Hydrobiologia (2010) 644:1–13
DOI 10.1007/s10750-010-0108-z

REVIEW PAPER

Data integration for European marine biodiversity research: creating a database on benthos and plankton to study large-scale patterns and long-term changes

Leen Vandepitte · Bart Vanhoorne · Alexandra Kraberg · Natalie Anisimova · Chryssanthi Antoniadou · Rita Araújo · Inka Bartsch · Beatriz Beker · Lisandro Benedetti-Cecchi · Iacopo Bertocci · Sabine Cochrane · Keith Cooper · Johan Craeymeersch · Epaminondas Christou · Dennis J. Crisp · Salve Dahle · Marilyse de Boissier · Mario de Kluijver · Stanislav Denisenko · Doris De Vito · Gerard Duineveld · Vincent Escaravage · Dirk Fleischer · Simona Fraschetti · Adriana Giangrande · Carlo Heip · Herman Hummel · Urszula Janas · Rolf Karez · Monika Kedra · Paul Kingston · Ralph Kuhlenkamp · Maurice Libes · Peter Martens · Jan Mees · Nova Mieszkowska · Stella Mudrak · Ivka Munda · Sotiris Orfanidis · Martina Orlando-Bonaca · Rune Palerud · Eike Rachor · Katharina Reichert · Heye Rumohr · Doris Schiedek · Philipp Schubert · Wil C. H. Sistermans · Isabel Sousa Pinto · Alan J. Southward · Antonio Terlizzi · Evagelia Tsiaga · Justus E. E. van Beusekom · Edward Vanden Berghe · Jan Warzocha · Norbert Wasmund · Jan Marcin Weslawski · Claire Widdicombe · Maria Wlodarska-Kowalczuk · Michael L. Zettler

Received: 28 August 2009 / Revised: 4 January 2010 / Accepted: 18 January 2010 / Published online: 4 February 2010
© Springer Science+Business Media B.V. 2010

Abstract The general aim of setting up a central database on benthos and plankton was to integrate long-, medium- and short-term datasets on marine biodiversity. Such a database makes it possible to analyse species assemblages and their changes on spatial and temporal scales across Europe. Data collation lasted from early 2007 until August 2008, during which 67 datasets were collected covering three divergent habitats (rocky shores, soft bottoms and the pelagic environment). The database contains a total of 4,525 distinct taxa, 17,117 unique sampling locations and over 45,500 collected samples, representing almost 542,000 distribution records. The database geographically covers the North Sea (221,452 distribution records), the North-East Atlantic (98,796 distribution records) and furthermore the Baltic Sea, the Arctic and the Mediterranean. Data from 1858 to 2008 are presented in the database, with the longest time-series from the Baltic Sea soft bottom benthos. Each delivered dataset was subjected to certain quality control procedures, especially on

Electronic supplementary material The online version of this article (doi:10.1007/s10750-010-0108-z) contains supplementary material, which is available to authorized users.

The first three authors were responsible for the coordination of writing this article; the first two authors have taken the lead on this article. Other authors have mainly contributed by making their data available and they are mentioned in alphabetical order.
D. J. Crisp and A. J. Southward are deceased.

Handling editor: T. P. Crowe

L. Vandepitte (✉) · B. Vanhoorne · J. Mees · E. Vanden Berghe
Vlaams Instituut voor de Zee, Wandelaarkaai 7, 8400 Oostende, Belgium
e-mail: leen.vandepitte@vliz.be

A. Kraberg · K. Reichert
Alfred Wegener Institute for Polar and Marine Research, Biologische Anstalt Helgoland, P.O. Box 180, 27483 Helgoland, Germany

the level of taxonomy. The standardisation procedure enables pan-European analyses without the hazard of taxonomic artefacts resulting from different determination skills. A case study on rocky shore and pelagic data in different geographical regions shows a general overestimation of biodiversity when making use of data before quality control compared to the same estimations after quality control. These results prove that the contribution of a misspelled name or the use of an obsolete synonym is comparable to the introduction of a rare species, having adverse effects on further diversity calculations. The quality checked data source is now ready to test geographical and temporal hypotheses on a large scale.

Keywords Macrobenthos · Plankton · Data acquisition · Quality control · Biogeography

Introduction

Globally, there is an increasing need to measure marine biodiversity and to quantify the rate at which it is changing (e.g. Gaston, 2000; Seys et al., 2004; Zeller et al., 2005). Additionally, there is a need to explore how diversity on a small scale is related to that on larger scales, as the knowledge of the role of patterns and processes at different scales forms the basis to understand global variation in biodiversity (Lawton, 1996; Gaston, 2000).

Different initiatives have aimed to map marine biodiversity by bringing together small to medium scale datasets so far mostly scattered throughout the scientific landscape. These initiatives pay special attention to capturing datasets and the corresponding metadata that have never been published, as there is an imminent danger that they will disappear from scientific memory (Zeller et al., 2005). Large temporal and spatial scale biological datasets are one of the most important tools in studying and understanding long-term distributions and abundances of marine life and their evolution. Since these are scarce, integrating and managing scattered biological data from local datasets into a central database are an alternative way to meet the need for data and information on a broad scale and to support global decision-making (Grassle, 2000; Seys et al., 2004). Such data compilations have never been of greater importance: climate change is altering marine systems at an unprecedented rate and the

N. Anisimova
PINRO, 6 Knipovich Street, Murmansk 183038, Russia

C. Antoniadou
Department of Biology, Laboratory of Zoology, Aristotle University of Thessaloniki, P.O. Box 134, 54124 Thessaloniki, Greece

R. Araújo · I. S. Pinto
Centre of Marine and Environmental Research, University of Porto, Rua do Bragas 298, 4050-123 Porto, Portugal

I. Bartsch · E. Rachor
Alfred Wegener Institue for Polar and Marine Research, AM Handelshafen 12, 27570 Bremerhaven, Germany

B. Beker · M. de Boissier · M. Libes
Centre National de la Recherche Scientifique, Centre d'Océanologie de Marseille, Station Marine d'Endoume, Rue de la Batterie des Lions, 13007 Marseille, France

L. Benedetti-Cecchi · I. Bertocci
Dipartimento di Biologia, Università di Pisa, Via Derna 1, 56126 Pisa, Italy

S. Cochrane · S. Dahle · R. Palerud
Akvaplan-Niva, Polar Environmental Centre, 9296 Tromsø, Norway

K. Cooper
Centre for Environment, Fisheries and Aquaculture Science, Pakefield Road, Lowestoft NR33 OHT, UK

J. Craeymeersch
Wageningen IMARES, Institute for Marine Resources and Ecosystem Studies, Korringaweg 5, 4400 AB Yerseke, The Netherlands

E. Christou
Hellenic Centre for Marine Research, 19013 Anavissos, Attika, Greece

D. J. Crisp · N. Mieszkowska · A. J. Southward
Marine Biological Association of the UK (MBA), The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK

M. de Kluijver
Expert Center For Taxonomic Identification (ETI), Mauritskade 61, 1092 AD Amsterdam, The Netherlands

S. Denisenko
Zoological Institute of the Russian Academy of Science, St. Petersburg, Russia

establishment of such large-scale integrated databases and subsequent analyses can help to document and explain the broad-scale spatial and temporal patterns in biodiversity and give scientists the opportunity to explore their implications (Gaston, 2000; Costello & Vanden Berghe, 2006; Vanden Berghe et al., 2007; Vandepitte et al., 2009). As large-scale investigations are a prerequisite to verify the influence of regional patterns and processes on populations and to determine to what extent small-scale patterns and processes can be generalised (Lawton, 1996; Fraschetti et al., 2005), they can also assist in profoundly documenting the implications and consequences of scale-dependent spatial heterogeneity (Gaston, 2000; Fraschetti et al., 2005; Terlizzi et al., 2007; Orlando-Bonaca et al., 2008). In addition, the establishment of large-scale integrated databases and informatics-supported analyses has allowed to unravel the global nature of various phenomena (Costello & Vanden Berghe, 2006) such as, e.g. the long-term changes in phytoplankton distribution in the Baltic Sea (Wasmund & Uhlig, 2003; Suikkanen et al., 2007) or the range shifts and northward migrations of several marine species (Mieszkowska et al., 2006; Philippart, 2007). A comparable exercise on European scale has recently been carried out using soft bottom benthos data in the MacroBen database (Vanden Berghe et al., 2009). Renaud et al. (2009) used this centralised MacroBen database to check the latitudinal clines for the first time on marine ecosystems. The uniqueness of the available data within this MacroBen database covering a latitudinal area from 36° to 81° north made it possible to demonstrate that the latitudinal clines present in terrestrial and limnic systems could not be detected within marine ecosystems. On the other hand, Webb et al. (2009) were able to prove that macro-ecological patterns from terrestrial ecology could also be found in marine ecosystems.

The development of the LargeNet—*Large-scale and long-term networking on the observation of Global Change and its impact on Marine Biodiversity*—integrated database to assess long-term changes in biodiversity and their possible causes, was funded within the EU-FP6 Marine Biodiversity and Ecosystem Functioning Network of Excellence (MarBEF NoE) which serves as a platform for the integration of

D. De Vito · S. Fraschetti · A. Giangrande · A. Terlizzi
Department of Biological and Environmental Science and Technologies, Laboratory of Zoology and Marine Biology (LZMB), University of Salento, Strada Provinciale Monteroni, 73100 Lecce, Italy

G. Duineveld · V. Escaravage · C. Heip · H. Hummel · W. C. H. Sistermans
Netherlands Institute of Ecology (NIOO), P.O. Box 140, 4400 AC Yerseke, The Netherlands

D. Fleischer · H. Rumohr · P. Schubert
Leibniz Institute for Marine Sciences, IFM-GEOMAR, Duesternbrooker Weg 20, 25105 Kiel, Germany

C. Heip
Royal Netherlands Institute for Sea Research (NIOZ), Landsdiep 4, PB 59, AB Den Burg, Texel, The Netherlands

U. Janas · S. Mudrak
Institute of Oceanography, University of Gdansk, Al. Pisudskiego 46, 81-378 Gdynia, Poland

R. Karez
State Agency for Agriculture, Environment and Rural Areas (LLUR), Hamburg Chaussee 25, 24220 Flintbek, Germany

M. Kedra · J. M. Weslawski · M. Wlodarska-Kowalczuk
Institute of Oceanology, Polish Academy of Sciences, Powstancow Warszawy 55, 81-712 Sopot, Poland

P. Kingston
Institute for Offshore Engineering, Heriot-Watt University, Edinburgh, Scotland, UK

R. Kuhlenkamp
Phycomarin, Hamburg, Germany

P. Martens · J. E. E. van Beusekom
Alfred Wegener Institue for Polar and Marine Research, Wadden Sea Station Sylt, Hafenstrasse 43, 25992 List/Sylt, Germany

I. Munda
Scientific Research Centre of the Slovenian Academy of Sciences and Arts, Novi trg 2, 1000 Ljubljana, Slovenia

S. Orfanidis · E. Tsiaga
National Agricultural Research Foundation, Fisheries Research Institute, 640 07 Nea Peramos, Kavala, Greece

M. Orlando-Bonaca
Marine Biology Station, National Institute of Biology, Fornace 41, 6330 Piran, Slovenia

D. Schiedek
National Environmental Research Institute, University of Aarhus, Frederiksborgvej 399, P.O. Box 358, 4000 Roskilde, Denmark

interdisciplinary marine research and disseminates knowledge on marine biodiversity to scientific, policy and end-user communities. LargeNet was one of the smaller research projects implemented within MarBEF. The specific aims of LargeNet were to integrate long-term datasets on marine biodiversity, so that species assemblages and their changes over a hierarchy of temporal and spatial scales—including latitudinal and longitudinal gradients—across Europe could be analysed and related to the variability and trends of the hydro-climatic environment. Subsequently, hypotheses should be formulated to reveal the causes and consequences of these changes in an ecosystem context allowing the development of new concepts and research approaches. This in order to detect and compare similar trends driven by global change in different regions of Europe.

This article describes the data management aspects of the project, its data policy, database architecture and functionalities. It also gives a brief overview of the content of the integrated database and a case study to underscore the importance and relevance of standardisation and quality control procedures. Currently, about 20 scientists from marine institutes across Europe are actively working with the integrated database, preparing at least six collaborative scientific papers. The results of these joint analyses will be published elsewhere (e.g. Terlizzi et al., 2009).

Data collation, management and availability

Data collation for the LargeNet project lasted from early 2007 until August 2008. During this period, 19 institutes provided 67 datasets covering three divergent habitats: (1) rocky shores, (2) soft bottoms and (3) the pelagic environment. All these datasets contained data and information on the spatial distribution of macrobenthos or plankton, obtained from a large number of small- to medium-scale studies.

Every contributing dataset was archived and described at the data centre of the Flanders Marine Institute (VLIZ). Describing each component dataset made it possible to create a searchable inventory which facilitated querying and sharing information. This metadata—or data explaining the data—gives a thorough description of the content of each dataset and is freely available on the MarBEF website (http://www.marbef.org/projects/largenet/data.php). To keep track of the datasets and to fully document them, the 'Integrated Marine Information System' (IMIS) was utilised (Cattrijsse et al., 2006). To ensure the continued existence of all collected data and associated metadata, the Marine Data Archive (MDA) was deployed (Claus et al., 2008).

When contributing data to the LargeNet project, data custodians agreed to accept the rules and agreements described in the LargeNet Declaration of Mutual Understanding (LDMU), also referred to as the LargeNet data policy. The full policy is online available http://www.marbef.org/projects/largenet/docs/DMULargeNet.doc. For an overview of the most important aspects of this policy, we refer to Box 1.

Box 1 Important highlights of the LargeNet data policy

All datasets remain the property of their data providers

Each contributing dataset undergoes quality assurance and quality control procedures, performed by the data management team

The data management team cannot distribute the delivered datasets to a third party without the explicit permission of the data contributor, unless stated otherwise in the freely available metadata

Co-authorship of the data provider(s) in all produced written scientific documents each time (part of) their dataset is used is an irrevocable right

Each data provider has the right to participate in joint analyses on the central database

Contributing datasets can be connected to EurOBIS, a long-term online repository for biogeographical data in Europe, thereby greatly increasing the visibility of the research

It is recommended to make the delivered datasets publicly available for two reasons:

Present Address:
E. Vanden Berghe
Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, NJ 08901, USA

J. Warzocha
Department of Fisheries Oceanography and Marine Ecology, Sea Fisheries Institute, Kollataja 1, 81-332 Gdynia, Poland

N. Wasmund · M. L. Zettler
Leipniz Institute for Baltic Sea Research, Seestrasse 15, 181 19 Rostock Warnemünde, Germany

C. Widdicombe
Plymouth Marine Laboratory (PML), Prospect Place, West Hoe, Plymouth, PL13 DH, UK

Box 1 continued

Contacting the original data provider(s) becomes harder as time passes: people change jobs or retire

Colleague-scientists can benefit from previous research and could come to new findings when combining or comparing their work with this formerly collected data and information

If the data are part of ongoing research such as, e.g. a PhD programme, a moratorium period of five years can be established, giving reasonable time to process the data and publish related findings.

In the process of data collation, 13 paper-based datasets have been digitised. Although digitising and standardising historical datasets was rather time consuming, it was ultimately very useful as the retrieved data greatly extended the temporal scope for analysis and might reveal some new scientific insights when combined with more recent data from the same area.

For a correct citation of each contributing dataset and link to the metadata description, we refer to Appendix I—Supplementary material.

Data standardisation and quality assurance

When bringing together several datasets from different sources and collected for various purposes and under very diverse circumstances, it is essential to harmonise these datasets towards the integration process. Data standardisation of the received datasets was done on three levels: (1) taxonomy, (2) geography and (3) units. Figure 1 gives an overview of the complete quality assurance algorithm.

All taxonomic names were matched against the European Register of Marine Species (ERMS), an authoritative list of taxa occurring in the European marine environment (Costello et al., 2001). Abundance information for four datasets was semi-quantitative, expressed as AFCOR-coding after Crisp & Southward (1958) and this was converted to presence–absence values. In all cases, the originally delivered data remain available within the database. Biomass expressed as ash-free dry weight (AFDW) was not available for all records, but where not provided, this was—where possible—calculated using conversion factors provided by The Netherlands Institute of Ecology (NIOO) (Sistermans & Hummel, 2009). All these operations made it possible to rationally compare the contributing datasets.

LargeNet database

Data model

The central LargeNet database was developed in MSAccess. The relational database contains eight different tables (Fig. 2) and its model is based on the MacroBen (Vanden Berghe et al., 2009) and MANUELA database (Vandepitte et al., 2009), both developed at the Flanders Marine Institute within the MarBEF framework.

The ‘meta’ table includes the dataset name, data providing institute, contact person, the broader geographical range and an indication of the nature of the sampled habitat. Checkboxes declare if the dataset contains abundances, biomass data and/or environmental readings. The ‘stations’ table reports on the exact location of the collected samples, whereas the ‘samples’ table gives more detail on sampling date, sampled area, equipment used, replicates and sampling depth. Depth information was split up in minimum and maximum depth, making it possible to capture the depth range of pelagic samples. If depth was fixed, this was assigned to the minimum depth field. In the ‘abundance’ table quantitative, semi-quantitative and qualitative distribution information is stored. A distinction between percent coverage values and actual counts was made, to avoid confusion in further calculations. For pelagic taxa, it was possible to indicate if the identified taxon was either auto-, hetero- or mixotrophic. Biomass values were listed in the ‘biomass’ table, expressed in wet weight and ash free dry weight. The ‘species’ table contains the originally received taxon names, combined with the valid taxon names from the European Register of Marine Species (ERMS). Environmental or abiotic information was stored in two separate tables: (1) abiotic parameters, defining the variable, its unit and a short description, and (2) abiotic readings, containing the actual values per parameter.

All datasets were delivered to the Flanders Marine Institute (VLIZ) as Excel sheets or Access databases. Each dataset was converted or adjusted to a separate Access database with a fixed structure on which

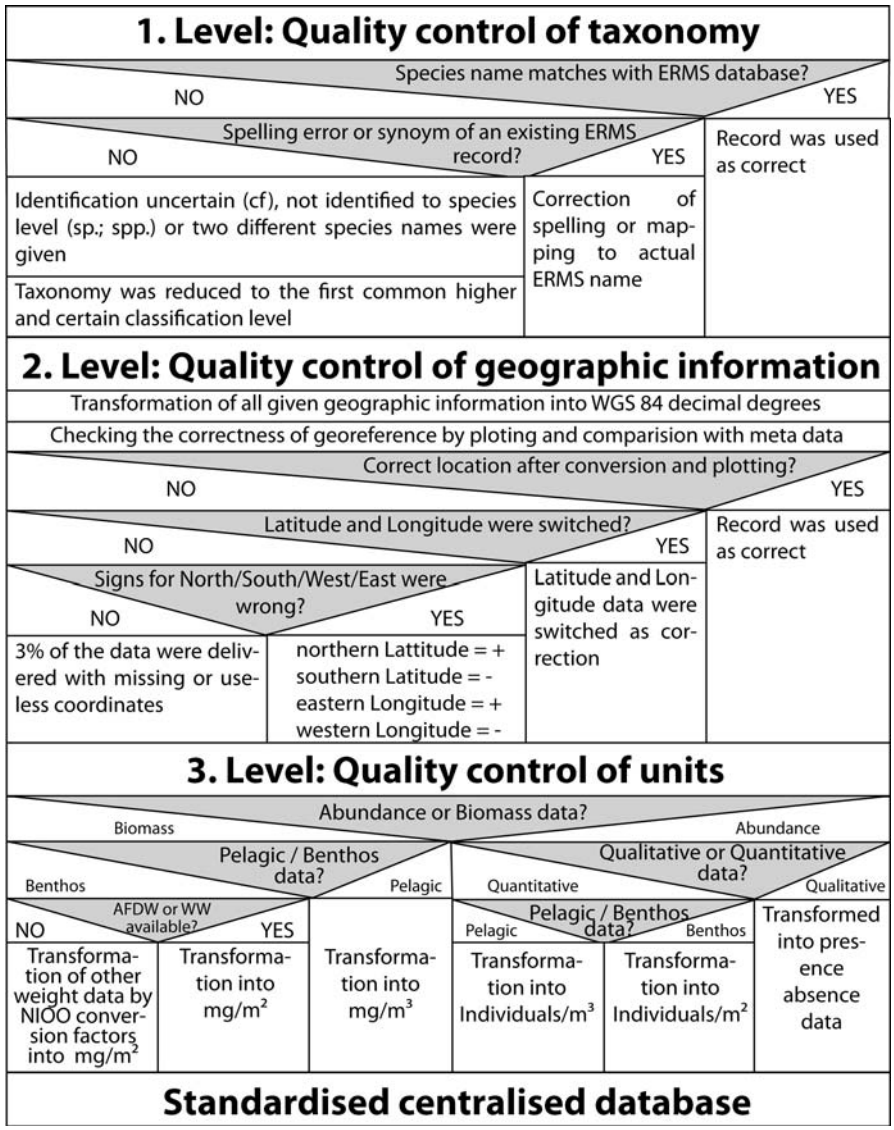

Fig. 1 Nassi-Shneiderman diagram (NSD) for the quality assurance algorithm applied to every record within the LargeNet database. All these tests were necessary to achieve the general comparability between the individual datasets

quality control procedures were performed. Subsequently, the separate datasets were uploaded into the central LargeNet database.

Functionalities

A number of built-in tools has been provided to help the user analyse (bespoke subsets of) the data. The user can select datasets, exclude certain sampling methodologies and define the desired temporal and/or spatial boundaries. Furthermore, it is possible to limit the data to a certain taxon or taxonomic rank. Keeping

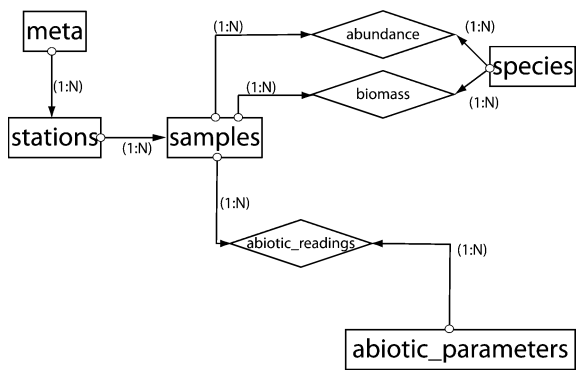

Fig. 2 Data model of the LargeNet database

in mind the presence of both complete and incomplete species identifications, it is possible to lump taxa to a certain taxonomic level (family, genus or species) prior to the analysis (see also Vandepitte et al., 2009). Data can be reduced to presence–absence information and/or utilised to analyse a certain life-stage (adults versus non-adults). Information on life-stage and replicates can be pooled. Taking into account the impact of rare taxa on certain analyses and their interpretation, such taxa can be excluded. As the collected data are expressed either in percentage cover or actual counts, one can also select the nature of the retained data (only percentage cover, only counts or a combination of both) depending on the research question to be answered.

Once the desired data matrix has been composed, densities can be calculated. As sampling sizes in the collected datasets were quite diverse, an option was provided to adjust the sampling size to the needs of the analyses, making it possible to re-calculate the original counts to densities varying from individuals per 0.05 m^2 to individuals per 1 m^2. Caution is necessary, however, as an extrapolation of data collected at a given resolution to another scale assumes linearity, and this assumption may be violated depending on the scale of aggregation of organisms. As organisms are not distributed randomly in space, a scale transition procedure that assumes linearity can affect the magnitude of differences amongst samples and variance estimates (Krebs, 1998; Benedetti-Cecchi, pers. comm.). Minor changes should, however, not result in a biased view.

Using density values, a number of taxonomic and diversity indices frequently used in marine macrobenthology studies can be calculated: Shannon's diversity index (H'; Shannon and Weaver, 1949), Simpson's diversity index (D; Simpson, 1949), Hill's numbers (N_1, N_2, N_∞; Hill, 1973), Margalef's diversity index (Dm; Margalef, 1958) and Hurlbert's diversity index for 50 individuals (ES50; Hurlbert, 1971). Hurlbert's diversity index is subsequently included in the calculation of the Benthic Quality Index, used to assess the benthic environmental quality (BQI; Rosenberg et al., 2004). Based on the ERMS taxonomic tree, it is also possible to calculate the following indices describing taxonomic diversity and distinctness on the selected data: Δ, Δ^*, Δ^+ and Λ^+ (Clarke & Warwick, 1998, 1999, 2001).

To apply more advanced statistical analysis techniques, it is possible to export the data selections from the database to commonly used data formats for Twinspan, Primer, PcOrd or the R package.

Results

General content

The 67 collected datasets represent 17,117 unique sampling locations. The exact geographic coordinates were available for 97% of these stations (Fig. 3), whereas only information on the broader geographical range could be recovered for the remaining ones. Over 45,500 samples were gathered, representing almost 542,000 distribution records. Table 1 provides a detailed overview of the number of datasets and distribution records per geographical area and habitat. One dataset contained only environmental data from the North Sea and the Baltic area and will further be referred to as 'environmental dataset'. Biomass information was available in 11 datasets, representing 163,028 records, the main portion coming from soft bottom datasets (8 datasets, 110,392 records). Besides biological data, 12 datasets also contained abiotic information: 13,096 abiotic readings were available from both the water column and the sediment. Most of these measurements (5,716) were related to the water temperature in the upper water layer (SST) and mainly originated from the environmental dataset.

Although originally only eight datasets did not represent a continuous time series, temporal discontinuities became visible when combining several datasets to the habitat and geo-region level (Table 2). Taking into account the aim of this integrated database—exploring temporal changes in species assemblages—this aspect has to be given due consideration in more detailed long-term trend analyses.

In total, 6,172 unique taxon names were submitted to LargeNet. After a thorough quality control, however, this number was reduced to 4,525, mostly due to spelling variations and synonymy. Such quality control is highly needed, since a misspelled or obsolete name could be compared to the introduction of a rare species, with adverse effects on further (biodiversity) calculations.

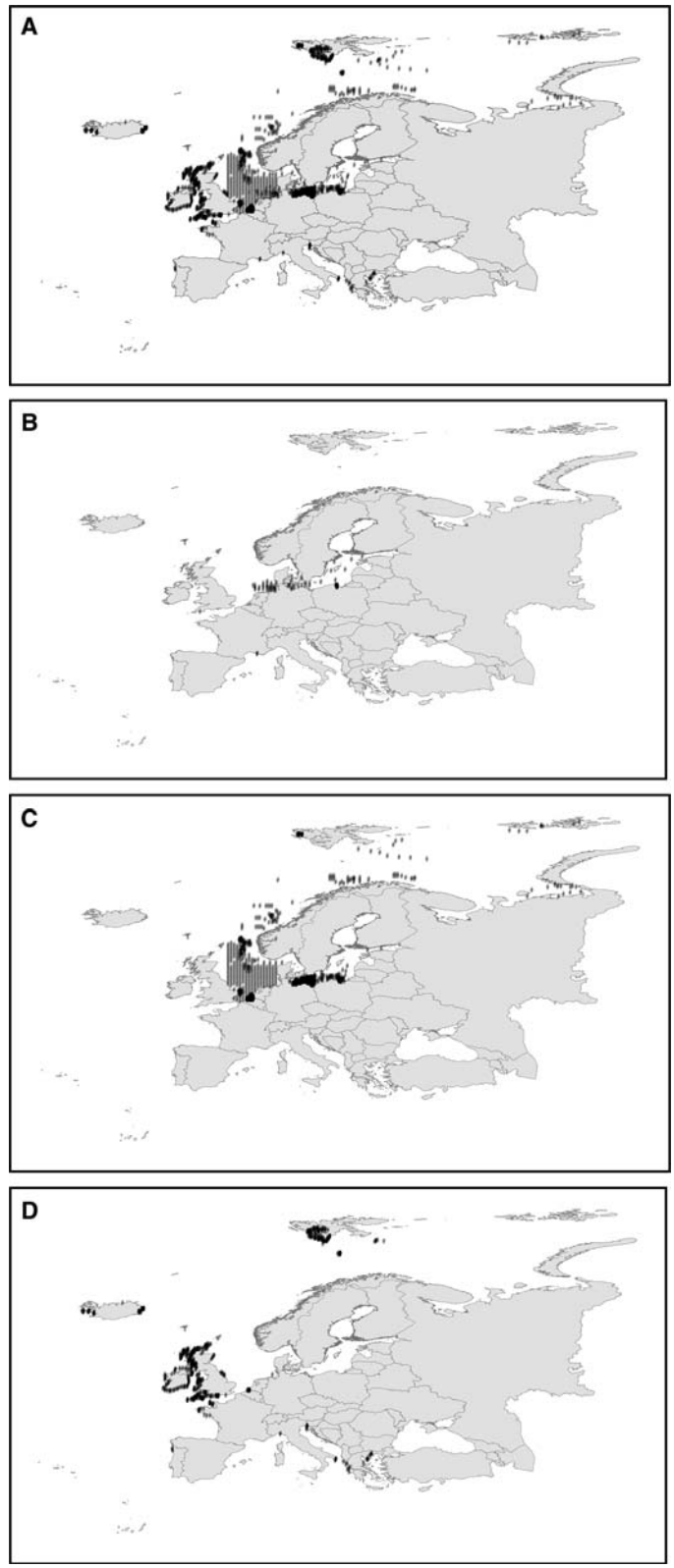

Fig. 3 Overview of sampling stations available in the LargeNet database. **A** All sampling locations, including environmental stations; **B** pelagic sampling locations; **C** soft bottom sampling locations; and **D** rocky shore sampling locations

Table 1 Number of datasets and distribution records in relation to the larger defined geographical areas (geo-regions)

Geo-region	Rocky shores			Soft bottoms			Pelagic		
	Datasets	Distribution records		Datasets	Distribution records		Datasets	Distribution records	
		#	%		#	%		#	%
Arctic	6	8,828	6.5	7	20,350	8.2	–	–	–
Baltic	–	–	–	13	22,899	9.2	3	65,549	41.7
North-East Atlantic	7	40,199	29.5	–	–	–	1	49,597	31.5
North Sea	3	39,884	29.2	3	148,494	59.9	2	33,073	21.0
Meditterranean	14	47,516	34.8	1	2,004	0.8	4	9,111	5.8
Mixed[a]	–	–	–	2	54,336	21.9	–	–	–
Total	30	136,427		26	248,083		10	157,330	

[a] Mixed indicates that samples have been collected in more than one region, being the North-East Atlantic and the Arctic. Samples for the environmental dataset were collected in the North Sea and the Baltic area (not shown). Allocation of datasets into geo-regions is based on the VLIZ Marine Gazetteer (VLIMAR, available at http://www.vliz.be/vmdcdata/vlimar/)

Table 2 Sampling range per geo-region and per habitat

Geo-region	Rocky shores		Soft bottoms		Pelagic	
	Range	Gaps	Range	Gaps	Range	Gaps
Arctic	1965–1997	24	1992–2006	4	–	–
Baltic	–	–	1858–2007	49	1979–2007	None
North-East Atlantic	1948–2007	45	–	–	1988–2007	None
North Sea	2004–2008	None	1974–2004	8	1975–2006	7
Mediteranean	1967–2006	28	2005	None	1988–2007	None
Mixed	–	–	1990–2002	None	–	–
Total	1948–2008	31	1858–2007	49	1975–2007	2

Range is expressed as the first and the last year sampled within the habitat and geo-region; 'gaps' indicates the number of years within the range that were not sampled. 'None' indicates that samples are present for all the years within the defined range. The dataset only containing environmental data was collected between 1861 and 2005 and represented an uninterrupted time-series (not in table)

Case study—calculating diversity indices

The validation of taxonomic species names is the most important procedure preceding the comparison of two independent datasets. It is most likely to have spelling errors and synonyms as variations in the distinct species list (Alroy, 2002). To demonstrate this, a case study was performed on the pelagic and rocky shore data in different geographical areas. Diversity indices were calculated using all originally delivered taxon names. These calculations used the original species identifications, thus including all combinations and derivatives with "sp.", affinities and doubtful determinations. The same exercise was then repeated with the species names retained after quality control, which passed the first level of quality assurance (see Fig. 1). The results of this exercise are presented in Table 3.

As expected, all indices show a higher diversity in each geographical area for the originally provided species names compared to those indices retained after quality control. Uncertain identifications, synonyms or misspelled names unintentionally increase species richness. To the knowledge of the authors, a study dealing with the artificial effects of species lists from combined datasets has so far not yet been published. Table 3 shows the overestimated index results such as Shannon's (H') and Simpons's (1-D) diversity indices or Hurlbert's diversity index for 50 individuals (ES50). As taxonomic precision is somewhat reduced by the quality control, the true diversity will be slightly underestimated. It is expected that the

Table 3 Diversity indices for rocky shore and pelagic data, per geographic region

	Species names before quality control					Species names after quality control				
	# Species	# Rare species	H'	$1 - D$	ES50	# Species	# Rare species	H'	$1 - D$	ES50
Rocky shore data										
ANE	219	15	4.63602	0.98777	38.11	187	11	4.45772	0.98509	36.25
Arctic	646	69	6.00024	0.99666	46.33	378	44	5.38261	0.99403	43.67
Mediteranean	1,120	238	5.74091	0.99342	43.35	834	159	5.49015	0.99105	41.74
North Sea	251	29	4.50662	0.98424	35.89	163	25	3.95956	0.97469	30.14
Pelagic data										
ANE	288	7	4.90740	0.98913	39.59	180	4	4.33821	0.97818	33.79
Baltic	592	94	4.95148	0.98361	38.13	483	82	4.76476	0.98216	37.13
Mediteranean	420	103	4.98571	0.98772	39.42	249	66	4.40238	0.97717	34.24
North Sea	118	15	3.41447	0.95754	23.20	64	9	2.06743	0.79005	10.80

Species = number of distinct species; # Rare species = number of distinct species with only 1 distribution record; H' = Shannon's diversity index; $1 - D$ = Simpson's diversity index; ES(50) = Hurlbert's diversity index for 50 individuals. ANE = North-East Atlantic

true diversity for the larger geographical areas will be situated between the values calculated for the originally provided species names and the species names after quality control.

Based on this analysis, diversity is the lowest in the North Sea with an expected number of species (ES50), ranging between 30 and 36 for the rocky shores and 11 and 23 for pelagic species. Largest differences between diversity indices before and after quality control are also found for this region, but differences appear in each studied area. The largest reduction in number of distinct species and number of rare species after taxonomic matching appeared in both the rocky shore and pelagic data from the Mediterranean, indicating that these data had the most diverse spelling variations, incomplete identifications and synonyms. This might be attributed to the large number of people and institutes who provided data from this area.

Data sharing beyond LargeNet

To date, the distributional information for the macrobenthos and plankton of 27 datasets has been made publicly available through EurOBIS, representing 314,914 distribution records. EurOBIS is an online, freely accessible long-term repository for biogeographical data in Europe that has been developed under the MarBEF umbrella (http://www.eurobis.org). This website integrates multiple datasets containing biogeographic information on marine organisms in Europe and can be explored through a dynamic search-interface. It acts as the European node of OBIS, the Ocean Biogeographic Information System (Costello et al., 2005). EurOBIS passes all its harvested distribution records onto OBIS, where they are stored together with marine data from all over the world in an online and freely accessible system. OBIS in its turn makes its data available to GBIF—the Global Biodiversity Information System (http://www.gbif.org)—which gathers biodiversity data from all over the planet, both marine and terrestrial.

LargeNet data available in (Eur)OBIS are less detailed than those available in the integrated database itself, e.g. presence values rather than actual counts. The harmonised database is still under a moratorium period and cannot yet be released to third parties. The LargeNet data that have been made available in EurOBIS can, however, be freely downloaded and used through the EurOBIS website, under the condition they are properly cited when used.

Conclusions

Collecting data from different sources and bringing them together in one central database was a time-consuming task and presented some significant challenges. The inclusion of historical data in the

LargeNet database provided a valuable contribution to the need for pristine reference conditions for comparison reasons. As all data have been quality controlled and standardised to European or international standards, scientists can rely on the correctness of all taxon names and measurement units in the database (Vandepitte et al., 2009). A centralised database like the one presented here is also a quality assurance for the contributing scientists and their field and lab work. The database also provides researchers with the opportunity to compare data on larger geographical and longer temporal scale than they can do themselves due to limitations in, e.g. infrastructure, time and money (Costello & Vanden Berghe, 2006; Renaud et al., 2009; Vandepitte et al., 2009). Moreover, the collected datasets were carefully selected based on their sampled habitat, region and timeframe and the way they could help answering the questions and hypotheses within the LargeNet project. This way, a very specific and focalised integrated database was created to serve as a working platform for scientists. Hence, they save time and effort searching for complementary data and therefore they can spend more time analysing data and testing hypotheses pertinent to current biodiversity issues, which can then be published in literature. This approach of collating specific datasets into an integrated, focussed and quality controlled database differs from other—more globally oriented—data collection activities. The latter cannot always guarantee thorough quality control, mostly due to the enormous amount and variety of data they receive. This imposes constraints on performing analyses on such data and can raise questions concerning the validity and standardisation. Moreover, the desired data or information may not even be present in the database.

Databases which integrate existing data also form an irreplaceable complement to (newly started) long-term monitoring activities, as they represent the only way to expand these recent time-series with their historical counterparts. The combination of historical data with more recent data makes it possible to perform reliable long-term trend analyses (Vanden Berghe et al., 2007; Vandepitte et al., 2009) and it can provide a fundamental baseline that assists in countering the 'shifting baseline syndrome' (Zeller et al., 2005). This allows a thorough assessment of the past and current biodiversity situation. The outcome of such analyses can then be utilised to inform decision makers on global issues such as global environmental change (Mieszkowska et al., 2005), the consequences of overfishing (Costello & Vanden Berghe, 2006), bio-security risks from the introduction of alien species (McNeely, 2001) or the (possible) causes of alternating abundances of key-species within a certain habitat.

Although compiling, exploring and analysing existing datasets cannot be seen as a replacement for carefully planned innovating research, its importance should not be underestimated. This process not only gives the data a second life, but can provide a better understanding of the large-scale patterns which might be slumbering in more locally oriented small-scale datasets (Vanden Berghe et al., 2007). As data were collected through a variety of sampling methods, one should be careful when comparing and interpreting the data. The creation of a more restricted dataset comprising stations that were sampled with the same sampling gear and comparable sample area, however, overcomes this issue and facilitates scientifically sound comparisons across datasets.

The success of compiling an integrated database ultimately depends on the willingness of scientists to share their data. In the case of LargeNet, data sharing is made beneficial by (1) offering co-authorship to each data provider every time (part of) their data is used in publications based on the integrated LargeNet database and (2) by guaranteeing that the data providers remain owner of their data and that they can decide independently if their data can be used by a third party or not. Despite this, many scientists are still reluctant to share their data. The most commonly quoted motive not to share data is they want to further explore their collected data in additional work—in the absence of any competition—followed by the belief that they will encounter insurmountable logistical difficulties when sharing or exchanging data (Parr & Cummings, 2005). It should be realised, however, that sharing data increases their value in time (Costello & Vanden Berghe, 2006) and—in proportion to their use by others—it can lead to more publications, a greater importance of the performed research and an increase in the visibility of the researchers and their institutes within their field of expertise (Parr & Cummings, 2005 and references therein; Costello, 2009). In this light, one should also consider the value and uniqueness of collected data:

recently collected data might seem unimportant or uninteresting at this time, but may turn out to be very valuable in the future (Zeller et al., 2005). Replicating the original conditions from, e.g. yesterday or ten or more years ago remains impossible, which again emphasises the need to safeguard data—especially historical data—and stresses the role they can play in long-term trend analyses.

The LargeNet initiative has brought together a wide variety of researchers from different countries and different fields of expertise and this has led to a very diverse central database in taxonomic coverage, space and time. The LargeNet integrated database is one of the most comprehensive databases on combined benthos and plankton data and currently holds the largest amount of rocky shore datasets ever integrated in Europe. The scientists involved in the project hope this initiative will attract other researchers to cooperate and share their data in the future.

Acknowledgements The first author would like to thank all LargeNet scientists for their patience in answering the many questions concerning the submitted datasets. The LargeNet project has been carried out in the framework of the MarBEF Network of Excellence 'Marine Biodiversity and Ecosystem Functioning' which is funded by the Sustainable Development, Global Change and Ecosystem Programme of the European Community's Sixth Framework Programme (Contract No. GOCE-CT-2003-505446). This publication is contribution number 09041 of MarBEF.

References

Alroy, J., 2002. How many named species are valid? Proceedings of the National Academy of Sciences of the United States of America 99: 3706–3711.

Cattrijsse, A., S. Claus, T. D'haenens, J. Haspeslagh, R. T'Jampens, E. Vanden Berghe & L. Vandepitte, 2006. IMIS Integrated Marine Information System Input Manual version 1.0 December 2006. Flanders Marine Institute (VLIZ), Oostende, Belgium.

Clarke, K. R. & R. M. Warwick, 1998. A taxonomic distinctness index and its statistical properties. Journal of Applied Ecology 35: 523–531.

Clarke, K. R. & R. M. Warwick, 1999. The taxonomic distinctness measure of biodiversity: weighting of step lengths between hierarchical levels. Marine Ecology Progress Series 184: 21–29.

Clarke, K. R. & R. M. Warwick, 2001. A further biodiversity index applicable to species lists: variation in taxonomic distinctness. Marine Ecology Progress Series 216: 265–278.

Claus, S., A. Vanhoorne, J. Mares, L. Vandepitte, K. Deneudt & F. Hernandez, 2008. Manual on How to Use the Marine Data Archive (MDA). Version 1.0. Flanders Marine Institute (VLIZ), Oostende, Belgium.

Costello, M. J., 2009. Motivating online publication of data. Bioscience 59: 418–427.

Costello, M. J. & E. Vanden Berghe, 2006. 'Ocean biodiversity informatics': a new era in marine biology research and management. Marine Ecology Progress Series 316: 203–214.

Costello, M. J., C. Emblow & R. White (eds), 2001. European Register of Marine Species: a check-list of the marine species in Europe and a bibliography of guides to their identification. Collection Patrimoines Naturels, 50. Muséum national d'Histoire naturelle, Paris, France. ISBN 2-85653-538-0.

Costello, M. J., J. F. Grassle, Y. Zhang, K. Stocks & E. Vanden Berghe, 2005. Where is what, and what is where? Online mapping of marine species. MarBEF Newsletter 2: 20–22.

Crisp, D. J. & A. J. Southward, 1958. The distribution of intertidal organisms along the coast of the English Channel. Journal of the Marine Biological Association of the United Kingdom 37: 157–208.

Fraschetti, S., A. Terlizzi & L. Benedetti-Cecchi, 2005. Patterns of distribution of marine assemblages from rocky shores: evidence of relevant scales of variation. Marine Ecology Progress Series 296: 13–29.

Gaston, K. J., 2000. Global patterns in biodiversity. Nature 405: 220–227.

Grassle, J. F., 2000. The Ocean Biogeographic Information system (OBIS): an on-line, worldwide atlas for accessing, modelling and mapping marine biological data in a multidimensional geographic context. Oceanography 13: 5–9.

Hill, M. O., 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54: 427–431.

Hurlbert, S. H., 1971. The non-concept of species diversity: a critique and alternative parameters. Ecology 52: 577–586.

Krebs, C. J., 1998. Ecological methodology. Addison Wesley Longman, Menlo Park, California.

Lawton, J. H., 1996. Patterns in ecology. Oikos 75: 145–147.

Margalef, R., 1958. Information theory in ecology. General Systems 3: 36–71.

McNeely, J. A., 2001. An introduction to human dimensions of invasive alien species. In McNeely, J. A. (ed.), The Great Reshuffling: Human Dimensions of Invasive Alien Species. IUCN Publishers, Gland, Switzerland: 5–22.

Mieszkowska, N., M. A. Kendall, S. J. Hawkins, R. Leaper, P. Williamson, N. J. Hardman-Mountford & A. J. Southward, 2006. Changes in the range of some common rocky shore species in Britain – a response to climate change? Hydrobiologia 555: 241–251.

Mieszkowska, N., R. Leaper, P. Moore, M. A. Kendall, M. T. Burrows, D. Lear, E. Poloczanska, K. Hiscock, P. S. Moschella, R. C. Thompson, R. J. Herbert, D. Laffoley, J. Baxter, A. J. Southward & S. J. Hawkins, 2005. Marine biodiversity and climate change: assessing and predicting the influence of climatic change using intertidal rocky shore biota. Occasional publications. Marine Biological Association of the United Kingdom 20: 53.

Orlando-Bonaca, M., L. Lipej & S. Orfanidis, 2008. Benthic macrophytes as a tool for delineating, monitoring and assessing ecological status: the case of Slovenian coastal waters. Marine Pollution Bulletin 56: 666–676.

Parr, C. S. & M. P. Cummings, 2005. Data sharing in ecology and evolution. Trends in Ecology and Evolution 20: 362–363.

Philippart, C. J. M. (ed.), 2007. Impacts of Climate Change on the European Marine and Coastal Environment: Ecosystems Approach. ESF Marine Board Position Paper 9. European Science Foundation, Marine Board, Strasbourg, France.

Renaud, P. E., T. J. Webb, A. Bjorgesaeter, I. Karakassis, M. A. Kendall, C. Labrune, N. Lampadariou, P. J. Somerfield, M. Wlodarska-Kowalczuk, E. V. Berghe, S. Claus, I. F. Aleffi, J. M. Amouroux, K. H. Bryne, S. J. Cochrane, S. Dahle, S. Degraer, S. G. Denisenko, T. Deprez, C. Dounas, D. Fleischer, J. Gil, A. Gremare, U. Janas, A. S. Y. Mackie, R. Palerud, H. Rumohr, R. Sarda, J. Speybroeck, S. Taboada, G. Van Hoey, J. M. Weslawski, P. Whomersley & M. L. Zettler, 2009. Continental-scale patterns in benthic invertebrate diversity: insights from the MacroBen database. Marine Ecology and Progress Series 382: 239–252.

Rosenberg, R., M. Blomqvist, H. C. Nilsson, H. Cederwall & A. Dimming, 2004. Marine quality assessment by use of benthic species-abundance distributions: a proposed new protocol within the European Union Water Framework Directive. Marine Pollution Bulletin 49: 728–739.

Seys, J., P. Pissierssens, E. Vanden Berghe & J. Mees, 2004. Marine data management: we can do more, but can we do better? Ocean Challenge 13: 20–24.

Shannon, C. E. & W. Weaver, 1949. The Mathematical Theory of Communication. University of Illinois Press, Urbana.

Simpson, E. H., 1949. Measurement of diversity. Nature 163: 688.

Sistermans, W. C. H. & H. Hummel, 2009. Conversion factors for lengths and weights of soft sediment macrozoobenthos in the south-west Netherlands. Monitor Taskforce reports 2009-01, NIOO reports February 2009.

Suikkanen, S., M. Laamanen & M. Huttunen, 2007. Long-term changes in summer phytoplankton communities of the open northern Baltic Sea. Estuarine, Coastal and Shelf Science 71: 580–592.

Terlizzi, A., M. J. Anderson, S. Fraschetti & L. Benedetti-Cecchi, 2007. Scales of spatial variation in Mediterranean subtidal sessile assemblages at different depth. Marine Ecology Progress Series 332: 25–39.

Terlizzi, A., M. J. Anderson, S. Bevilacqua, S. Fraschetti, M. Wlodarska-Kowalczuk & K. E. Ellingsen, 2009. Beta diversity and taxonomic sufficiency: do higher-level taxa reflect heterogeneity in species composition? Diversity and Distributions 15: 450–458.

Vanden Berghe, E., S. Claus, W. Appeltans, C. Arvanitidis, P. Somerfield, I. F. Aleffi, J. M. Amouroux, N. Anisimova, G. Bachelet, S. Cochrane, M. J. Costello, J. Craeymeersch, S. Dahle, S. Degraer, S. Denisenko, C. Dounas, G. Duineveld, C. Emblow, V. Escaravage, M.-C. Fabri, D. Fleischer, A. Grémare, M. Herrmann, H. Hummel, I. Karakassis, M. Kędra, M. Kendall, P. Kingston, L. Kotwichi, C. Labrune, J. Laudien, H. Nevrova, A. Occhipinti, F. Olsgard, R. Palerud, A. Petrov, E. Rachor, N. Revkov, H. Rumohr, R. Sardá, W. C. H. Sistermans, J. Speybroeck, U. Janas, G. Van Hoey, M. Vincx, P. Whomersley, W. Willems, M. Włodarska-Kowalczuk, A. Zenetos, M. L. Zettler & C. Heip, 2009. Macroben integrated database on benthic invertebrates of European continental shelves: a tool for large-scale analysis across Europe. Marine Ecology Progress Series 382: 225–238.

Vanden Berghe, E., H. L. Rees & J. D. Eggleton, 2007. North Sea Benthos Project 2000 Data Management. ICES Committee Meetings Documents CM 2007(A:18). ICES, Copenhagen, Denmark.

Vandepitte, L., J. Vanaverbeke, B. Vanhoorne, F. Hernandez, T. Nara Bezerra, J. Mees & E. Vanden Berghe, 2009. The MANUELA database: an integrated database on meiobenthos from European marine waters. Meiofauna Marina 17: 35–60.

Wasmund, N. & S. Uhlig, 2003. Phytoplankton trends in the Baltic Sea. ICES Journal of Marine Science 60: 177–186.

Webb, T. J., I. F. Aleffi, J. M. Amouroux, G. Bachelet, S. Degraer, C. Dounas, D. Fleischer, A. Gremare, M. Herrmann, H. Hummel, I. Karakassis, M. Kedra, M. A. Kendall, L. Kotwicki, C. Labrune, E. L. Nevrova, A. Occhipinti-Ambrogi, A. Petrov, N. K. Revkov, R. Sarda, N. Simboura, J. Speybroeck, G. Van Hoey, M. Vincx, P. Whomersley, W. Willems & M. Wlodarska-Kowalczuk, 2009. Macroecology of the European soft sediment benthos: insights from the MacroBen database. Marine Ecology Progress Series 382: 287–296.

Zeller, D., R. Froese & D. Pauly, 2005. On losing and recovering fisheries and marine science data. Marine Policy 29: 69–73.