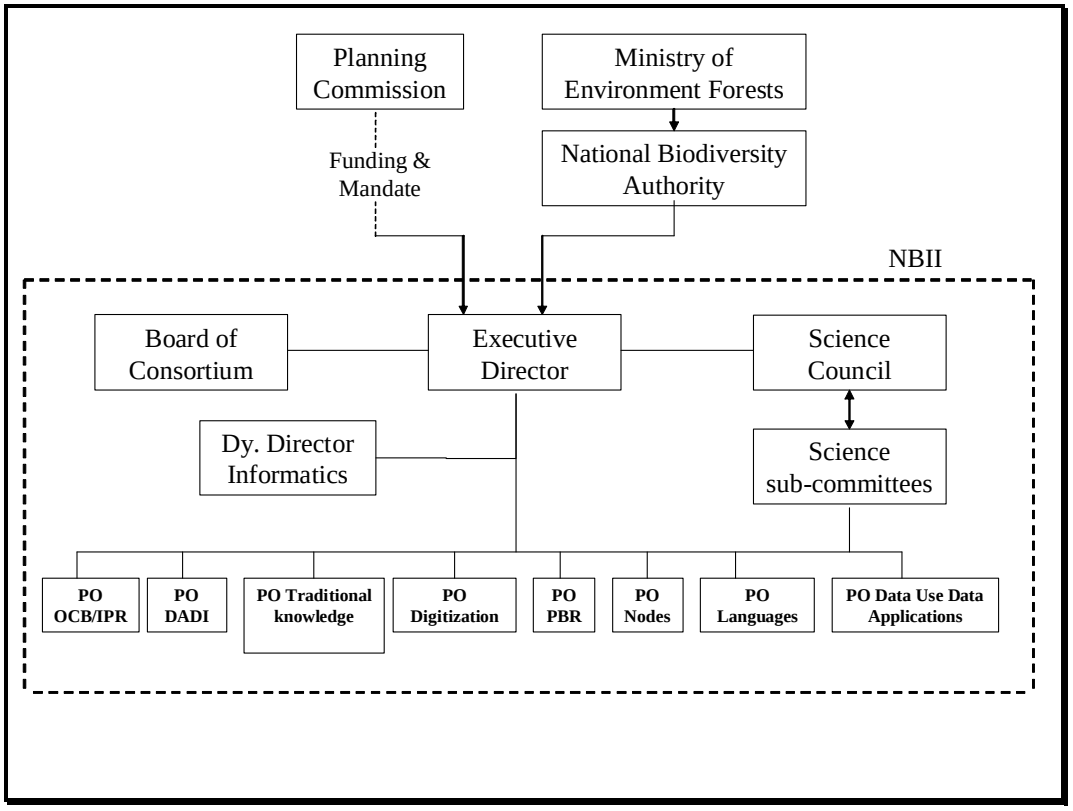


Figure 5.3: Governance structure of NBII

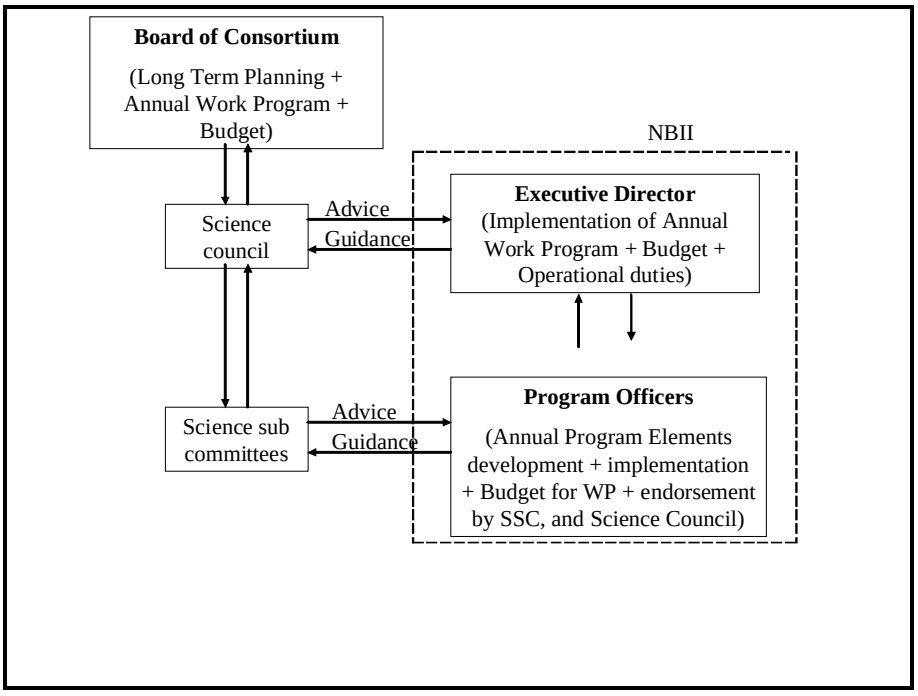


Figure 5.5: Operational Workflow of NBII

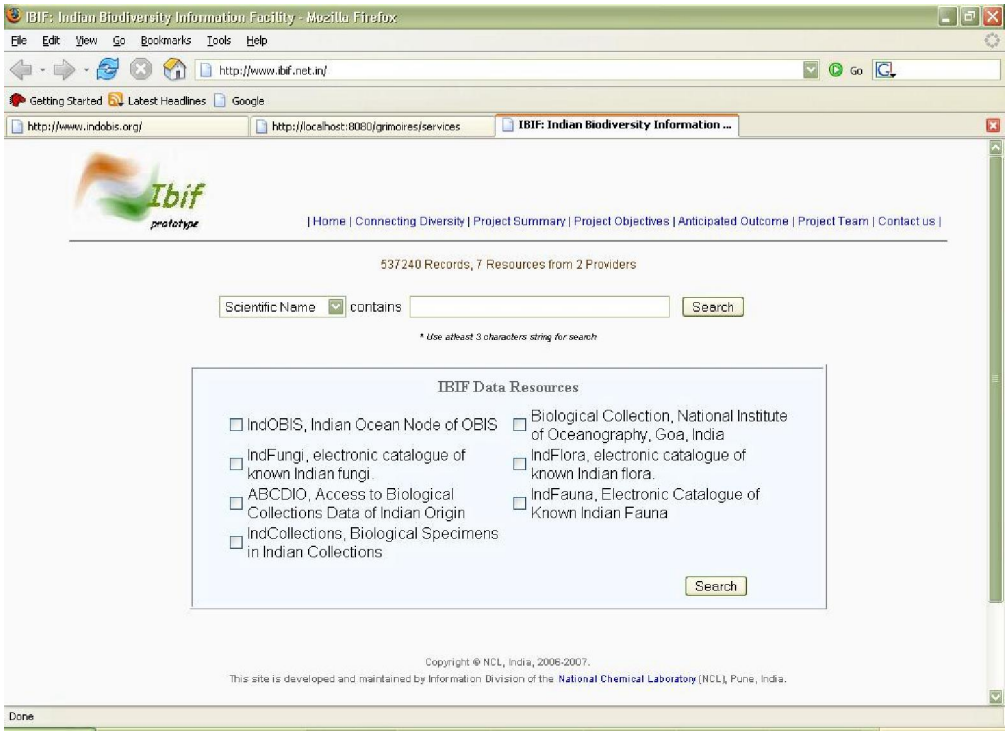


Figure 5.7: IBIF*prototype* uses open source UDDI registry to integrate over 500,000 species and specimen records from 7 distributed databases.

ANNEXURES: I -III

Annexure I

BIR: Biodiversity Information Resources Database

Introduction

Biodiversity and ecosystem informatics are emerging disciplines, which are both dynamic and demanding in nature. In recent past technical and political developments, have led to generation of vast amount of data about world's biodiversity and ecosystems that harbor these biotic resources. In order to manage, analyze, as well synthesize these flooding datasets, many data management and databas(e)ing, analysis, visualization and modeling tools have been developed or customized. Most of these tools are aimed at easy data management and improving usability of biodiversity and ecosystems related data leading to sustainable utilization of biotic resources.

However, similar to uneven distribution of biodiversity and biodiversity data (Gaikwad and Chavan, 2006), there is uneven development, as well usage pattern of these biodiversity and ecosystem data management and analysis tools. Thus, justifying Bisby (2000), that the massive development of biodiversity-related information systems on the Internet has created much that appears exciting but chaotic, a diversity to match biodiversity itself. On one end, it has led to much needed advancement in the area of biodiversity and ecosystem informatics so that community at large and resource managers in particular could take informed decisions with regard to sustainable management of the biological resources, while at other end it is leading to emerging of biological data management systems, analytical and visualization tools being developed by two geographically separated tools which are aimed at performing near-similar tasks. 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. For instance, Chavan and Krishnan, (2003) listed 29 software tools for biological collections management packages, of which 12 each are for taxon independent biological collections, and herbarium management respectively. This creates confusion in the minds of its potential users as to which tools should fulfill their objectives and purpose. Especially in the developing and under-developed mega-biodiversity world, where these disciplines are yet to receive much needed encouragement and support,

such a scene is used as excuse for not undertaking biodiversity and ecosystem data management, analysis, and prediction activities.

One of the major reasons for this growing catastrophe is unavailability of metadata about these resources at a single click of a mouse, as the development of these tools is often scattered geographically and isolated from one another. This results into development of information resources, which are non-interoperable with each other, do not comply with standards facilitating easy integration, thus hindering the process of data exchange and sharing as well across multiple resources data analysis, modeling, visualization, and synthesis. Such a state further encourages developers to undertake development of tools that are more and more customized, and not universally applicable, and thus further limiting their spectrum of usage. Thus, there is growing and urgent need for web based repository of biodiversity and ecosystem information resources. Such a resource would reveal the shortfalls of existing tools vis-à-vis immediate and long term requirements of the biodiversity and ecosystem data users community. This would facilitate humankind in learning how to exploit massive datasets, learning how to store and access them for analytical purposes, develop methods to cope with growth and change in data and make it possible to 'repurpose' previously existing data so as to comprehend and sustainably utilize the biodiversity resources of the world (Lane et al. 2000).

To address this issue, authors have developed BIR (Biodiversity Information Resources) database to collate metadata about distributed and isolated biodiversity and ecosystem information resources. Accessible at http://www.ncbi.org.in/BIR/ provides metadata information about 1383 such resources. In this Annexure, while describing development of BIR, authors have attempted to analyze these resources with respect to their taxon and geographic scope, and resource types.

BIR: Development and Features

In order to develop BIR that would fulfill the objective of providing single click of a mouse metadata about the scattered and isolated biodiversity and ecosystem information resources, a survey was commissioned to estimate the universe of such resources. It reveals that little over 1500 biodiversity and ecosystem information resources are currently available. These resources could be broadly categorized into four categories, viz., information system and networks, database/databanks, softwares, and standards/protocols. In order to collate metadata of these resources, and to use it for analysis purposes a data collection performa was developed, which was later used

in developing the database structure of BIR. As detailed in Table 1.3, BIR consists of one table with 17 parameters about which metadata can be collated. These data parameters and their sub-specifics were evolved to achieve classification of resources based on its type, geographic and taxonomic scope, functions, etc. As on date metadata of 1383 resources have been collected and manually entered into BIR. Detailed analysis and inferences are presented in subsequent sections.

- **Ecosystem/Habitat scope** – Attempt has been made to classify each of the resource documented in eBIRD based on ecosystem and habitat data that it documents and disseminates. Table 1.4, depicts the basis of the classification along with major categories and sub-categories.
- **Taxon Scope** – Taxonomic scope of a resource was determined using Calivlier-Smith's eight kingdom classification viz., Animalia, Plantae, Chromista, Fungi, Virus, Bacteria, Protozoa, and Archae. Those resources document data about all eight kingdom's were classified as "all biota". Few resources also cover more than one taxon.
- **Resource Type** – Based on the function of a resource, it is classified into one of the four resource types, viz. (a) information systems/networks, (b) database/databanks, (c) software tools, and (d) standards and protocols. As listed in Table 1.5, categories (a), (b), and (c) were further classified based on kind of data that they document or functions that they perform.

MySQL 4.1 was used as backend for developing BIR. Java Server Pages (JSP), together with JavaBeans, HTML, Javascripts, and Cascading Style Sheets (CSS) has been used to develop web based data entry and data retrieval modules of BIR. Tomcat 5.0 has been implemented as web server for BIR.

While documenting metadata details of 1383 biodiversity and ecosystem information resources, architecture and web based data entry modules of BIR have been vigorously tested. Though metadata of 1383 resources have been documented as initial exercise, authors were aware of their limitation to locate and acquire metadata of each of the existing resources, as well those developed in the future. Thus, "Suggest a Resource" module was developed, with an objective to facilitate developers, owners, and custodians of a resource to register their resource in BIR. This would ensure that BIR provides up-to-date and current metadata of increasing number of biodiversity and ecosystem information resources.

Moderate search module has been developed to facilitate retrieval of metadata of resources. Resources documented in BIR could be searched based on (a) any string in resources title, (b) URL of a resource, and (c) keywords as indexed from a resource title. As depicted in Fig. 1.5, detailed metadata of each resource could then be visited by clicking on resource name, where in URL of a resource links to homepage of a resource.

Summary

It is my belief that BIR would facilitate up-do-date, and current documentation of metadata of existing and new biodiversity and ecosystem information resources. As it would help in understanding areas of gap in biodiversity and ecosystem informatics, it would also encourage collaborations between research groups spread across the globe. BIR brings together disparate resources and analytical tools for biodiversity researchers to solve problems in biodiversity and analyze biodiversity patterns. However, metadata repository such as BIR needs to be constantly updated, if our goal is to bridge the imbalance between the biodiversity and ecosystem informatics products and distribution of biodiversity and its data.

Data Parameters	Description of data parameters
Resource Name	Title of the resource
Resource Host / Custodian	Custodian / Owner / Developer of the resource
Availability	Availability of resource (online / offline)
Availability Details	Modes of availability (Online – web /ftp, Offline – CD/DVD)
URL	Universal Resource Locate of the resource
Geographic Scope	Geographic coverage of the resource (Global/Regional/National/Local/Province/Area)
Geographic Scope Details	Name of the geographical areas that resource covers
Ecosystem / Habitat Scope	Ecosystem and habitat scope of the resource. Ecosystem and habitats are classified into three categories with each of them having sub-categories. (a) Ecosystem / Habitat types, (b) Nature of protection, (c) Biomes
Taxon Scope	Taxon scope of a resource. Cavilier Smith's 8 kingdom classification schema is adopted to categorize a resource, which few resources covering all 8 taxons are classified as "ALL BIOTA".
Resource Type	Functional type of resource. Resource types are classified into 4 categories with each of them having sub-categories. (1) Information System / Network, (2) Database / Databank, (3) Software Tools, and (4) Standards / Protocols.
Description	Description of a resource.
Citation	Citation details of a resource.
Remarks	Usability remarks by developers of eBIRD.
Reference	Publications that referred to a resource.
Copyright	Copyright status and details of a resource.
Current Status	Accessibility and functionality details of a resource.

Table 1.3: Descriptions of BIR Data Parameters

Ecosystem / Habitat types	Nature of protection	Biomes
q Forest q Savanna q Shrubland q Grassland q Desert q Rocky Areas q Caves and other Subterranean Habitats q Wetlands q Mangroves q Freshwater Ecosystems q Marine Ecosystems	q Strict Nature Reserve q National Park q Sanctuary q Biosphere Reserve q Natural Monument q Protected Landscape / Seascape q Sacred Groves q Biodiversity Hotspot / wilderness areas	q Rainforest q Grasslands q Desert q Taiga q Temperate q Tundra
Enter other Ecosystem scopes, if any: Other Scope:	Scope Details:	

Table 1.4: Sub-categories of Ecosystem/Habitat Scope in BIR

Information System / Network	Database / Databank	Software Tools	Standards / Protocols / Schema
q Global q Regional q National q Local q Thematic	q Taxonomic § Nomenclature Database § Checklist / Flora / Fauna / Catalogue / Index § Species Database q Natural History § Specimen Database § Fossil Database § Culture Collection Database q Bibliographic / Referral q Phyllogenetic q Genomic q Observation / Survey q Peoples Biodiversity Register (PBR) q Image Database q Geospatial Database q Climate Database q Educational Database	q Analytical softwares q Modeling softwares q Natural history softwares q Data management system q Electronic field guides q Expert system / artificial intelligence q Search engines	q
Enter other resource types, if any			

Table 1.5: Sub-categories of resource types in BIR

Biodiversity Information Resources (BIR)

| Administration | Home |

Resource Name	:	IndFauna-Electronic Catalogue of Known Indian Fauna
Resource Host / Custodian	:	Information Division of the National Chemical Laboratory (NCL), Pune, India.
Availability	:	Online
Availability Details	:	Web
URL	:	http://www.ncbi.org.in/biota/fauna/
Geographic Scope	:	National
Geographic Scope Details	:	India
Ecosystem / Habitat Scope , Ecosystem / Habitat types	:	
Ecosystem / Habitat Scope , Nature of protection	:	
Ecosystem / Habitat Scope , Biomes	:	
Taxon Scope - Taxon Details	:	**Animalia** - Animalia
Resource Type	:	Database / Databank
Resource Type Details	:	Taxanomic
Database Type	:	Species Database
Description	:	This Database is a completion of baseline information for over 94340 scientific names a son date,Synonyms as on date: 52894,Common names as on date: 14927, Localities as on date: 6520 known Indian faunal species.The data was pulled from over 7800 sources of literature such as Fauna of British India, and monographs of the Zoological Survey of India along with several reports and journal articles. Other than this the data from several online checklists and databases. Alongwith basic taxonomic information, IndFauna collates data regarding synonyms, common names, and occurance records as listed in literature over past 100+ years.
Citation	:	Retrieved [11/22/2006], from the NCBI - Indian Fauna (http://www.ncbi.org.in/biota/fauna).
Remarks	:	NCL Centre for Biodiversity Informatics (NCBI) is an effort to collect, collate, analyze, predict and disseminate knowledge about Indian biota and its environ. IndFauna, electronic catalogue of known Indian fauna was declared as India's best e-Content 2006. IndFauna was awarded with Manthan-AIF Award 2006 in e-Environment category. Manthan-AIF Award is India's national pre-selections for the World Summit Award 2006, and hence IndFauna, would be nominated as India's nomination in e-Environment category for World Summit Award 2006.
Reference	:	
Ownership	:	NCL, India 2001-2005.
Current Status	:	Last updated on 21/11/2006

| Administration | Home |

Figure 1.5: Metadata of a typical resource BIR.

Annexure II

Open Access Geospatial Data Repository (OAGDR)

In recent times, the need of having an easily accessible spatial data infrastructure (SDI) has been strongly emphasized (Ramachandran, 2000, and NSDI Task Force, 2001). Based on these recommendations, a "National Spatial Data Infrastructure" (NSDI, 2007) was formed with several participating agencies. Similar infrastructures in the developed part of our globe (FGDC, 2007) has resulted in availability of large volumes of data in public domain most of which is generated by government agencies. For example, NASA makes available Landsat Thematic Mapper data in UTM projection format (NASA, 2007) most of which has a resolution of less than 20 m. This complements digital elevation model (DEM) data generated through the Shuttle Radar Topographic Mission (SRTM), which is also in public domain.

Access to such geospatial data is critical for various GIS and ecological modeling applications in biodiversity studies (Murthy et al., 2003), leading to informed decisions on conservation and sustainable utilization of natural resources. However, free access to such data continues to be an impediment in majority of the developing and under-developed nations. For instance, most commonly required data, such as administrative boundaries, hydrology, lakes, water-bodies, road and rail networks and population statistics, are still unavailable in public domain. Thus, *open access* to spatial data is still a dream, although the data is available. In the year 2000, *Current Science* through a special issue (Ramachandran, 2000) made a strong case for public access to Indian geospatial data highlighting concerns, advantages and drawbacks. In most cases, non-accessibility to data generated through public funding including those by national agencies was seen to be a primary weakness (Gupta, 2000, and Srikantia, 2000). However, concrete actions are still awaited to make such data available in public domain.

To overcome this problem and foster a community-driven effort towards building a geospatial data infrastructure, an Open Access Geospatial Data Repository (OAGDR) has been developed. Accessible at http://www.ncbi.org.in/oagdr/ (Fig. 4.6), the need for such a repository arose from difficulties experienced while accessing such data for a web-GIS species-distribution mapping system. This experience is in tune with that of the other groups within the country. Thus, the primary objective of

OAGDR is to bridge this gap and improve public domain accessibility to spatial data generated and processed by various working groups. It is appealed to those involved in generation, processing, and employing geospatial data for various analytical and modeling studies to contribute to OAGDR along with necessary metadata. Once populated, such a repository would not only improve accessibility to public domain geospatial data, but also prevent duplication of efforts. This in turn would allow the biodiversity community to spend more time on ecological modeling, leading to better management of our natural resources. Such an open access model would not only satisfy our commitment towards "information commons" (Bollier and Watts, 2002), but also certainly enhance our Nation's ability to take environmentally sound informed decisions.

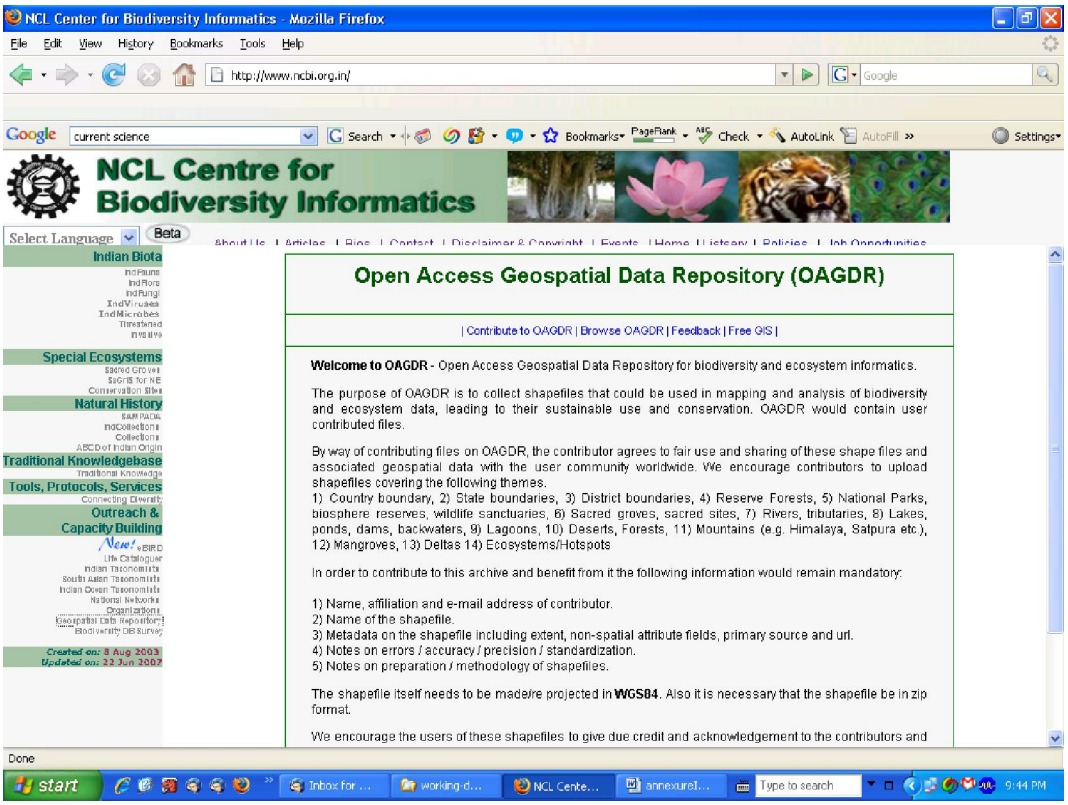


Figure 4.6: OAGDR facilitate contributing shapefiles which are often used in biodiversity research.

Annexure III

Connecting Diversity: Pilot project for development of an interoperable framework for connecting distributed and heterogeneous bioresourses databases

Project Summary

A wealth of information exists about Indian biodiversity and its biotic resources. However, this information is scattered in several structured and unstructured information sources, which are distributed, heterogeneous. These are developed on different platforms, using variety of database management systems. They also differ in their format, content types, database structure (data models), as well as access mechanism. Efficient, appropriate, authentic and current information retrieval is not only like re-searching, but also time consuming. This adversely affects the sustainable utilization of our biotic resources and their conservation.

Based on the proposal submitted by the author, Government of India's Department of Biotechnology (DBT) sanctioned one year pilot project to develop web based interoperable framework for interconnecting these distributed and heterogeneous databases by establishing resource discovery and access control mechanism. During the pilot phase experience open source UDDI registry was established to integrate species and specimen data. Through two DiGIR providers, 7 distributed databases were harvested and over 500,000 species and specimen data records were indexed into central registry. In order to discover and disseminate these records IBIF prototype portal (http://www.ibif.net.in/) was commissioned.

Thus, pilot project achieve its goal to demonstrate that UDDI registry based interoperable framework can be used for integrating both legacy and emergent data and services. It further demonstrated that it would aid in increased integrity, interoperability, scalability and extensibility amongst the data sources, we foresee that it would not comprise on autonomy of the database developers and owners, as they would not lose control over their own data resources. Most importantly it would achieve both flexible and optimal use of data sources, despite heterogeneity of different data sources.

Why Connecting Diversity: Need for a pilot project

During past decade several institutions and individuals in India have developed databases and information systems related to biodiversity and ecosystems

(Chavan and Krishnan, 2004). However, what is interesting and encouraging is that even some of the prominent non-biodiversity informatics initiatives have developed some good bioresources databases. For instance, closer review of the databases developed as part of the Department of Biotechnology (DBT) supported Biotechnology Information System (BTISNet) reveals that more than 50% of the databases are either biodiversity focused or associated with the biotic resources and its environ.

However, these databases are not only isolated, distributed, and exists independently, but they have been developed using heterogeneous techniques, and approaches. These databases cover diverse aspects such as bibliography and referral, experts and institutions, collections, taxonomic, observations, geospatial, conservation, ecological and environmental, and people or traditional knowledge digital libraries (Fig. 5.8). However, very few are available online and those offline needs to be launched online. Some of these are well-structured; others are largely project /species specific and/or unstructured. There is no framework to link the scattered data so as to facilitate exchange of data amongst the different databases.

Accessing these databases poses challenge due to their heterogeneous nature and incompatibility with each other. This is being seen as major obstacle in integrating information together from these data sources and makes them available through single portal to wide variety of user community anytime and anyplace. The challenge before us is to set standards and make technological choices that would facilitate networking of databases, and add real value to the information being brought together, while at the same time, (a) maintain the autonomy of the various databases and ensure that there is abundant scope for expression of creativity and originality of the designers and managers of different databases, as also (b) ensure the security of the data, and (c) protect all legitimate intellectual property rights.

This calls for coordinated efforts in networking of these existing biodiversity databases to take advantage of synergies, and to link all of these to activities leading to value addition. As a part of this process, the existing biodiversity databases will need to be considerably augmented and strengthened, and new ones created. We will have to come up with novel ways of bringing on board the substantial knowledge base of country's barefoot ecologists and grass-roots innovators. We also need to devise a country wide decentralized system of monitoring biodiversity. Such a decentralized system could serve to enhance the quality of education by engaging teachers and

students in first hand understanding of biodiversity and associated knowledge and in creating, using, and managing electronic databases, including those employing Indian languages.

To address these concerns and to demonstrate the technological feasibility, and ease of implementation in national scenario, it was felt necessary to commission a pilot project that could integrate few types of biodiversity data from distributed, heterogeneous databases.

Project Objectives

Mission of this pilot project was to achieve,

"development of an inteoperable framework for connecting isolated, distributed and heterogeneous, biodiversity and bioresources databases" using web services architecture.

This would be achieved through following objectives;

- Development of biodiversity and bioresources data providers' registry service.
- Development of Biodiversity Database Interoperability (data aggregation, indexing, transformation) tools.
- Biodiversity Data Access (Search engines, XML web services and HTML interface) schemas and portal.

Architectural Framework

Call as IBIF Prototype, information system of this project is based on an interoperable framework which used web services approach. This consists of distributed data providers, a central registry and a central portal to access the data (Fig. 5.8), all communicating using standardized XML messages.

The data providers are installed with wrapper software that maps the local database schema into the Darwin Core format, translates XML-encoded queries coming from Internet into SQL, and return data or metadata the same way. Currently supported protocol for this communication is DiGIR. DiGIR is a client/server protocol for retrieving information from distributed resources. It uses HTTP as the transport mechanism and XML for encoding messages sent between client and server. The DiGIR Provider is a service application that has no implicit user interface, but rather is intended for machine-to-machine communications. DiGIR protocol is implemented in "provider" software, works with collection database installations to make data

records searchable and accessible on the network. DiGIR Provider software using DiGIR protocol provides an efficient way to request and retrieve information from multiple, distributed databases, each with their own unique internal data structures or schema. The source for DiGIR Provider information and software is: available at http://digir.sourceforge.net/.

The data providers can announce their presence in the IBIF*prototype* Registry (http://registry.ibif.net.in/). This is a central service that is based on a registry tool called Grimoires. As an UDDIv2 compliant registry for Web Services, Grimoires is, itself, implemented as a Web Service. It was designed and developed in the context of the myGrid project (Stevens et al., 2003) and is now part of the OMII project seeks to support capabilities for semantic reasoning and inquiry, which subsequently increases its usability range (Tan et al., 2005). It is important for the registry to be persistent, so that the contents of the registry are not lost at the moment of a crash. The persistence of GRIMOIRES relies on that of the RDF (http://www.w3.org/RDF/) triple store used in GRIMOIRES. It uses Jena (http://jena.sourceforge.net/) as the triple store. Jena can be made persistent in different ways, e.g., by using a relational database, a file-based hash table, or simply a plain text file to store the triples. Currently, IBIF*prototype* uses a relational database called MySQL as RDF triple storage for Grimoires registry.

Crawler gets the list of network addresses for data providers from IBIF*prototype* Registry. It is a Java Program typically works within Java Runtime Environment. It uses MS-DOS commands to issue search requests that retrieve matching data from DiGIR providers. It can be used to query one or more of the Providers through a single query as well as to debug Providers. In the case of DiGIR, this query format or query schema is based on an information content standard called the Darwin Core. The DiGIR protocol uses the Darwin Core standard to define the content or semantics of data fields and employs XML as the language for the exchange of queries and resultsets between portals and providers. When a provider receives a DiGIR query, it translates the request from DiGIR XML format to the local query language of the target collection database and tells the database manager to run the query. The database returns any matching records to the DiGIR provider software, which then translates the database response into the community standard Darwin Core XML format and returns the data to the requesting Crawler. The Crawler simultaneously sends the data to the temporary database. The data from temporary database is then transferred to the main database.

However, after registration, a Node Manager from an existing IBIF*prototype* participant (coordinator from a state/ or organization) endorses the data provider as somebody that is indeed sharing scientific biodiversity data. This is a rudimentary quality assurance step, although the Node Manager does not currently scrutinize the actual data records. Before such an endorsement is received, IBIF*prototype* does not make the data available.

IBIF*prototype* operates a prototype data portal at http://www.ibif.net.in/ that can be used to search, browse and drill into the data of the endorsed data providers. This central gateway to IBIF maintains a central index of the most important data elements of all the records of the data providers, which can be accessed through the network frequently.

Fort the purpose of this pilot study two types of databases were selected (a) species, and (b) specimen databases. Species databases include – (i) IndFauna, electronic catalogue of known Indian fauna, (ii) IndOBIS Catalogue of Life, (iii) IndFungi, electronic catalogue of known Indian fungi, and (iv) IndFlora, electronic catalogue of known Indian flora. Specimen databses include – (i) ABCDIO, Access to Biological Collections Data of Indian Origin, (ii) IndCollections, biological specimens in Indian collections, and (iii) Biological Collections of the National Institute of Oceanography. Two providers were commissioned to pull together data from these databases.

Project Achievements

Over 550,000 data records were harvested from 7 resources through 2 providers. Two types of search features were developed.

- All the records from all resources could be search based on (a) scientific name, (b) genus, (c) locality, (d) identifier, (e) family, (f) species, (g) country, and (h) continent / oceans (Figure 5.9).
- Second search feature facilitate data retrieval from a specific resource ((Fig. 5.10).

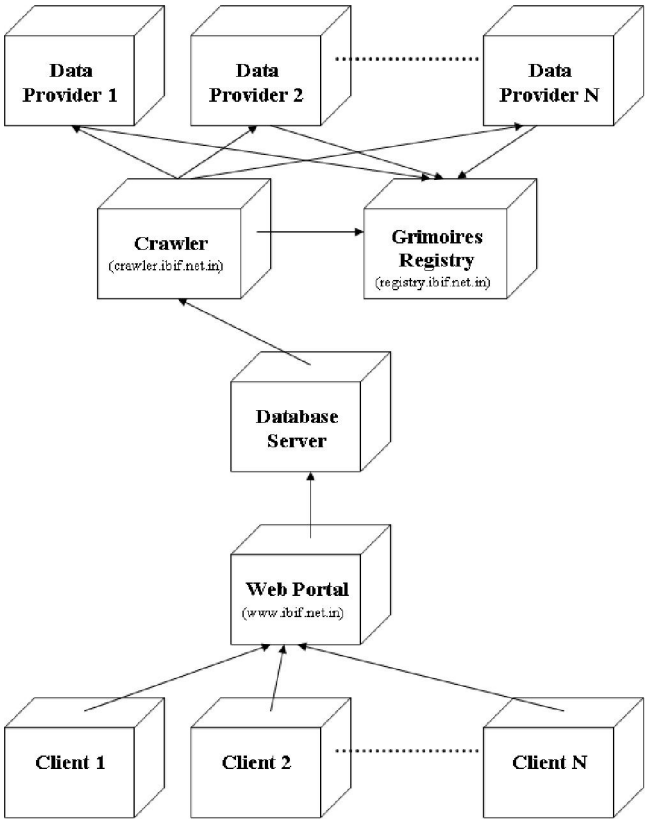


Figure 5.8: Framework, Architecture and Components of IBIF

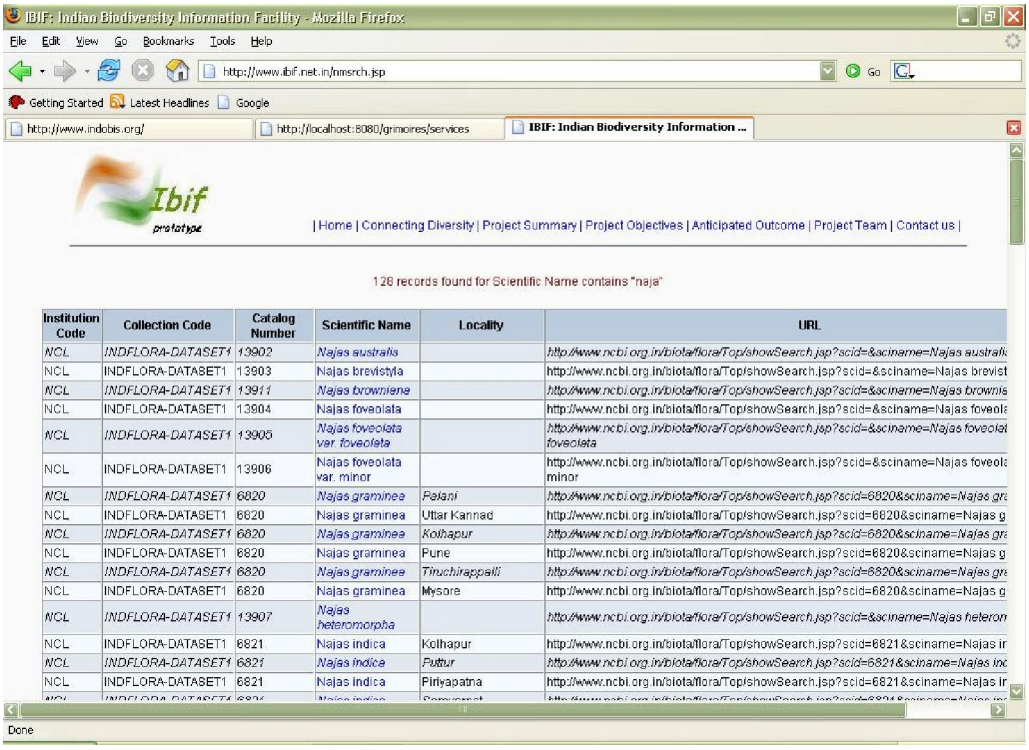


Figure 5.9: Search engine facilitate searching data across all resources.

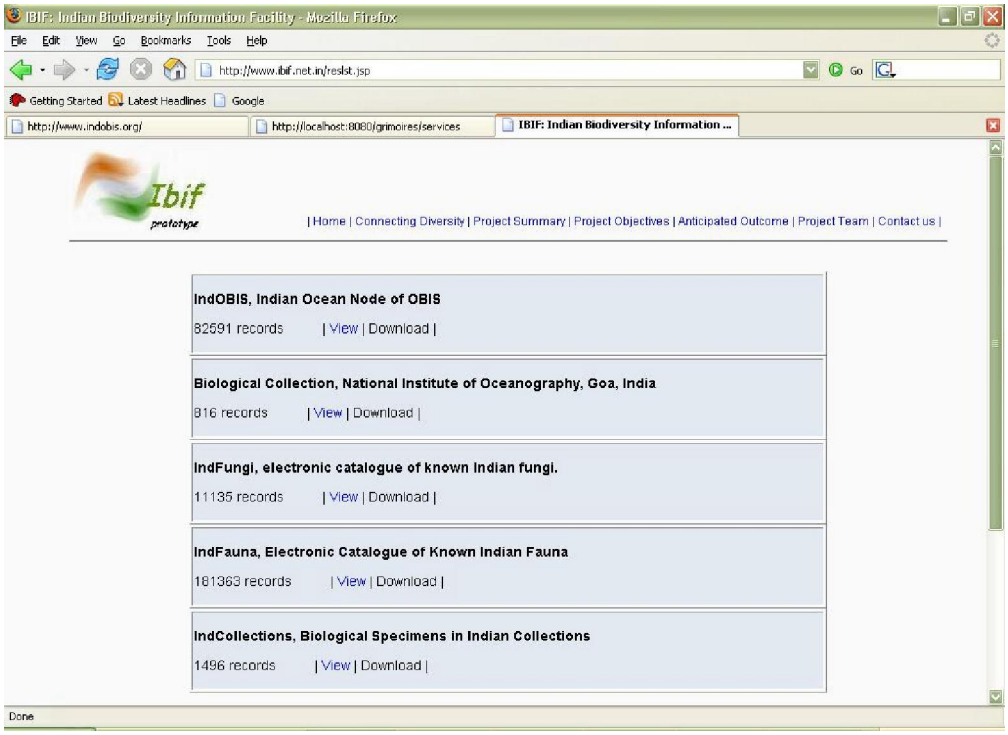


Figure 5.10: Retrieval of data from a specific resource.

REFERENCES

REFERENCES

§ Aditya, G. and S. K. Raut. (2001). Food of the snail, *Pomacea bridgesi*, introduced in India. *Current Science*, 80 (8): 919 – 920.

§ Agarwal, V. C. (1998). Mammalia. *In*: Alfred J. R. B., A. K. Das and A. K. Sanyal (eds.) *Faunal diversity in India*. Zoological Survey of India, Kolkata. 459 – 469.

§ Agosti, D. and N. F. Johnson. (2002). Taxonomists need better access to published data. *Nature* 417: 222.

§ Alfred, J. R. B. (1998). Faunal diversity in India: An overview. *In*: Alfred J. R. B., A. K. Das and A. K. Sanyal (eds.) *Faunal diversity in India*. Zoological Survey of India, Kolkata. 1-9.

§ Austin, M. P., Nicholls, M. D., Doherty, M. D., and J. A. Meyers (1994) Determining species response functions to an environmental gradient by means of a b-Fucntion. *Journal of Vegetation Science*, 5(2): 215-228.

§ Bawa, K. (2007). India biodiversity portal: meeting at the National Knowledge Commission. *Personal Communication dated 30 August 2007*.

§ Beaman, R. and B. Conn. (2003). Automated geoparsing and georeferencing of Malesian collection locality data. *Telopea*, 10: 43-52.

§ Beaman, R., Wieczorek, J., and S. Blum. (2004). Determining space from place for natural history collections in a distributed digital library environment. *D-Lib Magazine*, 10, accessible at http://www.dlib.org/dlib/may/04/beaman/beaman.html.

§ Berendsohn, W. G. (1997). A taxonomic information model for botanical databases: the IOPI model. *Taxon*, 46: 283-309.

§ Berendsohn, W. G. (2001) Biodiversity Informatics. *Proc. 2nd Natl. Colloquium on global change research, Bad Honnef, Jan 26-27, 2001*, Accessibleat http://www.bgbm.org/BioDivInf/def-e.htm.

§ Berendsohn, W., Guntsch, A. and Ropert, D. (2003). Survey of existing publicly distributed collection management and data capture software solutions used by the world's natural history collections. *Global Biodiversity Information Facility, Copenhagen*, 44pp.

§ Bhattacharya, S. B. 1998. Acanthocephala: In *Faunal Diversity in India*. Ed. by Alfred J.R.B., A. K. Das and A. K. Sanyal, Zoological Survey of India, Kolkata. 93 –98.

§ Bingham, C. T. (1907). The Fauna of British India. Butterflies. *Teylor and Francis, London*, Vol. II : 1 – 472.

§ Bisby, F. A. (2000). The Quiet Revolution: Biodiversity Informatics and the Internet *Science*. 289(5488): 2309-2312.

§ Bisby, FA, YR Roskov, MA Ruggiero, TM Orrell, LE Paglinawan, PW Brewer, N Bailly, J van Hertum, eds. (2007). Species 2000 & ITIS Catalogue of Life: 2007 Annual Checklist. *Species 2000: Reading, U.K*, digital resource at www.catalogueoflife.org/annual-checklist/2007/.

§ Bisby, F. A. (2007). Investments in Species2000. *Personal Communication dated September 18, 2007*.

§ Bollier, D., and T. Watts. (2002). Saving the Information Commons – A new public interest agenda in digital media. *New America Foundation, Washington DC*, 2002, 83 p. Accessible at http://www.publicknowledge.org/pdf/saving_the_information_commons.pdf.

§ Booth, T. H. (1996). Matching trees and sites. Proceedings ACIAR workshop held in Bangkok, Thailand, 27-30 March 1995, ACIAR Proceedings no. 63.

§ Borowiec L. and J. Swietojanska (2002) Available at http://www.biol.uni.wroc.pl/cassidae/katalog%20internetowy/aspidimorpha.htm.

§ Brands, S. J. (comp) (1989-2005). Systema Naturae 2000. Amsterdam, The Netherlands. Accessible at http://sn2000.taxonomy.nl/.

§ Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone, (1984). Classification and Regression Trees, *Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA.*

§ Camerano, L. (1912) Gordiens du musee Indien. *Records of the Indian Museum*, 7: 215-216.

§ Canhos, V. P., S. R. Gioanni, and D. A. L. Canhos (2004a). Global Biodiversity Informatics: Setting the scene for a "new world" of ecological modeling. *Biodiversity Informatics*, 1: 11-13.

§ Canhos, D.A.L., P. Uhlir, and J. M. Esanu (eds.) (2004b). Access to environmental data. *Summary of an Inter American Workshop, Committee on Data for Science and Technology, Paris*.

§ CBD (2004). Consideration of the results of the meeting on "2010 – The Global Biodiversity Challenge". *Convention on Biological Diversity, Report of the meeting of experts 21-23, May 2003, (UNEP/CBD/COP/7/INF/22,*. Accessibleat http://www.biodiv.org/doc/meetings/cop/cop-07/information/cop-07-inf-22-en.doc

§ CBIF (2003). Noctuoidea of Ontario. 2003, Accessible at: http://www.cbif.gc.ca/spp_pages/noctuoidea/provinces/onnoct1_e.php

§ CES (2007). Center for Ecological Sciences, ENVIS Centre web site. Accessible at http://ces.iisc.ernet.in/hpg/envis/lkwestern.htm.

§ Chapman, A. D. (2004). Technical report, March 2003-2004. *Biota/FAPSP, centro de Referencia em Informaco Ambiental, Caminas, Brazil*.

§ Chapman, A. D. (2005a). Uses of primary specues-occurrence data, version 1.0. *Report for the Global Biodiversity Information Facility, Copenhagen*, 106pp.

§ Chapman, A. D. (2005b). Principles of Data Quality, version 1.0. *Report for the Global Biodiversity Information Facility, Copenhagen*, 58pp.

§ Chapman, A.D. (2005c). Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0. *Report for the Global Biodiversity Information Facility, Copenhagen.* 72pp.

§ Chapman, A. D. and Busby, J. R. (1994). Linking plant species information to continental biodiversity inventory, climate and environmental monitoring. In: *Miller R.I., eds. Mapping the diversity of nature. London: Chapman & Hall*; 1994. pp. 177-195.

§ Chapman, A. D. and Milne, D. J. (1998). The impact of Global warming on the distribution of selected Australian plant and animal species in relation to soils and vegetations. *Environment Australia, Canberra*.

§ Chavan, V., and S. Krishnan, 2003. Natural history collections: A call for national information infrastructure. Current Science, 84(1): 34-42.

§ Chavan, V., A. Watve, and S. Krishnan (2003). Electronic catalogue of known Indian fauna. *Current Science.* 85(11): 101.

- Chavan, V and S. Krishnan (2004). Biodiversity Information in India: Challenges and Potentials. *In Building Capacity in Biodiversity Information Sharing* 2003. 114-120.
- Chavan, V., A. V. Watve, M. S. Londhe, N. S. Rane, A. T. Pandit, and S. Krishnan. (2004). Cataloguing Indian biota: the electronic catalogue of known Indian fauna. *Current Science*, 87(6): 749-763.
- Chavan V., N. S. Rane, H. V. Ghate and S. Krishnan (2005a). IndCollections: biological specimens in Indian collections. *Current Science*, 89(9): 1454-55.
- Chavan V., N. Rane, A. Watve and M. Ruggiero. (2005b) Resolving Taxonomic Discrepancies: Role of Electronic Catalogues of Known Organisms. *Journal Biodiversity Informatics* 2: 70-78.
- Chavan, V., A. V. Watve, N. S. Rane, and S. Krishnan (2005c) Response to: Chandra K. (2005) Cataloguing Indian biota. *Current Science*, 88(4): 532-533.
- Chavan, Vishwas and S. Krishnan (2006), Establishing and enriching collaborative framework for connecting biodiversity databases. Concept paper prepared for the 2nd meeting of the NBA Expert Group on Biodiversity and Traditional Knowledge Databases, held at the National Chemical Laboratory, Pune on Feberuary 3, 2006.
- Coleoptera. (2005). Coleoptera. Accessible at www.coleoptera.org
- Corey, M. J., M. Abbey, and I. Abramson (2002). Oracle9i: A Beginners guide, *McGraw Hill Professional*, 535pp.
- Costello, M. J. (2000). Developing species information systems: the Eurpoean Register of Marine Species (ERMS). *Oceanography*. 13: 48-55.
- Costello, M. J.; Bouchet, P.; Boxshall, G.; Emblow, C. and E. Vanden Berghe (2004). European Register of Marine Species. Accessible online at http://www.marbef.org/data/erms.php.
- Costello, M. J., and E. V. Berghe. (2006). Ocean biodiversity informatics: a new era in marine biology research and management. *Marine Ecology Progress Series*, 316: 203-204.
- Costello, M. J, P. Bouchet, C.S. Emblow, A. Legakis (2006). European marine biodiversity inventory and taxonomic resources: state of art and gaps in knowledge. *Marine Ecology Progress Series*, 316:257-268.
- Daily, G. C. (1997). Nature's Services: Societal Dependence on Natural Ecosystems. Daily, G. C. (eds.), Island Press, Washington DC, USA, 392pp.
- Dalcin, E. C. (2004). Data Quality Concepts and Techniques applied to taxonomic databases. *Thesis for the degree of Doctor of Philosophy, School of Biological Sciences, Faculty of Medicine, Health and Life Sciences, University of Southhampton*, November 2004. 226 pp. http://www.dalcin.org/eduardo/downloads/edalcin_thesis_submission.pdf
- Das, A. (2003). A Catalogue of New Taxa Described by the Scientists of the Zoological Survey of India during 1916-1991. *Records of the Zoological Survey of India, Occasional Paper*, 208, 1 - 530.
- Datta, R., Nouguier, C., Pascal, J.P. and B. R. Ramesh. (1997). Endemic tree species of the Western Ghats (India). CD ROM, *French Institute of Pondicherry, Pondicheery.*
- David, S. K. and A. K. Ghosh (1982). The Fauna of India and the adjacent countries. Homoptera : Aphidoidea. *Zoological Survey of India, Kolkata*. p. 167.
- DBT (2007). Biotechnology Information System (BTISNet). Accessible at http://www.btisnet.nic.in/.

§ Dhandapani, P. (1977). Descriptions of two new species of larvaceae with a list of other species collected from the Bay of Bengal. In: Proceedings of the Symposium on Warm Water Zooplankton. UNESCO/NIO, National Institute of Oceanography, Dona Paula, Goa, India, pp. 60-64

§ Distant, W. L. (1903). The Fauna of British India. Rhynchota. 2(1): 242.

§ Distant, W. L. (1904). The Fauna of British India. Rhynchota. (2)2: 243-503.

§ Distant, W. L. (1910). The Fauna of British India including Ceylon and Burma. (Rhynchota: Heteroptera) V.

§ Distant, W. L. (1916). The Fauna of British India including Ceylon and Burma. (Rhynchota : Homoptera Appendix) VI.

§ Dragicevic, S. (2004). The potential of Web-based GIS. *Journal of Geographical Systems*, 6: 79-81.

§ Edwards, J. L., M.A. Lane and E. S. Nielsen. (2000). Interoperability of Biodiversity Databases: Biodiversity Information on Every Desktop. *Science*, 289(5488): 2312-2314.

§ Emerton, Lucy (2000). Economics and the Covention on Biological Diversity. *IUCN, The World Conservation Union*, pp.4. Accessible at http://www.undp.org/bpsp/thematic_links/IUCN2.pdf

§ EoL (2007). A Leap for all Life: World's leading scientists announce creation of "Encyclopedia of Life". *EoL Press Release dated May 9, 2007*, Accessible at http://www.eol.org/press_release.html.

§ European Union (2007). EU to take part in global database of all life on earth. *EU Press release IP/07/647 dated May 9, 2007*, Accessibleat http://europa.eu/rapid/pressReleasesAction.do?reference=IP/07/647&format=HTML&aged=0&language=EN.

§ Evans, W. H. (1910). A list of butterflies of the Palani hills with the descriptions of two species. Journal of the Bombay Natural History Society, XX(1): 380-391.

§ Evans, W. H. (1932). The identification of Indian butterflies (2nd edition), *Bombay Natural History Society, Mumbai, India*. 464 pp.

§ Faith D. P., Walker, P.A., Margules, C. R., Stein J., and G. Natera (2001). Practical application of biodiversity surrogates and percentage targets for conservation in Papua New Guinea. *Pacific Conservation Biology*, 6: 289-303. http://wwwscience.murdoch.edu.au/centers/others/pcb/toc/pcb_contents_v6.html

§ FGDC (2007). US National Spatial Data Infrastructure. Accessible at http://www.fgdc.gov/nsdi/nsdi.html, Retrieved on June 23, 2007.

§ Flemons, P., R. Guralnick, J. Krieger, A. Ranipeta, and D. Neufeld. (2007). A Web based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA). *Ecological Informatics*, 2: 49-60.

§ FRLHT (2005). Foundation for Revitalization of Local Health Tradition web site. Accessible at http://www.frlht-india.org/.

§ FRLHT (2007). FRLHT's Encyclopedia of medicinal plants. Accessible at http://www.medicinalpnats.in/.

§ Gad, S. C. and S. M. Taulbee. (1996). Handbook of data receording, maintenance, and management for the biomedical sciences. Boca Raton: *CRC Press*, 85pp.

§ Gadgil, M., Sheshagiri, P. R., Utkarsh, G, Pramod, P., Chhatre, A., and Members of the People's Biodiversity Initiative (2000). New meaning for old

knowledge: the people's biodiversity registers program. *Ecological Applications*, 10(5): 1307-1317.

§ Gadgil, M. (2003). India's Biological Diversity Act 2002: An act for the new millennium. *J. BioSci*, 28(2): 145-147.

§ Gadgil, M., K.P.Achar, Harish Bhat, P.R. Bhat, Shubhada Deshmukh, Ajay Dolke, Yogini Dolke, N. Vijay Edlabadkar, A.K. Ghosh, Satish Gogulwar, Yogesh Gokhale, Shrikanth Gunaga, B.V. Gundappa, Nilesh Heda, Mohan H. Hiralal, N. Indiramma, Kailash C. Malhotra, M.B.Naik, M.P.Nair, N.H.Ravindranath, G. Nalini Rekha, Kaustubh Pandharipande, S.G.Patgar, Ramakrishnappa, P.R. Seshagiri Rao, V.V. Sivan, S. Srinidhi, S. Sujith, K.A. Subramanian, Devaji Tofa, C. Yathiraju (2006). Ecology is for the People: A methodology manual for peoples biodiversity registar, *National Workshop on People's Biodiversity Registar, Chennai, India, 22-23, June 2006*, pp. 237. Accessible at http://www.nbaindia.org/docs/ec_pbr_manual.pdf

§ Gaikwad, J. and V. Chavan (2006). Open access and biodiversity conservation: challenges and potentials for the developing world, *Data Science Journal*, 5: 1-17.

§ Ganeshaiah, K. N., Kathuria, S., and R. Uma Shaanker. (2002). Floral resources of Karnataka: A geographic perspective. *Current Science*, 83(7):810-813.

§ Ganeshaiah, K. N. (2003). Sasya Sahyadri CDROM, *University of Agricultural Sciences, Bangalore*.

§ GBIF (2004). GBIF and Catalogue of Life Partnership sign memorandum of cooperation. *GBIF press release dated 07, January 2004*, accessible at http://www.gbif.org/Stories/STORY1073492201.

§ GBIF (2005) The GBIF 3rd Year Review, pp. 234. Accessible at http://circa.gbif.net/Public/irc/gbif/pr/library?l=/review_documents/report_ph_pdf/_EN_1.0_

§ GBIF (2006). GBIF Plans 2007-2011: from prototype towards full operation, *Lane, M. (eds), Global Biodiversity Information Facility, Copenhagen, Denmark*, 85pp.

§ Ghosh A. K. (1982). The Fauna of India, Aphids. Part II. *Zoological Survey of India, Kolkata, India*. 167pp.

§ Government of India (2007). Indian Bioresources Information Network. *Copyright Government of India*, accessible at http://www.ibin.co.in/, retrieved on September 20, 2007.

§ Graham CH, Ferrier S, Huetteman F, Mortiz C, and A. T. Peterson. (2004) New developments in museum based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, 19: 497-503.

§ Gupta, I. J. and J. P. N. Shukla (1988) Studies on the butterflies of Arunachal Pradesh and adjoining areas, India (Lepidoptera: Acraeidae, Satyridae, Nymphalidae, Riodinidae and Lycaenidae). *Records of the Zoological Survey of India, Occassional paper* no. 109: 1-115.

§ Gupta, R. (2000). SWOT analysis of geographic information: The case of India. *Current Science*, 79: 489-498.

§ Guralnick R, J. Wieczorek, R. Beaman, R. J. Hijmans, the BioGeoMancer Working Group (2006). BioGeoMancer: Automated Georeferencing to map the world's biodiversity data. *PLoS Biology*, 4(11): e381. DOI: 10:1371/journal.pbio.0040381.

§ Guralnick, R., and D. Neufeld. (2005). Challenges building online GIS services to support global biodiversity mapping and analysis: lessons from the Mountain and Plains Database and Informatics Project. *Biodiversity Informatics*, 2: 56-69.

§ Hallan, J. (2003). Genera of Acari, Accessible at http://insects.tamu.edu/research/collection/hallan/acarallgen.html

§ ICZN (2003). International Code for Zoological Nomenclature, Accessible at http://www.iczn.org/iczn.htm/index.html

§ ION (2005). Index to Organism Names, Accessible at http://www.biosis.org.uk/ion/search.htm

§ IPNI (2005). International Plant Name Index, Accessible at http://www.ipni.org/

§ ITIS (2005). Integrated Taxonomic Information System, Accessible at http://www.itis.gov/.

§ IUCN (2005). 2005 IUCN Redlist of threatened species. Accessibleat http://www.iucnredlist.org/.

§ Javed, S. (2001). Current state of biodiversity information in India and need for an integrated biodiversity information system (IBIS). *White paper commissioned by the National Biodiversity Strategy and Action Plan, 2001*, accessible at http://sndp.delhi.nic.in/nbsap/.

§ Johnston, C. A. (1998). Geographic information systems in ecology. *Blackwell Science, Oxford, UK*, 235pp.

§ Jones, M. B., M. P. Schildhauer, O. A. Riechman, and S. Bowers. (2006). The New Bioinformatics: Integrating ecological data from the Gene to Biosphere. *Annual Review of Ecology, Evolution, and Systematics*, 37: 519-544.

§ Knapp, S. , Bateman, R. M. , Chalmers, N. R. , Humphries, C. J. , Rainbow, P. S. , Smith, A. B. , Taylor, P. D. , Vane-Wright, R. I. and M. Wilkinson. (2002). Taxonomy needs evolution, not revolution. *Nature* 419: 59

§ Kothari, A. (2003). Indian Biodiversity Information System (IBIS), *Personal Communication dated December 4, 2003*.

§ Kraak, M. (2004). The role of the map in a web-GIS environment. *Journal of Geographical Systems*, 6: 83-93.

§ Krishtalka, L. and P. S. Humphrey. (2000). Can natural history museums capture the future? *BioSciences*, 50(7) 611-617.

§ Kunte, K., Joglekar, A., Ghate, U., and P. Pramod (1999) Patterns of butterfly, bird and tree diversity in the Western ghats. *Current Science*, 77(4): 577-586.

§ Lane, M. A., J. L. Edwards, and E. S. Neilsen. (2000). Biodiversity informatics: the challenge of rapid development, large databases, and complex data. *In Proc. 26th Int. Conf. Very Large Databases, Cairo, Egypt, 2000*.

§ Leslie, M. (2005). Species master list hits milestone. *Science*, 308:609.

§ Longmore, R. (ed) (1986) Atlas of Elapid snakes of Australia. *Australian Flora and Fauna series no. 7*, Australian Government Publishing Service, Canberra.

§ Loreau M., Alfred Oteng-Yeboah, M. T. K. Arroyo, D. Babin, R. Barbault, M. Donoghue, M. Gadgil, C. Häuser, C. Heip, A. Larigauderie, K. Ma, G. Mace, H. A. Mooney, C. Perrings, P. Raven, J. Sarukhan, P. Schei, R. J. Scholes & R. T. Watson., (2006). Diversity without representation. *Nature*, 442: 245-246.

§ Lunetta, R. S. and Lyon, J.G.(eds.). Remote sensing and GIS accuracy. *Boca Raton, FL, USA: CRC Press*. 320pp.

- Maier, D., Landis, E., Cushing, J., Frondorf, A., Silberschatz, A., and J. L. Schnase. (2001). Research directions in biodiversity and ecosystems informatics, *Report of an NSF, USGS, NASA Workshop on Biodiversity and Ecosystems Informatics (NASA Goddard Space Flight Center, June 22-23, 2000, Greenbelt, Maryland)*, pp.30.
- Maletic, J. I. and Marcus, A. (2000). Data cleansing: Beyond integrity analysis. In: *Proceedings of the Conference on Information Quality (IQ2000), Massachusetts Institute of Technology, Boston, USA*, pp. 200-209.
- Margules, C.R. and R. L. Pressey. (2000) Systematic Conservation Planning, *Nature*, 405:243-253.
- Mathiyalagan, V., Grunwald, S., Reddy, K.R. and S. A. Bloom. (2005). A WebGIS and geodatabase for Florida's wetlands. *Computers and Electronics in Agriculture*, 47: 69-75.
- May, R. M. (1988) How many species are there on Earth?, Science, 247: 1441-49.
- May R. M. (1999). The dimensions of life of earth. In *Nature and Human Society, the Quest for Sustainable Worlds*, (ed. Raven, P. H.), National Academy Press, Washington DC, 1999, pp. 30-45.
- Mitra, B., P. Parui and D. Banergee. (2002). On the Diptera of Nayachar island, West Bengal. *Journal of Bombay Natural History Society* 99(2): 343-347.
- MoEF (2007). Environmental Information System. Accessible at http://www.envis.nic.in/.
- MSSRF (2007). MS Swaminathan Research Foundation web site. Accessible at http://www.mssrf.org/.
- Murphey, P.C., Guralnick, R.P., Glaubitz, R., Neufeld, D., and J. Ryan. (2004). Georeferencing of museum collections: A review of problems and automated tools, and the methodology developed by the Mountain and Plains Spatio-Temporal Database-Informatics Initiative (Mapstedi). *PhyloInformatics*, 3, 1-29.
- Murthy, M. S. R., A. Giriraj, and C. B. S. Dutt (2003). Geoinformatics for biodiversity assessment. *Biol. Lett.*, 40(2): 75-100.
- Nandi, N. C., S. R. Das, Bhuinya and J. M. Dasgupta. (1993). Wetland Faunal Resources of West Bengal, I., North And South 24-Parganas Districts. *Records of the Zoological Survey of India. Occasional Paper*, 150: 1 - 50.
- NASA (2007). Applied Research and Technology Project Office, John C. Stennis Space Center, National Aeronautic and Space Administration. Accessible at https://zulu.ssc.nasa.gov/mrsid/, Retrieved June 23, 2007.
- Nature (2002). Genomics and taxonomy for all. *Nature* 417, 573
- NBRI (2005). National Botanical Research Institute web site, Lucknow. Accessible at http://www.nbri-lok.org/bioinformatics1.htm.
- NBRI (2007) Legume database. Accessible at http://www.nbri-lko.org/bioinformatics1.htm.
- NCL (2005). NCL Centre for Biodiversity Informatics. Accessible at http://www.ncbi.org.in/.
- Nearctica (1998). Insects, Heteroptera and Homoptera, http://www.nearctica.com/nathist/insects/homops.htm.
- Nicholls, N. (1997). Increased Australian wheat yield due to recent climate trend. *Nature*, 387: 484-485.

§ NIO (2005). Gateway to Indian Ocean web site. National Institute of Oceanography, Goa. Accessible at http://www.indian-ocean.org/.

§ NIO (2007). National Institute of Oceanography Bioinformatics Center web site, Accessible at http://www.niobioinformatics.org/.

§ Nix, H. A. (1986). A biogeographic analysis of Australian elapid snakes. Pp.4-15 in Atlas of Elapid snakes of Australia. Australian Flora and Fauna Series *(R. Langmore, ed.), Number 7, pp. 4-15, Australian Government Publishing Service, Bureau of Flora and Fauna, Canberra*.

§ NLWRA (2003). Natural resources information management toolkit. Canberra: *National Land and Water Resources Audit* http://www.nlwra.gov.au/toolkit/content.html Retrieved on Spetember 21, 2007.

§ NSDI (2007). National Spatial Data Infrastructure Portal. Accessible at http://gisserver.nic.in/nsdiportal/, Retrieved on June 23, 2007.

§ NSDI Task Force (2001). National Spatial Data Infrastructure (NSDI): Strategy and Action Plan. *Taskforce on NSDI, DST, New Delhi*, 2001, 45 p.

§ Olden, J.D. and D.A. Jackson. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*. **154**, pp. 135–150.

§ Page, R.D.M.. 2005. A Taxonomic Search Engine : Federating taxonomic databases using web services. BMC Bioinformatics 2005, 6:48.

§ Patton, R. (2006). Software testing (2nd Edition), SAMS Publication, pp. 377.

§ PCAST (1998). Teaming with Life: Investing in Science to Understand and Use America's Living Capital. *Presidents Committee of Advisors on Science & Technology, Panel on Biodiversity and Ecosystems, Office of the President, Washington D. C., USA*. Accessible at http://www.ostp.gov/Environment/html/teamingcover.html.

§ Peterson, A. T., and Vieglais, D. A. (2001). Predicting species invasion using ecological niche modeling. *BioScience*, 51: 363-371.

§ Peterson, A. T., Navarro-Sieguenza, A. G., and Benitez-Diaz, H. (1998) The need for continued scientific collecting: A geographic analysis of Mexican bird specimens. IBIS, 140: 288-294.

§ Peterson, A. T., Ortega-Huerta, M. A., Bartley, J., Sanchez-Cordero, V., Soberon, J., Buddemeier, R. H., and D. R. B. Stockwell. (2002). Future projections for Mexican faunas under global climate change scenarios. *Nature* 416: 626-629.

§ Platypus. (2003). Platypus, ver. 3.3. Australian Biological Resources Study, Accessible at http://www.environment.gov.au/biodiversity/abrs/online-resources/software/platypus/.

§ Pouliquen-Young, O. and Newman, P. (1999) The implications of Climate change for land based Nature Conservation Strategies. *Final Report 96/1306, Australian Greenhouse Office, Environment Australia, Canberra, and Institute for Sustainability and Technology Policy, Murdoch University, Perth, Australia*, 91pp.

§ Pushpangadan, P. and K. N. Nair. (2001). Future of systematics and biodiversity research in India: Need for a National Consortium and National Agenda for systematic biology research. *Current Science*. 80(5), 631-635.

§ Ramachandran, R. (2000). Public access to Indian geographical data. *Curr. Sci.*, 2000, 79, 450-459.

- Rao, D. V. Kamla Devi and P.T.Rajan.. 2000. An account of Ichthyofauna of Andaman and Nicobar islands, Bay of Bengal. *Records of the Zoological Survey of India*, *Occasional Paper* 178:1-434.
- Ritchie, J. (1910). The Hydroids of the Indian Museum. I. The deep sea collection. *Records of the Indian Museum,* V (I): 1 – 30.
- Roy, P. S. and S. Ravan. (1996). Biomass estimation using satellite remote sensing data – an investigation on possible approaches for natural forest. *Journal of Biosciences*, 21(4): 535-561.
- Roy, P.S. and S. Tomar. (2000). Biodiversity characterization at landscape level using geospatial modeling technique. *Biological Conservation*, 95(1): 95-109.
- Roy, P.S., Saran, S., Ghosh, S., Prasad, N., Karnatak, H., and G. Talukdar. (2002). Development of biodiversity information system for the North East India using Internet GIS. *Proc. Symp. Geospatial Theory, Processing and Applications, Ottawa, Canada, 2002*.
- SACON (2007). Wetlands of India website, Salim Ali Center for Ornithology and Natural History, Coimbatore. Accessible at http://www.wetlandsofindia.org/.
- Saha, S. K., Saha S. K., Mukherjee A. K. and T. Sengupta. (1992). Carabidae (Coleoptera: Insecta) of Calcutta. *Records of Zoological Survey of India*. *Occasional Paper* 144: 1 – 63.
- Salem, B.B. (2003). Application of GIS to biodiversity monitoring. *Journal of Arid Environments*, 54, 91-114.
- Santana, F. S., Fonseca, R. R., Saraiva, A. M., Corrêa, P. L. P., Bravo, C., and R. Giovanni. (2006). OpenModeller - an open framework for ecological niche modeling: analysis and future improvements. *World Conference on Computers in Agriculture and Natural Resources (WCCA)*.
- Sante, I., Crecente, R., Miranda, D., Tourino, J., Canzobre, F., Doallo, R., 2004. A GIS web-based tool for the management of the PGI potato of Galicia. *Computers and Electronics in Agriculture*, 44: 161-171.
- Sanyal, A. K. and A. K. Bhaduri (1986). Check list of Oribatid Mites (Acari) of India. *Records of The Zoological Survey of India, Occasional Paper* 83: 1 – 79.
- Sanyal, A. K. Saha, S. and S. Chakraborty. (2003). Three new species of the genus Chaunoproctus Pearce (1906) (Acarina : Oribatida) from India. *Records of the Zoological Survey of India*, 101 (1-2): 57-66.
- Sarkar, I. (2007). Grand challenges in biodiversity informatics. *Asia Pacific Biotech News*, 11: 15-18.
- Schalk, P. H. (1998). Management of marine natural resources through by biodiversity informatics. *Marine Policy*. 22(3): 269-280.
- Schmidt-Rhaesa, A. and A. K. Yadav (2004). On the occurrence of Chordodes cf. furnessi (Nematomorpha) from praying mantis in India, and a note on Indian nematomorph species. *Current Science*, 86(7): 1023-1027.
- Schnase, J. L. (2000). Research directions in biodiversity informatics. In *Proc. 26th Int. Conf. Very Large Databases*, Cairo, Egypt, 2000.
- SEC (2002). Final data quality assurance guidelines. *United States Securities and Exchange Commission*. http://www.sec.gov/about/dataqualityguide.htm.
- SNMNH (2007). Smithsonian National Museum of Natural History. http://goode.si.edu/mcs/iz/Query.php, retrived on September 21, 2007.

§ Soberon, J. and A. T. Peterson (2004). Biodiversity informatics: managing and applying primary biodiversity data. *Philosophical Transactions: Biological Sciences, Royal Society of London*, 359: 689-698.

§ Soubadra Devi, M. and P. Davidar (2001) Response of wet forest butterflies to selective logging in Kalakad-Mundanthurai tiger reserve: implications for conservation. Current Science, 80(3): 400-405.

§ Sreenivasulu, V. and H. B. Nandwana. (2001). Networking of agricultural information systems and services in India. *INSPEL*, 35(4):226-235.

§ Srikantia, S.V. (2000). Restriction on maps: A denial of valid geographic information. *Current Science*, 79: 484-488.

§ Srivastava, G. K. (2003). Fauna of India Dermaptera. *Zoological Survey of India, Kolkatta*, 235pp.

§ Stevens, R.; A. Robinson, and C. A. Goble. (2003). MyGRID: Personalised bioinformatics on the information grid. In *proceedings of 11th International Conference on Intellignet Systems in Molecular Biology, 29th June – 3rd July 2003, Brisbane, Australia, published Bioinformatics*, vol. 19, suppl. 1: i302-i304.

§ Stockwell, D. R. B., and D. P. Peters. (1999). The GARP modelling system: Problems and solutions to automated spatial prediction. *International Journal of Geographic Information Systems*, 13:143-158.

§ Talwar, P. K. and R. K. Kacker (1984). Commercial Sea Fishes of India. *Published by Zoological Survey of India, Kolkata*., 997pp.

§ Tan, V., W. Frang, S. C. Wong, S. Miles and L. Moreau. (2005). A security architecture for a semantic grid registry. In *proceedings of 4th UK e-Science All Hands Meetings (AHM), Nottingham, UK*, Accessible at http://eprints.ecs.soton.ac.uk/11240/.

§ Thackway, R. and Cresswell, I. (eds.) 1995. An interim biogeographic regionalization for Australia: A framework for setting priorities in the Natural Reserves System Cooperative Program (version 4.0), *Australian Nature Conservation Agency, Canberra*.

§ Thorne J. (2003). Zoological record and registration of new names in zoology. *Bulletin of Zoological Nomenclature*,60(1): 7-11.

§ Thuiller, W. (2003). BIOMOD ? optimizing predictions of species distributions and projecting potential shifts under climate change. *Global Change Biology*, 9: 1353-1362.

§ Tikader B. K. (1982). Fauna of India: Arachnid (Vol 2): Spiders. *Zoological Survey of India*, 536pp.

§ Tikader, B. K. and D. B. Bastawade. (1983). Fauna of India: Scorpions (Scorpionida : Arachnida). *Zoological Survey of India, Kolkata*, 671pp.

§ van Emden, F. I. (1965). The Fauna of India and Adjacent Countries. Diptera. 7(1): 647.

§ Wadsworth, R. and Treweek, J. (1999) Geographic information systems for ecology: an introduction. *Longman, Essex, UK*. 208pp.

§ Wieczorek, J., Guo, Q., and R. J. Hijmans. (2004). The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, 18, 745-767.

§ WII (2005). Wildlife Institute of India web site. Accessible at http://www.wii.gov.in/new/nwdc/index.html

§ Williams, B.K. (1996). Assessment of accuracy in the mapping of vertebrate biodiversity. *Journal of Environmental Management*, 47, 269-282.

- Wilson, E. O. (2003). The Encyclopedia of Life. *Trends in Ecology and Evolution*, 18(2): 77-80.
- Wynter-Blyth, M. A. (1957). Butterflies of the Indian region. *Bombay Natural History Society, Mumbai, India*. 523 pp.
- Yee, T. W. and N. D. Mitchell. (1991). Generalized additive models in plant ecology. *Journal of Vegetation Science*, 2(5): 587-602.
- Zhang, J. and M. F. Goodchild. (2002) Uncertainty in Geographic Information. *Taylor and Francis, London*, 288pp.
- Zoo Outreach (1995) The ABC of the Indian wildlife (protection) act schedules as amended in 1991. *Zoo's Print Journal* (February 1995), X(2): 21-32.
- ZSI (1991) Animal resources of India (Protozoa to Mammalia). *(Eds – Director, Zoological Survey of India)*, pp. 694.
- ZSI (1997) Fauna of West Bengal, Part 7. *(Eds – The Director, Zoological Survey of India)*, p. 755.
- ZSI (2000) Fauna of Meghalaya : Part – 5 : Insecta : State Fauna Series – 4, *edited by The Director Zoological Survey of India. Calcutta, Zoological Survey of India*, 2000, 665 pp.