Sample database of the Centre for Polar Ecology - Database design and data management

Jana Kvíderová[1,2*]

[1]*Faculty of Science, University of South Bohemia, Branišovská 31,* 370 05 *České Budějovice, Czech Republic*
[2]*Institute of Botany AS CR, Dukelská 135,* 379 82 *Třeboň, Czech Republic*

Abstract

The increasing number of observations and samples led to development of systems for data storage and management. In this paper, design and experience with data management of the Sample database (SampleDTB) used in the Centre for Polar Ecology, Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic, is presented. The SampleDTB was designed for microbiological, phycological or hydrobiological data. The SampleDTB consists of data tables including defined lists of climatic zones, habitats, communities and taxons, specific queries for datasets determination and searches, forms for filling in samples and reports. The data tables contain detailed information on site, its environment, types of habitats and communities, including data on taxonomic diversity. The queries provide source data for reports or serve for searches for specific taxon, sample *etc*. Forms are used primarily for data entry or modifications. The reports provide summaries and charts for export, either for whole data set or for specific datasets. Data management resulted in system of sample numbering, site specification, and system for photographs storage. Possible future development will be focused on on-line data access, biovolume and diversity indices calculation, laboratory sample processing, and connection to culture collection database.

Key words: database, species diversity, ecology

DOI: 10.5817/CPR2014-2-14

Introduction

Since large amounts of diverse ecology data from various scientific disciplines are collected and analyzed by many institutions, the funding agencies and leading journals are becoming to require access to published data for other scientists in order to maximize data utilization (Khalsa et Yarmey 2014). Application of molecular biology methods together with informatics approaches to ecological data management and analyses has started approximately decade ago (Jones et al. 2006), and resulted in

Received October 30, 2014, *accepted* December 29, 2014.
*Corresponding author: Jana Kvíderová <jana.kviderova@ibot.cas.cz>
Acknowledgements: The study was realized within the project *Creating of Working Team and Pedagogical Conditions for Teaching and Education in the Field of Polar Ecology and Life in Extreme Environment*, reg. No. CZ.1.07/2.2.00/28.0190 co-financed by the European Social Fund and the state budget of the Czech Republic. The research was also supported by projects LM2010009 *CzechPolar - Czech polar stations: Construction and logistic expenses* (MŠMT) and a long-term research development project no. RVO 67985939 (IB).

software and database developments (Cole et al. 2009, Schloss et al. 2009). However, these data are very heterogeneous due to differences in methods used, variable spatial and temporal scales of sampling, units of measurement used, and taxonomical identification problems of genera and species observed (Jones et al. 2006). For example, Kol (1968) distinguished two different species of snow algae *Chlamydomonas nivalis* and *Scotiella nivalis* which were later identified as two different stages of one species, *Chlamydomonas nivalis* (Hoham et Mullet 1978).

In general, the databases used in ecological research could be divided into two types (Jones et al. 2006). The project-specific databases are usually based on relational database systems and are designed for specific needs (Jones et al. 2006), *e.g.* CLO-PLA database (Klimešová et De Bello 2009). Databases of culture collections could be also considered as typical project-specific ones (Watanabe et al. 1992). The second type of databases are data warehouses to which many investigators could freely contribute (Jones et al. 2006), for example GenBank for sequence databases (Benson et al. 1993) or VegBank for vegetation plot data (Peet et al. 2012).

For sample management, the project specific databases seems be more comfortable (Jones et al. 2006). In polar microbiology, an ideal sample database for should store environmental as well as microbial diversity data. These data usually include some basic information on site sampled, macro- and micro-photographs of sampled site, together with physico-chemical data measured *in situ*, *e.g.* pH or temperature. If performed, results of laboratory analyses, like nutrient concentrations should be added later concerning particular sample. Species diversity data like list of species or genera observed or pyrosequencing results should be specified for each sample. The database should provide basic data evaluation and export data for further analyses.

In this paper, the structure of the sample database used by the Centre for Polar Ecology (CPE), Faculty of Sciences, University of South Bohemia, České Budějovice, Czech Republic (SampleDTB), is described and experiences with data management are discussed. The database structure should reflect the need to characterize specific datasets (*e.g.* samples collected by students in given years) expressed as number of samples collected, number of sites visited, number of taxons (genera or species) observed, lists of sites visited and taxon observed, and finally, to provide sample catalogue and individual sample datasheets.

Material and Methods

About the database

The database is a project-specific relational database in MS Access 2013® environment for phycological, hydrobiological and microbiological types of samples. The language used is English.

Data Entry - Sampling Protocol

During field observation and measurements, data on particular samples are recorded in Sampling Protocol notebook (Fig. 1). The fields in the Sample Protocol book correspond to those in Sample Entry form in the SampleDTB (Fig. 2) in order to keep integrity of database fields. In the book, additional information could be recorded like drawings, photograph numbers or further sample processing.

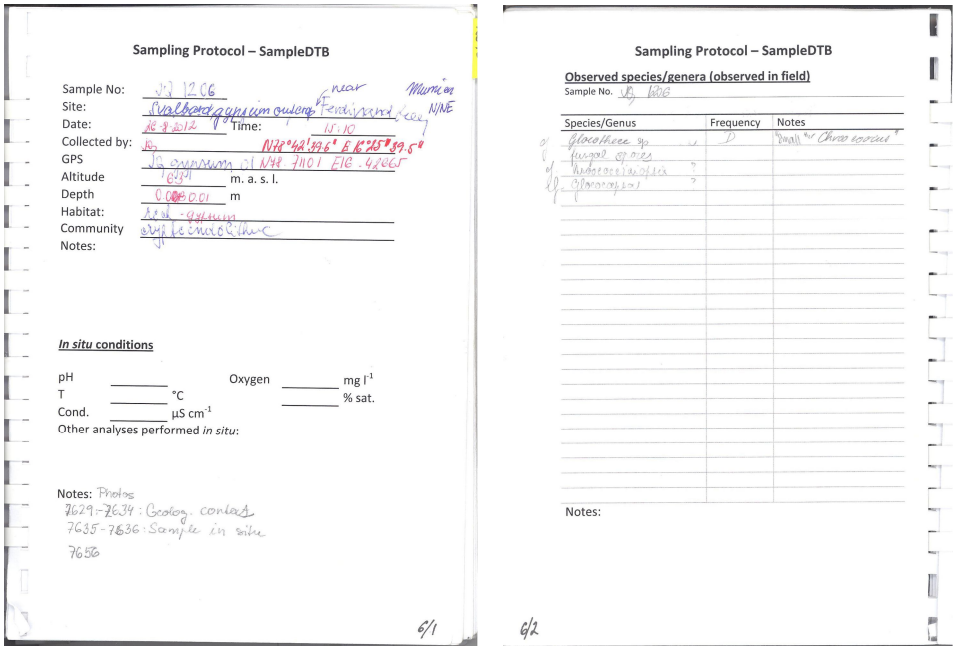


Fig. 1. The example of filled sample data in Sampling protocol notebook, in this case sample no. JQ1206, endolithic community in a gypsum outcrop at NE slope of Mumien Mt., central Svalbard. (a) site and physico-chemical data (b) species observed.

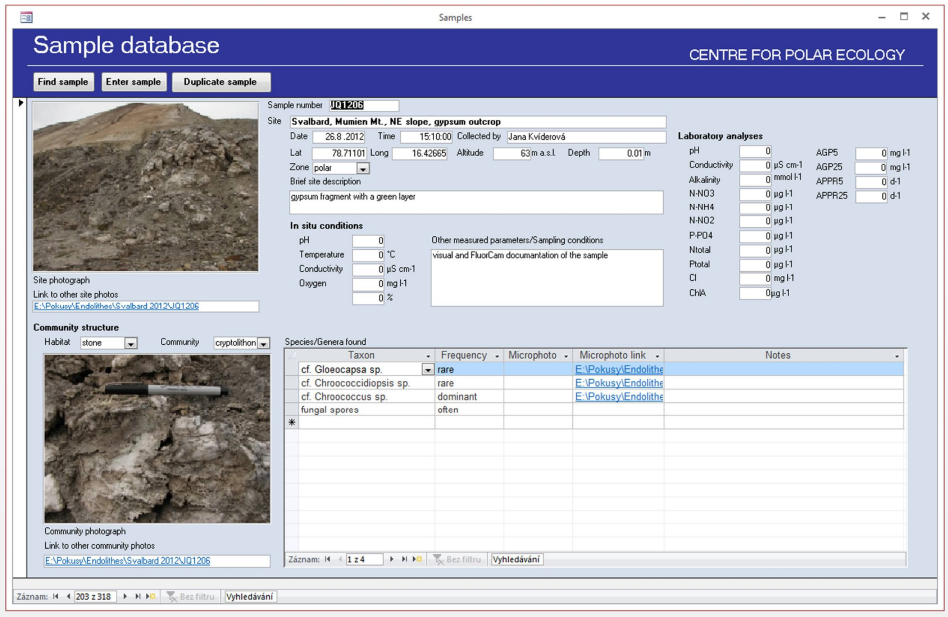


Fig. 2. The same sample as in Sampling Protocol book (Fig. 1) as filled in Sample Entry form in the SampleDTB.

Tables

Source data for the SampleDTB include following fields in the main table

- Sample number: unique alphanumeric code; mandatory
- Site data: site description, date and time of sample collection, names of collecting persons/groups, climatic zone GPS coordinates, elevation, sample depth
- Environmental data from *in situ*: pH, temperature, conductivity and oxygen concentration expressed in mg l^{-1} and % of saturation
- Habitat type: list of habitats
- Community type: list of communities
- Species composition: species/genus with relative abundances
- Laboratory analyses: pH, alkalinity, nutrient concentration (nitrogen and phosphorus, chlorine), chlorophyll a concentration, algal growth potential, algal primary production rate at 5 and 25 °C (for justification of evaluation of algal growth potential and algal primary production rate at 5 and 25 °C refer to Kvíderová et Elster (2013).
- Photographs (site, community, microphotographs) and links to photographs

Several fields are used for detailed description of the samples and analyses, *e.g.* geological setting or datalogger number positioned at the site.

Several supporting tables/lists define types of climatic zones, habitat types, community types, taxons (in SampleDTB *sensu* species or genus) with taxonomical remarks (class, order and family) and abundances on relative scale (dominant – often – rare). The relations between the tables are shown in Fig. 3.

Table Reference includes all literature used for data entry. At present, the literature includes determination keys to cyanobacteria and terrestrial algae (Ettl et Gärtner 2014, Komárek et Anagnostidis 1999, 2005, Komárek 2013).

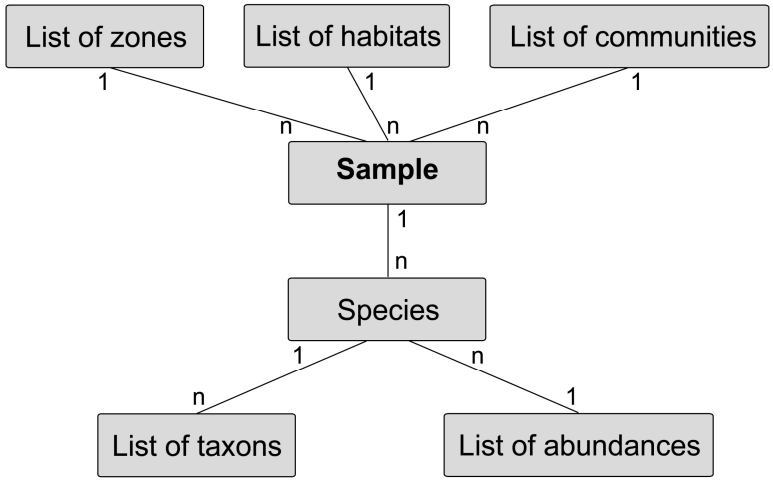


Fig. 3. The relations among the tables.

Queries
The queries are based on SQL languages. Simple queries are used for summaries or whole dataset or for specific parameter, *e.g.* collecting persons/group names, or for looking up specific samples and taxon. Queries are also used for selection of data for export to other applications like GIS or multifactorial analyses. More complicated queries can be defined by a user skilled in MS Access® or SQL language.

Forms
The forms are used for query dialogs and navigation in the SampleDTB, for data entry and data modification (Fig. 3). To facilitate the data entry from nearby sites, sample duplication was included. However, the sample number must be modified to keep uniqueness. If not, an error prompt will appear. To modify data, the sample can be accessed directly in the form. A special form is used to enter a new algal or cyanobacterial taxon (species, genera) into Taxon table. The fields concerning class, order and family are included.

Reports
Reports are based on the tables or the queries. The reports may include all data in database, or may be focused on specific region or group according to user's definition. As in queries, reports may be used for data export, especially for publications or annual reports. Reports allow creating basic charts, however, export of values is necessary for more sophisticated charts.

Database administration
In order to keep database integrity, only administrator can modify the structure. To avoid entering incomplete or false data, Administrator can enter new sample or modify individual data fields exclusively. The user can look into samples, generate reports and export the data.

Results

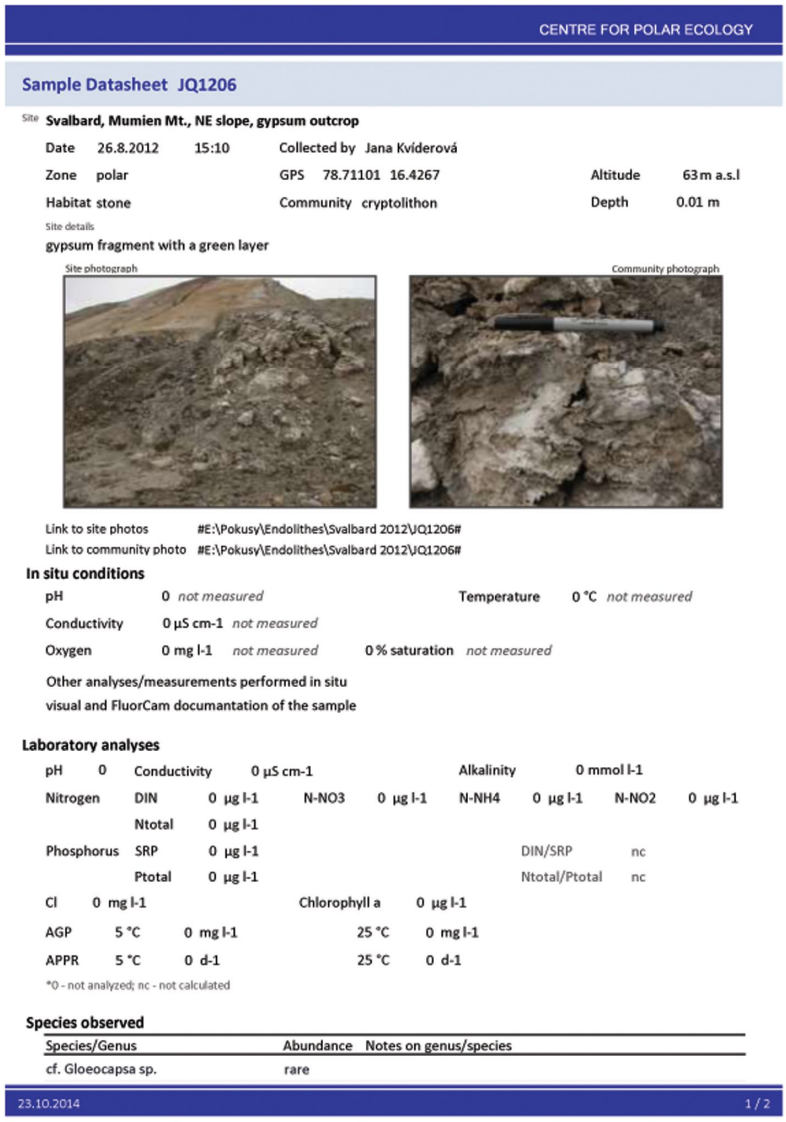
The database provides several outputs options – datasheets, summaries, lists and direct exports.

Datasheets
For each sample, a datasheet for sample catalogue is generated (Fig. 4.) and can be printed as hardcopy. Simplified datasheets without site/community photographs may be summarized in sample lists and may include list of taxons observed. In Sample Datasheets, additional fields were introduced for characterization of locality according to pH (acidic-neutral-alkaline) and algal growth potential according to Žáková (1980).

Summaries
The database provides number of samples collected, number of sites visited and number of taxons observed in complete dataset or of selected sample sets. The sample sets could be selected according to name of collector primarily, however queries for sample selections may be modified according to user's needs.

Lists
Lists are used to cataloged sites visited, samples collected or taxons observed, either in whole dataset or in selected samples.

Exports and further data evaluation
The datasheets can be exported to a text processor to become a part of an article or a report. The queries and reports can be exported to a table processor for more sophisticated calculations and analyses.



CENTRE FOR POLAR ECOLOGY

Sample Datasheet JQ1206

Site **Svalbard, Mumien Mt., NE slope, gypsum outcrop**

Date 26.8.2012 15:10 Collected by Jana Kvíderová

Zone polar GPS 78.71101 16.4267 Altitude 63 m a.s.l

Habitat stone Community cryptolithon Depth 0.01 m

Site details

gypsum fragment with a green layer

Site photograph

Community photograph

Link to site photos #E:\Pokusy\Endolithes\Svalbard 2012\JQ1206#

Link to community photo #E:\Pokusy\Endolithes\Svalbard 2012\JQ1206#

In situ conditions

pH 0 *not measured* Temperature 0 °C *not measured*

Conductivity 0 µS cm-1 *not measured*

Oxygen 0 mg l-1 *not measured* 0 % saturation *not measured*

Other analyses/measurements performed in situ

visual and FluorCam documantation of the sample

Laboratory analyses

pH 0 Conductivity 0 µS cm-1 Alkalinity 0 mmol l-1

Nitrogen DIN 0 µg l-1 N-NO3 0 µg l-1 N-NH4 0 µg l-1 N-NO2 0 µg l-1

Ntotal 0 µg l-1

Phosphorus SRP 0 µg l-1 DIN/SRP nc

Ptotal 0 µg l-1 Ntotal/Ptotal nc

Cl 0 mg l-1 Chlorophyll a 0 µg l-1

AGP 5 °C 0 mg l-1 25 °C 0 mg l-1

APPR 5 °C 0 d-1 25 °C 0 d-1

*0 - not analyzed; nc - not calculated

Species observed

Species/Genus	Abundance	Notes on genus/species
cf. Gloeocapsa sp.	rare	

23.10.2014 1 / 2

Fig. 4. The first page of Sample datasheet for the same sample as in Figs. 1 and 2. The list of species continues on the second page (not shown).

Discussion

Management of the sample database

The first task was to develop robust and logical sample ID system. At present, the Sample number consists of two to four letters indicating person(s)/groups ("call signs" in capital letter, specific for each key member of CPE and group) that collected the sample. These "call signs" are followed by two digits indicating year of sample collection and two numerical symbols indicating sample number in each year. For example, JQ1206 indicates a sample collected by Jana Kvíderová ("call sign" JQ) in 2012 (12) and the sixth sample collected in that year (06). In order to save space in Sampling Protocol notebook, samples from the same site may be distinguished by adding one alphabetical symbol after the sample number. For example, samples of cryptoendolithic community JQ1205 and JQ1205a were collected at the same stone, but the first one on August 8, 2012, and the second one on August 26, 2012.

Robust and logical system also had to be developed for site identification in order to avoid inconsistences. The template for site is hierarchical starting from large structures (*e.g.* country) to more specified levels (region, town, mountain…), for instance, Svalbard, Mumien Mt, NE slope, gypsum outcrop. Some un-official local names had to be introduced like "Oblík" (island/peninsula in front of the Nordenskiöldbreen, probably Retrettøya) or "Roklinka" (small narrow valley with high waterfall on the east bank of Petuniabukta in Wordiekammen range; between Fortet and Skottehytta) in order to facilitate communication in the field and search functions in the SampleDTB.

Photographs (macro- as well microphotographs) of a sample were stored in separate folders with the same name as a sample. The microphotographs files are named according to taxon and objective magnification used.

Since the database was also designed to track algal and cyanobacterial diversity, the taxonomical remarks (class-order-family) were introduced only for algae and cyanobacteria. For other microorganisms like fungal hyphae, "n/a" (not applicable) was added. When no (micro)organism had been observed in given sample, "none" was filled in. Adjective "uncertain" meant that further analyses are required like long-term cultivation for life cycle observation or sequencing for identification. Especially in case of green coccal algae, unidentified taxons were recorded as LGB, *i.e.* "little green balls".

During database operation, the human factor caused main inconsistences. Providing incomplete data for SampleDTB from field measurements reduced the availability data for further analyses, and hence, the quality of outputs. If the sampling site specifications did not include GPS coordinates, the data could not be used by any GIS, or later re-sampling is not possible. If some *in situ* analyses are omitted, *e.g.* physico-chemical parameters measurements in liquid samples, the environmental characteristics cannot be include into multivariate analyses and only unconstrained methods can be applied (Ter Braak et Šmilauer 2012). If only minor attention is paid to photodocumentation of site and community, it may cause problem during re-sampling of a site.

Future development

In future, the database could be improved by several ways. At first, database should become accessible on-line together with image archive. The on-line access will allow basic search within data and photographs and some basic lists will be generated, like sites visited. General remarks on individual taxon should be shown as a taxon datasheet. For un-official site names introduced during the field activities, a map including these names should be generated by some GIS software. The position of these un-official sites should be determined by GPS coordinates of a given locality. For small areas; *i.e.* sampling sites located several meters apart, mean GPS coordinates should be sufficient. Large areas, *i.e.* sampling sites located several tens to hundreds of meters apart, should be defined by several border GPS points.

At second, the database should allow selected types of calculations according to users' requirements. Considering that there is Taxon List defined, biovolume calculation seems to be easy to add. For each taxon, a specific equation for its shape is defined and the size data (diameter, length, width…) should be added to calculation protocol together with cell counts (Sun et Liu 2003). When introducing cell counts for each taxon, these data could be also used for calculation of various diversity indices for individual samples, communities, habitats or sites (Magurran 1988).

The data fields should be extended. The taxonomical diversity table of each sample should be modified to include pyrosequencing data. The sample data should include tables with description of sample manipulation in laboratory like storage information, sub-samples provided, *etc.* Additional fields should be provided for scanned images of drawings. The community selection should be dependent on habitat type, for instance there is no plankton in stone.

Finally, as algal and cyanobacterial strains are being isolated from samples, a culture collection of algal and cyanobacterial strains should be established with culture-collection specific database. A link between a strain data in Culture Collection DB and original sample in the SampleDTB could avoid duplicity of data in both databases and reduce the number of errors during data entry.

References

BENSON, D., LIPMAN, D. J. and OSTELL, J. (1993): GenBank. *Nucleic Acids Research,* 21: 2963-2965.

COLE, J. R., WANG, Q., CARDENAS, E., FISH, J., CHAI, B., FARRIS, R. J., KULAM-SYED-MOHIDEEN, A. S., MCGARRELL, D. M., MARSH, T., GARRITY, G. M. and TIEDJE, J. M. (2009): The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research,* 37: D141-D145.

ETTL, H., GÄRTNER, G. (2014): Syllabus der Boden-, Luft- und Flechtenalgen. Springer, Berlin Heidelberg, 773 p. 2nd edition.

HOHAM, R. W., MULLET, J. E. (1978): *Chloromonas nivalis* (Chod.) Hoh. & Mull. comb. spec. nov., and additional comments on the snow alga, *Scotiella*. *Phycologia,* 17: 106-107.

JONES, M. B., SCHILDHAUER, M. P., REICHMAN, O. J. and BOWERS, S. (2006): The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics,* 37: 519-544.

KHALSA, S. J. S., YARMEY, L. (2014): Data management challenges in polar ecology. *In*: Kavan, J. and Bernardová, A. (eds.): Polar Ecology Conference 2014. Faculty of Science, University of South Bohemia in České Budějovice, České Budějovice, pp. 70.

KLIMEŠOVÁ, J., DE BELLO, F. (2009): CLO - PLA: the database of clonal and bud bank traits of Central European flora. Journal of Vegetation Science, 20: 511-516.

KOL, E. (1968): Kryobiologie. Biologie und Limnologie des Schnees und Eises I. Cryovegetation. E. Schweizerbart'sche Verlagbuchhandlung, Stuttgart, 220 p.

KOMÁREK, J., ANAGNOSTIDIS, K. (1999): Süßwasserflora von Mitteleuropa 19/1. Cyanoprokaryota. 1.Teil: Chroococcales. Gustav Fischer Verlag, Jena, 548 p.

KOMÁREK, J., ANAGNOSTIDIS, K. (2005): Süßwasserflora von Mitteleuropa 19/2. Cyanoprokaryota. 2.Teil: Oscillatoriales. Elsevier/Spektrum, Heidelberg, 759 p.

KOMÁREK, J. (2013): Süßwasserflora von Mitteleuropa 19/3. Cyanoprokaryota. 3.Teil: Heterocytous genera. Springer, Heildelberg, 1131 p.

KVÍDEROVÁ, J., ELSTER, J. (2013): Standardized algal growth potential and/or algal primary production rates of maritime Antarctic stream waters (King George Island, South Shetlands). *Polar Research,* 32: 11191, 17p. DOI: 10.3402/polar.v32i0.11191.

MAGURRAN, A. E. (1988): Ecological diversity and its measurement. Springer-Science+Business Media, B.Y., 179 p.

PEET, R. K., LEE, M. T., JENNINGS, M. D. and FABER-LANGENDOEN, D. (2012): VegBank: a permanent, open-access archive for vegetation plot data. *Biodiversity & Ecology,* 4: 233-241.

SCHLOSS, P. D., WESTCOTT, S. L., RYABIN, T., HALL, J. R., HARTMANN, M., HOLLISTER, E. B., LESNIEWSKI, R. A., OAKLEY, B. B., PARKS, D. H., ROBINSON, C. J., SAHL, J. W., STRES, B., THALLINGER, G. G., VAN HORN, D. J. and WEBER, C. F. (2009): Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Applied and Environmental Microbiology,* 75: 7537-7541.

SUN, J., LIU, D. (2003): Geometric models for calculating cell biovolume and surface area for phytoplankton. *Journal of Plankton Research,* 25: 1331-1346.

TER BRAAK, C. J. F., ŠMILAUER, P. (2012): Canoco reference manual and user's guide: software for ordination, version 5.0. Microcomputer Power, Ithaca, USA, 496 p.

WATANABE, M., SHIMIZU, A. and SATAKE, K. (1992): NIES-Microbial Culture Collection at the National Institute of Environmental Studies: Cryopreservation and database of culture strains of microalgae. Proceedings of Symposium on Culture Collection of Algae. NIES, Tsukuba, Japan, pp. 33-41.

ŽÁKOVÁ, Z. (1980): Trofický potenciál a jeho aplikace ve vodním hospodářství [Trophic potential and its application in water management]. VÚV, Praha, 116 p.