Creation of the information retrieval system for collections of the marine animals (fish and invertebrates) at the Zoological Institute of the Russian Academy of Sciences

Igor S. Smirnov, Andrei L. Lobanov, Alexei A. Golikov, Elena P. Voronina and Alexei V. Neyelov

Zoological Institute
Russian Academy of Sciences 199034, St. Petersburg, Russia
E-mail: smiris@zin.ru

Abstract

The collection of the Zoological Institute RAN (ZIN) is one of the largest in the world and contains over 100,000 samples of 26,000 species of marine invertebrates and over 160,000 specimens of 8,700 species of marine and freshwater fishes and fishlike vertebrates of the world. Digitizing these catalogue data and creating a virtual library is now one of the main objectives of ZIN. The creation of the collection database on fish and invertebrates started in 1987 and is now very prospective for studying biocoenotic relationships and marine fauna ecosystems. In the course of the organization of our faunistic, ecological and collection data, the informational retrieval systems OCEAN and ECOANT (in standard ZOOCOD) were designed at the ZIN. The databases comprise Pleuronectiformes and Scorpaeniformes, as well as chiton, bivalve and brittle star collections. Different international projects are in development now, and it will be necessary to examine the experiences of these teams and from here make the attempt to create not ideal but optimal systems for the input and treatment of marine data.

Keywords: databases; zoological collection; marine fauna.

The unique collection of different animal groups, kept at the Zoological Institute RAN (over 60 million items in total), including type specimens and series, is worldwide known and is of great interest to zoological research. Today, for example, the ichthyological collection contains over 160,000 catalogued specimens (over 53,000 catalogued items) of 8,700 species of marine and freshwater fishes and fishlike vertebrates of the world. The marine invertebrate collection contains over 100,000 samples; some of them include tens and hundreds of specimens of 26,000 species. The scientific collection of ZIN is permanently supplemented and makes the number of specimens grow. The species diversity of Russian seas and adjacent waters is almost entirely represented in the collection and in large series from many localities. The participation of ZIN specialists in Russian and foreign expeditions has allowed us to obtain material from various distant areas of the world. For instance, the ZIN actively participates in research of the Southern Ocean and Antarctic biota, since the First Soviet Antarctic Expedition in 1955. Thanks to this participation a huge amount of material on

the fauna of this region has been collected and for the most part catalogued (Atlas of Antarctica, 1969; Smirnov and Neyelov, 1996).

Digitizing the catalogue data and creating a virtual library has now become one of the main objectives of the largest museums of the world (Smirnov, Lobanov and Dianov, 1999). The experience of using high technology and databases in foreign countries started much earlier and was more intensive than in Russia. Today some of the largest natural history museums of the world (Natural History Museum, London; Museum National d'Histoire Naturelle, Paris; California Academy of Sciences, San Francisco; National Museum of Natural History, Washington; National Science Museum, Tokyo etc.) have web sites with electronic catalogues or collection data bases on several animal groups. Creating virtual natural history museums is promoted by these electronic catalogues and libraries, such as the electronic catalogue of invertebrates on the web site of the United States Antarctic Program http://www.nmnh.si.edu/iz/usap/usapdb.html, and FishBase: http://www.fishbase.org, a large information system with key data on all marine and freshwater fishes of the world, as well as collection data of different world museums.

The information retrieval systems and the geographic information systems not only make the work of the zoologists easier, gathering the data on species from the collection catalogue and field books and manually mapping these data, also allow us to quickly visualize the information on the occurrences of the animals from the collections that are kept at the different museums during years and centuries. This kind of software and databases will be helpful in the analysis of long-term changes of fauna composition among different regions. Together with paleontological material and geographical data about the changing of the boundaries and the position of the continents, it will allow a quick analysis of the various hypotheses on the distribution of taxa, using maps with geological reconstructions. It would be possible to retrace the history of the faunal formation and to study the influences of both climate and geological changes onto biota.

Except for the historical and cognitive value, this work has significant ecological importance. With the accumulation of the zoological samples during a long period, so-called monitoring collections, it becomes possibly to trace the alteration of the marine ecosystems under global climatological, local hydrological and anthropogenic influences.

The creation of electronic databases, firstly on marine invertebrates, started at the Zoological Institute in 1987. Since 1989, PCs helped us in resolving some problems of building and updating data bases and information retrieval systems and allowed us to use them more efficiently.

The lack of a universal international approach to the management of collection data and a number of existing software on the basis of the different computer models, along with some specific problems with data input (*e.g.* Cyrillic symbols), did not allow us to already use compiled foreign software.

The databases on different groups of animals are often interactive and successfully supplement each other. Such combined databases on parasites and their hosts (mammals

and fish), insects and food plants etc. were developed at the Zoological Institute. Another example of such a combination is the databases on marine fishes and invertebrates. The material on both groups of animals can be collected as one sample at the same stations and by the same gear. This is the basis for working towards a combined strategy of data input for marine hydrobiology and ichthyology. In 1991, work on ichthyological databases was started (Voronina *et al.*, 1999). Fishes and invertebrates are the main components of any marine biocoenosis and therefore parallel research using joint databases is highly promising for the study of biocoenotic relationships and marine ecosystems.

Designed at the Zoological Institute, the information retrieval system "OCEAN" consists of four main tables: the taxonomic table containing the name and nomenclature of the taxa, the geographical table including the data of field books and catalogue of museum collections (locality of sampling: coordinates of stations, gear etc.), the ecological table (biomass, depth, temperature, salinity, oxygen etc.) and a bibliographic table. The system was improved by a new method of data input with the help of a thesauri system, developed by A.A. Golikov in FoxPro for Windows that minimizes the number of errors.

In spite of fast evolving information systems, standardization and digitizing of biological, in particular zoological, investigations is very slow, partly because of the complications of nomenclature and taxonomical relations. In the course of the organization of faunistic and ecological data and the information retrieval system, two serious problems appeared:

- the input and use of scientific names, especially synonyms
- the formalization of geographical data. The first problem is being solved by using the classifier of scientific names of animals based on the user-friendly and periodically updated ZOOCOD standard, popular among Russian biological institutions dealing with biodiversity research (Table I). The standard was developed in the late 1980s at the Zoological Institute RAS to transform the hierarchical classifications into a relational table (Lobanov and Zaitsev, 1993; Lobanov and Smirnov, 1997; Lobanov *et al.*, 1999).

Coordinates and the developed geographical information system were used to solve the second problem (Dianov and Lobanov, 1995).

The databases comprise information on field stations – localities of collection of marine invertebrates and fish, *i.e.* coordinates, depth, type of bottom, as well as method, gear, date of collecting and collector's name.

The system of the geographical data input takes into account the data standard Darwin Core. In combination with the taxonomic table (information on structure of fauna of certain region) and the collection table (place and method of storing collected specimens), the station data base allows creating different analytical queries to the derivative tables with consideration for hierarchical relations of fish and invertebrate taxa and geographic regions.

Table I. A structure of a ZOOCOD's classificatory system.

GENUS Latin name of genus (not obligatory)	**LATNAM** latin name of taxon	**SYN** code of synonymy	**RANCOD** taxonomic code of a rank	**ABBR** unique mnemonic code of taxon	**SYSCOD** digital systematic code
	Animalia		1	AN	100
	Arthropoda		10	AR	110
	Crustacea		20	CR	120
	Insecta		20	IN	130
	Coleoptera		40	INCO	13010
	Diptera		40	INDI	13013
	Chordata		10	CH	140
	Mammalia		20	MA	150
	Primates		40	MAPR	15010
	Pongidae		50	MAPRPO	15010100
	Hylobatidae	=	50	MAPRHY	15010100
	Gorilla		70	MAPRPOGOR	150101001000
Gorilla	gorilla		90	MAPRPOGORGOR	1501010010001000
	Pan		70	MAPRPOPAN	150101001010
	Hominidae		50	MAPRHO	15010105
	Homo		70	MAPRHOHOM	150101051000
Homo	sapiens		90	MAPRHOHOMSAP	1501010510001000
Homo	recens		94	MAPRHOHOMSAPRE	150101051000100010

The creation of the electronic databases and the design of the information retrieval systems OCEAN and ZOOINT (in standard ZOOCOD) carried out at the Zoological Institute, have allowed us to receive support for the project entitled "ECOANT" - "Creation of an information retrieval system on ECOlogy of benthos of the ANTarctica". The information retrieval system "ECOANT" can promote the resolution of the following problems:

- to refine the faunal structure of biota and its taxonomic features for the different areas
- to obtain ecological information
- to reveal changes in the structure of fauna in the investigated regions under influence of climatological and anthropogenic factors, which is one of the aims of the global ecological monitoring. The realization of the project is based mainly on the Russian biological data of Antarctic Regions and, first of all, on the unique benthic collections of the Antarctic and Sub Antarctic Seas. The preliminary information about this project is available on the web (http://www.zin.ru/projects/ecoant/index.html).

Using the Active Server Pages technology the database on the Antarctic seabirds, brittle stars and chitons are presented on the server of the Zoological Institute (http://www.zin.ru/projects/ecoant/eco1form.asp). The list of fishes and fishlike vertebrates of the Antarctic Region is prepared to come online.

The results of developing and updating the information retrieval system "OCEAN" on fauna of the Arctic, Antarctic, Far East Seas and inland seas of Russia will be unique

since the collections of marine fishes and invertebrates, accumulated and kept by generations of scientists during almost two ages, serves as a unique source of information.

By now the ichthyological part of the taxonomical table contains over 5,800 records: all high-level taxa including families, taking into consideration modern fish taxonomy; species of fish and fishlike vertebrates of the Antarctic Region, flatfishes (order Pleuronectiformes - 13 families) and scorpaeniform fishes (suborders Scorpaenoidei - 10, Cottoidei - 3 and Platycephaloidei - 2 families) of the world. The collection table includes data on fish specimens of the orders Pleuronectiformes and Scorpaeniformes as well as some fish species of the Antarctic Region kept at the Zoological Institute RAS. The database on marine invertebrates contains over 15,000 station records and information on chiton, bivalve and brittle star collections. Some characteristics of the collection database on marine fish and invertebrate collections are presented in Table II.

The information of the collection database is, however, still incomplete, because it covers information only for some taxa. Nevertheless up to now it is already possible to use the collection databases in analysis of secondary information and it will show some characteristics of the fish and invertebrates collection.

Table II. Some characteristics of databases on fish and invertebrates collections.

Group of animals	**Number of stations**	**Number of taxa**	**Number of inventory units**	**Specimens**
Fish	5,845	272 genera, 710 species	9,083 (including 157 types)	26,524 (including 23 stuffed fishes)
Invertebrates (Arctic)	14,897	62 genera, 110 species	11,913	82,851 wet, 19909 dry
Invertebrates (Antarctic)	2,520	64 genera, 136 species	2,887	2,619 wet, 1,608 dry

A great part of the material of the Zoological Institute was collected during the well-known Russian expeditions as well as foreign ones (Table III). Only few expeditions published their route and station data in the special issues and for many years these works were not available, even for specialists (Lindberg, 1954). Many other expeditions have only handwritten diaries. Sometimes the label data, and therefore catalogue data, are very fragmentary. The creation of the joint international expedition database would be historically very interesting and also very useful for further input of collection data in helping to unify data and in avoiding errors during input. The example of such a useful information source is the Challenger Expedition 1873-1876 database on the site "Biogeoinformatics of the Hexacorals: http://www.kgs.ku.edu/Hexacoral/" (Fautin and Buddemeier, 2003).

Table III. The most extensive collected material of Russian and foreign expeditions kept at the ZIN.

Name and abbreviation of expedition	Number of "ichthyological" stations	Number of "invertebrate" stations	Date
Polar Exp. of K. Baer	12		1840
Murman Scientific-Fishery Exp. (ENPIM)	482	2896	1880-1915
Spitsbergen Exp.	34	2	1899-1901
Russian Polar Exp. (RPE)	44	107	1900-1902
Novaya Zemlya Exp.	35	13	1901-1935
Baltic Exp.	23		1907-1908
Far East Exp. (FEE)	388	81	1908-1915
Hydrographic Exp. To East Ocean (HEEO)	128	279	1908-1927
Exp. ZIN to Japan Sea	63	32	1934
VNIRO Kara Sea Exp.	46	19	1945-1946
Kuril-Sakhalin Exp. (KSE)	314	1156	1947-1949
Soviet-Chinese Exp. (SCE)	37	16	1956-1959
Southern Sakhalin Exp.	40	12	1946
TINRO Exp.	68	173	1928-1978
Tropic Exp.		147	1974-1975
Arctic Exp. "Polarstern" (Germany)	18	222	1985-1998
MERA-95		194	1995
Severnyi Polyus (North Pole)	19	404	1946-1948
Shantar Exp.		300	1978
Antarctic expeditions			
Antarctic Exp. "Polarstern" (Germany)		111	1972-1978
AzCherNIRO Exp.		336	1969-1976
Soviet Antarctic Exp. (SAE)		843	1956-1989

An example of one of the most intensive expansions of collections, in relation with the long and extensive expeditions such as ENPIM (1880-1915) is given in Fig. 1. It is expected that the periods of severe social circumstances (1917-1920 and 1941-1945) are characterized by very few samplings of zoological material.

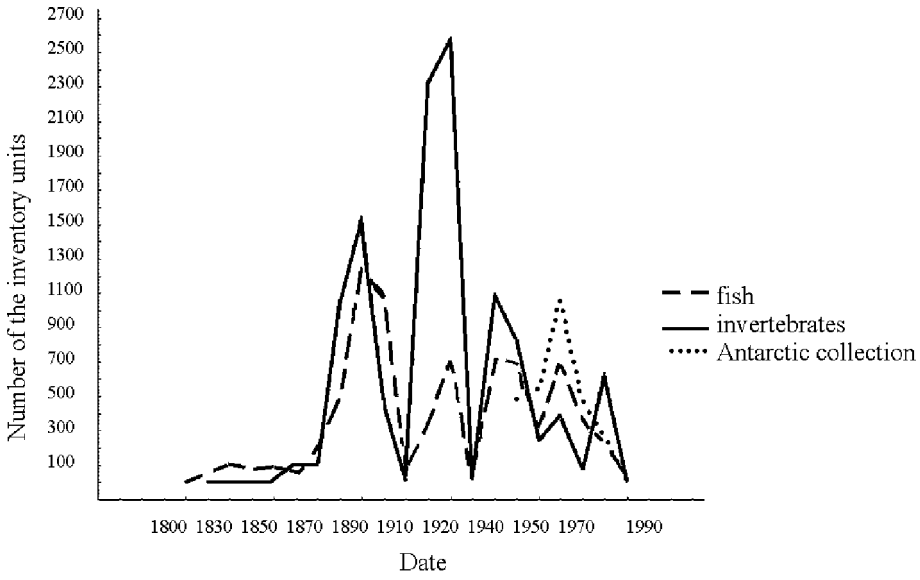

Fig. 1. Chronological chart of the fish and invertebrates collecting (according to data bases today).

In total about 50 expeditions and 850 collectors sampled zoological material since 1828 (Table III, IV) aboard of 180 vessels, among them some well-known scientific vessels such as - "Andrei Pervozvannyi"; "Vityaz"; "Akademik Knipovich" (VNIRO); "Ob" (AARI); "Skif" and "Aelita" (AzCherNIRO); "Zund" and "Evrika" (AtlantNIRO) as well as Marine Fishery Fleet vessels, and occasionally even military vessels. The part of the samples collected by the academician vessels is 5-10%. The biggest part of the benthic collection (65-70%) was sampled during expeditions on board Marine Fishery Fleet vessels, despite of defects of the sampling and labels, difficulties of preservation and storage etc. Each cruise, even those with little material, contributed to faunistic investigations of this unique world of marine life.

Table IV. The names of the collectors of the largest number of samples recorded in our current databases.

Name of collector	Number of ichthyological stations	Number of invertebrate station
Andriashev A.P.	96	
Arngold E.E.	26	114
Averintsev V.G.	12	270
Barsukov V.V.	127	
Brazhnikov V.K.	71	
Bryazgin V.F.	5	395
Bunge A.A.	47	
Byalynitskii-Birulya A.A.	10	135
Bykhovskii B.E.	66	
Derbek F.A.	92	
Fedorov V.V.	106	
Foroshchuk V.P.	97	
Golikov A.N.	20	303
Gorbunov G.P.	55	447
Gruzov E.N.	7	89
Gurjanova E.F.	43	
Herzenstein S.M.	20	296
Knipovich N.M.	27	481
Kobyakova Z.I.	4	119
Koltun V.M.	12	513
Kondakov A.N.	5	83
Legeza M.I.	90	
Lindberg G.U.	93	
Neyelov A.V.	45	
Petryashov V.V.		256
Rutenberg E.P.	78	
Shmidt P.Ju.	200	
Sideleva V.G.	97	
Sirenko B.I.	5	557
Smirnov A.V.	20	197
Soldatov V.K.	364	
Starokadomskii L.M.	36	123
Ushakov P.V		333
Vagin V.L.	65	368
Vinogradov L.G.		310
Voznessenskii I.G.	31	

Some material (*e.g.* 178 inventory numbers of the ichthyological collection and more than hundreds invertebrates) has been received in exchange with foreign museums.

The text catalogues of the collections of the Zoological Institute were published only for type specimens. The catalogue of all flatfish collections (Pleuronectiformes) has been

compiled on the basis of the information retrieval system OCEAN (Voronina and Volkova, 2003) and published recently. It is planned to launch the collection databases directly on the internet.

Table V. Comparison of the information retrieval systems OCEAN and ARTEDIAN.

Advantages of OCEAN	**Advantages of ARTEDIAN**
Station unique codes are generated automatically.	By default more fields describing the object can be filled in
Leading spaces are being deleted automatically.	The collectors are treated in separate fields.
The important fields (vessel and others) are filled in Russian, in addition to English.	The IRS is conform the Darwin Core in coordinate notation (degrees, minutes and seconds separately)
Some more additional fields (EXPEDITION, GEAR, STATION NUMBER etc.) provide an opportunity to verify location with a route table for marine expeditions.	Additional fields are designed for freshwater stations, conform the Darwin Core (lakes, rivers, states, provinces, county etc.)

International projects to create data base and information retrieval systems for sharing biodiversity information on a global scale are in development. Some examples of the comparison of the data input systems of OCEAN and Artedian, the system used in creating the collection databases FishBase, is presented in Table V. The main point is that in foreign projects the system of data input is exclusively based on Latin symbols, but not Cyrillic. This restricted approach leads to a considerable reduction of information originating from the actual labels that are often hand-written in local national languages, *e.g.* Russian. In addition, it is sometimes difficult to translate these data equivalently and completely. Therefore we feel it is important to consider special fields in the tables of the information retrieval system to allow input of data as well as to perform the queries in Cyrillic symbols together with the Latin ones. The information retrieval system OCEAN provides this possibility. It is also worth noting the necessity to examine the experiences of different teams and try to create not ideal but optimal systems for input and treatment of marine data.

Acknowledgements

Supported came from the Project N11 “Exploration and research of the Antarctic Region” of the Federal Program “World Ocean”, grants № 05-07-90354, 04-04-49300, NSh 1668.2003.4 and program “Information system on a biodiversity of Russia”.

References

Atlas of Antarctic. 1969, v. 2. Gidrometeoizdat, L. 598p.

Dianov M.B. and A.L. Lobanov. 1995. Computerized geographical system ZOOMAP for mapping of plants and animals areas. In: Abstr. II Soveshchanie “Kompyuternye bazy dannykh v botanicheskikh issledovaniyakh”. St. Petersburg, April 17-19 1995:16-17

Fautin, D.G. and R.W. Buddemeier. 2003. Biogeoinformatics of the Hexacorals: http://www.kgs.ku.edu/Hexacoral/

Lindberg G.U. 1954. Obzor rabot Kurilo-Sakhalinskoi morskoi kompleksnoi ekspeditsii Zoologicheskogo Instituta I Tikhookeanskogo instituta rybnogo khozyaistva (Review of the works of the KSE). Trudy Kurilo-Sakhalinskoi Ekspeditsii ZIN-TINRO, 1947-1949. Vol. 1. M.-L.:7-100.

Lobanov A.L., I.S. Smirnov and M.B. Dianov. 1999. ZOOCOD – conception of representation of zoological hierarchical classifications in relational databases. In: Abstracts of the Int. Symposium "Information retrieval system in biodiversity research", Proceedings of the Zoological Institute, V. 278. St. Petersburg:65-66.

Lobanov A.L. and I.S. Smirnov. 1997. Principles of arrangement and using of classifiers of animals in the standard ZOOCOD. In: Data bases and computer graphics in zoological investigations. Proceedings of the Zoological Institute, V. 269. St. Petersburg:66-75.

Lobanov A.L. and M.V. Zaitsev. 1993. Creation of computer data bases on the systematics of mammals on the basis of classificatory of animals names "ZOOCOD". In: Questions of systematics, faunistics and paleontology of small mammals, Proceedings of the Zoological Institute. V. 243. S.-Petersburg:180-198.

Smirnov I.S., A.L. Lobanov and M.B. Dianov. 1999. Zoological digital (virtual) museums. In: Scientific service in network Internet. Abstract of All-Russian scientific conference in Novorossiysk on September 20-25, Publ. of Moscow University:185-187.

Smirnov I.S. and A.V. Neyelov. 1996. Studying of Antarctic bottom fauna in cruises of fishery vessels of USSR and Russia. In: Abstr. of Int. conference: The history of native oceanology. Kaliningrad, 28 October – 1 November 1996:106-107.

Voronina E.P. and G.A. Volkova. 2003. Catalogue of specimens in the collection of the Zoological Institute, Russian Academy of Sciences. Osteichthyes, Pleuronectiformes. Explorations of the fauna of the seas. Vol. 55(63). St. Petersburg, Zoological Institute RAS. 198p.

Voronina E.P., I.S. Smirnov and A.A. Golikov. 1999. The computer approaches to the ichthyological studies in Zoological Institute RAS. In: Abstracts of the Int. Symposium "Information retrieval system in biodiversity research", Proceedings of the Zoological Institute, V. 278. St. Petersburg: 116-117.