REBECCA databases: experiences from compilation and analyses of monitoring data from 5,000 lakes in 20 European countries

S. Jannicke Moe · Bernard Dudley · Robert Ptacnik

Published online: 22 April 2008

Abstract Chemical and biological data from more than 5,000 lakes in 20 European countries have been compiled into databases within the EU project REBECCA. The project's purpose was to provide scientific support for implementation of the EU Water Framework Directive (WFD). The databases contain the biological elements phytoplankton, macrophytes, macroinvertebrates and fish, together with relevant chemistry data and station information. The common database strategy has enabled project partners to perform analyses of chemical–biological relationships and to describe reference conditions for large geographic regions in Europe. This strategy has obvious benefits compared with single-country analyses: results will be more representative for larger European regions, and the statistical power and precision will be larger. The high number of samples within some regions has also enabled analysis of type-specific relationships for several lake types. These results are essential for the intercalibration of ecological assessment systems for lakes, as required by the WFD. However, the common database approach has also involved costs and limitations. The data process has been resource-demanding, and the requirements for a flexible database structure have made it less user-friendly for project partners. Moreover, there are considerable heterogeneities among datasets from different countries regarding sampling methods and taxonomic precision; this may reduce comparability of the data and increase the uncertainty of the results. This article gives an overview of the contents and functions of the REBECCA Lakes databases, and of our experiences from constructing and using the databases. We conclude with recommendations for compilation of environmental data for future international projects.

Keywords Database management · EU Water Framework Directive · Macroinvertebrates · Macrophytes · Monitoring data · Phytoplankton

S. J. Moe (✉) · R. Ptacnik
Norwegian Institute for Water Research (NIVA),
Gaustadalléen 21, 0349 Oslo, Norway
e-mail: jannicke.moe@niva.no

R. Ptacnik
e-mail: robert.ptacnik@niva.no

B. Dudley
Centre for Ecology and Hydrology (CEH), Bush Estate,
Penicuik, Midlothian EH26 0QB, UK
e-mail: bedu@ceh.ac.uk

Abbreviations

BQE	Biological quality element
GIG	Geographical Intercalibration Group
IC	Intercalibration
REBECCA	RElationships Between Ecological and Chemical stAtus in surface waters
TOC	Total organic carbon
WFD	Water Framework Directive

Introduction

The Water Framework Directive (WFD) requires that European water bodies are classified according to their ecological status. Ecological classification systems for rivers are already proposed or in use by EU countries, based on, for example, macroinvertebrates (Hering et al. 2004; Verdonschot and Moog 2006) and fish (Degerman et al. 2007; Pont et al. 2007). For lakes, however, ecological classification systems are less developed. An important task for the EU-funded project REBECCA (http://www.environment.fi/syke/rebecca) was therefore to analyse relationships between chemical pressures and ecological responses in lakes. The aim of this project was to provide scientific support for the development of new ecological classification systems and for validation of existing systems. For this purpose, we have collated available monitoring data from all projects partners, as well as from external data providers. The data from all countries have been compiled into common databases for each major taxonomic group: phytoplankton, macrophytes, macroinvertebrates and fish. These taxonomic groups will be referred to as "biological quality elements" (BQEs), as defined by the WFD. Altogether there are more than 30,000 samples of these biological elements, representing more than 5,000 lakes in 20 countries (Table 1). Most of the biological samples are identified to species level. In addition there are >80,000 chlorophyll *a* samples, representing total phytoplankton biomass. Most of the samples are from the period between 1988 and 2003.

An important motivation for developing the common databases was that the larger datasets would enable us to analyse pressure–response relationships for different lake types separately. A set of lake types based on geological and chemical properties (see Table 2) has been defined for five groups of countries within Europe (Geographical Intercalibration Groups; GIGs) by the pan-European WFD Common Implementation Strategy (European Commission 2003). These lake types are expected to have specific ecological reference conditions (i.e. community composition in non-disturbed conditions) and specific ecological responses to pressures. This lake typology will not serve as an optimal categorisation for all biological elements and for all pressure types (see e.g. Verdonschot 2006b), but we expect that type-specific analyses will at least reduce some of the unexplained variation in the ecological responses. In REBECCA, we have been able to characterise ecological relationships for all lake types separately: chlorophyll (Carvalho et al. 2008; Ptacnik et al. 2008a, b), phytoplankton (Ptacnik et al. 2008b); or for combined groups of lake types: macrophytes (Penning et al. 2008a, b) and macroinvertebrates (Schartau et al. 2008).

Another motivation for the common database approach was to assist the Intercalibration process, i.e. the intercalibration of class boundaries of existing national ecological classification systems within the GIGs (European Commission 2005). Datasets compiled within the GIGs were provided to the REBECCA databases, and results of data analyses (or data tables formatted for analysis) were returned to the GIGs. This collaboration gave synergies for both parties: the REBECCA project obtained a considerably larger empirical foundation for characterisation of pressure–response relationships, and the GIGs obtained more precise results for the intercalibration of their classification systems (Lyche Solheim et al. 2008).

A database strategy was not actually planned from the beginning of the project, nor was a trained database manager involved, but the need for proper relational databases became obvious after the start of project. Two main factors necessitated the databases: the amount of data that we received was far greater than expected, and the data formats were more heterogeneous than foreseen. An explanation for this development is that the external interest in the REBECCA databases grew throughout the project: as preliminary results from the project were presented in meetings, we were offered more data, both from REBECCA partners and from institutions that were not project partners. In particular, the Intercalibration project contributed a substantial amount of phytoplankton data from large parts of Europe. Although a template was developed for data submission by the project partners, we eventually decided to accept data in any format from external data providers (and to some degree from partners). Therefore, the earlier versions of the databases had to be modified in order to accommodate this variety of data formats.

The strategy employed for REBECCA Lakes of holding data in common databases had obvious benefits compared with single-country analyses, but it also involved considerable efforts and challenges.

Table 1 Overview of contents of the REBECCA Lakes databases per country: number of lakes and samples for chlorophyll (as a proxy for phytoplankton abundance), and number of lakes, samples and taxa per biological quality element

Country	Chlorophyll		Phytoplankton			Macrophytes			Macroinvertebrates			Fish		
	No. of lakes	No. of samples	No. of lakes	No. of samples	No. of taxa	No. of lakes	No. of samples	No. of taxa	No. of lakes	No. of samples	No. of taxa	No. of lakes	No. of samples	No. of taxa
AT (Austria)	3	3	–	–	–	–	–	–	–	–	–	–	–	–
BE (Belgium)	2	4	–	–	–	2	22	18	–	–	–	–	–	–
CY (Cyprus)	3	12	3	12	2	–	–	–	–	–	–	–	–	–
DE (Germany)	197	800	179	704	662	–	–	–	–	–	–	–	–	–
DK (Denmark)	21	108	9	137	196	–	–	–	–	–	–	66	66	36
EE (Estland)	12	12	–	–	–	4	114	34	–	–	–	–	–	–
ES (Spain)	22	173	21	90	105	–	–	–	–	–	–	–	–	–
FI (Finland)	3,281	39,638	162	757	603	527	6,643	73	152	397	184	104	104	24
FR (France)	3	3	–	–	–	–	–	–	–	–	–	–	–	–
HU (Hungary)	7	268	–	–	–	–	–	–	–	–	–	–	–	–
IE (Ireland)	388	1,697	–	–	–	117	601	40	220	769	269	–	–	–
IT (Italy)	106	1,957	42	570	416	–	–	–	13	68	90	6	21	38
LT (Lithuania)	13	39	–	–	–	5	114	39	–	–	–	–	–	–
LV (Latvia)	180	231	182	321	256	144	1,047	44	–	–	–	–	–	–
NL (The Netherlands)	312	25,648	129	1,701	863	46	344	45	–	–	–	–	–	–
NO (Norway)	472	3,060	513	3,401	679	269	2,607	83	56	513	152	–	–	–
PL (Poland)	69	347	6	97	271	5	72	25	–	–	–	–	–	–
RO (Romania)	27	460	15	184	271	19	211	33	18	514	164	60	60	30
SE (Sweden)	132	1,540	98	520	407	254	3,699	88	674	1,185	395	2,115	5,044	38
UK (United Kingdom)	242	4,978	91	232	285	50	548	66	33	355	348	–	–	–
Sum	5,492	80,978	1,450	8,726	–	1,442	16,022	–	1,166	3,801	–	2,304	5,248	–

This article does not attempt to describe the optimal way of constructing and operating ecological databases, but rather to share the experiences of researchers who were faced with the challenge of handling vast amounts of ecological data. Thus, the aims of this article are:

1. To give an overview of the contents of the REBECCA Lakes biological databases (Phytoplankton, Macrophytes, Macroinvertebrates and Fish). The purpose is both to provide more background information for the results presented in the other REBECCA articles of this special issue, and to inform about the availability of these data for future projects.
2. To share our experiences from the database processes, from data submission through standardisation to extraction for data analysis.
3. To discuss the cost and benefits with of the common database approach, and give recommendations for compilation of ecological monitoring data for future projects.

We believe it is likely that other European environmental research projects will run into similar problems, and we hope that our experiences regarding data compilation can become useful. Our experiences should be especially relevant for projects addressing the WFD and assessment of ecological status. Challenges regarding the analysis and interpretation of these large datasets will be addressed in the subsequent REBECCA articles in this special issue.

REBECCA database contents

Abundance data were collated for the biological quality elements phytoplankton, macrophytes, macroinvertebrates and fish, together with accompanying chemistry data and geo-referenced station information. Data on phytoplankton, macroinvertebrates and fish were compiled and managed at the Norwegian Institute for Water Research (NIVA). Macrophyte data were compiled and managed at the Centre for Ecology and Hydrology (CEH, UK). In addition, chlorophyll was used as a proxy for phytoplankton biomass, because the number of chlorophyll observations is almost an order of magnitude higher than the number of phytoplankton abundance observations (Table 1). The number of lakes with observations of the different BQEs is shown per country in Table 1, whereas Table 2 shows the number of different lake types per GIG. Reference lakes are by definition those with insignificant anthropogenic pressures, and the reference status of lakes is assigned by the data providers. Lakes belonging to the same lake type are assumed to have similar ecological reference conditions. The composition of lakes is further characterised by the range of each typology factor for each database (Table 3). A similar overview of the chemical determinands associated with three of the biological databases is given in the appendix (Table A1).

Examples of multi-national ecological databases that have been compiled within other EU projects are given below. Most of these databases contain data from rivers or coastal zones, and they usually focus on one main taxonomic group only. To our knowledge, the REBECCA Lakes databases are currently the most extensive databases for biological data from lakes that are compiled at EU level.

- Data from eight river basins across Europe were collected within the EU FP5 project HarmoniRiB (http://workplace.wur.nl/harmonirib). Although some biological data are available, these databases contain mostly physical, chemical and hydrological data (see e.g. Refsgaard et al. 2007). This project is mainly focussed on quantifying and storing information on uncertainty associated with the data (Refsgaard et al. 2005).
- Several other projects have compiled large-scale data from catchments, but with a lesser focus on biological data than in REBECCA, for example EUROHARP (http://www.euroharp.org) and Eurolimpacs (http://www.eurolimpacs.ucl.ac.uk).
- Large databases on marine phytoplankton from the Baltic Sea have been compiled within e.g. the project CHARM (http://www2.dmu.dk/1_Viden/2_Miljoe-tilstand/3_vand/4_Charm/charm_main.htm). Data from this database has also been used in the coastal part of the REBECCA project (Carstensen and Heiskanen 2007), and in the ongoing project THRESHOLDS (http://www.thresholds-eu.org).
- Data on macroinvertebrates in rivers were collected by the EU projects AQEM (http://www.aqem.de; Hering et al. 2004) and STAR (http://www.eu-star.at; Furse et al. 2006). The AQEM/STAR databases contain 1,660 samples representing 16 countries and 48 stream types. They contain data on

Table 2 Definition of lake types used in the REBECCA Lakes databases, based on the Intercalibration typology developed by ECOSTAT

GIG region	Countries	Lake type	Altitude[a]	Mean depth[b]	Surface area[c]	Alkalinity level[d]	Humic level[e]	No. of lakes	No. of ref. lakes
Atlantic	IE, UK	L-A1	Low	Shallow	Small	High		14	5
		L-A2	Low	Shallow	Large	High		56	7
		L-A3	Low	Shallow	Small	High	Humic	8	2
		Other						6	1
Alpine	AT, DE, IT	L-AL3	Low	Deep	Large	Medium/high		20	11
		L-AL4	Medium	Shallow	Large	Medium/high		30	2
		Other						26	8
Central-Baltic	BE, DE, DK, EE, FR, HU, LT, LV, NL, PL, RO, UK,	L-CB1	Low	Shallow		High		258	10
		L-CB2	Low	Very shallow		High		201	2
		L-CB3	Low	Shallow		Medium		20	4
		Unknown						427	20
		Other						23	2
Mediterranean	CY, ES, IT, RO	L-M1	Low	Shallow	Large	High		15	0
		L-M5	Low	Deep	Large	Low/medium		2	0
		L-M7	Medium	Deep	Large	Low/medium		12	2
		L-M8	Medium	Deep	Large	High		32	3
		Unknown						38	5
		Other						34	1
Northern	FI, IE, NO, SE, UK	L-N1	Low	Shallow	Large	Medium	Clear	143	53
		L-N2a	Low	Shallow	Large	Low	Clear	257	195
		L-N2b	Low	Deep	Large	Low	Clear	99	75
		L-N3a	Low	Shallow	Large	Low	Humic	327	208
		L-N3b	Low	Shallow	Large	Low	Very humic	137	68
		L-N5	Medium	Shallow	Large	Low	Clear	86	65
		L-N6a	Medium	Shallow	Large	Low	Humic	55	33
		L-N6b	Medium	Shallow	Large	Low	Very humic	5	5
		L-N7	High	Shallow	Large	Low	Clear	3	2
		L-N8a	Low	Shallow	Large	Medium	Humic	165	42
		L-N8b	Low	Shallow	Large	Medium	Very humic	18	2
		Unknown						1,451	384
		Other						1,806	568

GIG = Geographical Intercalibration Group. The columns “No. of lakes” and “No. of ref. lakes” give the number of all lakes and reference lakes, respectively, with data on chlorophyll a. Lake type “unknown” means that information was missing for one or more typology factors. Lake type “other” means that these lakes do not belong to an intercalibration type

[a] Altitude intervals (m): low [0–200); medium [200–800); high [800–)

[b] Mean depth intervals (m): very shallow [0–3); shallow [3–15); deep [15–)

[c] Surface area intervals (km^2): small [0–0.5); large [0.5–)

[d] Alkalinity intervals (meq/l): low [0–0.2); medium [0.2–1.0); high [1.0–); Alkalinity intervals, Ca (mg/l): low [0–4); medium [4–20); high [20–)

[e] Humic intervals, colour (mg/l Pt): clear [0–30); humic [30–90); very humic [90–); Humic intervals, TOC (mg/l; for reference lakes only): clear [0–5); humic [5–15); very humic [15–)

Table 3 Characterisation of lakes in the REBECCA databases: range of values of typology factors (25 percentile, median and 75 percentile)

Typology factor	Unit	Database	Count	25%	Median	75%
Alkalinity	meq l^{-1}	Phytoplankton	5,041	0.091	0.078	0.29
		Macrophytes	5,925	n.a.	0.34	n.a.
		Macroinvertebrates	1,137	−0.059	0.08	0.25
		Fish	301	0.05	0.13	0.37
Altitude	m	Phytoplankton	7,332	65	91	135
		Macrophytes	n.a.	n.a.	n.a.	n.a.
		Macroinvertebrates	1,441	28	105	246
		Fish	288	89.05	170.9	242.2
Colour	mg l^{-1} Pt	Phytoplankton	6,153	10	34	71
		Macrophytes	1,645	n.a.	44	n.a.
		Macroinvertebrates	545	24.2	39	60
		Fish	273	22	57	90
Mean depth	m	Phytoplankton	7,244	1.9	3.6	7.6
		Macrophytes	n.a.	n.a.	n.a.	n.a.
		Macroinvertebrates	1,376	1.9	3	5.3
		Fish	289	3	4.5	7
Surface area	km^2	Phytoplankton	7,969	0.39	1.6	15
		Macrophytes	n.a.	n.a.	n.a.	n.a.
		Macroinvertebrates	1,658	0.067	0.47	3.9
		Fish	289	0.0136	0.23	1.53

The values are given for each biological quality element, because these elements are often sampled from different lakes. The fish lake statistics does not include data from Romania

both occurrence and autecology of species (Schmidt-Kloiber et al. 2006), as well as a software tool for ecological assessment of rivers.

- A dataset on benthic invertebrates in coastal waters has been compiled for intercalibration of coastal classification systems: 589 abundance samples from different locations in seven countries along the European Atlantic coasts (Borja et al. 2007).
- Existing data on fish in streams from 12 countries were compiled in the project FAME (http://fame.boku.ac.at/). These data have been used for correlating fish metrics used in the European Fish Index with environmental component scores (Beier et al. 2007).
- The Modelkey Database (http://www.modelkey.ufz.de) contains monitoring data including macroinvertebrates and fish from three river basins. The data have been used for identification of probable cause–effect relationships on the basis of data on chemical pollution, habitat, toxicity and biological inventories (Brack et al. 2005) and comparison of ecological assessments methods for environmental pollution (Ohe et al. 2007). Examples of use of monitoring data on environmental pollution and ecological responses can also be found in Schriever and Liess (2007) and Schafer et al. (2007).

Publications from other projects that are based on existing monitoring data often give a good overview of the database contents, but less information on the construction and use of the databases, and on the challenges and solutions (but see Beier et al. 2007).

REBECCA database structure

When developing a structure for the databases, we tried to meet two conflicting needs. On one hand, the database structure should be both detailed and flexible enough to accommodate the different data formats and the many updates and corrections. Moreover, since the aim of the project was to analyse biological responses to chemical pressures, an important requirement to the databases was to allow the linking of chemical and biological data in various ways. On the other hand, because the project did not have resources for a professional database manager who could extract tables for data analyses, it was desirable to have a relatively simple database structure so that at least some project partners were able to extract their own tables.

The phytoplankton database, being the largest and most frequently updated, needed to have a flexible construction (Fig. 1; see further description below). For consistency, the macroinvertebrate database was constructed in the same format. We chose to store the data in a form that was close to the original, to facilitate data updating and checking by the providers. However, this structure made it difficult to use for most project partners. The macrophyte database was initiated later in the project, when some lessons had already been learned from the work on phytoplankton and macroinvertebrates. For this database, the data were standardised as much as possible prior to import to the database. Data providers were requested to provide the data in a standard format (with partial success), and the remainder of the data was standardised by the database manager. The fish datasets, which were simpler (no taxonomic information) and more homogenous, were stored in Microsoft Excel.

We did not attempt to combine and harmonise the station lists for the databases among different biological quality elements, because this would be very time-consuming. Many datasets did not contain a unique station code, only station names, for which the spelling was not always consistent among datasets. It was therefore demanding enough to combine the stations for biological and chemical samples within the same database. Thus, for the time being, we were not able to analyse the combined responses of two or more different biological quality elements. Such a combined analysis might nevertheless be possible in a future project.

Phytoplankton and macroinvertebrate databases (NIVA)

In each database, the data were organised into five main tables (Fig. 1): station information, chemistry sample information, biology sample information, chemistry values (incl. pressure variables such as pH or phosphorus) and biology values (such as

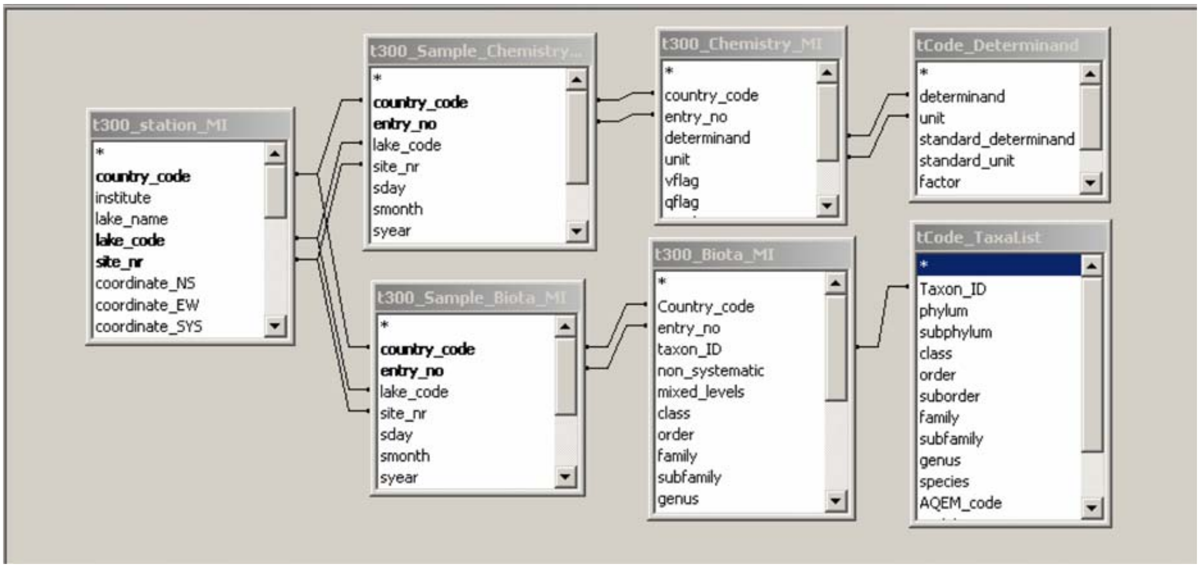

Fig. 1 Illustration of database structures: main tables and relationships between fields within tables. The example shows the macroinvertebrate database. The following properties were found to be particularly useful: (1) Two levels of station identity (lake and site within lakes); (2) separate sample tables for chemistry and biology; (3) separate sample identities (country code + entry no.); (4) tables for standardisation of determinands and taxa

biomass or abundance per taxon). Chlorophyll values were stored in the chemistry table, even though it represents a biological element, because it is usually measured together with chemical parameters, and it does contain any taxonomic information. Separate tables for identifying chemistry samples and biology samples were necessary because the biology samples did not always have corresponding chemistry samples from exactly the same station and date. More chemistry samples than biology samples were provided (see Appendix 1). All unique combinations of original chemistry determinand names and units were stored in a separate table for standardisation. A total of 527 unique combinations of original names for determinands and units were reduced to 139 unique determinands with standardised names and units. Storing the data with their original units and values provided better traceability of the original data. It also allowed extraction of data into tables that had similar format as the original data supplied, and this facilitated data verification by the data providers. However, storing data with their original units also implied that the data must be linked to a standardisation table in order to harmonise the names and the units for each data extraction.

Uniqueness of records was determined by multiple fields both in the station table and in the sample table (Fig. 1). For example, the uniqueness of samples was defined by country code and a sample code that was unique within countries. Defining unique samples by these two fields facilitated the addition and numbering of new samples from countries that were already represented in the database. At the same time, this multiple-field definition of relationships between tables made it more difficult for other project partners to extract data from the database.

Macrophyte database (CEH)

The "Determinands" table was somewhat simpler than for the NIVA databases, as all physical/chemistry data were standardised before importation to the database. A table of "Sources" was kept with a source code and description. These sources related to the data provider and allowed traceability of individual records to their provider. The source information from this table was used in multiple tables, wherever it was possible to attribute a record to a single data provider. Macrophyte abundance data were stored using various provided abundance measures. These included the ECOFRAME scale (a categorical scale from 1 to 3), the DAFOR scale (another categorical scale from 1 to 5), Relative Point Frequency (a continuous scale between 0 and 1) and the Finnish Vegetation Index (a semi-continuous scale with values from 2 to 8,192). These data were stored in their original form, but later converted to common measures (further described in Penning et al. 2008b).

In contrast to the NIVA databases, uniqueness of records in all tables was often determined by an arbitrarily constructed field, e.g. for lake station codes, from the country code concatenated with the data provider's own lake code. This structure was easier to use by project partners, but required more work in its construction. It also made data quality checking more difficult, as the data had been altered from their original form.

REBECCA database processes

The main steps from receiving data until the extraction of tables for use in analyses are summarised below. These steps were, in principle, similar for all of the databases presented here (see Fig. 2).

Data cleaning

Checking and correction of data were usually required before the raw data could be used. A common problem was erroneous units, such as mg l^{-1} instead of μg l^{-1}. (Note that if "μ" is typed as "m" with symbol font in one software, it may be changed into "m" in a different software). Moreover, both the comma "," and the period "." are used as decimal symbols in Europe; data with a decimal symbol that does not match the computer's settings may be interpreted as text. Missing values were coded in many different ways in the raw data. Plotting of coordinates on maps revealed that longitude and latitude were sometimes mixed (e.g. when lakes appeared to be positioned in the Mediterranean Sea). There were variations in spellings of physical/chemical determinands and in the names of biological taxa. Numerous other irregularities were encountered. This process of data checking often revealed inconsistencies and errors the data providers themselves were not aware of. Despite our initial screening and correcting of irregularities before importing the data, new errors were often discovered by the data analysts, or by the data providers themselves when preliminary results were presented.

Data reorganisation

A template for data collation was initially developed and distributed to the partners. However, many partners experienced the reorganisation of their data into the specified format as a very time-consuming job. We therefore decided to accept data in any format also from project partners. The raw data were usually organised in so-called cross-tabular or pivot format, i.e. with samples arranged in rows and each physical, chemical and biological determinands arranged in separate columns. In a database, on the other hand, all determinand values for a particular type of data (chemical/physical or biological) are listed in the same column. This avoids empty cells, so the space is used more efficiently, and facilitates extraction of data into various table formats. Additional information such as flags to denote reliability of the measurement (e.g. "<", meaning "below detection limit") is also stored more appropriately in a separate field. In many cases text data, such as "less than 3", was stored with numeric data. We used a Microsoft Excel^{TM} macro for the reorganisation of data into database format (B. Bjerkeng, unpubl.), combined with extensive manual checking.

Import to Access database

Although the databases managed by NIVA and CEH differed in some aspect (cf. Fig. 1), the key aspects are common to both. Data were generally separated into location, chemistry and biology. The location data comprised the name of the lake (and sometimes sampling stations within the lake), reference status (reference lake or not), typology factors such as size, depth and altitude and geographical coordinates of the lake. The sample data (in the NIVA databases) contained sampling location and date, as well as information about sampling method, where available. The chemistry data, as well as containing chemical data such as concentrations of nutrients and pH, also included some physical determinands, such as Secchi depth (transparency), turbidity and temperature. "Chemistry" data also included chlorophyll concentrations (cf. explanation in the previous section). The biological data generally consisted of a list of taxa recorded at a site with some measure of abundance (count of cells/individuals and/or estimated biomass per unit volume, length of filamentous algae or some type of abundance class). Each database also included a species list, linked to the biological abundance data table. Data were added and adjusted to the main databases by a series of so-called append queries and update queries.

Typification (assignment of lake types)

Lakes were assigned to lakes types according to the Intercalibration Lake Typology, as used in the WFD Common Implementation Strategy process of intercalibration of assessment systems (see Table 2). Since analysis of lake-type-specific relationships was a highly prioritised issue in REBECCA, we

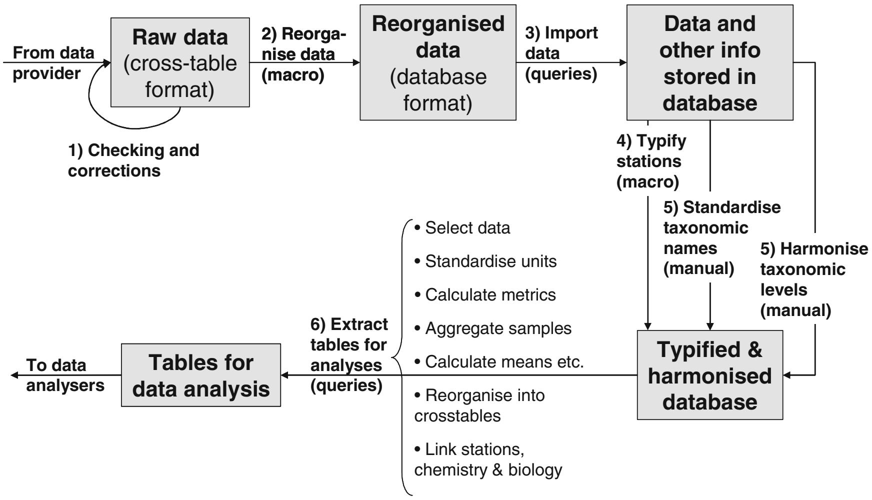
From data provider
Raw data (cross-table format)
1) Checking and corrections
2) Reorganise data (macro)
Reorganised data (database format)
3) Import data (queries)
Data and other info stored in database
4) Typify stations (macro)
5) Standardise taxonomic names (manual)
5) Harmonise taxonomic levels (manual)
Typified & harmonised database
• Select data
• Standardise units
• Calculate metrics
• Aggregate samples
• Calculate means etc.
• Reorganise into crosstables
• Link stations, chemistry & biology
6) Extract tables for analyses (queries)
Tables for data analysis
To data analysers

Fig. 2 The data pathway in the REBECCA: from raw data via databases to reformatted tables for statistical analyses. Each step is described in detail in the section "REBECCA database processes"

aimed at typifying as many stations as possible. We therefore used not only the information on lake types given by the data providers, but also all available station and chemistry information. Nevertheless, a large proportion of the lakes could still not be typified due to lack of data. Eventually a request was sent to all data providers for expert judgment on the levels of typology factors, according to the categories agreed in the intercalibration process (Table 2).

The designation of each station to one or more lake types was thus a very elaborate process, because it was necessary to combine up to 20 fields of information from each data provider. For each of the main typology factors that were stored in the station table (altitude, mean depth, surface area, alkalinity (or calcium) and colour (or TOC, for reference lakes), we used primarily numeric values (if available), and secondarily the information on typology categories, as provided by expert judgement. In addition to this information from the station table, we used the available chemistry values for alkalinity (or calcium, if alkalinity values were not available) and colour (or, for reference lakes, TOC if colour was not available). Finally, in cases where there were not sufficient data for typification, we used information on IC types as given directly by the data providers (where this was available). Most countries belong to only one of the five current GIG regions (see Table 1), but four countries (Ireland, Italy, Romania, UK) belong to more than one region. For these countries, alternative IC types were also designated where possible.

Reference status was set solely by the data provider, according to either pressure criteria, impact criteria, expert judgement or a combination of these.

Taxonomy: standardisation and harmonisation

For taxonomy, we distinguish between the standardisation of taxonomic names, and the harmonisation of taxonomic levels. The standardisation of names was a crucial and time-consuming task. The phytoplankton taxonomy was elaborated in collaboration with data providers (P. Brettum, NIVA and L. Lepistö, SYKE, Finland). The number of taxa was thus reduced from 5,500 unique names (including various spellings) to 1,900 unique taxa codes. The macroinvertebrate taxonomy is based on the EU projects AQEM and STAR (Schmidt-Kloiber et al. 2006). The macrophyte taxonomy was based on a species list provided by for the Central GIG (J. Hanganu, DDNI, Hungary) and extended by species recorded from the Northern and Atlantic GIGs. For the phytoplankton and macroinvertebrate elements, all observations were stored with their original names, and linked to the standardised taxonomy tables by a unique species code, allowing us to look up the original names if needed. Macrophyte names were standardised before importation into the database, and stored as species codes, which were linked to a master species table. Details of the standardisation were kept separately for each dataset imported.

Harmonisation of taxonomic levels was necessary because different datasets could have different degree of taxonomic resolution (e.g. some identified to species level, others to genus or higher levels). This may result in spurious country-wise differences in the number of taxa observed. Moreover, a mix of taxonomic levels within samples may artificially increase the number of recorded taxa. For example, individuals of the species *Baetis rhodani* can be recorded as *Baetis rhodani*, *Baetis* sp., family Baetidae, and order Ephemperoptera—apparently four different taxa. Moreover, different taxonomic levels may lead to apparent absence of certain taxa in certain regions. For example, *Cryptomonas* (a cryptophyte alga) was not split into species in Sweden and Finland, but was analysed to species level in Norway. Thus, a comparison of number of cryptophyte taxa across these countries is not feasible on species level. The biological data were therefore coded at all possible taxonomical levels (species, genus, family and order), allowing the data users to perform analyses at the appropriate taxonomic levels.

Extraction of tables

In order to be used for statistical analyses, the data had to be rearranged into single files, usually as cross tables. Practically any kind of table can be extracted by a combination of so-called select queries and cross-table queries. The queries can be designed in a graphical interface, which are interchangeable with both datasheet format and SQL format (programming language). The main steps were as follows.

- Selection of data (e.g. reference lakes; Northern GIG; summer samples).

- Standardisation of units and taxonomy (link value tables to standardisation tables and multiply values by standardisation factors).
- Calculation of metrics per sample (e.g. proportion of cyanobacterial biomass or counts per sample).
- Aggregation of chemistry samples and biology samples at the same time unit (e.g. per season or month), so that chemical and biological data can be linked.
- Calculation of summary statistics for data (e.g. mean total phosphorus concentration, mean proportion of cyanobacteria biomass).
- Reorganisation into cross tables (if more than one chemistry determinand per station is wanted).
- Linking (aggregated) biology samples with corresponding chemistry samples, via the station table.

The database thus allowed reorganisation and aggregation of data in ways that would be virtually impossible with so-called flat files (such as Excel).

Costs and benefits of the REBECCA databases

The main cost of our common lakes database approach was the vast amounts of time required for data inspection, correction and standardisation. Requests to the different data providers for explanations and for missing information were an unavoidable and time-consuming task. A better planning in advance of the data compilation procedures would probably have saved much time. In particular, the database structures and functions should have been discussed with a professional database manager before the data templates were developed. Nevertheless, some modification of the database structure and instructions to data providers were unavoidable, since the data sources and the heterogeneity of data increased during the project.

Since we eventually accepted data in any format, it was necessary to standardise the station, chemistry and biology data. The standardisation of names and conversion to common units for physical and chemical determinands was relatively trivial. However, taxonomic standardisation and harmonisation of biological data were considerably more demanding.

For constructing the databases we chose the software Microsoft Access, which is commonly available to researchers and is relatively easy to use also for beginners. However, as the complexity of the databases grew (in order to accommodate the various formats of the raw data), it became increasingly difficult for the project partners to extract their own tables. Table extractions were also done by the database manager upon request from the data analyst, but this procedure required precise communication and could be inefficient. Hence, the more complicated table extraction was a significant additional cost of the increased data intake.

The benefits of our common-database approach should be reflected in several other REBECCA Lakes publications in this issue of *Aquatic Ecology* (e.g. Carvalho et al. 2008; G.-Tóth et al. 2008; O'Toole et al. 2008; Penning et al. 2008a, b; Phillips et al. 2008; Ptacnik et al. 2008b; Schartau et al. 2008). We have been able to analyse reference conditions and pressure–response relationships across a range of countries, which has made the results more representative within each GIG region. For some biological elements we have also been able to analyse lake type-specific relationships, particularly within the Northern GIG (chlorophyll—Phillips et al. 2008; phytoplankton composition—Ptacnik et al. 2008a, b; Schartau et al. 2008). For other elements, one or more typology factors have been used in the analyses, either as covariates in the model (macroinvertebrates—Schartau et al. 2008; macrophytes—Penning et al. 2008a, b), or to split the dataset into groups of lake types (chlorophyll—Carvalho et al. 2008). These type-specific results were essential information for the GIGs as a basis for boundary setting between the different ecological status classes. The assessment of type-specific reference conditions has also been made possible by these databases particularly for chlorophyll (European Commission 2003). Defining reference conditions is a critical first step for setting Ecological Quality Ratio values, and is thus also important for assessment of ecological status (Ptacnik et al. 2008b).

Another benefit of analysing combined datasets was that the data covered a larger range of the pressure gradient, and a more complete picture of the pressure–response relationship could be described. In fact, an apparent lack of significant relationships within a national dataset might be due to analysis on a too narrow pressure range. On the other hand, if different ranges of the pressure gradient are dominated by data from different countries, it may be difficult to separate the real effect of the pressure from spurious effects of, e.g., differences in national sampling methodology. For the fish analyses, however, the large-region analyses were usually not in conflict with the single-country analyses (T.O. Haugen pers. comm.).

As a side effect, analyses of the multi-national data also resulted in interesting discoveries that were not directly related to REBECCA. For example, analyses of phytoplankton data within the Northern GIG revealed hitherto unknown geographical trends in phytoplankton species richness (see Ptacnik et al. 2008a, b). There is a large potential for more interesting results from further analyses of these data within other projects, for example related to large-scale patterns in biodiversity.

The large number of observations should generally increase the precision of the estimates, and thus make the results more statistically reliable. For the fish data, for example, variance explained by the combined datasets decreased to about 50% of the variance explained at country level (T.O. Haugen pers. comm.). However, because of the vastness of data there is also greater heterogeneity, which can be difficult to disentangle or to reduce (discussed below). These uncertainties made it more difficult to interpret the large-scale results. For example, apparent geographical trends for macroinvertebrates metrics may have been blurred by country-specific factors (Schartau et al. 2008). More robust results may be obtained if the biological responses are analysed as presence/absence or proportions instead of absolute abundances (Schmidt-Kloiber and Nijboer 2004; Verdonschot 2006a; Moe et al. 2007). In some cases, local knowledge about lakes and their biota might be required for the interpretation of the results (E. Penning pers. comm.).

When taxonomic resolution varied among the datasets, we had to exclude the datasets with too low resolution, or to aggregate all data to the lowest common level (e.g. from species to family or order). Taxonomic aggregation apparently did not influence results for macrophytes (E. Penning pers. comm.), while the impact was more variable for phytoplankton (R. Ptacnik unpubl.). Other studies have demonstrated that certain metrics do not perform properly if family-level data are used instead of species level (Schartau et al. 2008).

The flexible structure of the databases implied that it was relatively easy to aggregate data in different ways, e.g. taxonomically, temporally or geographically. Thus, this structure enabled testing of metrics calculated for different taxonomic levels, or to aggregate chemistry and biology samples for different time periods (see also Borja et al. 2007).

The process of data compilation may have positive side effects beyond those originally intended. The construction of a database develops a feedback process between data providers, researchers and end-users (Beier et al. 2007). The process can contribute to organising large amounts of data that are otherwise not easily accessible. For example, the extensive Norwegian dataset on macroinvertebrates in lakes consisted of >200 Excel sheets of somewhat varying formats. The process of standardisation provides a mechanism for quality control, thus making each national dataset more valuable (Beier et al. 2007). In REBECCA, the analysis and plotting of national data in a larger context made it easier to identify errors and outliers in individual datasets.

There is a large potential for further use of the REBECCA databases in research projects. Most of the data providers have given consent to further use of the data after the end of the projects, although with some restrictions (e.g. requiring co-authorship). Data on phytoplankton, macrophytes and macroinvertebrates have already been used in the Intercalibration process within the Northern GIG, and will also be used in the next intercalibration exercise by this GIG. Although there exists a European intercalibration register, it has been recognised by both the European Commission and member states that additional data from non-intercalibration sites may be required to progress the intercalibration exercise (Refsgaard et al. 2007). Experiences from the REBECCA databases will also be used by the European Environment Agency, who intend to start compiling biological data from all EU member states for State of Environment information in WISE (Water Information System for Europe; http://water.europa.eu) (A. Künitzer pers. comm.). The databases provide an opportunity for analysing combined pressure–response relationships for two or more biological quality elements, provided that the station lists of the different biological databases are harmonised. More generally, the data should be valuable for research on large-scale patterns in biogeography and biodiversity. However, one should keep in mind that each dataset is originally collected for a specific purpose (e.g. presence of acid-sensitive macroinvertebrate taxa), and that it may not contain information that would be required for a different purpose (e.g. temporal trends in abundances).

Challenges with compilation of ecological data

There is a need for further data collection to fulfil the WFD requirements, since much of the characterisation and classification has been carried out based on expert knowledge (Refsgaard et al. 2007). However, since collection of data from the field is very resource-demanding, new environmental research projects will often be based on existing data. Environmental data are becoming more accessible, for example through EU initiatives such as WISE and INSPIRE, following the Aarhus Convention on access to information in environmental matters. Nevertheless, there are still large problems with regard to data access in practice. The main problem is not necessarily the data availability, but accessibility, quality, and relevant information about the data (including uncertainty). The constraints can be of different types: economic, political, data formats, fragmented databases or transboundary barriers (harmonising or exchanging data across national barriers) (Refsgaard et al. 2007). As reported by the HarmoniRiB project (Refsgaard et al. 2007): "In projects where existing data are used the data collection is often cumbersome and requires a lot of resources, because the data access is difficult with many practical and economic constraints." "Often data collected in one research project is not used in many other projects due to lack of proper data documentation and dissemination after the termination of old projects. The same data are therefore often collected several times by different research projects. This is obviously non-optimal and requires a lot of research resources both in terms of costs and manpower that could have been utilised much better." Moreover, scientists who produce data are often unwilling to share them, due to strong traditions, competition for funding or other circumstances (Beier et al. 2007). Other practical problems have been reported (Lorenz et al. 2004; Vlek et al. 2006): data are not always available digitally; different institutes have been collecting the same data without cooperation; closely related data are stored in different databases at different institutes or even in private companies.

Biological data, in particular, are often collected by different experts, and data for different taxonomic groups are stored separately by the researchers. For biological data, the practical problems regarding formats and standards are also likely to be even greater than for other environmental data, for several reasons. (1) Biota are usually heterogeneously distributed in the water body, both in space and time. (2) Sampling methods are more difficult to standardise. (3) Taxonomic systems are changing continuously, resulting in numerous synonyms. This causes problems when combining datasets from different researchers and/or countries. Important properties of the samples, such as number of species recorded, can be affected by sample size (Clarke and Hering 2006). (4) Methods for quantification of abundance are more variable and more imprecise (e.g. estimates of biomass, density of individuals or coverage of surface or semi-quantitative scales). (5) Additional sources of variation arise from sample processing and taxonomic identification error, and from effects of environmental stress on the biota (Refsgaard et al. 2005, 2007). This implies that detailed sampling information may be even more critical for biological data than for other environmental data.

All data have some degree of associated uncertainty, let alone biological data. Refsgaard et al. (Clarke and Hering 2006; Haase et al. 2006) recommend that information on data quality and uncertainty is stored as a part of the data documentation. This implies a need for modification in database structure, as standard databases today are not designed to enable storage of data uncertainty. However, the data quality may not have equal importance for all purposes. Exploratory analyses of large-scale patterns (e.g. comparing a pressure–response relationship for different lake types) may be more robust to data uncertainty than predictive modelling, which has higher demands for accuracy (e.g. trying to predict the amount of cyanobacteria for a given phosphorus concentration level). Thus, the results of the REBECCA analyses, which are mostly exploratory, may not be critically dependent on information on data uncertainty.

Recommendations

Based on our experiences within the REBECCA project, we would like to give some recommendations for data compilation within other research projects. Our recommendations are not meant to be general guidelines for database management; they apply to the particular challenges of compiling multinational ecological data.

Planning and resource allocation

Sufficient time should be allowed within the project for the necessary data processing. A trained database manager should be allocated to data-handling tasks, especially with respect to the complexity and quantity of data that will be involved in the project. A combination of data-processing skills and ecological knowledge is required for this task. The database manager should be informed about the needs and planned uses of the database, and be involved in designing templates for data request, database structures, and data transfer tools in close collaboration with the project leaders. If there are not enough resources for having a data manager to provide extracts for the data users, then key users should be trained in basic database skills for extracting their own tables.

Organisation of files by data providers

Data providers should be given precise instructions on the required data format, or at least be informed on the most important aspects of data organisation. Data should be stored in raw format (not aggregated). Unique standardised codes should be used for all localities and sampling stations, and be used to identify both chemistry and biology samples. All dates should be recorded as day, month and year separately, because date formats in different softwares are not always compatible. There must be a unique identifier for missing values, to avoid artefacts by numeric codes like "-999". For sampling information, as many details as possible should be requested. The data templates should also make room for additional, potentially relevant information at all levels. For biological data, the best solution may be that all data providers add a common taxonomic code to the observations (e.g. the AQEM/STAR code for macroinvertebrates, and the so-called REBECCA code for phytoplankton). In addition, a complete taxonomic list (with spell-checked names) should be provided. Data providers should be requested to check for suspicious values before submitting the data, for example by box-and-whisker plots or at least by checking minimum/maximum values. A high rate of errors in environmental data has been discovered in other projects (Beier et al. 2007) as well as in REBECCA.

Data submission and sharing

All REBECCA data were submitted by e-mail or by direct transfer from computer to computer. The REBECCA toolbox (http://www.rbm-toolbox.net/rebecca) was also used for returning data extracts to providers after the database compilation, but only as a tool for document sharing. We recommend the use of more efficient tools to facilitate data management and sharing. For example, Beier et al. (Refsgaard et al. 2005) report on the use of an input database and manual, automated quality control tools and a series of input and export queries developed using Data Transformation Services (DTS). Guidelines or templates should be developed for reporting of suspicious values and for providing corrected values or additional information. All updates and corrections in to the database should be logged.

Database construction

The database software Microsoft Access is relatively easy to use and mostly worked well for our purposes, except that the table size is limited to 256 columns, which can cause problems for extracting tables with either species or samples in columns. Although a database structure should be planned from the beginning of the project, it should also be possible to change the structure during the project if this turns out to be favourable. For example, when it turned out that our phytoplankton database had a large number of chemical samples that did not match with the dates of the biological data, we decided to split the common sample table into separate sample tables for chemistry and biology, which enabled a higher match of chemical and biological samples (after temporal aggregation). Information on data sources and data providers should be stored in the database. The database should include complete taxonomy for each biological observation, so that aggregation at any taxonomic level is possible. If possible, information on data quality and uncertainty should also be stored.

Data analysis and interpretation

Interpretation of results requires some special considerations when the data are compiled from many different countries. Some of the differences can be standardised as described above, while other differences are inherent to the data. A typical inherent problem is differences in geological, geographic and climatic conditions. Dividing the data into IC lake types may account for some of this variation, but including typology factors as continuous covariables might increase precision further. Another inherent problem is differences in methodology. For example, different mesh size for macroinvertebrate samples may result in different number of taxa as well as individuals. For macrophytes, different semi-quantitative abundance measures were used. A short-term solution might be to analyse responses within countries and compare results qualitatively. For example, one can check whether different abundance measures show breakpoints or abrupt changes in the same interval along the pressure gradient. Sampling information can also to some degree be used to standardise data. For example, coastal benthic invertebrate samples were standardised for sample area, sieve mesh size and sediment type in the Intercalibration (Borja et al. 2007). A longer-term solution would obviously be standardisation of sampling and analysis methods, as is being initiated by CEN (Comité Européen de Normalisation).

Acknowledgements We thank the all data providers and contact persons for their contributions, and especially those from institutes external to the REBECCA project. *REBECCA partners*: DDNI: Orhan Ibram, Jenica Hanganu; IRSA: Gianni Tartari, Elena Legani; NINA: Ann Kristin Schartau; CEH: Laurence Carvalho; SLU: Leonard Sandin, Mats Wallin, Eva Willén; SYKE: Arjen Raateland, Olli-Pekka Pietiläinen, Seppo Hellsten; TCD: Constanze O'Toole, Ian Donohue; WL|Delft: Ellis Penning; NIVA: Anne Lyche Solheim, Pål Brettum, Marit Mjelde. *External data providers*: JRC (CB-GIG): Sandra Poikane; Finland: Jukka Aroviita, Heikki Hämäläinen; Germany: Ute Mische, Rita Adrian, Stella Berger, Peter Krause, Ursula Gaedke, Tom Kunz; Ireland: Deirdre Thierny, Ruth Little, Alice Wermer, John White; Norway: Gunnar Raddum; Poland: Malgorzata Golub, Agnieska Kolada; Portugal: Joao Padua; Spain: M^{a} Luisa Serrano; Sweden: Mikaela Gonczi, Eva Willén; UK: Geoff Phillips. Finally, we thank Gunnar Severinsen for enormous help with constructing the databases at NIVA and for training and assistance, Thrond O. Haugen for providing information on the fish data and Birger Bjerkeng for providing the Microsoft Excel^{TM} macro for reorganising data. REBECCA is jointly funded by the EC 6th Framework Programme (Contract number SSPI-CT-2003-502158), various national research organisations and the project partners.

Appendix 1

Table A1 Overview of major chemical determinands in the REBECCA databases: (A) Phytoplankton, (B) Macrophytes, (C) Macroinvertebrates

Determinand	Unit	No. of observations	10% percentile	Median	90% percentile
(A) Phytoplankton					
Al3+	μg l^{-1}	654	18.4	66	190
Alkalinity	meq l^{-1}	25,979	0.04	0.17	1.53
ANC	μekv l^{-1}	980	8.5	121	390
BOD5	mg l^{-1}	459	2	5.1	12
BOD7	mg l^{-1}	67	1.2	2.2	3.6
Ca	mg l^{-1}	8,163	0.95	3.7	59
Chlorophyll-a	μg l^{-1}	80,978	2.2	11	97
Cl	mg l^{-1}	5,078	3.5	11	202
COD	mg l^{-1}	25,165	5.1	9.9	21
Colour	mg Pt l^{-1}	31,991	15	50	150
Conductivity	μS cm^{-1} 25°C	41,575	25	52	252
DIN	μg l^{-1}	24,898	5	28	263
DIP	μg l^{-1} P	59,204	1	9	280
DOC	mg l^{-1}	1,262	3.4	6.3	11.5

Table A1 continued

Determinand	Unit	No. of observations	10% percentile	Median	90% percentile
DOM	mg l^{-1}	104	0.09	0.29	0.99
Fe	µg l^{-1}	450	2	43	310
K	mg l^{-1}	944	0.11	0.52	3.0
Kjeldahl N	mg l^{-1} N	27,246	0.9	2.1	4.2
Mg	mg l^{-1}	3,413	0.34	1.7	5.2
Na	mg l^{-1}	943	0.65	2.6	9.9
NH3	mg l^{-1}	4,631	0.008	0.02	0.061
NH4	µg l^{-1} N	33,587	20	100	680
NO2	mg l^{-1} N	14,339	0.005	0.02	0.1
NO3	µg l^{-1} N	14,656	10	124	2,600
Oxygen	mg l^{-1}	34,150	7.7	9	11.2
pH	unit	45,459	6.2	7	8
PO4_f	µg l^{-1} P	1,305	0.5	1	5
POM	µg l^{-1} N	104	473	991	2,011
Si	mg l^{-1}	10,399	0.19	0.95	3.1
SO4	mg l^{-1}	1,637	2.1	13.5	175
SS	mg l^{-1}	3,569	1.3	2.31	11
Temperature	°C	45,828	8	16.7	21.5
TN/F	µg l^{-1}	910	34	111	446
TOC	mg l^{-1}	5,746	2.9	7.5	15
TON	mg l^{-1}	3,076	0.04	0.15	1.3
TotN	µg l^{-1}	45,097	280	515	1,100
TotP	µg l^{-1}	82,564	8	36	400
Transparency	m	68,317	0.3	1.2	3.4
Turbidity	FTU	25,946	0.5	1.7	7.2
(B) Macrophytes					
Absorbance (420 nm)	Au	20	0.01	0.05	0.16
Alkalinity	meq l^{-1}	5,925	0.04	0.34	2.37
Aluminium (3+)	µg l^{-1}	493	29	83	197
Ammonia	µg l^{-1} N	2,487	10	30	90
Ammonium	µg l^{-1} N	644	3	12	263
Calcium	mg l^{-1}	4,044	0.9	3.8	47.1
Chlorine	mg l^{-1}	3,781	3.7	9.4	44.0
Chlorophyll-a	µg l^{-1}	5,621	0.8	4.9	34.7
Colour	mg Pt l^{-1}	1,645	14	44	105
Conductivity	µS/cm	6,666	31	72	444
Dissolved inorganic phosphorus	µg l^{-1}	171	4	10	20
Dissolved organic carbon	mg l^{-1}	1,040	3.4	6.0	11.9
Iron	mg l^{-1}	59	0.05	0.49	1.40
Magnesium	mg l^{-1}	92	0.45	0.85	1.61
Manganese	mg l^{-1}	67	0.01	0.02	0.05
Nitrate	µg l^{-1} N	2,230	28	150	579
Nitrite	mg l^{-1} N	902	0.002	0.007	0.021
Oxygen (% saturation)	%	2,809	81.9	94.6	106.7
Oxygen (absolute)	mg l^{-1}	3,044	8.9	10.5	12.5

Table A1 continued

Determinand	Unit	No. of observations	10% percentile	Median	90% percentile
pH	unit	7,233	5.8	7.1	8.3
Phosphate	μg l^{-1} P	4,044	1	8	97
Potassium	mg l^{-1}	91	0.21	0.47	0.90
Silicate	mg l^{-1}	4,801	0.4	1.4	5.7
Sodium	mg l^{-1}	92	0.69	1.21	2.59
Strontium	mg l^{-1}	16	0.010	0.015	0.046
Sulphate	mg l^{-1}	86	1	2.2	6.2
Suspended solids	mg l^{-1}	2,325	1.4	3.1	13.0
Temperature	°C	2,510	4.1	11.0	17.8
Total nitrogen	μg l^{-1} N	2,264	200	580	2,690
Total organic carbon	mg l^{-1}	667	2.3	5.1	9.7
Total oxidised nitrogen	mg l^{-1}	2,319	0.06	0.21	1.71
Total phosphorus	μg l^{-1} P	6,161	5	20	119
Transparency (Secchi)	m	1,406	0.8	2.1	4.9
Turbidity	NTU	1,165	0.9	2.7	10.5
(C) Macroinvertebrates					
Absorbance	420/5	1,350	0.017	0.10	0.33
AL_ICP	μg l^{-1}	1,350	17	120	250
Alkalinity	meq l^{-1}	4,130	0.0076	0.19	2.57
ANC	μeq l^{-1}	1,776	5.5	174	576
Ca	mg l^{-1}	3,047	0.67	2.1	20.5
Chlorophyll_a	μg l^{-1}	3,004	1.23487	5.4	41.3832
Cl	mg l^{-1}	2,613	0.016	0.44	8.0
COD	mg l^{-1}	511	4.5	9.8	15.1
Colour	mg l^{-1} Pt	1,940	12	38	90
Conductivity	μS cm^{-1}	4,493	13	60	368
DIN	mg l^{-1} N	706	0.004	0.13	0.63
DIP	μg l^{-1} P	210	3	5	17
DOC	mg l^{-1} C	482	3.4	5.7	10.6
Fe	μg l^{-1}	1,356	17	173	1,100
K	mg l^{-1}	1,731	0.005	0.015	0.19
Mg	mg l^{-1}	1,737	0.3	1.6	5.1
Na	mg l^{-1}	1,735	0.04	0.11	2.4
NH3	mg l^{-1} N	1,634	0.01	0.03	0.06
NH4	μg l^{-1} N	1,390	4	13	77
NO2	mg l^{-1} N	473	0.01	0.01	0.01
NO3	mg l^{-1} N	1,299	0.013	0.112	0.32
O2	mg l^{-1}	1,541	8.1	10.3	13.3
O2_sat	%	470	85	93	105
O2_summer_min	mg l^{-1}	166	1.5	7.5	9.8
O2_summer_min_sat	mg l^{-1}	95	11.4	70	85
O2_winter_min	mg l^{-1}	158	0.69	4.6	9.0
O2_winter_min_sat	mg l^{-1}	87	5.8	35	69
pH	.	3,628	5.2	6.7	8.2
PO4	mg l^{-1} P	1,234	0.001	0.003	0.015

Table A1 continued

Determinand	Unit	No. of observations	10% percentile	Median	90% percentile
Si	mg l^{-1}	3,092	0.2	1.1	3.04
SO4	mekv l^{-1}	1,350	0.023	0.089	2.2
SO4	mg l^{-1}	363	0.023	0.089	2.2
SS	mg l^{-1}	701	0.96	2.0	9.1
Temperature	°C	3,210	4.0	9.2	20.2
TOC	mg l^{-1}	2,002	1.5	6.1	14
TON	mg l^{-1} N	1,333	0.10	0.24	0.66
TotN	µg l^{-1}	3,998	190	510	1,181
TotP	µg l^{-1}	4,672	4.0	14	69
Transparency	m	1,310	0.6	2.3	4.8
Turbidity	FTU	1,344	0.53	1.6	7.0

The range for each determinand is represented by the 10% and 90% quantiles (The databases contain also other chemical determinands with a lower number of observations)

References

Beier U, Degerman E, Melcher A, Rogers C, Wirlöf H (2007) Processes of collating a European fisheries database to meet the objectives of the European Union Water Framework Directive. Fish Manag Ecol 14:407–416

Borja A, Josefson AB, Miles A, Muxika I, Olsgard F, Phillips G, Rodriguez JG, Rygg B (2007) An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. Mar Pollut Bull 55:42–52

Brack W, Bakker J, de Deckere E, Deerenberg C, van Gils J, Hein M, Jurajda P, Kooijman B, Lamoree M, Lek S, de Alda MJL, Marcomini A, Munoz I, Rattei S, Segner H, Thomas K, Von der Ohe PC, Westrich B, de Zwart D, Schmitt-Jansen M (2005) MODELKEY—Models for assessing and forecasting the impact of environmental key pollutants on freshwater and marine ecosystems and biodiversity. Environ Sci Pollut Res 12:252–256

Carstensen J, Heiskanen AS (2007) Phytoplankton responses to nutrient status: application of a screening method to the northern Baltic Sea. Mar Ecol Prog Ser 336:29–42

Carvalho L, Solimini A, Phillips G, Berg Mvd, Pietiläinen O–P, Solheim AL, Poikane S, Mischke U (2008) Chlorophyll reference conditions for European lake types used for intercalibration of ecological status. Aquat Ecol. doi:10.1007/s10452-008-9189-4

Clarke R, Hering D (2006) Errors and uncertainty in bioassessment methods—major results and conclusions from the STAR project and their application using STARBUGS. Hydrobiologia 566:433–440

Degerman E, Beier U, Breine J, Melcher A, Quataert P, Rogers C, Roset N, Simoens I (2007) Classification and assessment of degradation in European running waters. Fish Manag Ecol 14(6):417–426. doi:10.1111/j.1365-2400.2007.00578.x

European Commission (2003) Common Implementation Strategy for the Water Framework Directive (2000/60/EC). Guidance document no. 6. Towards a guidance on establishment of the intercalibration network and the process on the intercalibration exercise. Produced by Working Group 2.5 Intercalibration, pp. 54

European Commission (2005) Common Implementation Strategy for the Water Framework Directive (2000/60/EC). Guidance document no. 14. In guidance on the intercalibration process 2004–2006, pp. 26

Furse M, Hering D, Moog O, Verdonschot P, Johnson R, Brabec K, Gritzalis K, Buffagni A, Pinto P, Friberg N, Murray-Bligh J, Kokes J, Alber R, Usseglio-Polatera P, Haase P, Sweeting R, Bis B, Szoszkiewicz K, Soszka H, Springe G, Sporka F, Krno I (2006) The STAR project: context, objectives and approaches. Hydrobiologia 566:3–32

G.-Tóth L, Poikane S, Penning WE, Free G, Mäemets H, Kolada A (2008) Comparing national assessment methods for macrophytes as a biological quality element for the WFD: results from the first steps of the Central-Baltic intercalibration exercise. Aquat Ecol. doi:10.1007/s10452-008-9184-9

Haase P, Murray-Bligh J, Lohse S, Pauls S, Sundermann A, Gunn R, Clarke R (2006) Assessing the impact of errors in sorting and identifying macroinvertebrate samples. Hydrobiologia 566:505–522

Hering D, Moog O, Sandin L, Verdonschot P (2004) Overview and application of the AQEM assessment system. Hydrobiologia 516:1–20

Lorenz A, Kirchner L, Hering D (2004) "Electronic subsampling" of macrobenthic samples: how many individuals are needed for a valid assessment result? Hydrobiologia 516:299–312

Lyche Solheim A, Rekolainen S, Moe SJ, Carvalho L, Phillips G, Ptacnik R, Penning E, Toth LG, O'Toole C, Schartau AKL, Hesthagen T (2008) Ecological threshold responses in European Lakes and their applicability for WFD implementation—synthesis of REBECCA Lakes results. Aquat Ecol. doi:10.1007/s10452-008-9188-5

Moe SJ, Ptacnik R, Penning E, Kuikka S, Malve O (2007) Statistical and modelling methods for assessing the relationships between ecological and chemical status in lakes. REBECCA Deliverable 12. NIVA report nr. 5459–2007, pp. 38

O'Toole C, Donohue I, Moe SJ, Irvine K (2008) Relationships between nutrient status and benthic invertebrate communities: application of the REBECCA database. Aquat Ecol. doi:10.1007/s10452-008-9185-8

Ohe PCvd, Prüß A, Schäfer RB, Liess M, Deckeree Ed, Brack W (2007) Water quality indices across Europe – a comparison of the good ecological status of five river basins. J Environ Monitor 9:970–978

Penning WE, Dudley B, Mjelde M, Hellsten S, Hanganu J, Ecke F, Willby N, Phillips G (2008a) Using aquatic macrophyte community indices to define the ecological status of European lakes. Aquat Ecol. doi:10.1007/s10452-008-9183-x

Penning WE, Mjelde M, Dudley B, Hellsten S, Hanganu J, Ecke F, Willby N, Phillips G (2008b) Classifying aquatic macrophytes as indicators of eutrophication in European lakes. Aquat Ecol. doi:10.1007/s10452-008-9182-y

Phillips G, Pietiläinen O-P, Carvalho L, Solimini A, Solheim AL, Cardoso AC (2008) Chlorophyll—nutrient relationships of different lake types using a large European dataset. Aquat Ecol. doi:10.1007/s10452-008-9180-0

Pont D, Hugueny B, Rogers C (2007) Development of a fish-based index for the assessment of river health in Europe: the European Fish Index. Fish Manag Ecol 14(6):427–439. doi:10.1111/j.1365-2400.2007.00577.x

Ptacnik R, Andersen T, Solimini AG, Brettum P, Lepistö L, Willén E, Rekolainen S, Tamminen T (2008a) Diversity predicts stability and resource use efficiency in natural phytoplankton communities. Proc Natl Acad Sci 105: 5134–5138

Ptacnik R, Lepistö L, Willén E, Brettum P, Andersen T, Rekolainen S, Solheim AL (2008b) Phytoplankton classes sensitive to eutrophication. Aquat Ecol. doi:10.1007/s10452-008-9181-z

Refsgaard JC, Nilsson B, Brown J, Klauer B, Moore R, Bech T, Vurro M, Blind M, Castilla G, Tsanis L, Biza P (2005) Harmonised techniques and representative river basin data for assessment and use of uncertainty information in integrated water management (HarmoniRiB). Environ Sci Policy 8:267–277

Refsgaard JC, Jørgensen LF, Højberg AL (2007) Data availability and accessibility. State of the art on existing data required for modelling for research purposes and for the implementation of the Water Framework Directive: Geological Survey of Denmark and Greenland

Schafer RB, Caquet T, Siimes K, Mueller R, Lagadic L, Liess M (2007) Effects of pesticides on community structure and ecosystem functions in agricultural streams of three biogeographical regions in Europe. Sci Total Environ 382:272–285

Schartau AK, Moe SJ, Sandin L, McFarland B, Raddum G (2008) Macroinvertebrate indicators of lake acidification: testing on data from UK, Norway and Sweden. Aquat Ecol. doi:10.1007/s10452-008-9186-7

Schmidt-Kloiber A, Nijboer R (2004) The effect of taxonomic resolution on the assessment of ecological water quality classes. Hydrobiologia 516:269–284

Schmidt-Kloiber A, Graf W, Lorenz A, Moog O (2006) The AQEM/STAR taxalist—a pan-European macro-invertebrate ecological database and taxa inventory. Hydrobiologia 566:325–342

Schriever CA, Liess M (2007) Mapping ecological risk of agricultural pesticide runoff. Sci Total Environ 384: 264–279

Verdonschot P (2006a) Data composition and taxonomic resolution in macroinvertebrate stream typology. Hydrobiologia 566:59–74

Verdonschot P (2006b) Evaluation of the use of Water Framework Directive typology descriptors, reference sites and spatial scale in macroinvertebrate stream typology. Hydrobiologia 566:39–58

Verdonschot P, Moog O (2006) Tools for assessing European streams with macroinvertebrates: major results and conclusions from the STAR project. Hydrobiologia 566: 299–310

Vlek H, Šporka F, Krno I (2006) Influence of macroinvertebrate sample size on bioassessment of streams. Hydrobiologia 566:523–542