Information in the Biological Datasets and Biodiversity Estimation on the Basis of the Peridat On-line Database

Éva Hajnal[1], Gábor Teke[2], Csilla Stenger-Kovács[2] and Judit Padisák[2]
1 Óbuda University Alba Regia University Center H-8000 Székesfehérvár Budai str. 45.
2 University of Pannonia Department of Limnology, H-8200 Veszprém Egyetem str. 10.
e-mail:hajnal.eva@arek.uni-obuda.hu

***Abstract*—As the result of the diatom research of the Department of Limnology of the University of Pannonia, thousands of data were accumulated. For the storage and queries of these data the Peridat on-line database was developed. The aim of this study is to describe the database and to investigate with help of this database, how we can compare the biodiversity of three periphyton datasets. The other question was the efficiency of the periphyton species inventories of a waterflow at different sampling frequencies. We used the datasets of Torna stream at Devecser from two, one year long period from 2004 to 2005 and from 2008 to 2009, and the dataset of Csigere stream at Devecser from 2008 year. We compared the taxon lists, and calculated the cumulative number of species. We queried the relative abundance values of diatom species from all the three datasets. The cumulative numbers of species have an asymptote (~75 species) at the data of Torna stream, but data of Csigere stream are not asymptotic, however their maximum is above 100. By the help of EstimateS software we calculated the whole number of species for the whole succession period by different nonparametric formulas from the relative abundance values. The asymptote of the second order Jackknife index was 104 at the samples of Torna stream in the period of 2004-2005 and 94 at the samples of Torna stream 2008-2009 and 141 at Csigere stream 2008. As we determined, the taxon lists of Torna stream differed at 20-30 taxa, but these differences can be attributed to sampling. One sampling during a year in random time makes possible finding of the 30% of the whole taxon list, the monthly sampling expolores ~50% and the biweekly sampling ~70-80% of the whole taxon list.**

I. Introduction

Database usage in biological research is widespread, because the projects are lasting for few decades during which large amount of expert data accumulate. These data after standardization can be stored in. Database queries make possible comparative studies, and large scale statistical calculations which use thousands of data, or model calculations. So these databases are meaningful background of ecology [1, 5, 6, 8, 10], as were proved in some studies. In phytoplankton studies, database helped us to assess water quality changes of Lake Balaton during the last thirty years. It also demonstrated that database is not only the aid of data process but a way of widening scientific knowledge [6]. An important ecological survey of water bodies in Hungary was the ECOSURV, which extended to the whole area of the country. Its purpose was the state assess by single samples and the survey comprised five indicator groups as recommended by the Water Framework Directive of the European Community. It was executed during the 2004 and 2005, and the generated large amounts of data were organized into a specially developed database. Its results were important in the implementation of the Water Framework Directive in Hungary [1]. Later this database became accessible for the experts.

The Department of Limnology of the University of Pannonia has been dealing with diatom research of the Hungarian streams for about a decade. Consequently a large number of benthos and water samples were collected, and analyzed. The benthos samples were investigated to determine their diatom species and the relative abundance values of each diatom species. The water samples were analyzed for chemical compounds and some physical factors were measured during the sampling [2, 9, 10]. The systematical storage of these data required to construct a scientific database. The purpose of our paper is to demonstrate the database structure and functions and a case study of the application of our database.

II. Results

After the examination of the available data we decided to develop an on-line database which can be accessible on the Internet from any location of the world for a registered user. It can be a starting point of a larger database system, which accepts voluntary users. The basic rule is in our system is that, who provides the system with his/her own data, can access automatically the other datasets too. We hope that we can collect enough data for the diversity calculations of a larger region. The database was developed with standard methods for RDBMS, preparing the Entity-Relationship model, Schemes of the Relations and the physical plans [7].

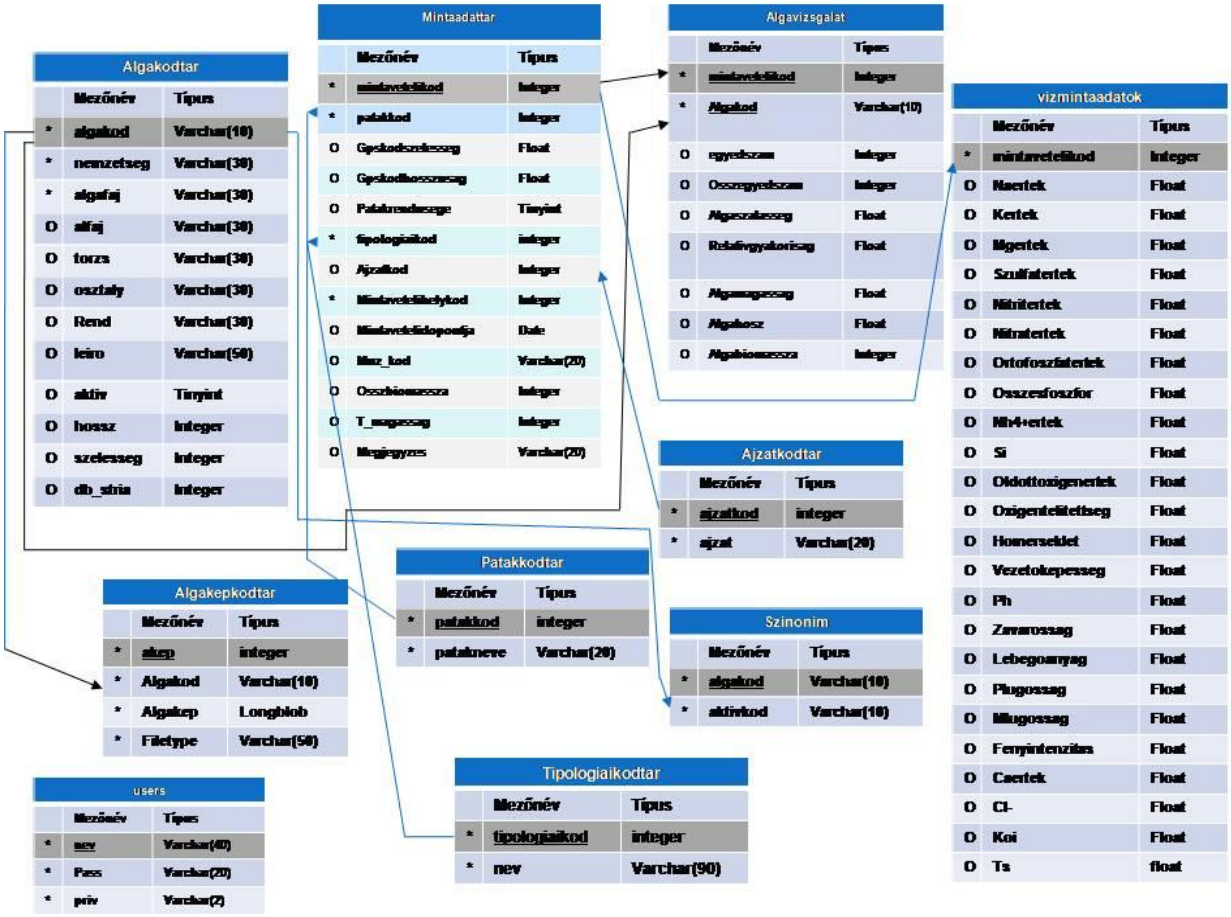

Figure II-1. The scheme of Peridat database

The physical plan, shown in figure II-1, contains only few tables. The database stores the basic data of diatom species, which are necessary for the high level scientific data processing, and this database function is a valuable supply for an expert and a student to become absorbed in taxonomy of periphyton species. The data of algal species can be supplemented with their photos, because the modern taxonomy requires light and in some cases electron microscopic phototechnique. As a result of the taxonomical research the diatom species are often reclassified, and get new taxon names. Therefore synonyms rise and are used parallel. These names need to be connected to each other with a relation.

The database contains separate tables for the data of sampling (date and location of sampling, remarks etc. see figure II-1.), data of water samples (chemical data completed few environmental factors measured in the ground) and data of periphyton samples. Tables of substrata types and typological types of water flows according to the recommendation of the Water Framework Directive of the European Community are also parts of the database. The user table, which helps in organizing the access levels and the rights of the different users, is worth mentioning. The implementation was accomplished with MySQL database engine and InnoDB tables. The user interface software was developed in PHP programming language. The PERIDAT ON-LINE database can be accessed and tried in the http://www.limnologia.hu/perifiton URL, and after logging in, you can use it.

The user interface is shown in figure II-2. Users can access the database at three levels. Before login you can visit the presented photo collection and list the name and classification of each species. During the registration users get a login name and a password to use the database at two levels. The “visitors” allowed using query forms for listing basic data of species and basic data of water and periphyton samples. The “data owners” additionally may load data and photos onto the database, and are allowed to modify basic data, and use any SQL queries, which are not related to the user table. Database rights modifying are allowed (including inserting new user) only for the database administrator. It does not mean further workload, according to the limited number of users.

After the development of the data structure and software we need to load data into the Peridat on-line database. The own datasets of the Department Limnology of University of Pannonia were loaded primarily onto the database. One part of data was stored already electronically, but some data were only printed, and we had to store these in sheets, in csv format files. Now the Peridat On-Line database contains data of 21 water flows, which means datasets of about 700 samples. The basic datasheet contains the records of about 600 species. In summary, now there are ~23 000 data in the database, which are partly data of water samples and partly periphyton samples, as it is shown in figure II-3.

The database is useful for ecological calculations, as we demonstrate here on the diversity estimation of some small water flows. For calculations the datasets of Torna stream from 2004 year until 2005 year and from 2008 until 2009 and data of Csigere stream from 2008 were chosen, because in these datasets the sampling frequency was adequate for the investigation. The sampling period in all three cases held over a year and the number of samples is between 20 and 80. The datasets were queried by SQL, and results were exported into Excel sheets for further statistical analysis. Firstly, we compared the taxon lists, and calculated the cumulative number of species. After all the relative abundance values of diatom species from all the three datasets were queried. The cumulative numbers of species have an asymptote (~75 species) at both data of Torna stream, but data of Csigere stream are not asymptotic, however, its maximum is above 100. If we compare the two species lists of Torna stream, we could find several differences. The taxon lists of Torna stream differed at 20-30 taxa, but this phenomenon does not concern the frequent species. From this point of view, our purpose was to calculate the local diversity (now diversity means the number of species) on the basis of the three datasets. The question was if the taxon list differences are only the results of the sampling methods, or reflect real ecological changes? The further question was the estimation of the sampling efficiency. For calculations among others the second order Jackknife formula was used [3,4],

$$S_P = S_o + \left[\frac{a_1(2n-3)}{n} - \frac{a_2(n-2)^2}{n(n-1)} \right]$$

where S_p is the calculated inventory of species (items), S_0 is the total number of species counted in peryfiton samples, a_1 the number of species that occurred in only one sample, a_2 the number of species that occurred in two samples and n is the number of periphyton samples. The calculations were made with the help of the EstimateS 7.0 software [4], its input was the Excel sheets exported from the database queries. The results can be seen in Table II-1., and it predicts substantial difference between the counted and estimated numbers of species. For further understanding we calculated the ratio between the counted and the estimated total number of species as function of number of samples (see Figure II-4.). This ratio characterizes the sampling method. The efficiency of sampling is between 30% and 90% in the real ground work. One

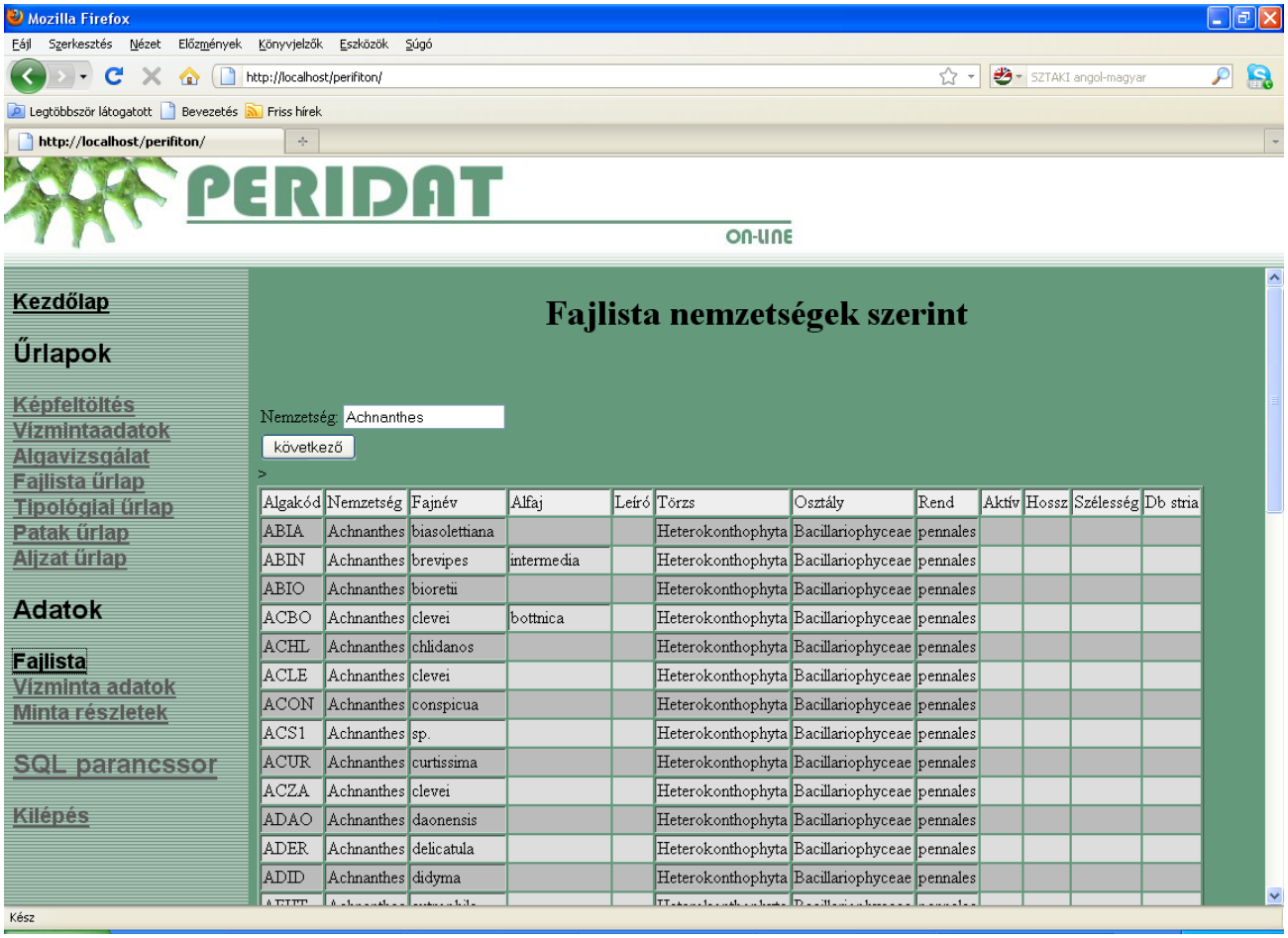

Figure II-2. The user interface of the Peridat on-line database

sampling during a year in random time makes possible the finding of the 30% of species of the whole taxon list, the monthly sampling causes the finding of ~50%, biweekly sampling the ~70-80% of the whole taxon list.

III. CONCLUSION

Datasets, that accumulated as a result of the continuous research of Department of Limnology of University of Pannonia were collected, standardized and organized into a database. A case study was executed to investigate the diversity of two streams, the changes of the diversity, and its statistical background. The basic question was the efficiency of periphyton sampling.

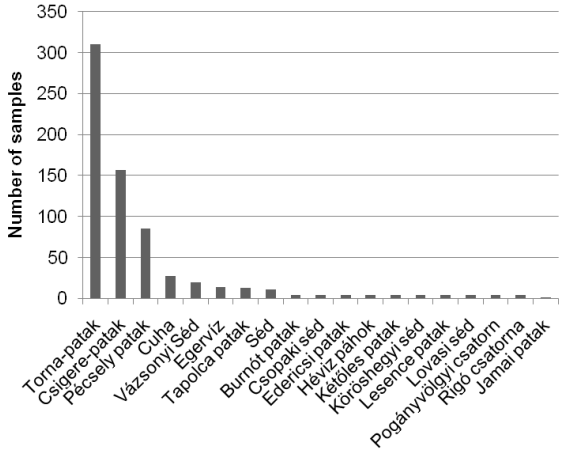

Figure II-3. Number of samples by streams

We established that the really found local diversity is much smaller than the calculated one. This result does not question the water qualifying methods, which are based on calculating indices from the abundance values of periphyton species. These calculations use the weighted relative abundance values, their results are not affected by the not found rare species. We must bear in mind these results when comparing the species list of different habitats.

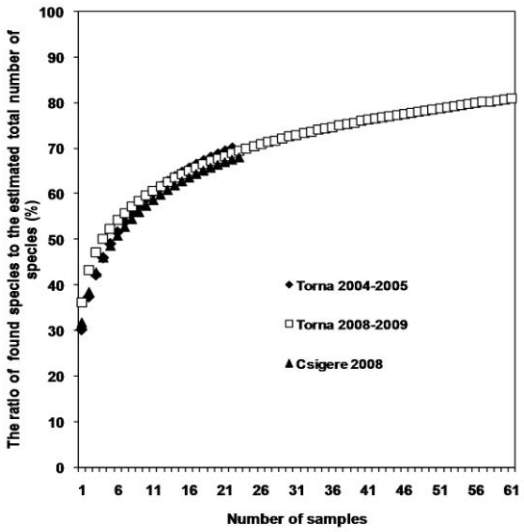

Figure II-4. The sampling efficiency by the second order Jackknife estimations.

Furthermore, it is worth comparing the species lists, which contain only the species that occurred at least in three or more samples. In local diversity analyses we need to take into account the number of samples, and in comparative studies we need to convert our dataset into standard number of samples.

II-1. Table: Measured and estimated diversity values by three perifyton datasets.

Dataset	Cumulative species number	Estimated whole species number (Sp_max)
Torna stream 2004-2005	73	94
Torna stream 2008-2009	75	104
Csigere stream 2004-2005	110	141

IV. ACKNOWLEDGMENT

We thank Gábor Kövér, student of Óbuda University, for his help in programming work and Csaba Kálmán, for the graphical design. The sampling and taxonomical research was supported by OTKA (K 75552).

V. REFERENCES

[1] Arcadis 2005. ECOSURV Felhasználói kézikönyv, Vízügyi Minisztérium

[2] Bíró, P. 2007. Az Achnanthidium minutissimum (Kützig) Czarnecki szezonális dinamikája és annak összefüggése a Torna-patak fizikai-kémiai paramétereivel. Diplomadolgozat, PE Limnológia Intézeti Tanszék könyvtára, Veszprém. 138 pp.

[3] Colwell and Coddington 1994. Estimating terrestrial biodiversity trough extrapolation. Phil. Trans. Roy. Soc. London B 345, 101–118.

[4] Estimates: http://viceroy.eeb.uconn.edu/estimates

[5] Hajnal, É. & J. Padisák, 2006. Balatoni fitoplankton adatbázis (ALMOBAL) létrehozása és alkalmazhatósága vízminőségi monitorozásra. [The ALMOBAL DataBase for water quality monitoring of Lake Balaton]. Hidrológiai Közlöny 86: 149-150. [in Hungarian with English summary]

[6] Hajnal, É. & Padisák, J. (2008): Analysis of long-term ecological status of Lake Balaton based on the ALMOBAL phytoplankton database. - Hydrobiologia 599: 227-237.

[7] Halassy B., 2002. Adatmodellezés Nemzeti Tankönyvkiadó Rt., Budapest

[8] Lakner J., Hajnal É., Lakner G., Padisák J Statistical mathematical modelling for multitude number estimation of rare and frequent species. Ecological Modeling under submission

[9] Lengyel E (2009) A köztes diszturbancia és az egyensúlyi állapot vizsgálata a Torna-patak kovaalga közösségeiben 3. szaklabor beszámoló PE Limnológia Intézeti Tanszék könyvtára, Veszprém. 138 pp.

[10] Teke, G., Lengyel, E., Bíró, R.,Stenger-Kovács Cs., Padisák, J and É., Hajnal (2011) Fajgazdagság és mintavétel összefüggésének vizsgálata a PERIDAT on-line perifiton adatbázis segítségével. Hidrológiai Közlöny in press