Hydrobiologia (2005) 542:151–163
H. Segers & K. Martens (eds), Aquatic Biodiversity II
DOI 10.1007/s10750-004-7423-1

A large-scale stream benthic diatom database

Véronique Gosselain[1,]*, Michel Coste[2], Stéphane Campeau[3], Luc Ector[4], Claude Fauville[1], François Delmas[2], Markus Knoflacher[5], Magdalena Licursi[1], Frédéric Rimet[4], Juliette Tison[2], Loïc Tudesque[4] & Jean-Pierre Descy[1]

[1]*Facultés Universitaires N-D de la paix, Department of Biology, URBO, rue de Bruxelles, 61, 5000 Namur, Belgium*
[2]*Cemagrel, Bordeaux (France)*
[3]*Université du Québec à Trois-Rivières, Section of Geography Trois-Rivières (Quebce, Canada)*
[4]*Centre de Recherche Public Gabriel Lippmann, CREBS, Luxembourg (Luxembourg)*
[5]*Austrian Research Center Seibersdorf, Vienna (Austria)*
*(*Author for correspondence: E-mail: veronique.gosselain@fundp.ac.be)*

Key words: diatom taxonomy, environmental conditions, relational database, auto-ecology, diatom typology

Abstract

A relational database linking benthic diatom records, taxonomic nomenclature including synonyms, and corresponding environmental data has been built in MS Access. It allowed flexible and long-term use of a relatively important amount of data (~3000 records) gathered in the framework of the EC-funded PAEQANN project, gathering precise and documented information both about benthic diatoms and quantitative or semi-quantitative environmental data. Such a database has been shown to be a useful tool for the definition of benthic diatom typology at a multi-regional scale, the prediction of the impact of environmental characteristics on the structure of diatom communities, and additionally for a new insight on the auto-ecology of some taxa. This database could serve as a template for further work on diatoms and, after some implementation, on other freshwater communities. It could also be the basis for wider typology of stream diatoms, extended to other regions.

Introduction

In a context of environmental changes, there is an increasing need to organise information about biodiversity and community structure in natural or near-natural conditions, and to identify changes due to natural factors from those driven by changes from human activities. Benthic diatoms have long been recognised as excellent indicators of ecological status of water bodies (e.g. Descy, 1979; McCormick & Cairns, 1994; Prygiel et al., 1999). Therefore, diatoms have been used in water quality monitoring programs, in which there is, however, a demand from managers to simplify and reduce as much as possible identification level (e.g. Prygiel et al., 1996), to make the techniques accessible to non-specialists having received minimal, but adequate training.

Most studies on benthic diatoms have been carried out at a regional level, and only in the US variation of diatom composition along various gradients at a continental scale has been addressed (Pan et al., 1996, 2000; Potapova & Charles, 2002). Both for scientific and applied issues, structuring relevant and quantitative information about auto-ecology of diatom at a multi-regional level would certainly be valuable. This would allow, for instance, gathering in a single data matrix a large number of diatom records and corresponding environmental information from various regions, allowing statistical analysis and development of predictive models, and making information available for progress in ecological research.

Generalised databases have been developed for various purposes: paleo-environmental reconstruction (EDDI: Battarbee et al., 2000, 2001; DPDC: Sullivan & Charles, 1994), taxonomy (e.g. Kusber & Jahn, 2003; Index Nominum Algarum), collections (Alga Terra: Jahn et al., 2004; HANNA; The UCMP Collection Catalogue), images (e.g. ANSP; PID: O'Kelly & Littlejohn, 1994–2004; BGSU Algae Image Laboratory), and identification (OMNIDIA: Lecointe et al., 1993, 1999; Joynt & Wolfe, 1999).

In the framework of a European research program aiming at predicting aquatic communities in order to assess aquatic ecosystem quality and define river restoration objectives, we developed a diatom relational database at a European multi-regional scale. This database, built in MS Access, allowed flexible use and processing of a rather important amount of data (~3000 records). The choice to build a relational database instead of using spreadsheets arose when facing the two following methodological aspects, which were identified as needing particular attention: (1) the uneven quality of the environmental data within the database, and (2) the different nomenclature used for diatom data originating from different institutes and collected at different times, as well as the different taxonomical precision achieved by different institutes. General advantages of databases as compared to spreadsheets are, first, that information is partitioned over different tables, in order to be stored only once and not repeated for each record, and in a sequential format, avoiding empty cells in tables. This leads to a considerable reduction of disk space, in addition to reduction of errors. Second, queries allow an easy extraction of information, and are stored instead of the resulting tables. In addition, data can be stored along with meta-data about their origin and quality, and a link between biotic (diatom records) and environmental data can easily be established. Finally, in an Access database, all raw data can be stored together, which is not always possible in an Excel spreadsheet due to row and column number limitation. MS Access has been chosen as relatively easy to learn and use by non-IT-specialists.

The aim of this paper is to present the general structure of an Access database on stream diatoms and environmental conditions, which could serve as template for further applications in algal ecology, but also for other freshwater communities. A brief presentation of the actual dataset is given, as well as applications carried out. Limitations, and further possible and/or needed implementations and uses are discussed.

The database presented here was used as a scientific tool and it was not intended to put it on the Internet for public use. No user interface was built, neither to add nor to extract data. A lighter version was nevertheless created for use in the PAEQANN tool that was developed in the project, which is available for downloading on the Internet (http://aquaeco.ups-tlse.fr/).

Materials and methods

Benthic diatom records and corresponding environmental data under consideration have been gathered during the EC-funded PAEQANN project (EVK1-CT1999-00026). Part of the records and data were already available from previous studies carried out in several regions of Belgium, France, Luxembourg, and Austria. Another part was obtained by sampling new river sites, mostly located in regions which had not, or incompletely, been sampled in past studies. Diatom sampling, slide preparation, and counting under the microscope followed standard procedures (AFNOR, 2000; CEN 2002, 2004).

As diatom records originated from different laboratories and periods of time, an important harmonisation of the taxonomy had to be done. In addition, some slides were re-examined in order to take into account taxonomical updates, and some records counted with low level of discrimination between taxa were checked in order to distinguish morphologically close taxa with different ecology. Diatom identifications were based mainly on the Süßwasserflora von Mitteleuropa (Krammer & Lange-Bertalot, 1986, 1988, 1991a, b). Harmonisation of taxonomy and identification level was carried out at the scale of the entire database (see below).

Diatom records were all characterised by PSI index (Polluo-Sensitivity Index; Coste *in* Cemagref, 1982), as this index was used in further analyses. PSI is a water quality index, which is calculated from relative abundances of benthic diatoms collected in a given site. In the PSI system, a large number of stream diatoms have an indicator score, according to their sensitivity to pollution and ecological amplitude. PSI has been tested several times in different countries as, Finland, Germany, Poland, Portugal (Prygiel et al., 1999), and is usually considered as a reference method for water quality assessment using diatoms (Descy & Coste, 1991). It has been calculated using the OMNIDIA software (version 3.2.; Lecointe et al., 1993, 1999). When not available initially, records were re-encoded in OMNIDIA in order to generate the PSI. PSI was considered as an independent estimate of water quality, more reliable than physical and chemical water analysis, and was used to select records for defining reference conditions (see details and full discussion in Gosselain et al., in press).

Environmental data were collected at the time of sampling and/or provided by authorities in charge of monitoring and management of the sampled rivers. We estimated that the best expression of the water quality data to be considered when dealing with benthic diatom assemblages is a 3-month average of the measurements made at the sampling site. Whenever detailed data were available, environmental data were averaged over the 3-month period before sampling. However, the number of data taken into account varied greatly, depending on monitoring frequency. Due to practical issues, it was not possible however to store original values of environmental data in the database and process them through queries. Only averages were kept with information about the origin of data (see below).

The database was built using the MS Access 2000 software, after drawing a logical model.

The database structure

The database consists of 11 main tables and 13 dictionaries. Relationships have been created between and within tables to facilitate the organisation of information. In most of the tables, new IDs have been created to identify unequivocally the records. Introduction of possible duplicates was checked through appropriate queries. Referential integrity has been applied to relationships between tables to guarantee correct links between parent and child data.

The overall structure of the database, i.e. its logical model, is presented on Figure 1. Each table and each relationship have been defined in order to describe the most precisely as possible the content of the table or the nature of the relationship between two tables (or within a table), respectively (see example on Table 1). The full list of tables (entities) and fields (attributes) with their definition and description is available on the PAEQANN web site at http://aquaeco.ups-tlse.fr/Results/Data/DiatomsDatabaseAceess.htm. The main features of the diatom database are described hereafter.

Tables of the main path

The main path corresponds to a succession of tables from the river to the diatom counted, going through the site visited, the visit(s), i.e. the sampling occasions, the diatom samples taken on those occasions, the slides prepared from those samples, the diatom records obtained from microscope observation and counting, and the detail of diatom taxa counted.

Two different tables allowed defining the sampling position: SITE and STATION. The site corresponds to a certain area, homogenous for all environmental characteristics (water quality, habitat,...), which can contain several sampling points, the stations. The station is a precise sampling point, identified in a measurement network and/or precisely defined by geographic coordinates. The distinction was mainly needed as sampling points for diatoms might be slightly different than sampling points for water chemistry while both are perfectly compatible and could be used as corresponding data. In addition, a same station could be part of more than one measurement network. The station table was thus linked to the STATION_CODE table through a one-to-many relationship. Each station code ID was related to the dictionary of STATION NETWORK. Theoretically, at least some of the environmental variables of sites (see below) should have been related to the station instead, e.g. the geographical coordinates. Nevertheless, as the station concept was not taken into account when building the database and was added afterward, the rule was to describe the diatom sampling point as the 'site'. The whole

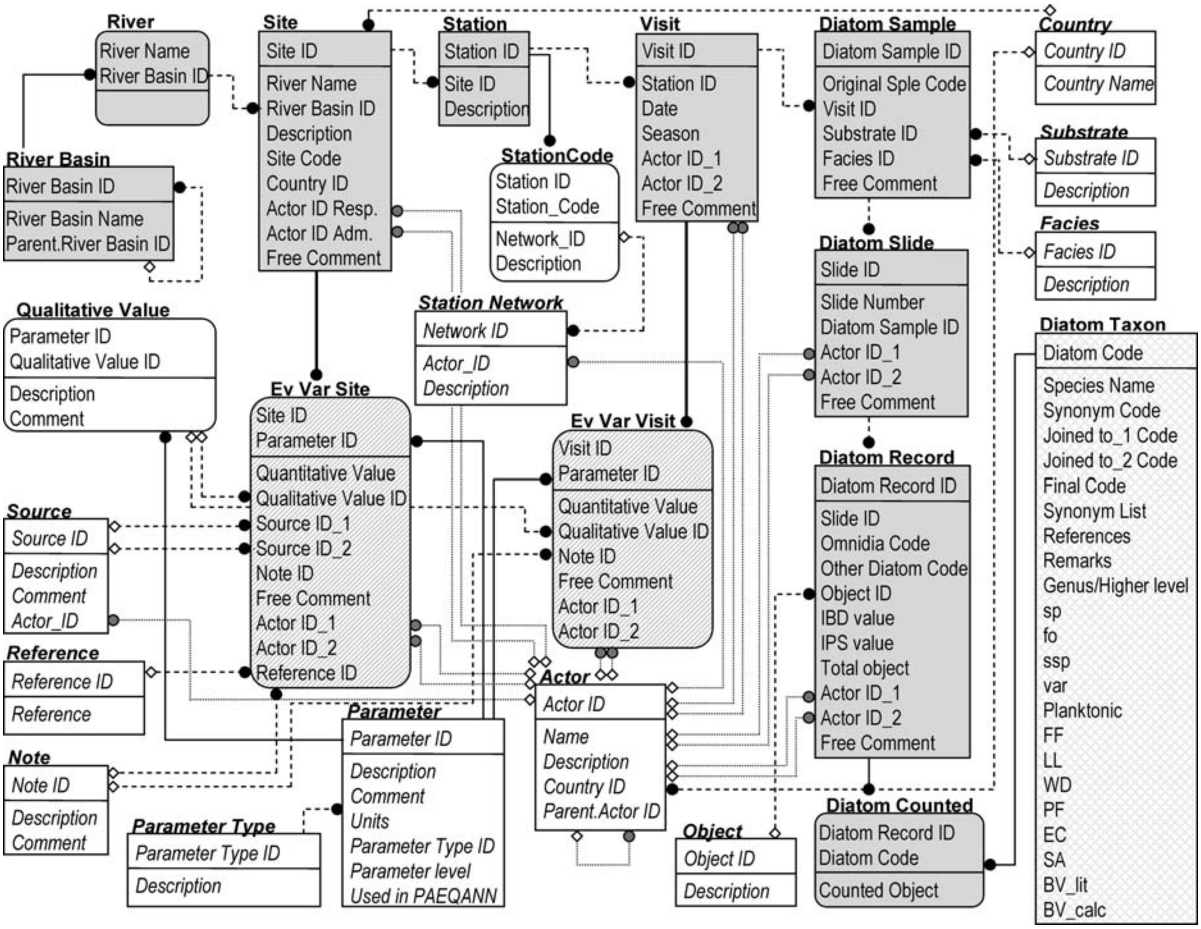


Figure 1. Logical model of the Diatom PAEQANN database. Legend: Ev_Var_Site: environmental variables related to site; Ev_Var_Visit: environmental variables related to a visit to the site; sp: indicates if the taxon is noted sp.; fo: form; ssp: sub-species; var: variety; Planktonic: indicates if the taxon is planktonic; FF: geometric shape; LL: length; WD: width; PF: depth; EC: fourth dimension; SA: surface area; BV_lit: biovolume as found in the literature; BV_calc: biovolume as calculated through the macro F_Algamica.

potential of the distinction between site and station has thus been by-passed in the present database, and could be easily implemented in further applications.

The Actor dictionary was linked to most of the tables. It contains name and contact information about all people and/or entities that were involved in a specific action in the project. This includes partners of the project, people actually involved in each step of the sample processing (visiting and collecting, mounting slides, counting diatoms), entities providing environmental data, administrative authorities from which depends a site or a sampling network.

Environmental data

Two different tables gathered the environmental data. One was linked to the site (Ev_Var_Site) and comprised environmental features of the site, i.e. characteristics that are not expected to change with time, provided that major physical changes (regulation, dredging,...) are not made. The second table was linked to the visit (Ev_Var_Visit), containing environmental data that usually vary with time. Typically, environmental data associated with the visit were water quality data. Nevertheless, as the database comprised data ranging over more than 10 years, data as

Table 1. Example of (panel a) entity and (panel b) attribute definition in the Diatom PAEQANN database
Panel a

Entity name	Entity definition
ACTOR	An ACTOR is a person or an entity involved in a specific ACTION in the project
DIATOM TAXON	The DIATOM TAXON is the dictionary of all the species encountered in the PAEQANN project, that indicates synonyms and joined species, as well as dimensions and biovolumes, as found in the literature or calculated by the macro F-Algamica
NOTE	A NOTE is an information about the origin of the value, indicating whether the value has been calculated or estimated and whether it is an a unique value or an average
PARAMETER	A PARAMETER is a physical or chemical parameter
PARAMETER TYPE	A PARAMETER TYPE classifies parameter according to the type of information it gives, i.e. General, Stream or Water characteristics
VISIT	A VISIT made to the site at a specific date. Measurements and sampling(s) for further analyses and diatom counting are carried out. The visit always corresponds to the visit for diatom sampling.

Panel b

Attribute name	Attribute definition	Attribute required	Attribute entity name
Actor ID	An ACTOR ID is a code given to a person or an entity involved in a specific action in the project	No	ACTOR
		No	DIATOM RECORD
		No	EV VAR SITE
		No	EV VAR VISIT
		No	STATION NETWORK
		No	SOURCE
	ACTOR ID of the person or entity who carried out the diatom sampling	No	VISIT
Actor ID Adm	ACTOR ID Adm identifies the administrative entity which the site belongs to	No	SITE
Actor ID Resp	ACTOR ID Resp is a the code of the person in charge of the PAEQANN sub-database comprising this site	No	SITE
Diatom Code	A DIATOM CODE is a 4 letter code given to a diatom taxon, according to the OMNIDIA species list	No	DIATOM COUNTED
		Yes	DIATOM SPECIES LIST
Diatom Record ID	A DIATOM RECORD ID is a unique number of diatom record for the entire database	Yes	DIATOM COUNTED
		Yes	DIATOM RECORD
Parameter ID	A PARAMETER ID is a string abbreviation for a parameter, if necessary informing of the unit used when several units are possible in the database	Yes	EV VAR SITE
		Yes	EV VAR VISIT
		Yes	QUALITATIVE VALUE
		Yes	VISIT
		No	PARAMETER
		Yes	EV VAR VISIT

‘presence of hydropower plant in the upstream 10 km’ were encoded in the Ev_Var_Visit table.

The Parameter dictionary listed all parameters; the parameter_level field indicated the table in which data would be recorded, Ev_Var_Site or Ev_Var_Visit. When entering new data in those tables, a control query was checking that they were entered in the right table. In addition, as environmental data could be quantitative or qualitative, the Qualitative_Value table listed all parameters and the qualitative values they can take; zero was used for quantitative or missing value. Additional information on data (origin, quality, …) were given in the fields Note, Source, Reference, and Free Comments in one or both environmental tables. The first three fields were linked to the corresponding dictionaries (see Fig. 1). Information gathered in those fields was thus clearly defined and systematic; it could be used in a query or be sorted (see example in Table 2). In particular, the Note table contained information about the origin of values, indicating whether they had been calculated or estimated and whether it was a single value or a mean. In fact, as the more relevant environmental data chosen to be used with diatom data were means over 3 months, only values as close as possible as 3-month means were entered in the database.

Diatom tables and associated dictionaries

Diatom river samples, slides and records were placed in three different tables in order to allow multiple or sub-samples in each cases. In each

Table 2. Example of environmental data from the Ev_Var_Visit table, with information about their origin and quality

VISIT_ID	NH4 (mg/l N)	Note	Free comment	Date	Actor
2542	1.040	ASY	Tilleur, Sept. 27, 1979 & Sept. 20, 1980	15-sept-79	RW
2701	3.000	ASY	Sept. 1978	15-sept-79	RW
2685	7.500	ASY	Roselies, Sept. 18, 1978 & Oct. 19, 1978	15-sept-79	RW
2541	0.710	ASY	Ombret, Sept. 27, 1979 & Sept. 20, 1980	15-sept-78	RW
2532	0.265	ASY	Dinant, Sept. 27, 1979 & Sept. 20, 1980	15-sept-79	RW
2546	1.280	ASY	Cheratte, Sept. 27, 1979 & Sept. 20, 1980	15-sept-79	RW
379	0.070	ASY	1979	17-juin-80	AERMC
922	5.462	AY		10-sept-97	AERMC
921	0.056	AY		10-sept-97	AERMC
918	0.031	A2M		25-sept-92	AERM
917	0.023	A2M		02-juil-92	AERM
916	0.290	A2M		26-sept-92	AERM
923	0.150	AY		09-sept-97	AERMC
914	0.155	A2M		25-sept-92	AERM
926	0.079	AY		09-sept-97	AERMC
913	0.125	A2M		02-juil-92	AERM
912	0.405	A2M		26-sept-92	AERM
881	0.050	M1		25-juil-01	CMGRF
1031	0.350	A3M		26-août-97	AELB
1030	0.198	A3M		06-sept-96	AELB
1029	1.167	A3M		30-août-99	AELB
2734	0.120	A4F		01-juin-99	RW
2735	0.120	A4F		15-oct-99	RW
2764	0.020	A4F		24-sept-99	RW

Note description: ASY: average on available values of the same season of other years; AY: yearly average = average on every month available for the sampling year; A2M: average on 2 months, the sampling month and the previous one; A3M: average on the 3 months before and including diatom sampling; A4F: average on 4–5 weeks, up to 7 (rarely more) before and up to 10 days after the diatom sampling date (or sometimes up to the next month), corresponding to an average on 1–5 values + value measured on the field at the occasion of the diatom sampling; M1: measured once;

Actor description: AELB: Agence de l'Eau Loire-Bretagne; AERMC: Agence de l'Eau Rhône-Méditerrannée-Corse; CMGRF: Cemagref; RW: Région Wallonne.

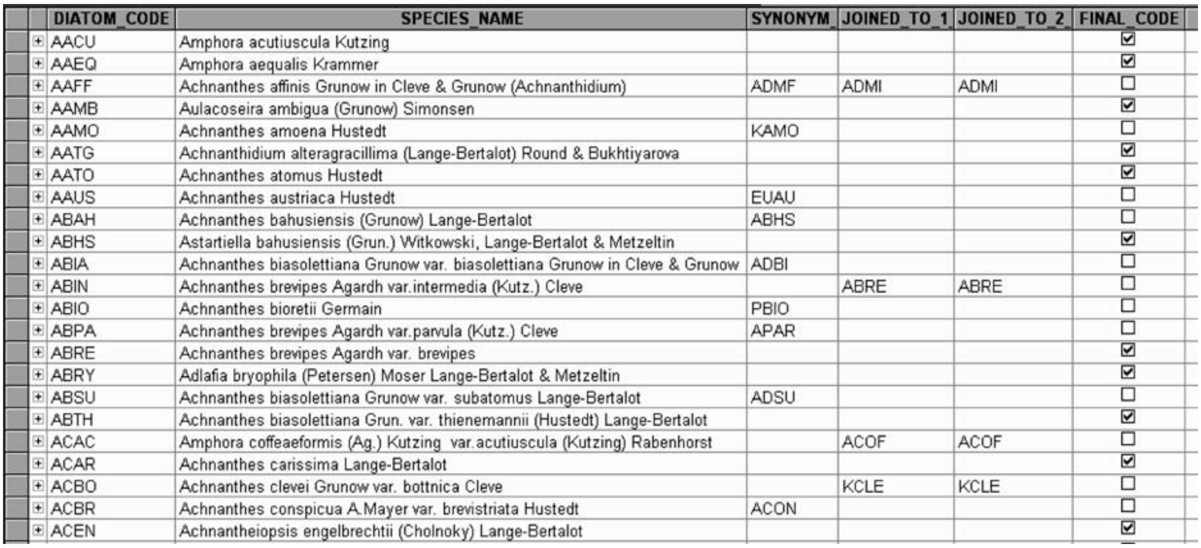
table, a field has been allocated to the original code. This would allow an easy identification of the actual sample when needed. A unique code for the database was added. In the DIATOM_SAMPLE table, information of substrate and facies were noted. In fact, samples, while mainly collected on rocks, could also have been originating from other substrates (plants, sediments,...), which could have an impact on further analysis of assemblages. In addition, lotic and/or lentic facies were sampled. Again information was kept in order to allow further selection of cases through queries. As an example, in further analysis (e.g. Gosselain et al., 2003, in press), only slides from samples taken from rocks in lotic facies were considered.

The DIATOM_RECORD table provided general information about the record. It also indicated objects actually counted: single diatom valves, entire frustules [2 valves], or indifferently single valves and frustules. Additionally, PSI and IBD (Prygiel & Coste, 1999) indexes were given, for each record and when available, respectively. Actually, a more flexible and generalised system would be to create a dictionary of indication methods to which to refer. This would prevent loosing information from the originally available diatom records, when values for other indexes were available.

The DIATOM_COUNTED table was created to solve a one-to-many relationship from both DIATOM_RECORD and DIATOM_TAXON tables (resolution table).

Diatom taxon dictionary

One of the most important and useful operations carried out when building the database was the diatom dictionary. Each taxon was entered in the table using its name as in the initial record. The codes (DIATOM_CODE field) used to identify the taxa followed the codes defined in the OMNIDIA software 3.2. When this code was used originally, it was entered as recorded initially; this led to a few cases where a single taxon was identified by two different codes. The coding system consists of four letters that indicate genus (one letter), species and varieties (the 3 last letters). At the present stage, the SPECIES_NAME field contains both the taxon name and authorities, due to encoding in the original file. A second code field (SYNONYM_CODE) indicates taxonomic transfer to the code associated with the most current taxon name. This nomenclature mainly followed recent updates of diatom taxonomy (e.g. Round et al., 1990) compiled from recent journals like Diatom Research, Diatom Monographs, or taxonomic listings (Kusber & Metzeltin, 2001; Kusber & Jahn, 2003), as provided in the OMNIDIA software 3.2. For example (Fig. 2 & Table 3), *Achnanthes biasolettiana* Grunow var. *biasolettiana* Grunow in Cleve & Grunow, which has the code ABIA, has the associated SYNONYM_CODE ADBI, indicating that *Achnantidium biasolettianum* (Grunow in Cl &



DIATOM_CODE	SPECIES_NAME	SYNONYM_	JOINED_TO_1	JOINED_TO_2	FINAL_CODE
⊞ AACU	Amphora acutiuscula Kutzing				☑
⊞ AAEQ	Amphora aequalis Krammer				☑
⊞ AAFF	Achnanthes affinis Grunow in Cleve & Grunow (Achnanthidium)	ADMF	ADMI	ADMI	□
⊞ AAMB	Aulacoseira ambigua (Grunow) Simonsen				☑
⊞ AAMO	Achnanthes amoena Hustedt	KAMO			□
⊞ AATG	Achnanthidium alteragracillima (Lange-Bertalot) Round & Bukhtiyarova				☑
⊞ AATO	Achnanthes atomus Hustedt				☑
⊞ AAUS	Achnanthes austriaca Hustedt	EUAU			□
⊞ ABAH	Achnanthes bahusiensis (Grunow) Lange-Bertalot	ABHS			□
⊞ ABHS	Astartiella bahusiensis (Grun.) Witkowski, Lange-Bertalot & Metzeltin				☑
⊞ ABIA	Achnanthes biasolettiana Grunow var. biasolettiana Grunow in Cleve & Grunow	ADBI			□
⊞ ABIN	Achnanthes brevipes Agardh var.intermedia (Kutz.) Cleve		ABRE	ABRE	□
⊞ ABIO	Achnanthes bioretii Germain	PBIO			□
⊞ ABPA	Achnanthes brevipes Agardh var.parvula (Kutz.) Cleve	APAR			□
⊞ ABRE	Achnanthes brevipes Agardh var. brevipes				☑
⊞ ABRY	Adlafia bryophila (Petersen) Moser Lange-Bertalot & Metzeltin				☑
⊞ ABSU	Achnanthes biasolettiana Grunow var. subatomus Lange-Bertalot	ADSU			□
⊞ ABTH	Achnanthes biasolettiana Grun. var. thienemannii (Hustedt) Lange-Bertalot				☑
⊞ ACAC	Amphora coffeaeformis (Ag.) Kutzing var.acutiuscula (Kutzing) Rabenhorst		ACOF	ACOF	□
⊞ ACAR	Achnanthes carissima Lange-Bertalot				☑
⊞ ACBO	Achnanthes clevei Grunow var. bottnica Cleve		KCLE	KCLE	□
⊞ ACBR	Achnanthes conspicua A.Mayer var. brevistriata Hustedt	ACON			□
⊞ ACEN	Achnantheiopsis engelbrechtii (Cholnoky) Lange-Bertalot				☑

Figure 2. Example of data records of the DIATOM_DICTIONARY table.

Table 3. Example of synonyms and joined taxa as extracted from the diatom dictionary of the Diatom PAEQANN database

F_CODE	F_SYN	DIATOM_CODE	SPECIES_NAME
ADBI	ADBI	ABIA	*Achnanthes biasolettiana* Grunow var. *biasolettiana* Grunow in Cleve & Grunow
		ADBI	*Achnanthidium biasolettianum* (Grunow in Cleve & Grunow) Round & Bukhtiyarova
	ADBT	ADBT	*Achnanthidium biasolettianum* (Grunow) Round & Bukhtiyarov fo. teratogene
ADMI	ADMF	AAFF	*Achnanthes affinis* Grunow in Cleve & Grunow
		ADMF	*Achnanthidium minutissima* (Kützing)Czarn. var. *affinis* (Grunow) Bukhtiyarova
		AMAF	*Achnanthes minutissima* Kützing var. *affinis* (Grunow) Lange-Bertalot
	ADMI	ADMI	*Achnanthidium minutissimum* (Kützing) Czarnecki
		AMIC	*Achnanthes microcephala* (Kützing) Grunow
		AMIN	*Achnanthes minutissima* Kützing var. *minutissima* Kützing
	ADMT	ADMT	*Achnanthidium minutissimum* (Kützing) Czarnecki fo.teratogene
	ADSA	ADSA	*Achnanthidium saprophila* (Kobayasi & Mayama) Round & Bukhtiyarova
		AMSA	*Achnanthes minutissima* Kützing var. *saprophila* Kobayasi & Mayama
	AMJA	AMJA	*Achnanthes minutissima* Kützing var. *jackii* (Rabenhorst) Lange-Bertalot
		AMRO	*Achnanthes minutissima* Kützing var. *robusta* Hustedt
ADMS	ADMM	ADMM	*Adlafia minuscula* var. *muralis* (Grunow) Lange-Bertalot
		NMMU	*Navicula minuscula* Grunow var. *muralis* (Grunow) Lange-Bertalot
	ADMS	ADMS	*Adlafia minuscula* (Grunow) Lange-Bertalot
		NMIS	*Navicula minuscula* Grunow in Van Heurck
	CMNO	CMNO	*Craticula minusculoides* (Hustedt) Lange-Bertalot
		NMNO	*Navicula minusculoides* Hustedt
ADSU	ADSU	ABSU	*Achnanthes biasolettiana* Grunow var. *subatomus* Lange-Bertalot
		ADSU	*Achnanthidium subatomus* (Hustedt) Lange-Bertalot

DIATOM_CODE: code given to a taxon, following the Omnidia software 3.2.; F_SYN: code associated to the "final synonym", the most current name, to which the taxon is transferred; F-CODE: "final code" to which the taxon is associated, in the framework of the PAEQANN project, in order to harmonise level of identification at the scale of the whole database.

Grun.) Round & Bukhtiyarova is the current name for the diatom in question.

Identification levels were sometimes highly different between institutes, according to the purpose of the original countings. Consequently, both taxonomic and identification levels had to be harmonised prior to analysis. When allowed by the ecology of taxa, merging of distinct taxa was proposed in order to reach a single level of identification for the entire database. On the basis of expert knowledge, two levels of merging were proposed, in the fields (1) Joined_To_1, corresponding to the harmonised level for all but the Austrian data, and (2) Joined_To_2, corresponding to the more severe and common level at the scale of the entire database. In some cases, slides were re-examined in order to refine the taxonomy when dominant taxa of different ecology had been counted together initially. That was the case to make the distinction between *Achnanthes biasolettiana* Grunow var. *biasolettiana* Grunow in Cleve & Grunow and *Achnanthes biasolettiana* Grunow var. *subatomus* Lange-Bertalot, which were counted together by one of the laboratories. In some cases, nevertheless, recounting was not considered, when a taxon was rare, or when forms possibly corresponding to distinct taxa were not unanimously recognised. This was the case for *Achnanthidium saprophila* (Kobayasi et Mayama) Round & Bukhtiyarova that was joined to *Achnanthidium minutissimum* (Kütz.) Czarnecki (Table 3). *A. saprophila*, counted by one of the laboratories, however accounted for only 1.2% of all objects counted as *A. minutissima*, in 7.5% of the corresponding records, while true *A. minutissima* accounted for 97.9% of objects, in 85.9% of corresponding records. The 0.9% counted objects remaining were either *Achnanthes minutissima* Kützing var. *jackii* or *Achnanthidium minutissima* (Kützing) Czarn. var. *affinis*; they were present in 6.6% of the records. It is to be noted that decision about taxa to be joined should be reconsidered each time new sets of data would be added in the database.

Two fields gave the list of synonyms of the taxon, and the references and date of publication, respectively. A few additional fields allowed to indicate if the taxon was (1) a genus or a higher taxonomical level, (2) noted sp., (3) a form, (4) a sub-species, (5) a variety, (6) planktonic. Finally, the closest geometrical shape of the taxon was identified and values of linear dimensions, surface area and biovolumes were provided using published size data (Krammer & Lange-Bertalot, 1986–1991; OMNIDIA software 3.2, *op. cit.*), at least for taxa to be used in further analysis. The fields SA and BV_lit gave the surface area and biovolume as given in the literature, while the BV_calc field gave the biovolume as calculated through the F_Algamica macro, following calculation provided in the counting program Algamica (Gosselain & Hamilton, 2000; http://Algamica.ibelgique.com). In fact, diatom biovolumes spanned at least three orders of magnitude and, as long recognised by planktonologists, biovolume of an algal unit is directly related to its carbon biomass, as well as its nutrient uptake and growth rates. Therefore biovolume is particularly relevant from a functional and ecological point of view. However, despite their significance, biovolume and carbon biomass have not commonly been used in studies on benthic algae (see nevertheless Ghosh & Gaur, 1998; Sabater et al., 1998; Wargo & Holt, 1998; Mayer & Galatowitsch, 2001; Peterson et al., 2001; Gosselain et al., 2003).

Actual dataset

The database presented here contains 2847 diatom records associated with corresponding environmental variables, from 1472 sites and 696 rivers, covering 118 river basin systems and 4 countries (see http://aquaeco.ups-tlse.fr/Results/Data/Diatomsmain.htm for details). It comprised 59 variables in addition to geographic coordinates, among which 23 were actually used for the benthic diatom application (Table 4; http://aquaeco.ups-tlse.fr/Results/Data/DiatomEnvVar.htm, for the complete list of variables). As some water quality data were far from the ideal 3-month averages (Table 5), information about the values was helpful for further interpretation of results. The diatom dictionary presently contains a total of 1719 different codes and names, corresponding to 1255 different taxa.

Queries were run in order to retrieve data for further analysis, in particular to put together data constituting a data matrix. For diatoms, the queries allowed to carry out a first pre-treatment of raw data: selection of records comprising enough counted objects, selection of species by rejecting too rare taxa or taxa with too low frequency of occurrence.

In order to allow averaging diatom samples but only from same substrate and facies, a new ID had to be created from queries. This ID identified cases as used in further analyses, where a single record corresponded to the mean diatom record (practically one or rarely two) from a single visit, in a single facies and on a single substrate; mean values of PSI and IBD were calculated.

Discussion and conclusion

The Diatom PAEQANN database has been shown to be useful to tackle multiple practical issues both about diatom taxonomy, and multiple origins and references of related environmental data. The database was thus the primary tool that allowed further analyses at a multi-regional scale while keeping track of all original information. This was needed in a concern of reference and quality control of the data, e.g. allowing checking

Table 4. List of main environmental variables collected

Var.	Description (units)	Basic statistics/categories								
		Min.	Max.	Median	Mean	75%ClUp	75% ClLo	SD	Var	*n*
Quantitative variables										
ALT	Altitude (m)	1	2660	203	257	265	248	276	76152	1472
SLOPE	Slope (m km^{-1})	0.0	133.3	1.6	4.23	4.51	3.95	9.25	85.49	1472
DIST	Distance from source (km^{-1})	0.0	964.42	29.80	66.97	70.26	63.68	109.66	12025.80	1471
CAreaS	Catchment surface area up to the site (km^2)	0.0	115413	241	2299	2558.04	2039.98	8633.0	7.4529 10^7	1471
ALK_meq	Alkalinity (meq l^{-1})	0.03	12.84	2.20	2.65	2.70	2.61	1.92	3.69	2612
pH	Water pH	3.8	10.04	7.76	7.72	7.73	7.71	0.50	0.253	2755
COND_20	Conductivity at 20 °C (μS cm^{-1})	7.65	24500	383.33	493.50	513.66	473.33	919.88	846171	2755
TEMP	Water temperature (°C)	2.3	27.9	16.0	15.9	16.0	15.8	4.2	17.64	2755
DO	Dissolved oxygen (mg l^{-1})	0.10	26.45	9.35	9.14	9.19	9.09	2.26	5.12	2749
DOC	Dissolved organic carbon (mg $^{-1}$)	0	153.75	2.90	3.83	3.95	3.71	5.24	27.42	2701
NO_3	Nitrate (mg NO_{3^-}-N l^{-1})	0	37.00	2.50	3.36	3.43	3.28	3.31	10.97	2742
NO_2	Nitrite (mg NO_{2^-}-N l^{-1})	0	3.028	0.03	0.14	0.10	0.09	0.20	0.04	2740
NH_4	Ammonium (mg NH_{4^+}-N l^{-1})	0	35.93	0.08	0.65	0.74	0.56	2.38	5.66	2744
PO_4	Phosphate (mg PO_4^{3-}-P l^{-1})	0	14.03	0.07	0.26	0.29	0.23	0.78	0.61	2738
Semi-qualitative or qualitative variables		Categories								
Season	Season	SP = spring, SA = autumn, SW = winter; coded as 2 dummy variables								
Geol	Geology	‘mudstone’, ‘limestone’, ‘sandstone’, ‘granitic’, ‘quaternary’, ‘mixed and other’; coded as 5 dummy variables								
Morph	River morphology	1 = natural, 2 = partly channelized, 3 = totally channelized								
Level	Water level	1 = lowest water levels, 2 = mid levels, 3 = flood levels								
Shad	Shading at the sampling site	1 = closed, 2 = mid, 3 = opened								
Hydropwr	Hydropower installation within 10 km upstream the sampling site	Yes or no								
RedFlow	Reduction of flow installation within 10 km upstream the sampling site	Yes or no								
Vel	Water velocity	1: < 0.2 m s^{-1}, 2: 0.2–0.5 ms^{-1}, 3: > 0.5 m s^{-1}								

Note: The zero (0) value for minima could either be actual zero or mean ‘below detection limit.’

Table 5. Summary of data available for water quality in the PAEQANN database, according to the NOTE given to the data

Average on 3 months	731	27%
Other average	1186.4	43%
Single value	562.56	21%
Estimated value	256.67	9%
Without note	14	1%
	2741.3	100%

Numbers given here are mean on all parameters.

outliers data in analysis (Table 5). However, due to practical issues, it was not possible, at this stage, to gather and process through queries original values of water quality data. This is now arising as the main weakness of the actual dataset.

The long-term use of the database has been guaranteed by some choices about its structure. The reference to dictionaries of parameters instead of limited lists of parameters included into environmental data tables, while requiring more complex queries to retrieve data, allows the introduction of values for new parameters in the future. In fact, they can be added without limitation in corresponding dictionaries. Information has been split into numerous tables in order to anticipate as most as possible the different concrete cases that could arise (duplicate samples, slides, counts, …). In order to avoid the creation of a new ID for cases through queries (see above), a table should be included between the Visit and Diatom_Sample tables, to first define the characteristics of the sample, in term of facies and substrate.

Further work on this database should deal with the development of an easy update procedure of the diatom dictionary and a more flexible taxonomic system. In addition, taxon names and authorities should be split in two different fields. Finally, the distinction between site and station, and related environmental data, should be fully implemented.

Data extracted from the database, analysed through artificial neural networks, allowed the definition of a typology of benthic diatom for near-natural conditions at a European multi-regional scale (Gosselain et al., in press), and analysis of diatom records originating from both undisturbed and disturbed conditions, providing a fresh insight about the changes of diatom assemblages along disturbed ecological gradients (unpublished). One of the objectives was to design a tool for prediction for water quality management (http://aquaeco.ups-tlse.fr/). Those analyses also offered new insight on the auto-ecology of some diatom taxa (Gosselain et al., 2003, in press). Establishing a correspondence between biotic and environmental data also allowed prediction of diatom assemblages from environmental conditions as well as the identification of main environmental conditions driving the occurrence of specific biotypes (Gosselain et al., in press). Similar application was also carried out at a regional level comprising relatively diverse environmental conditions but few cases for which both diatoms and corresponding environmental records were available (<100). This was possible due to the existence of the multi-regional database, providing extra cases for similar environmental conditions (Darchambeau et al., submitted).

In auto-ecological studies, taxonomical revisions are a common problem hampering the use of ‘old’ ecological and ecophysiological data. Therefore, a database designed for storing ecological records has to include precise and harmonized taxonomy and possibilities for updating, along with data on environmental conditions (Gosselain et al., in press; Darchambeau et al., submitted). At a time of high concern about assessment of ecological status of surface water bodies and identification of reference conditions for the various freshwater biota (Wallin et al., 2003), the development of databases gathering precise and documented information about aquatic communities, and corresponding high quality environmental data, becomes of prime interest. We suggest that scientists involved in ecology of freshwater communities should pay more attention to such problems, in order to save relevant ecological information in well-structured databases.

Acknowledgements

This work has been carried out in the framework of the PAEQANN project [standing for ‘Predicting Aquatic Ecosystem Quality using Artificial Neural Networks: impact of environmental characteristics on the structure of aquatic communities (algae, benthic and fish fauna)’], supported by the European Commission under the Fifth Framework Programme and contributing to the implementation of the Key Action ‘Sustainable Management and Quality of Water’ within the Energy, Environment and Sustainable Development [Contract no: EVK1-CT1999-00026]. The authors wish to thank French Water Agencies and Water Authorities, the Ministry of Environment of Luxembourg, the Ministry of the Wallonia Region (Belgium), and the Upper Austrian Water Authority, for providing physical and chemical data, diatom data, and for supporting a great number of ecological and monitoring studies that allowed gathering this enormous amount of data. VG also warmly thanks A. Audet and K. Teymouri, who introduced her to the world of databases and accompanied her first steps in the PAEQANN database. The authors finally thank three anonymous reviewers for their comments and suggestions that helped improving an earlier version of the manuscript.

References

AFNOR, 2000. Qualité de l'Eau. Détermination de l'indice biologique diatomées (IBD) – Norme NF T90-354, 63 pp.

ANSP Algae Image Database from the Phycology Section, Patrick Center for Environmental Research, The Academy of Natural Sciences at http://diatom.acnatsci.org/AlgaeImage/

BGSU Algae Image Laboratory: http://www.bgsu.edu/departments/biology/facilities/ algae/html/Image_Archive.html

Battarbee, R. W., S. Juggins, F. Gasse, N. J. Anderson, H. Bennion & N. G. Cameron, 2000. European Diatom Database (EDDI). An Information System for Palaeoenvironmental Reconstruction. European Climate Science Conference, Vienna City Hall, Vienna, Austria, 19–23 October, 1998: 1–10.

Battarbee, R. W., S. Juggins, F. Gasse, N. J. Anderson, H. Bennion, N. G. Cameron, D. B. Ryves, C. Pailles, F. Chali & N. Telford, 2001. European Diatom Database (EDDI). An Information System for Palaeoenvironmental Reconstruction. ECRC Research Report, 81, 210 pp.

Cemagref, 1982. Etude des méthodes biologiques d'appréciation quantitative de la qualité des eaux. Rapport Q.E. Lyon A.F. Bassin Rhône-Méditérannée-Corse, 218 pp.

CEN, 2002. Water quality – Guidance standard for the routine sampling and pretreatment of benthic diatoms from rivers, prEN13946, Final draft, 14 pp.

CEN, 2004. Water quality – Guidance standard for the identification, enumeration and interpretation of benthic diatom samples from running waters, EN14407: 2004, Final draft.

Darchambeau, F., V. Gosselain, C. Fauville & J.-P. Descy, in prep. Definition of regional reference conditions for diatoms based on a multi-regional typology. To be submitted to Freshwater Biology.

Descy, J.-P., 1979. A new approach to water quality estimation using diatoms. Nova Hedwigia, 64: 305–323.

Descy, J.-P. & M. Coste, 1991. A test of methods for assessing water quality based on diatoms. Verhandlungen der Internationalen Vereinigung für theoretische und angewandte Limnologie, 24: 2112–2116.

Ghosh, M. & J. P. Gaur, 1998. Current velocity and the establishment of stream algal periphyton communities. Aquatic Botany 60: 1–10.

Gosselain, V. & P. Hamilton, 2000. Algamica: revisions to a key-based computerized counting program for free-living, attached, and benthic algae. Hydrobiologia 438: 139–142.

Gosselain, V., C. Fauville, S. Campeau, M. Gevrey & J.-P. Descy, 2003. Typology and prediction of diatom assemblages in rivers: building of database and first predictive model. In Symoens, J.-J. & K. Wouters (eds), Biological Evaluation and Monitoring of Surface Water Quality. National Committee of Biological Sciences and National Committee SCOPE, Brussels: 45–57.

Gosselain, V., S. Campeau, M. Gevrey, M. Coste, L. Ector, Y. S. Park, S. Lek & J.-P. Descy. Diatom typology of reference situations at a large multi-regional scale: combined results of multivariate analysis and SOM. In Lek, S., M. Scardi, P. Verdonschot, Y. S. Park & J.-P. Descy (eds), Modelling Community Structure in Freshwater Ecosystems, Springer-Verlag, in press.

HANNA Database, California Academy of Sciences Diatom Collection, http://www.calacademy.org/research/diatoms/index.html#collection

Index Nominum Algarum, University Herbarium, University of California, Berkeley. Compiled by Paul Silva. Available online at http://ucjeps.berkeley.edu/INA.html

Jahn, R., W.-H. Kusber, L. K. Medlin, R. M. Crawford, D. Lazarus, T. Friedl, D. Hepperle, B. Beszteri, K. Hamann, F. Hinz, S. Strieben, V., Huck, J., Kasten, A. Jobst & K. Glück, 2004. Taxonomic, molecular and ecological information on diatoms: the information system AlgaTerra. In Poulin, M. (ed.), Seventeenth International Diatom Symposium 2002. Biopress, Bristol: 121–128. [AlgaTerra Homepage at www.algaterra.net]

Joynt, E. H. III & A. P. Wolfe, 1999. An image database for diatom identification and nomenclature. Journal of Paleolimnology 22: 109–114.

Kusber, W.-H. & D. Metzeltin, 2001. Checklist of new diatom combinations, replaced names and validations published by Horst Lange-Bertalot until the year 2000 and additional validations. In Jahn, R., J. P. Kociolek, A. Witkowski & P. Compère (eds), Studies on Diatoms, Lange-Bertalot – Festschrift, Gantner, Ruggell: 585–633.

Kusber, W.-H. & R. Jahn, 2003. Annotated list of diatom names by Horst Lange-Bertalot and co-workers – Version 3.0. [http://www.algaterra.org/Names_Version3_0.pdf] pdf-file, Version 3.0, (23 June 2003).

Krammer, K. & H. Lange-Bertalot, 1986. Bacillariophyceae. 1. Teil: Naviculaceae. In Ettl, H., J. Gerloff, H. Heynig & D. Mollenhauer (eds), Süsswasserflora von Mitteleuropa, Band 2/1. Gustav Fischer Verlag, Stuttgart, New York: 876 pp.

Krammer, K. & H. Lange-Bertalot, 1988. Bacillariophyceae. 2. Teil: Bacillariaceae, Epithemiaceae, Surirellaceae. In Ettl, H., J. Gerloff, H. Heynig & D. Mollenhauer (eds), Süsswasserflora von Mitteleuropa, Band 2/2. VEB Gustav Fischer Verlag, Jena: 596 pp.

Krammer, K. & H. Lange-Bertalot, 1991a. Bacillariophyceae. 3. Teil: Centrales, Fragilariaceae, Eunotiaceae. In Ettl, H., J. Gerloff, H. Heynig & D. Mollenhauer (eds), Süsswasserflora von Mitteleuropa, Band 2/3. Gustav Fischer Verlag, Stuttgart, Jena: 576 pp.

Krammer, K. & H. Lange-Bertalot, 1991b. Bacillariophyceae. 4. Teil: Achnanthaceae, Kritische Ergänzungen zu Navicula (Lineolatae) und Gomphonema, Gesamtliteraturverzeichnis Teil 1-4. In Ettl, H., G. Gärtner, J. Gerloff, H. Heynig & D. Mollenhauer (eds), Süsswasserflora von Mitteleuropa, Band 2/4. Gustav Fischer Verlag, Stuttgart, Jena: 437 pp.

Lecointe, C., M. Coste & J. Prygiel, 1993. 'OMNIDIA' software for taxonomy, calculation of diatom indices and inventories management. Hydrobiologia 269/270: 509–513.

Lecointe, C., M. Coste, J. Prygiel & L. Ector, 1999. Le logiciel OMNIDIA version 2, une puissante base de données pour les inventaires de diatomées et pour le calcul des indices diatomiques européens. Cryptogamie Algolologie 20: 132–134.

Mayer, P. M. & S. M. Galatowitsch, 2001. Assessing ecosystem integrity of restored prairie wetlands from species production–diversity relationships. Hydrobiologia 443: 177–185.

McCormick, P. V. & J. Cairns, 1994. Algal as indicators of environmental change. Journal of Applied Phycology 6: 509–526.

O'Kelly, C. J. & T. Littlejohn, 1994–2004. PID: Protist Image Database. Distribution: http://megasun.bch. umontreal.ca/protists/protists.html

Pan, Y., R. J. Stevenson, B. H. Hill, A. T. Herlihy & G. Collins, 1996. Using diatoms as indicators of ecological conditions in lotic systems: a regional assessment. Journal of North American Benthological Society 15: 481–495.

Pan, Y., R. J. Stevenson, B. H. Hill & A. T. Herlihy, 2000. Ecoregions and benthic diatom assemblages in Mid-Atlantic Highlands streams, USA. Journal of North American Benthological Society 19: 518–540.

Peterson, C. G., H. M. Valett & C. N. Dahm, 2001. Shifts in habitat templates for lotic microalgae linked to interannual variation in snowmelt intensity. Limnology and Oceanography 46 : 858–870.

Potapova, M. G. & D. F. Charles, 2002. Benthic diatoms in USA rivers: distributions along spatial and environmental gradients. Journal of Biogeography 29: 167–187.

Prygiel, J. & M. Coste, 1999. Progress in the use of diatoms for monitoring rivers in France. In Prygiel, J., B. A. Whitton & J. Bukowska (eds), Use of Algae for Monitoring Rivers III. Agence de l'Eau Artois-Picardie, Douai: 165–179.

Prygiel, J., L. Leveque & R. Iserentant, 1996. L'IDP: Un nouvel Indice Diatomique Pratique pour l'évaluation de la qualité des eaux en réseau de surveillance. Revue des Sciences de l'Eau, 9 : 97–113.

Prygiel, J., B. A. Whitton & J. Bukowska, 1999. Use of Algae for Monitoring Rivers III. Agence de l'Eau Artois-Picardie, Douai.

Round, F. E., R. M. Crawford & D. G. Mann, 1990. The Diatoms. Biology and Morphology of the Genera. Cambridge University Press, Cambridge, 747 pp.

Sabater, S., S. V. Gregory & J. R. Sedell, 1998. Community dynamics and metabolism of benthic algae colonizing wood and rock substrata in a forest stream. Journal of Phycology 34: 561–567.

Sullivan, T. J. & D. F. Charles, 1994. The feasibility and utility of a paleolimnology/ paleoclimate data cooperative for North America. Journal of Paleolimnology 10: 265–273.

The UCMP Collection Catalogue : http://www.ucmp.berkeley.edu/collections/micro.html

Wallin, M., T. Wiederholm & R. K. Johnson, 2003. Guidance on establishing reference conditions and ecological status class boundaries for inland surface waters, final draft, version 7.0, produced by CIS working group 2.3. – REFCOND, 5 March 2003. 93 pp.

Wargo, M. J. & J. R. Holt, 1998. Determination of stream reaches in a ridge and valley creek using diatom periphyton communities. Journal of Freshwater Ecology 13: 447–456.