A DATA-RICH APPROACH TO ASSESS BIODIVERSITY

Rainer Froese

ICLARM
Makati City, Philippines

ABSTRACT

It is suggested that four bits of information form the core of biodiversity data: 1) the scientific name of an observed species, 2) the locality, 3) the date and 4) the source. If properly standardized, they link to related information such as taxonomy, biology, human uses, indigenous knowledge, etc., through the species name, to habitat, ecosystem, land use, administrative units, etc. through the locality, to environmental conditions through the combination of locality and date, and to persons and institutions through the source. Existing tools can analyze these data to estimate, for example, areas of high endemism, high species richness, and special threat, or areas most suitable for reserves. A step-by-step plan to establish a national biodiversity database is suggested and recommended fields for a biodiversity database are listed.

INTRODUCTION

Current definitions of biodiversity include diversity at the level of genes, populations, species, ecosystems, and more recently, also large ecosystems. There is an emerging consensus that the critical unit to preserve genes and species is the population, and that populations are best protected by sustainable, precautious use and management of their respective: ecosystems. However, there is no consensus yet as to what data should be collected to monitor and assess the diversity in a given ecosystem. Currently several hundred types of data are collected by numerous programs, projects, initiatives, agencies, as well as interested lay persons. Unfortunately, these data are largely incompatible because of lacking standards. Further, these data are practically not widely available for analysis because they are reported in often inaccessible reports and articles, most often in summary form only. In this article we will argue that there is a small set of biodiversity data that forms a natural core of biodiversity information and gives it an efficient structure. This minimum set can be standardized with reasonable effort, it is already available in the form of millions of records, it is collected daily by numerous projects, and it can be readily compiled and made available through intelligent use of modem information technology.

RECENT ADVANCES IN INFORMATION TECHNOLOGY

We are living in the age of the information technology revolution. The Notebook computer of 1995 cost about US$ 3,000 and had the processing power of the Workstation computer of 1990, which cost about USS 30,000 and had the processing power of the mainframe of 1985, which cost about US$ 300,000. Similar, standard storage space on a Notebook computer is now more than 500 megabytes. CD-ROMs are very cheap to produce and can store 600 megabytes of information. The forthcoming standard for multimedia CD-ROMs will store 5 gigabytes of information. This technology is available in developed and developing countries alike, providing equal access to processing power. The Internet connects laypersons as well as universities and ministries throughout the world and provides the infrastructure for real-time data transfer as well as access to on-line databases. Thus, the necessary building blocks for national and global biodiversity data systems, consisting of quick, easy and cheap contribution, verification, analysis, and distribution of data, are in place.

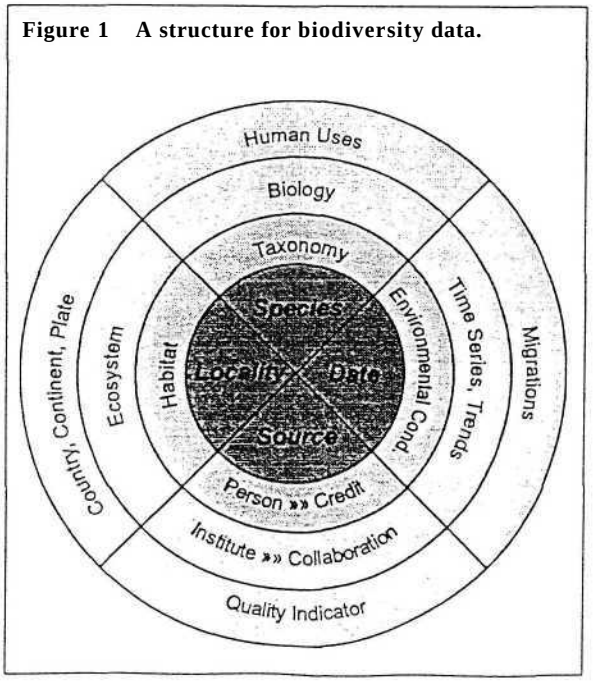
BIODIVERSITY CORE DATA

This then brings us to the question of "What data should be collected?" Common sense suggests that the selected data should fulfil the following requirements:

1. easy to collect, standardize, and verify;
2. allow detailed analyses;
3. applicable to all organisms; and
4. available for past, present, and, most likely, future.

We propose that a minimum of four related pieces of information fulfil all the above criteria. These pieces of information are: 1) the scientific name of a species; 2) the locality where a specimen has been collected or observed; 3) the date of the encounter; and 4) the source of this information. This type of information has been collected for centuries and is available in museum collections, survey reports, the scientific literature, but also more unconventional sources such as angling records, photos and films. Many of these records are already available in digitized form. Also, these data are collected daily in numerous projects all over the world. It can be assumed that most projects would be very willing to share this information, if only to make their work known and ensure that the records are stored safely, if there would be an easy mechanism for contributing the data to a central, trust-worthy focal point, with due acknowledgment of the source. Such contribution could be easily organized using the Internet for direct file transfer, e-mail for transfer of smaller amounts of data, or by sending diskettes for those who do not have access to Internet or e-mail.

Figure 1 A structure for biodiversity data.



STANDARDIZATION

A precondition for data compilation is the standardization of formats and the definition of acceptable contents. Such standardization allows to cross-reference independent data sets and to analyze the data. It also makes sure that data sets do not contain duplicates and that a search will find all relevant records. The preferred model for standardized data sets is the relational database. In contrast, text databases such as hypertext or bibliographic databases are difficult to standardize and a search usually results in too many or too few retrieved records, typically including a number of unrelated records. Lack of standardization is the main reason why the huge amount of information that is accessible through the Internet cannot be used for rigorous analyses. See UNEP/WCMC (1995) for a comprehensive and recent overview of existing and missing environmental standards. Means for standardizing the suggested four bits of information are discussed below.

SPECIES

The binomial scientific name is the globally accepted identifier for biological species. To describe taxa below the species level, descriptors for subspecies; varieties, strains, or genotypes can be appended to the binomen. International Codes of Nomenclature have been established for all major phyla and are updated and enforced by their respective Commissions under the umbrella of the International Union of Biological Sciences (IUBS). The codes are elaborated bodies of regulations and recommendations which define what format a scientific name must follow and what criteria it has to fulfil to be available for use. Efforts are underway to harmonize and simplify the codes. However, interpretation of the codes remains difficult and is often beyond the scope of non-taxonomists. Also, the fact that a name is available does not imply any statement about its status and classification. To enable biodiversity researchers to deal with thousands of species names, a computerized system is needed against which names can be matched to identify misspellings and synonyms and to derive the current classification. SPECIES 2000 is an initiative that has embarked on the enormous task of creating such a reference file for earth's 1.75 million recognized species and to make this information available on CD-ROM and on the Internet (Bisby, this volume).

LOCALITY

A locality is the name of a geographic entity such as a river, lake, mountain, valley, village etc. It is typically bound into a political (e.g., county, province, state, country) and geographic (e.g., river, river system, drainage, continent) hierarchy. For reasons of precision, verification, and usefulness, a reported locality should be complemented by the geographic coordinates. For microbes and parasites, the "locality" must include the host organism on which they were found.

Gazetteers are books that contain locality names, usually with type (river, lake, etc.), political and/or geographic hierarchy, other names in use for the same locality, and the coordinates. Gazetteers exist for many countries and can also be derived from computerized Geographic Information Systems (GIS) which are now developed in many national and international agencies. Global gazetteers are prepared for release on CD-ROM by institutions in the US and in the UK. However, there are still no global standards for names of political units below the country level or for geographic units. An initiative similar to SPECIES 2000 is urgently needed to harmonize the many existing gazetteers and standards and consolidate them into one global reference file. Computerized gazetteers can be used to automatically verify locality names, compare known coordinates with reported coordinates, or assign coordinates to records that only give locality names. They also can assign the respective political and geographic units to accepted locality names.

For living genetic resources the locality often is the institution, reserve, farm or gene bank where sperms, seeds, or organisms are maintained *ex situ.* In these cases the locality should include a contact address. It is still useful to provide geographic coordinates, e.g., to create a map of all institutions holding *ex situ* germplasm of a given species.

DATE

Standardization of the date format has recently become an issue with many banks and insurance companies realizing that the current format cannot distinguish between centuries, and that existing routines that include dates will produce erroneous results from 1 January 2000 onwards. Thus, the scientific date in the form *dd/mm/yyyy* should be used in biodiversity databases. Further standardization is needed to deal with date ranges (e.g., give start and end date) and incomplete dates such as month or year only. Additional standardization is needed for dates before 0001 and for fossil records.

The time of encounter is an important data for many species with diurnal cycles. The format should be local time in *hh:mm* in 24 hours notation. Provision should be made for time ranges and qualitative descriptors such as morning, noon, afternoon, evening, night, and time ranges that span into the next day.

SOURCE

Some standardization is also needed for the name(s) of the person(s) who reported a species, the names of their institutions, etc. The person's family name should be recorded first, followed by the initial(s), as is the *de facto* standard in bibliographic databases. Several lists of accepted acronyms of museums and herbaria have been published (e.g., Leviton *et al.* 1985). Names and acronyms of other institutions should be standardized as far as possible.

Verification

For reliable analyses of biodiversity data it is important to know the quality of the underlying data. We suggested that each record should have a qualifier indicating the probability that the species in question was identified correctly and did indeed occur at the specified locality at the specified date. Given the huge amount of data, such a qualifier has to be derived automatically by comparing the reported data with existing information. Ideally, the computer should reject questionable records and provide a printout with all cases that need further study, together with an indication of the problem. Reasons for rejection could be:

- provided scientific name cannot be assigned to a valid scientific name (i.e., either a misspelling or a new species);
- species has not yet been reported from the stated country, province, region, or locality (i.e., either a new record or a misidentification or an error in the locality);
- species has not yet been recorded from the provided locality in the provided season (i.e., either a new record or a misidentification, an error in the locality or in the date);

Data providers could use this feedback to flag their records accordingly, correct mistakes or provide additional information, and resubmit these records. Alternatively, accepted records could be flagged as "compatible with existing knowledge." Thus, matching their data against an independent biodiversity database gives the data providers valuable assistance in cleaning up their databases and might for some be the primary reason to participate in the exercise.

CROSS-REFERENCING

Once standardized, the suggested four items provide reliable links to many other databases with information that are important for biodiversity management. The four items can actually be seen as a core around which all relevant information can be conveniently grouped. For example, the scientific name leads to all available information on taxonomy, biology, "indigenous" or "local" knowledge, human uses, economic value, etc. If the scientific name is supplemented with information on subspecies, varieties, races, strains, or genotypes it links to information such as breed performance, population sizes, prices, phenotypes, genetic markers, and DNA sequences.

The locality links to geo-referenced information such as habitat type and status, altitude, land uses, ecosystem, sympatric species, human population, responsible authorities, etc.

The date in combination with the locality provides information about physical parameters such as temperature, humidity, photo period, season, and about other related events such as floods or draughts.

ANALYSES

The suggested four items in combination with related data allow a wide range of powerful analyses, including the following frequently requested outputs:

- areas of high species richness;
- areas of high endemism;
- hot spots (which can be defined by a variety of combined criteria);
- areas under special threat (disappearance of key species);
- most suitable locations for protected areas (by various criteria);
- monitoring of "high impact" or "flagship" species;
- estimation of status of threat;
- time series / trend analyses;
- suitable areas for relocation of species;
- suitable areas for re-establishment of locally extinct species;
- status of biodiversity for a country, a geographic unit, and the earth.

The data also allow for scientific studies aimed to increase our understanding of biodiversity, for example:

- compare species communities and identify key stone species;
- study trends in species composition over time, in relation to environmental factors;
- study co-occurrence and alternate occurrence of species;
- test zoogeographic and evolutionary hypotheses.

INDIGENOUS KNOWLEDGE

The value of "indigenous" or "local" knowledge on occurrence, behaviour, interactions, and possible uses of native species is now widely recognized (Pauly *et al.*, 1993; elsewhere, this volume). It is high time to record this knowledge before it is lost. In the context of biodiversity databases it has to be stressed that a "local knowledge database" must, in addition to the knowledge, include at least four pieces of information:

1. the local name of the species;
2. the country and language of the local name, which in combination determine the group of indigenous peoples;
3. the scientific name of the species; and
4. the source of information, i.e., who, where and when provided the information.

The local or common name allows indigenous people to access the information. The correct scientific name provides the link to biodiversity databases and connects the local with the scientific knowledge. Country, language and source are essential to properly identify and acknowledge the providers of the information.

HOW TO COMPILE NATIONAL BIODIVERSITY DATABASES

Signatory countries to the Convention on Biological Diversity are obliged to "Maintain and organize, by any mechanism, data derived from identification and monitoring activities" (Article 7d). Possible approaches to establish biodiversity information systems have been discussed in Canhos *et al.*, (1992) and at a recent workshop for Tropical Islands in the Pacific Region (Anon, 1995). We have updated and generalized the suggested approaches:

A first step is obviously to take stock of existing institutions, knowledge and activities related to biodiversity. The United Nations Environment Programme (UNEP) has established a Biodiversity Data Management Project (BDM) which provides countries with an elaborated framework and some financial support for this important step (Crain; Duff, this volume).

As a second step, a National Biodiversity Network has to be established, effectively linking the players identified in step one. It is important that all participants clearly and visibly benefit from sharing their data with others. A Network Coordinating Centre will be put in charge of designing the Biodiversity Database. Suitable designs have been published in Froese and Pauly (1994) and Froese and Palomares (1995). The Coordinating Centre will provide channels for data contribution of network members. Such data channels could be file transfer through electronic networks, e-mail, or the sending of diskettes through regular mail. Hardcopies should not be accepted unless the Coordination Centre has sufficient personnel to encode data. A simple software for data encoding should be made available to network members. A suitable data model is presented in Table 1. Note that entries in the Genus, Species, Locality, Date and Collector fields are mandatory, while entries in the other fields are optional. Incoming data have then to be subjected to standardization and verification as described above. Questionable records are returned through the appropriate channels for verification and correction.

The Coordinating Center takes a lead to provide access to national and international databases that can be meaningfully linked with the National Biodiversity Database. Examples of such databases are FishBase for finfish (Froese and Pauly, 1995), ILDIS for legumes (Bisby, 1993) and CoralBase for corals (Navin and Veron, 1995). Similarly, it provides access to a range of analytical tools that can be used to analyze the collected data.

For the success of the network it is essential that all participants have full access to the National Biodiversity Database as well as to related databases and tools. We suggest that the Coordinating Centre regularly releases a CD-ROM containing the complete National Biodiversity databases and all related databases and tools. Also, the Coordinating Centre will regularly organize training courses on how to analyze and interpret the collected biodiversity information, and on related issues.

Network members regularly offer training courses on how to analyze the collected biodiversity information.

Table 1 Recommended fields for a Biodiversity Database. Information in fields marked with an X is mandatory.

	Field	Type	Length	Remarks
	Family	[Text	30	I Family of species.
X	Genus	Text	30	Generic name of species.
X	Species	Text	40	Specific name of species.
	Strain	Text	60	Subspecies, variety, breed, population, genotype
	Length	Number	0,000.000	Length in cm from 0.001 cm to 99 m.
	LengthType	Text	30	Type of length measurement (TL = total length)
	Weight	Number	00,000,.000.000	Weight in gram from 0.001 g to 99 tons.
	WeightType	Text	30	Type of weight measurement (TW = total weight)
	LifeStage	Text	30	Sperm, egg, larvae, fry, juvenile, adult, seed, fruit, plant,.....
	Sex	Text	20	Female, male, unsexed
	Specimens#	Number	000,000	Number of specimens observed during time interval.
	Abundance	Text	30	very rare, rare, common, very common
X	Locality	Text	60	Name of locality.
	County	Text	40	Smallest governmental unit of locality.
	Province	Text	40	Name of province.
	Country	Text	30	Name of country.
	Latitude	Number	000.000	Number from 0.000-90.000; negative numbers for South; minutes expressed as decimal degrees.
	Longitude	Number	0000.000	Number from 0.000-180.000; negative numbers for West; minutes expressed as decimal degrees.
X	DateFrom	Date	dd/mm/yyyy	Date of observation or beginning of date range.
	Date To	Date	dd/mm/yyyy	For date ranges, end of range.
X	Collector	Text	40	Name of collector. Family name first, followed by initials (e.g. Myers, R.F.).
	Identifier	Text	40	Name of person who identified the specimen. Family name first, followed by initials.
	Remarks	Text	255	Additional remarks, e.g., gear used, environmental parameters, etc.

SUMMARY

It is suggested that four bits of information form a core of biodiversity data: 1) the scientific name of an observed species, 2) the locality, 3) the date and 4) the source. These pieces of information exist already in millions of records and are collected daily by numerous projects. If properly standardized, they link to related information such as taxonomy, biology, human uses, indigenous knowledge, etc. through the species name, to habitat, ecosystem, land use, administrative units, etc. through the locality, to environmental conditions through the combination of locality and date, and to persons and institutions through the source. Existing tools can analyze these data to estimate, for example, areas of high endemism, high species richness, and special threat, or areas most suitable for reserves. Other possible analyses include monitoring of "high impact" species, estimation of status of threat, areas for relocation or re-establishment of species, and status of biodiversity for any given area.

ACKNOWLEDGEMENT

Thanks are due to Daniel Pauly and Maria L. Palomares for providing valuable comments on the manuscript. This paper is a ICLARM contribution 1246.

REFERENCES

Anon. 1995. Notes from the convenors and summary of the working group sessions. p. 405-417 in Maragos, J., M.N.A. Peterson, L.G. Eldredge, J.E. Bardach and H.F. Takeuchi (eds.) Marine and Coastal Biodiversity in the Tropical Island Pacific Region. East-West Center, Honolulu, Hawaii.

Bisby, F. 1993. Species Diversity Knowledge Systems: The ILDIS Prototype for legumes. Annals of the New York Academy of Sciences, Biotechnology R and D Trends 700:449-454.

Canhos, V., D. Lange, B.E. Kirsop, S. Nandi and E. Ross (eds.). 1992. Needs and Specifications for a Biodiversity Information Network. Proceedings of an International Workshop held at the Tropical Database, Campinas, Brazil, 26-31 July, 1992. UNEP, Nairobi, Kenya. 16p.

Froese, R. and M.L.D. Palomares. 1995. FishBase as part of an Oceania biodiversity information system. p. 341-348 in Maragos, J., M.N.A. Peterson, L.G. Eldredge, J.E. Bardach and H.F. Takeuchi (eds.) Marine and Coastal Biodiversity in the Tropical Island Pacific Region. East-West Center, Honolulu, Hawaii.

Froese, R. and D. Pauly. 1994. A strategy and a structure for a database on aquatic biodiversity. p. 209-220 in J.-L. Wu, Y. Hu, and E.F. Westrum, Jr. (eds.) Data Sources in Asian-Oceanic Countries. Committee on Data for Science and Technology, Ann Arbor, Michigan.

Froese, R. and D. Pauly, (eds.) 1995. FishBase: A Biological Database on **Fish.** Concepts, Design and Data Sources. ICLARM, Manila, Philippines. 146p.

Leviton, A.E., R.H. Gibbs, Jr., E. Heal, and C.E. Dawson. 1985. Standards in herpetology and ichthyology; Part 1. Standard symbolic codes for institutional resources collections in herpetology and ichthyology. **Copeia**: 802-832

Navin, K.F. and J.E.N. Veron. 1995. CoralBase: a taxonomic and biogeographic information system for scleractinian corals. p. 327-332 in Maragos, J. M.N.A. Peterson, L.G. Eldredge, J.E. Bardach and H.F. Takeuchi (eds.) Marine and Coastal Biodiversity in the Tropical Island Pacific Region. East-West Center, Honolulu, Hawaii.

Pauly, D., M.L.D. Palomares and R. Froese. 1993. Some prose on a database of indigenous knowledge on fish. Indigenous Knowledge and Development Monitor 191:26-27

UNEP/WCMC 1995. Electronic Resource Inventory: A Searchable Resource for Biodiversity Data Management. WCMC, Cambridge, UK. [Windows 3.1 or higher].

BIODIVERSITY IN ASIA : CHALLENGES AND OPPORTUNITIES FOR THE SCIENTIFIC COMMUNITY

Proceedings of a Conference on Prospects of Cooperation on Biodiversity Activities

Chiang Rai, Thailand
15-19 January 1996

Edited by

Jeffrey A. McNeely
Chief Scientist IUCN - The World Conservation Union
Rue Mauverney 28
1196 Gland - Switzerland

and

Sunthad Somchevita
Office of Environmental Policy and Planning

with the assistance of S. Bunpapong, T. Pookpadi and P. Tongsom

---

Sponsors

The Royal Thai Government

Danish Cooperation on Environment and Development

Queen Sirikit Botanic Garden

Economic Evaluation of Biodiversity Project

Biodiversity Data Management Project

Published by

Office of Environmental Policy and Planning
Ministry of Science, Technology and Environment
Bangkok, Thailand

ISBN : 974-7575-67-1