TAXEX: TAXonomic EXpert system. History of development and technology of identification

Sergey Lelekov and Anton Lyakh

Dept. of Biophysical Ecology
Institute of Biology of the Southern Seas, Nakhimov av. 2, Sevastopol, 99011, Ukraine
E-mail: antonlyakh@gmail.com

Abstract

TAXEX is a series of taxonomic expert systems, which are developed to help scientists to professionally identify living organisms. They provide scientists with different taxonomic information, including taxon descriptions and diagnosis, geographic distributions, scientific nomenclature, identification keys and illustrations; it creates a tool for interactive identification of living organisms and trains new taxonomists. The main goal of TAXEX is to give public access to taxonomic and expert knowledge of the Black and Azov Sea biota. These systems can be used in interdisciplinary sciences like biological oceanography, biophysics, landscape ecology, bioecology, etc., in which specialists from different scientific fields are needed. Using taxonomic expert systems instead of high-paid taxonomists will reduce costs of scientific research and will allow many scientists without a specific biological education to work independently.

Keywords: TAXEX; Taxonomic expert system; Identification; Taxonomists training; Taxonomic knowledge base.

Introduction

The loss of biological diversity, our genetic heritage and the loss of habitats is accelerating in many parts of the world. That loss, exacerbated by our incomplete knowledge of the earth's biota, diminishes stewardship, restricts management, and imperils conservation of biological resources. Two components of this global problem are:

- loss in expertise necessary for the identification and inventory of biota
- poor state of knowledge of many aquatic and terrestrial organisms.

TAXEX (an abbreviation of TAXonomic EXpert system) is a series of taxonomic systems created to solve the problems mentioned above. They provide scientists with various taxonomic information, including taxon descriptions and diagnosis, geographic distribution, scientific nomenclature, identification keys, illustrations; it gives them a tool for interactive identification of living organism and allows training of new taxonomists. The main focus of the TAXEX Project is the Black Sea and Azov Sea region. The biodiversity of the Black and Azov Seas is well documented in local monographs and scientific papers, but there are no public Internet resources for those

regions. In the past few years, interest in this region has increased, due to unique discoveries of species previously unknown to science, inhabiting the hydrogen-sulphide zone of the Black Sea (Sergeeva, 2003). Previous considerations stated that there was no life in that particular zone except for anaerobic bacteria. Besides traditional Black Sea species, some unique fish, which previously did not occur in the Black Sea, have now been observed, for example, *Sphyraena obtusata*, *Micromesistius poutassou*, and *Heniochus acuminatus*. New migrant species have been found and described as well. The creation of publicly accessible Internet resources will help enhance conservation of biodiversity of the Black Sea and Azov Sea region, it will enlarge our knowledge of life in the region and it will open a way to gather new information (Tokarev *et al.*, 2002).

Objectives

The objectives of the TAXEX Project are:

- to collect, describe and classify broad taxonomic resources of the Black and Azov Sea
- to maintain a vast knowledge of Black and Azov Sea biota in databases and knowledge bases and save it for future generations of scientists
- to store the knowledge of expert taxonomists in the right format within taxonomic expert systems
- to give public access on taxonomic and expert knowledge of the Black and Azov Sea biota to the scientific community and to fill the gap of non-accessible information about Black and Azov Sea flora and fauna
- to develop software for expert taxonomic identifications of the Black and Azov Sea biota, and for training new generations of taxonomists

History and enhancement of TAXEX

TAXEX is a taxonomic expert system – an interactive computer identifier of biological species – and a knowledge base including specific taxonomic information, a glossary of terms, references, etc. TAXEX has been developed since the end of 1980s at the Institute of Biology of the Southern Seas (Sevastopol, Ukraine). During the development of TAXEX, the algorithms to taxonomic identification and the interface to present this knowledge has changed since, but the traditional dichotomous taxonomic keys were never used. We tried to model the behavior of the expert, when he is identifying a taxon.

The first system's versions were working under MS DOS, but this had many limitations. The first version of the TAXEX Expert System was based on the conception of the frame – which is a computerized presentation of the expert's idea about the identified object. Defining the frame properties allowed to identifying taxa. These principles were the basis of a computer identifier of the Black Sea Isopoda (Lelekov *et al.*, 1996; Butakov *et al.*, 1997b). Drawbacks of this approach had become apparent under attempts to create new identifiers. The description of frames and the different rules on how to use them were so specific for every group of organism that the creation of a common method to forming frame descriptions became difficult. Hence, further developments were

concentrated on the attempt to create a more universal model on the process of taxonomic identifications. Such a mechanism was developed (Lelekov, 1994) and used in computer identifiers for the Black Sea Fish Larvae, Black Seas Bivalvia, Black Sea Gastropoda, and Fishes of the Black and Mediterranean Seas (Butakov *et al.*, 1996; Butakov *et al.*, 1997a; Butakov *et al.*, 1998). These expert systems work under MS Windows and manage taxonomic knowledge that is stored in a database.

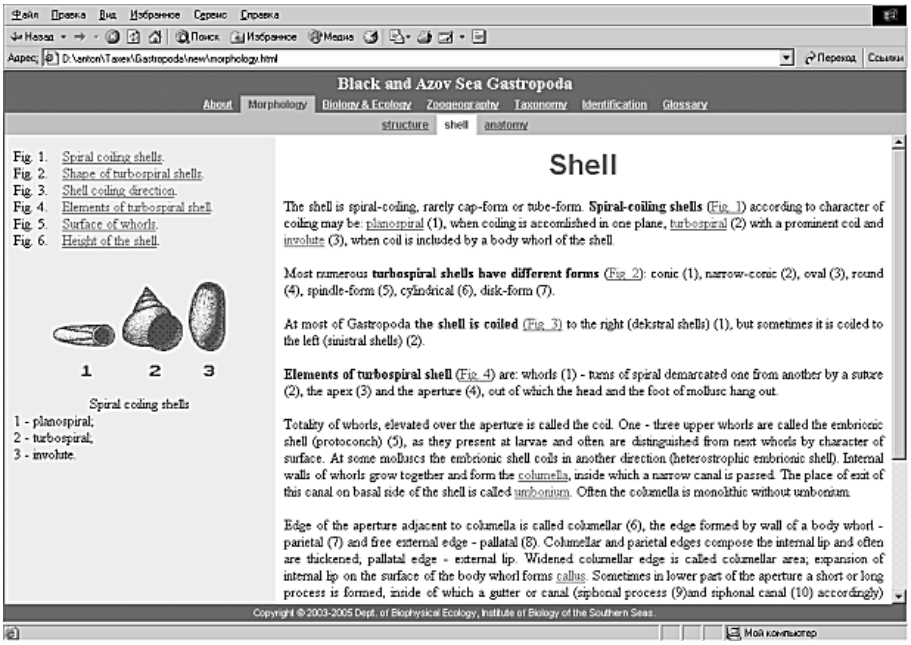

Fig. 1. TAXEX System of the Black and Azov Sea Gastropoda.

Providing public access to taxonomic and expert information needed the creation of a new generation of online software tools, which can work on both the Intranet and Internet. The Java version of our TAXEX Expert System was developed at the end of the 1990s, and now is a Java applet that uses information stored in identification tables, and the taxonomic knowledge base consisted of a set of linked HTML pages (Fig. 1). TAXEX can now be easily distributed and it is accessible for everyone.

Identification algorithms

To describe the identification algorithms used in TAXEX we first considered a classical taxonomic identification scheme based on a dichotomous key. It can be presented as a binary tree, where the nodes contain the descriptions of the taxonomic characters, and the leaves contain the description of taxa. To identify an organism it is necessary to consecutively traverse the identification tree from its root till one of the leaves (Fig. 2, A). For every step you need to choose the state of a character appropriate to your taxon and select the next direction of movement.

This dichotomous identification process has the following disadvantages:

- at every step of the identification you can only choose between two variants that increases the number of necessary steps
- the path is fixed: you move from tree up till a tree leaf and no character can be omitted
- the states of some characters are undefined for some taxa, if the characters and their taxa lie in different branches
- if you cannot find the state of a character, when an organism for instance is damaged, further diagnosis becomes impossible.

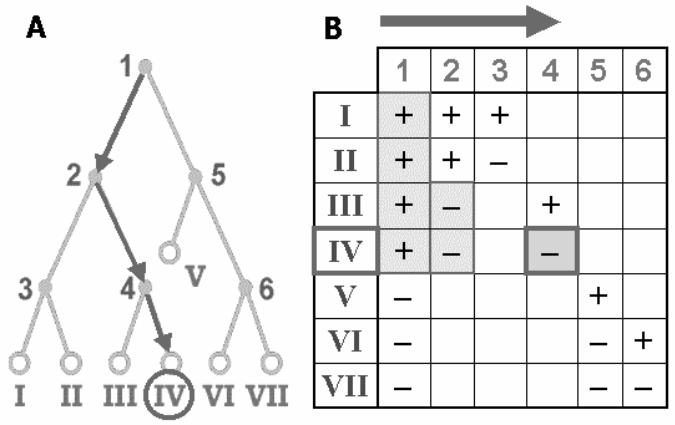

Fig. 2. (A) Representation of a classical identification key in the form of a binary tree. Arabic numerals are identification characters. Roman numerals are identified taxa. Arrows show one of the possible ways of identification. (B) Identification table, which corresponds to identification tree. The arrow above of the table shows the path of diagnosis.

Identification trees can be converted into an identification table – a matrix, where the rows contain the states of all the identification characters for a given taxon, and the columns contain the state of a given character for every taxon. Using an identification table, the organism's identification procedure can be presented as the consecutive division of the group of taxa into subgroups while a minimal element – a taxon – would be reached. The path of diagnosis is determined by the order of columns (Fig. 2, B).

The identification table for a binary taxonomical key has many empty cells, and the filled cells can only contain two different values, + or – (yes or no). To improve this situation we propose to use a new identification table, called improved identification table in which:

- there are no empty cells, that is: no undefined characters
- characters can have more than two states
- cells can keep more than one character state.

We have built our identification process on the improved identification table that allows any user of the TAXEX System to fill in the cells more freely, when he/she is diagnosing an organism. At every step of the identification, the user has to answer the questions about some taxon characters, and he can:

- omit the character, if he could not determine its state, for example when the organism is partially damaged
- choose more than one answer, when he/she notices that the taxon in question has multiple character states, or he/she is not sure of the accuracy of the character determination.

To identify a taxon in such non-rigorous conditions, the TAXEX System operates by a hypothesis concerning the taxonomical position of the organism. In the beginning all taxa have maximal probabilities, and the system supposes that the diagnosed taxon belongs to the higher taxonomic rank kept in the system knowledge base. Through the selection of states, the probability of taxa that do not have selected states decreases. (Fig. 3) Note, that probabilities of the hypothesis can be decreased only.

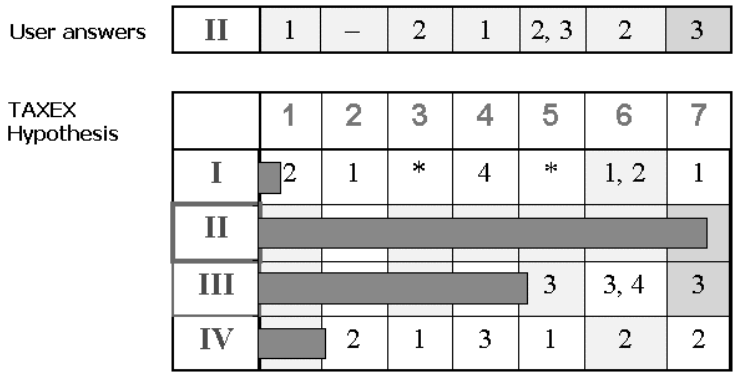

Fig. 3. Hypothesis of TAXEX after a user has answered the system's questions. Above are the states of the characters of the identified object. Below is the advanced identification table with horizontal bars, which shows the TAXEX hypothesis. In this case, the most probable hypothesis is that the identified object is taxon II.

When the identification is finished, the user obtains the taxon with the maximal probability as the result of the diagnosis. Certainly, he can view all other system versions with lower probabilities. If the user knows the taxonomic rank of the organism, he can alter the current system hypothesis, selecting the name of order, family, genus, etc., this allows omitting questions concerning other taxonomic groups and reducing the number of identification steps.

To reduce the number of necessary steps TAXEX tries to choose that character which could confirm the most probable hypothesis. This character has to divide the group of the most probable taxa into subgroups. At every step TAXEX tries to choose the best dividing character according to the current hypothesis and remaining not identified characters. Moreover the best divider has to satisfy on two criteria:

- the first criterion takes the probability of the occurrence of a taxon into account. Some species are common species. You observe them very often, and sometimes are present in every sample you study. Other species are rare, you only encounter them in few samples from limited regions, or you almost never find them. Therefore, in the beginning, it is useless to ask the user about the characters of rare species. Firstly, the system supposes that it is a common taxon and tries to identify it. When the diagnosis does not give any appropriate results, the system will ask you about the rare species characters
- the second criterion takes into account the cost of taxon characters determination. Usually, characters with small determination cost do not need special instruments; they are external and can be easily distinguished. Characters with large determination costs usually are internal; you need special instruments, such as a microscope or scalpel to examine them. Therefore, TAXEX first asks questions with small determination costs and gives you the possibility to identify a taxon without additional operations, like microscopic observations or dissections.

Costs of character determinations and probabilities of sampling a species in nature are set up by taxonomists, who developed the identifier. The best divider is chosen by the system at each step of identification.

As a result a new version of TAXEX identifiers has been developed. Its functioning is based on the following information:

- identification table, which set the correspondence between taxa and their characters
- costs of character determination
- probability of sampling of a certain taxon in nature.

The identification information is stored in a knowledge base, which also keeps data on species and taxonomic group descriptions, biology, ecology, biogeography, drawings and photographs, glossary of terms, bibliography, etc. The knowledge base and computer identifiers together constitute the Taxonomic Expert System.

Training

Training of object identification consists of the following elements:

- studying the relation of specimens to all object classes
- the study of distinctive characters of specimens from different classes
- becoming familiar with identification procedures.

Instructors with excellent expertise and good quality training capabilities, including access to biological collections, atlases of animals and plants, identification keys, are necessary. For young universities, where scientific schools are just about getting started, the availability of both the instructors and the training equipment is often a problem. In

these cases, TAXEX can help to solve this lack of resources. TAXEX Expert Systems allow accomplishing professional object identification, and they have special tests for the teaching and training of users. Any expert system includes four tests:

- How good do you know a specimen? The test is used to develop a pupil's ability to identify a specimen by using specific characters. During the test the pupil has to select the correct states of these characters.
- How good do you know taxon characters? This test is used for training the pupil to distinguish taxa within a taxonomic rank. The system shows characters, points out their states and asks the pupil to choose the taxa to which these characters are related.
- Determine a taxon by its characters. This test is used to gain knowledge about an identified taxon. During the test the TAXEX system enumerates states of the taxon characters and at the end asks the pupil to indicate the taxon, to which these characters are related.
- Determine a taxon by images. The test is used to train the pupil's ability to recognize a taxon by the use of images. The system shows one or more images of possible taxa and the pupil has to indicate its taxonomic name.
-

These four tests can serve as a good methodical basis for preparing new specialists. Moreover, interactive training tools stimulate the pupil to actively learn how to solve questions, hereby making use of the TAXEX knowledge base, in addition to other sources of information and expert knowledge. One of the important advantages of the system is the possibility to put it on the Internet. This allows organizing of distance learning.

Conclusion

By using TAXEX you will get access to expert knowledge and will be able to identify taxa like an expert. These systems can be used in interdisciplinary sciences like biological oceanography, biophysics, landscape ecology, bioecology, etc., in which specialists from different scientific fields are needed. Using taxonomic expert systems instead of high-paid taxonomists will reduce the costs of scientific research and will allow many scientists without a specific biological education to conduct their research more independently.

Identification tools are also useful for young taxonomists, who just begin to learn to identify species. Four tests, included in the expert system, can serve as good methodical basis for preparing new specialists. Interactive training tools stimulate the pupil to actively learn how to solve questions, hereby making use of the TAXEX knowledge base, other sources of information and expert knowledge. One of the important advantages of the system is the possibility to put it on the Internet, which allows organizing of distance learning.

References

Butakov E.A., S.G. Lelekov and V.D. Chuhchin. 1996. Kratkiy opredelitel dvustvorchatyh molluskov Chernogo moray treh nadsemeystv *Cardioidea*, *Veneriidea*, *Tellinoidea*. [Brief identifier of Black Sea molluscs of superfamilies *Cardioidea*, *Veneriidea*, *Tellinoidea*]. Institute of Biology of the Southern Seas. Sevastopol. 46p.

Butakov E.A., S.G. Lelekov and V.D. Chuhchin. 1997a. Opredelitel bruhonogih molluscov *Gastropoda* Azovo-Chernomorskogo basseyna. [Identifier of Black and Azov Sea *Gastropoda*]. Institute of Biology of the Southern Seas. Sevastopol. 127p.

Butakov E.A., S.G. Lelekov, E.B. Makkaveeva and V.F. Zhuk. 1997b. Opredelitel rakoobraznih otryada *Isopoda* Chernogo morya. [Identifier of Black Sea *Isopoda*]. Institute of Biology of the Southern Seas. Sevastopol. 79p.

Butakov E.A., E.Yu. Georgieva, S.G. Lelekov, N.P. Pakhorukov and L.P. Salekhova. 1998. Opredelitel semeystv ryb Sredizemnogo morya. [Identifier of Black Sea fish families]. Institute of Biology of the Southern Seas. Sevastopol. 483p.

Lelekov S.G. 1994. K voprosu vybora posledovatelnosty priznakov v diagnosticheskih expertnih systemah. [To the question of choosing characters in diagnostic expert systems]. Kibernetica 4:167-173.

Lelekov S.G. 1995. PC-based identification of species for ecological monitoring. p.179-183. In: ECOSET'95. International conference on ecological systems enhancement technology for aquatic environment. The Sixth International Conference on aquatic habitat enhancement. Tokyo.

Lelekov S.G. and E.B. Makkaveeva. 1996. Komputernoye opredeleniye rakoobraznih otryada *Isopoda* Chernogo morya. [Computer identification of Black Sea order *Isopoda*]. Zoological Journal 75:35-44.

Sergeeva N.G. 2003. Meiobenthos of deep-water anoxic hydrogen sulphide zone of the Black Sea. p.880-887. In: Oceanography of the Eastern Mediterranean and Black Sea. Similarities and differences of two interconnected basins. Yilmas A. (Ed.). Tubitak Pube.

Tokarev Yu.N., S.G. Lelekov, V.V. Melnikov and A.M. Lyakh. 2002. Perspectivy ispolzovaniya computernih opredeliteley v oblasty taxonomii. [Perspectives of using computer identifiers in taxonomy]. Ecologiya moray 61:95-98.