Collaborative Project (large-scale integrating project)
Grant Agreement 226273
Theme 6: Environment (including Climate Change)
Duration: March 1st, 2009 – February 29th, 2012

SEVENTH FRAMEWORK PROGRAMME

WISER

Water bodies in Europe: Integrative Systems to assess Ecological status and Recovery

DELIVERABLE

Deliverable 2.2-2: Guidelines for indicator development

Lead contractor: University of Duisburg-Essen (UDE)

Contributors: Daniel Hering, Sebastian Birk (UDE), Anne Lyche Solheim, Jannicke Moe (NIVA), Laurence Carvalho (NERC), Angel Borja (AZTI), Peter Hendriksen, Dorte Krause-Jensen, Torben Lauridsen, Martin Sondergaard (AU), Didier Pont (CEMAGREF), Richard Johnson (SLU), Agnieszka Kolada (IEP), Gwendolin Porst (IGB), Nuria Marba (CSIC), Peeter Noges, Ingmar Ott (EMU), Joao Carlos Marques (IMAR), Ken Irvine (TCD), Alberto Basset (USALENTO)

Due date of deliverable: Month 12

Actual submission date: Month 12

Project co-funded by the European Commission within the Seventh Framework Programme (2007-2013) Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Content

Non-technical summary

This guidance provides “cook books” for the development of common metrics and assessment systems to be applied for different Biological Quality Elements and water types. It is for internal use within the WISER project and might in a later stage be extended by best practise examples and be made available to the Geographical Intercalibration Groups.

The first purpose of the guidance is to develop common metrics, i.e. common yardsticks (“international currencies”) against which national assessment systems can be compared. The common metrics which will be developed by the WISER project will support the intercalibration process for the Water Framework Directive. Due to the strict time schedule of the intercalibration exercise the common metrics must be based on preliminary data evaluation; this guidance outlines the procedure to ensure that common metrics will be developed in a comparable way for different organism groups (Biological Quality Elements) and water types.

Second, the guidance outlines a methodology for developing assessment systems. This methodology has several commonalities with the common metric development, but is based on a more sophisticated data evaluation.

Guidelines for indicator development

Document history (and future)

- Version 1 produced by Daniel Hering and Sebastian Birk (UDE) by 15/12/09.
- Commented by the members of the WISER Steering Group by 4/1/10.
- Version 2 produced by UDE based on these comments.
- Commented by the relevant workpackage leaders and workpackage scientists (WP3.1 to 4.4) (by 31/1/10), individually:
  - WP3.1 (Laurence Carvalho)
  - WP3.2 (Agnieszka Kolada, Peeter Noges, Ingmar Ott, Martin Sondergaard, “Katrit”)
  - WP3.3 (Ken Irvine, Gwendolin Porst)
  - WP3.4 (Torben Lauridsen)
  - WP4.1 (Peter Hendriksen)
  - WP4.2 (Dorte Krause-Jensen, Nuria Marba)
  - WP4.3 (Angel Borja, Joao Carlos Marques, Alberto Basset)
- Version 3 produced by UDE based on these comments. Most of the minor comments have been taken on board and not been documented in detail. For the major comments an overview was compiled listing whether or not the comments were included and the resulting changes (see Annex).

Introduction

The development of WFD-compliant assessment systems is a pivotal aim of WISER. Assessment systems (often referred to as “classification systems”) translate biological information of a water body to an ecological status class (ranging from high status to bad status). Within the WISER project assessment systems will be developed for different water types (lakes, transitional and coastal waters) and different Biological Quality Elements (BQEs). The development of assessment systems is part of Modules 3 (lakes; workpackages 3.1-3.4) and Module 4 (coastal and transitional waters; workpackages 4.1-4.4).

Phytoplankton, macrophytes, macroalgae and angiosperms, benthic invertebrates and fish are sampled with different methods and devices and the resulting data are thus differently structured; there are also differences in data generated for lakes and transitional and coastal waters. Some differences among assessment systems developed in WISER are unavoidable owing to the individual requirements of the Biological Quality Elements (BQE) or water types. However, certain features of the development process and, thus, of the resulting assessment systems should be similar and provide a harmonized WISER assessment methodology to be adopted. Wherever possible, the process for developing assessment systems, therefore, needs to be harmonized and applied in a similar way by the workpackages within Modules 3 and 4. All WISER assessment systems will be based on metrics, either as single metrics or as multimetric indices. A “metric” is defined as a measurable part or process of a biological system empirically shown to change in value along a gradient of human influence (Karr and Chu 1999). It reflects specific and predictable responses of the biological community to human activities, either to a single impact factor or to the cumulative effects of multiple human impairments within a catchment. Metrics address comparable ecological aspects of a community, regardless of the stressor they are responding to.

Another important aim of WISER is to support the intercalibration process. The guidelines for the second phase of the intercalibration process are now finalized (Schmedtje et al. 2009), and they include a strict time plan. One of the first steps is to derive “common metrics”, i.e. biological measures created for benchmarking[1] and comparison of national assessment systems. The WISER workpackages 3.1 to 4.4 have agreed to support the development of common metrics and to suggest a first set of common metrics by end of March 2010. As final and validated results will not be available by then, the development of common metrics will, necessarily, be based on preliminary data evaluation and expert knowledge. Also the process of developing common metrics needs to be harmonized among WISER workpackages. In this context it must be clearly stated that common metrics are not meant as pan-European assessment systems replacing national methods, which are usually much better adapted to the regional situation. Common metrics are a common yardstick for comparing national assessment systems and their classification of ecological status.

Consequently, the aims of this guidance are twofold: (1) to guide and harmonize the rapid and preliminary development of common metrics by March 2010; and (2) to guide and harmonize the development of assessment methodologies among the relevant WISER workpackages. The guidance is structured accordingly, with one chapter dealing with common metrics and one with assessment systems. Each chapter covers criteria of the methods to be developed (e.g. applicability, statistical features), the development process (e.g. data sources and statistical methods to be used) and a brief description of the envisaged product. While the guidance strives for a harmonized approach it still allows for flexibility; it is generally difficult to transform biota and their response to stress into simple numbers and, therefore, different problems will appear for the individual Biological Quality Elements and water types.

The two main chapters overlap considerably. They represent “cook books” for slightly different purposes and we strived for a complete description of each procedure within a single chapter which can be applied without consulting other chapters or documents.

This guidance is mainly for internal use within the WISER project. After a practical test within the WISER consortium it might in future be extended by “best practise” examples and be made available to the Geographical Intercalibration Groups (GIG).

Brief introduction to the intercalibration process

The European Union Water Framework Directive (WFD) commits the EU member states to achieve good ecological status of all surface waters. The ecological status is classified by evaluating diverse parameters of the BQE. The member states have developed biological

[1] Definition of trans-national (absolute) reference points in intercalibration based on data from near-natural reference sites or sites impacted by similar levels of impairment.

assessment systems by which the water bodies are classified. In the intercalibration exercise these national methods are harmonized to ensure that good status is consistent with the directive's requirements and comparable among countries. The intercalibration exercise sets the common level of ambition to protect and restore water bodies across Europe.

The intercalibration process is organized separately for each water category: rivers, lakes, coastal and transitional waters. Within each category countries that share regions of similar biogeographical setting belong to a Geographical Intercalibration Group (GIG). The intercalibration exercise is performed between countries that hold similar water types, i.e. the broadly defined common intercalibration types, within a GIG. The GIGs set up BQE expert groups composed of national specialists that carry out the technical intercalibration work. In the current intercalibration phase the entire exercise comprises more than 50 expert groups that intercalibrate about 400 biological assessment methods for almost 70 common intercalibration types of 28 European countries.

In the intercalibration exercise the differences between national assessment methods are attributable to three domains:

- Data acquisition, i.e. the field sampling and sample processing to yield biological information;
- Numerical evaluation, i.e. the selection and combination of biological metrics used;
- Classification, i.e. the quality rating depending on reference definition and boundary setting.

These domains represent subsequent steps in the national assessment process and are critically important for successful intercalibration (Figure 1).

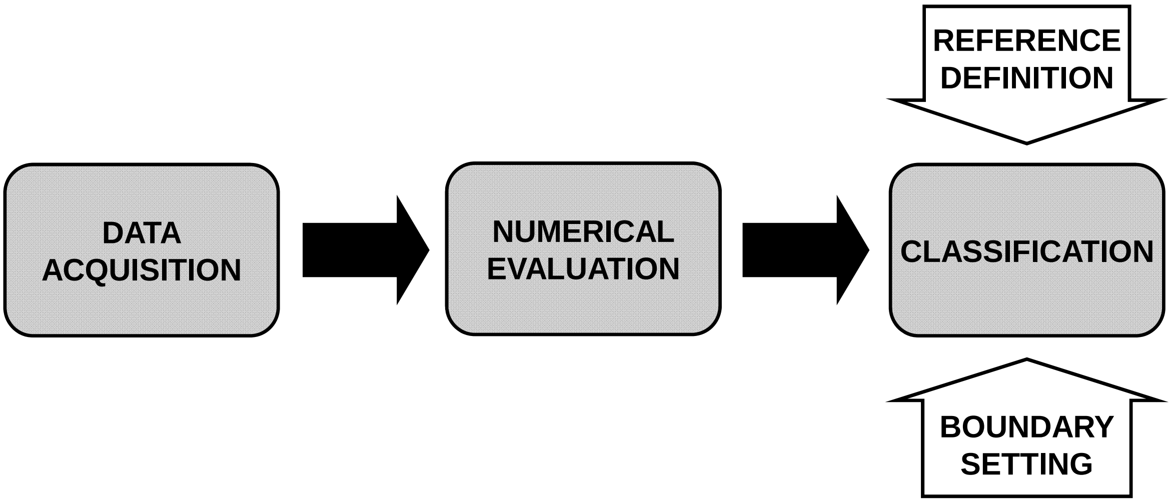

Figure 1: Main elements of national assessment methods relevant in the intercalibration process.

Intercalibration aims at harmonizing the classification of good ecological status. Discrepancies often result from differences in the national data acquisition and numerical evaluation. Here, the Guidance on the Intercalibration Process (Schmedtje et al. 2009) prescribes two intercalibration options: (1) The use of common metrics in case of differing data acquisition and numerical evaluation; and (2) direct comparison if the numerical evaluation is different, but similar data acquisition allows for a combined analysis of national data. For the latter, the use of common metrics is a supporting approach if the biogeographical differences are large. Schmedtje et al. (2010) explicitly state: "Common metrics can be used as 'international currencies' to which common boundary setting (including harmonized reference definition) and the GIG-wide descriptions of reference and 'borderline' conditions can be related […] . The[ir] ecological relevance further enhances the transparency of the intercalibration process." Among assessment methods that are conceptually different, or focus on dissimilar pressures or water type's intercalibration cannot be accomplished. In these cases the guidance requests the use of alternative approaches such as on-site comparisons, i.e. comparing the classification results of the various national methods applied to the same water bodies. A calibration of national class boundaries against comparable gradients of pressure may also be a feasible option.

The first phase of intercalibration was completed in 2008 with the results (i.e. harmonized national boundaries of good ecological status) published by the European Commission (2008). In the first phase the following BQE's were intercalibrated at least for some indicative parameters: phytoplankton, benthic invertebrates, angiosperms and macroalgae in coastal waters, benthic invertebrates and diatoms in rivers, macrophytes and phytoplankton in lakes. Major steps of the ongoing phase of intercalibration (2008-2011) comprise the development of common metrics (due by April 2010), the definition of benchmarks for intercalibration (due by October 2010) and the harmonisation of national class boundaries (due by April 2011).

Development of draft common metrics

Criteria

Common metrics play a prominent role in the WFD intercalibration exercise, as they are the principal mean of comparing the results of national assessment methods. The term "common metric" was first used by Buffagni et al. (2005) in their proposal to harmonize the national classification schemes for river invertebrates. They defined the Intercalibration Common Metric (ICM) as "A biological metric widely applicable within a Geographical Intercalibration Group (GIG) or across GIGs, which can be used to derive comparable information among different countries/stream types." Following this definition the important characteristics of common metrics are their wide geographical applicability for comparability of national assessment methods. Common metrics, thus, hold the pertinent properties of indices for bioassessment (e.g. Hering et al. 2006, Breine et al. 2007) while meeting the specific requirements of the intercalibration process (Schmedtje et al. 2009). We specify the main features of common metrics as follows:

Compliance with ecological concepts and legal requirements

Common metrics quantify the structural or functional attributes of biological communities, allowing for an assessment of ecological quality. They need to be based on ecological concepts, such that there is a rationale why a metric increases or decreases with the degradation of a water body. The WFD specifies various biological parameters for the ecological status classification of individual BQEs (Table 1). Common metrics derived from taxonomic as well as non-taxonomic data shall cover all required parameters combined to a multimetric index to allow for the intercalibration of the full BQE (Schmedtje et al. 2009), but may also reflect single parameters only, if the BQE currently cannot be fully intercalibrated. In the following the term "common metric" also relates to common multimetric indices.

Table 1: Indicative parameters to be included in biological assessment methods for the surface water categories and BQEs ([a] or depth distribution/cover for macroalgae and angiosperms, [b] only lakes, [c] bioaccumulation-bioassays). The table gives an overview of the normative definitions in the WFD and of the parameters mentioned in the CIS Guidance No 7 - Monitoring (WG 2.7) (optional issues are within brackets) (from Schmedtje et al. 2009).

Surface Water Category	Biological Quality Element	Taxonomic and functional composition	Abundance [a]	Disturbance sensitive taxa	Diversity	Age structure	Frequency and intensity of algal blooms	Biomass	Absence of major taxonomic groups	Taxa indicative of pollution
Rivers and Lakes	Phytoplankton	x	x				x	x[b]		
	Macrophytes and Phytobenthos	x	x							
	Benthic invertebrate fauna	x	x	x	x				x	
	Fish fauna	x	x	x		x				
Transitional Waters	Phytoplankton	x	x				x	x		
	Macroalgae	x	x							
	Angiosperms	x	x							
	Benthic invertebrate fauna	x	x	x	x					x
	Fish fauna	x	x	x						(x[c])
Coastal Waters	Phytoplankton	x	x		(x)		x	x		
	Macroalgae and Angiosperms		x	x	(x)					
	Benthic invertebrate fauna	x	x	x	x			(x)		x

Relationship to national methods

Common metrics need to relate to the results of the national assessment methods used in the particular intercalibration exercise: there should be a high correlation between the common metrics and each national method addressed in the intercalibration exercise. Criteria for this relationship which have been specified for the linear regression analysis using common multimetric indices are (Kelly et al. 2009, Schmedtje et al. 2009): Coefficient of determination (R^2) $\geq$ 0.5, root mean square error $\leq$ 0.15, slope of the regression line $\geq$ 0.5 and $\leq$ 1.5 (standardized values' scale such as EQR). Furthermore, the relationships should be inspected visually for heteroscedasticity, i.e. an inhomogeneous variance of the residuals that can simply be observed in the residuals' plot.

Robustness

In bioassessment, “robust” metrics reflect the effects of the stressor to be assessed (i.e. signal, for example the impact of eutrophication) while other sources of variability (i.e. noise, for example caused by natural variability or sampling effects) should have a relatively minor impact. However, the level of noise differs between water categories and types and is, for instance, relatively high for transitional waters that show large natural variability. The common metrics used in intercalibration show specific features of robustness: they are calculated from heterogeneous data sources, and, thus, need to be sufficiently robust across space (e.g. typological and biogeographical differences), time (e.g. seasonality, interannual variability), and scope of data acquisition (field sampling and sample processing).

A common metric is robust in **space** if the gradients of biological quality are equally well reproduced along the whole region where the metric is applied, and the indicator taxa on which the metric is based are present and hold similar bioindicative values across water types and geographical regions.

A common metric is robust in **time** if the effects of seasonality or interannual variability on the metric results are low. Stoddard et al. (2008) used the ratio of the variance among all sites in a large dataset (i.e. signal s) to the variance of repeated visits to the same sites (i.e. noise n) to quantify the reproducibility of metrics (s/n ratio). A ratio of s/n ≤ 1, for instance, indicates that visiting a single site twice yields as much metric variability as visiting two different sites. The authors used a ratio of s/n ≥ 2 to select candidate metrics for an US-wide assessment index using river invertebrates. Temporally robust metrics are less prone to seasonal dynamics caused by patterns of migration or emergence. The evaluation of robustness using the s/n ratios generally needs to take into account the intrinsic spatial-temporal heterogeneity of the different water types. And if the BQE shows strong spatial and/or temporal variability (e.g. phytoplankton) the common metric may be water type-specific and/or calculated from fixed sampling season data only.

A common metric is robust regarding the **scope of biological data** if it is fully applicable to the data that countries acquire in their monitoring programmes, i.e. the national sampling design yields all information relevant for calculating the common metric. The selection of common metrics shall account for the details of field sampling like the choice of sampling devices (e.g. mesh size), the sampled habitats or the time of survey. The metric should further be robust against aspects of sample processing, such as recording of abundance and level of taxonomic resolution. Common metrics consider the least common denominator of available data among countries. In the intercalibration exercise of river invertebrates, for instance, the common metric was operated at family level data (van de Bund 2009).

Sensitivity

Common metrics must respond to the stressor (or combination of stressors) addressed. The response may be positive (value increases with stressor intensity) or negative. If only a limited data source is available the response should be monotonous, i.e. constantly increasing or decreasing across the gradient of stress.

Process of developing common metrics

Due to the tight intercalibration schedule the identification of common metrics by the relevant WISER workpackages needs to be performed in a very short period of time, in many cases before the workpackage databases have been finalized. The common metrics suggested will, therefore, be preliminary and only partly based on data evaluation; in addition, literature data, expert judgement, and ecological theory and modelling will be used. The process should be composed of the following steps:

Step 1: Setting

Before selecting common metrics some background data must be gathered:

- For which GIGs is the intercalibration exercise in question relevant?
- Which stressor(s) will be addressed?
- Which national assessment methods are already available in the individual GIGs?

This information can be obtained from the method overview of the WISER workpackage 2.2 (Birk 2010a, 2010b) and the scientists responsible for the particular intercalibration exercises.

From the national assessment methods that are part of the relevant intercalibration exercise a common denominator of biological information will be defined. For example, if some assessment methods operate at species level and some at family level, the common metric needs to be defined at family level. However, the implications on the quality and precision of the assessment need to be evaluated thoroughly.

Step 2: Identification of candidate metrics

From the data gathered in the first step “setting”, from literature and using expert judgement, a list of candidate metrics will be compiled. Metrics will be selected according to the metric types (“indicative parameters”) listed in Table 1 (see also Hering et al. 2006). For each “indicative parameter” several metrics will be selected. Priority will be given to metrics which have a proven indicative potential of the stressors addressed, which are easy to calculate and/or which are already included into national assessment methods (see WISER overview for a list of national metrics). The common denominator in terms of biological information (see step 1) will be followed. Candidate metrics should also span a sufficiently wide range of values along a stressor gradient. For example, proportional metrics ranging from 0.1% to 0.5% or the number of sensitive taxa ranging from 0 to 3 along the whole stressor gradient should be avoided.

Step 3: Testing the relationship of candidate metrics and national assessment methods

The candidate metrics shall be related to all metrics (or the overall multimetric index) used in the national assessment methods. Ideally, these national metrics belong to the group of candidate metrics (see step 2). Common metrics should be well correlated with the complete national methods with correlation coefficients > 0.7 (coefficient of determination ≥ 0.5), i.e. almost half of the variability of the national method is explained by the common metric.

Step 4: Testing the relationship of candidate metrics and stress gradients

The relation of candidate metrics to stress gradients will be documented from literature and, where possible, be tested with preliminary versions of the relevant WISER workpackage databases or other relevant databases.

From the literature, data on the following will be documented for each metric on which data are available:

- Geographical region and water type
- Short description of the data source
- Explanatory (i.e. pressure) variables reflecting stress intensity, against which the metric was tested
- Strength of correlation / regression and statistical method used

The overview on national assessment methods collated in WISER also provides information on dose-response-relationships.

To test the relationship of candidate metrics to stress gradients with preliminary versions of the WISER workpackage databases or other relevant databases the following steps will be taken:

- Definition of explanatory (i.e. pressure) variables reflecting stress intensity
- Allocation of site groups based on classes of stress intensity, e.g. stressed and unstressed
- Metric calculation
- Quantitative relationship: Correlation or regression analysis (e.g. Spearman Rank correlation) between metrics and stress variables
- Qualitative relationship: Testing for differences in the metric distributions between groups of stressed and unstressed sites (visual: box-whisker-diagrams, statistical: t- or U-test)
- Estimation of Type I and II errors based on assessment of groups of stressed and unstressed sites

Step 5: Testing the robustness of candidate metrics

Most aspects of metric robustness can be judged from the literature and expert knowledge, e.g. including the validity of the indicator across regions. To evaluate the spatial robustness the candidate metrics will be applied to databases covering large geographical areas. The dose-response relationships will be analysed per geographical region and water type, or analysed by a model that accounts for natural environmental variability. If for some regions or types no significant correlation is found the spatial robustness of the metric is questionable. The WISER methods' overview may additionally be consulted to obtain information about the regions and water types where metrics are applied.

Robustness in time will be tested using data from sites sampled twice or more. The s/n ratio quantifies the metric variability between seasons or years at a site in relationship to the overall variability of the entire dataset. The ratio is then compared among metrics to select suitable common metrics with the highest ratio. This may require knowledge of the pressure history and evolution to avoid that changes in time from metrics are not due to changes in pressure.

If the biological data were acquired very differently, it is necessary that the quality gradient is reflected consistently, i.e. the dose-response patterns are not significantly different. This can be analysed by testing if the dose-response relationship of the common metric is comparable among the differing datasets (e.g. regression model is similar). In these cases, normalizing the common metric with reference / benchmark values derived from a homogeneous dataset (i.e.

data sampled with similar technique) as described, for instance, in Birk and Hering (2009) may not guarantee that in the intercalibration analyses the boundary values of the ecological status classes are sufficiently comparable.

Step 6: Metric selection

From as many metric types (“indicative parameter”; Table 1) as possible metrics will be selected for the common multimetric index. Usually this will be the metric with the highest correlation to stress gradients and/or with the highest degree of robustness and, ideally, with a high correlation to selected national assessment methods. In case two or more metrics of the same type perform similarly well, the metric with the lowest correlation to other metrics of the multimetric index will be selected. Note: This step is not applicable if the investigators aim for a single metric.

Step 7: Normalisation of metrics

The normalisation of common metrics is a crucial step of intercalibration, as it establishes the harmonized basis for comparing the national class boundaries. In the intercalibration exercise normalisation is carried out against common benchmarks based on reference sites or alternative approaches. Both require the identification of the natural “background” conditions of the water type, characterised by low levels of man-made stress without any anthropogenic impact on the biological parameters. Harmonised criteria to define these reference conditions for the intercalibration exercise are currently established for the various water categories in the cross-GIG activity on reference conditions (see Pardo et al. 2010). These criteria are intended to allow for screening “true” reference sites, and, a sufficient number of reference sites provided (e.g. n=10), their biological parameters are used to set the type-specific reference values (e.g. median of common metric distribution) and the reference variability (standard deviation of common metric distribution). Furthermore, the data on “true” reference sites can be analysed for possible biogeographical differences among GIG-regions and for variances in the national data acquisition.

If no or only very few sampling stations meet the reference criteria, the normalisation of common metrics is carried out against alternative benchmarks. Depending on the available data sampling stations in least disturbed conditions (LDC, Stoddard et al. 2006) are defined. Their actual distance from “true” reference sites should be quantified, for instance by a Principal Component Analysis that integrates all available pressure variables and “virtual” reference sites, i.e. sites not existing in reality but conceived as the potential components that should be present (Borja et al. 2004). These data can be based upon experience gained of the area, or the very few “true” reference sites still existing. The actual distance of the LDC sites from the virtual references allows evaluating the quality status of the available sampling stations in terms of their level of pressure. The biological parameters of these sites are used to establish biological benchmarks for intercalibration, i.e. the condition of the biological community that represents the trans-national reference point for harmonization (Birk and Hering 2009). Note: this reference point is different from the reference concept given by the WFD, i.e. the undisturbed, near-natural conditions. These can be derived from extrapolating the dose-response relationship (e.g. combined pressure gradient versus common metric) to the biological values at virtual reference sites, or predicting the reference values of the common metric in a multiple regression analysis using the individual pressures as independent variables (using the thresholds defined in the reference criteria).

Step 8: Combination to a multimetric index

Unless there is a good rationale, no weighting factors will be applied to enhance the influence of single component metrics over others. In this case the multimetric index will be calculated as the mean of all normalised metrics. If one metric represents more than one “indicative parameter” it may be weighted stronger. The same applies if one metric has a much stronger relationship to the stress gradients and / or national methods than others. In general, the multimetric index is more sensitive and robust than the single metrics.

Step 9: Documentation

A form on the resulting multimetric index and the process to develop it will be completed (for details see next section).

Product

The product will be a multimetric index ideally composed of one metric per metric type (“indicative parameters”, see Table 1). The common metrics need to be documented in a standard way. On a standard form, which will be placed on the WISER website and be communicated to the relevant GIGs, the following will be recorded:

- Intercalibration exercise addressed (e.g. macrophytes in lakes)
- Geographical intercalibration group addressed (e.g. Central-Baltic GIG)
- Main stressor addressed (e.g. eutrophication)
- Common metric suggested
    - Components of the common metric (i.e. individual metrics) and how they relate to the indicative parameters (Table 1)
    - Definition of the individual metrics, if relevant including formula to calculate the metric
    - Definition of the common metric, if relevant including formula to calculate the metric
    - Excel spreadsheet or other software used for calculation, if relevant and available
    - Online sources for further information
- Relationship of the common metric to the stressor addressed
    - Description of the data source used to derive the dose-response relationship
    - Statistical method (e.g. linear regression, generalised linear model, generalised additive model, ...)
    - Explanatory variables used (e.g. total phosphorus or chlorophyll-a in the case of eutrophication)
    - Statistical parameters describing the relationship of the common metric and the stressors (e.g. regression coefficients, effect size)
- Justification of the selected common metric (why it is better than possible alternative common metrics in case any such exist)
- List of assessment methods, which will participate in the intercalibration exercise

  - Name of the method (e.g. to be taken from the database established in the WISER workpackage 2.2)
  - Relationship between the common metric and the national methods: description of the data source; scatterplot of national method against common metric; coefficient of determination; slope of regression line; root mean square error; check for heteroscedasticity

Development of assessment systems

The WISER suite of assessment systems will be developed in eight workpackages (3.1 to 4.4). The individual requirements of the different BQEs and water types, which are also reflected by the workpackage descriptions in the project's Description of Work, demand for a certain degree of freedom in the guidelines. The following criteria and steps, however, need to be applied in a harmonized way, to obtain comparable products within the project.

The assessment methodologies developed in WISER could be the same as the common metrics suggested by the project (see previous chapter). In this case they need to be refined, amended and validated according to the criteria and steps outlined below. Alternatively, different assessment systems could be suggested; in this case a comparison with the common metrics should be performed.

The criteria and steps of development outlined below are based on Borja and Dauer (2008), Breine et al. (2007), Carstensen (2007), Hering et al. (2006), Herlihy et al. (2008), Pont et al. (2006, 2009) and Stoddard et al. (2008).

Criteria

The WISER suite of assessment systems need to comply with the following criteria:

- The methods must be **type-specific**, but as broadly applicable as possible. Ideally, the method should be applicable in a broadly defined water type, e.g. “shallow lakes in the Northern GIG”. Alternatively, assessment systems for more specific types could be defined, which should however always follow the same scheme. In any case, the range of applicability of the method must be defined.
- An alternative way is to not use type-specific metrics but **to correct metric results for the effects of natural environmental variability**, which allows the use of the metrics on a broader range of natural environmental variability (e.g. upstream to downstream, or along a thermal gradient; see Pont et al. 2009).
- The method must be **based on indices (metrics)** – optionally as a multimetric system or a prediction system. In total, the metrics must reflect the criteria defined by the WFD for the BQE and water type addressed, e.g. “diversity” or “ratio of tolerant and sensitive taxa” (see Table 1). A single metric can reflect more than one of these criteria, but this needs to be justified in each case.
- The method must be based on statistically proven **dose-response relationships**. For each type addressed and each metric used the relationship to stress-indicating variables (usually environmental variables) needs to be documented.
- The method must reflect the **impact of well-defined stressor types**. It can be restricted to a single stressor (such as eutrophication); in this case the dose-response relationship

must be based on environmental parameters reflecting the intensity of this stressor. Alternatively it can reflect the effect of multiple stressors, in which case the dose-response relationships must be based on environmental parameters reflecting the intensity of different stressors.

- Based on the results of the WISER field sampling campaign (in some cases also on additional existing long-term data) the effects of different sources of **uncertainty** on the assessment result need to be quantified.
- The **development process of the assessment system** will be documented in a standardised way. The databases used to derive the dose-response relationships and the uncertainty estimation will be stored on the project's website. Statistical parameters characterizing the method (e.g. correlation coefficients between metrics and environmental parameters) will be documented on a standardized form.

Process

All WISER assessment systems will be based on metrics, either as single metrics, as multimetric indices or as prediction systems. The procedure of data analysis during the development of metric-based assessment systems typically involves the following steps. All steps, ideally, should be performed for the entire workpackage dataset first (including different types and geographic regions), and then subsequently be refined (analyses for individual types). The aim should always be to derive a method as broadly applicable as possible.

Metric selection

- *Metric calculation*: For all data sets included in the workpackage database, metrics need to be calculated. In most cases the relevant datasets will be exported from the databases and metrics will be calculated with external software packages. In any case, metrics representing the relevant "indicative parameters" (Table 1) need to be calculated. The selection of metrics to be calculated will be task of the workpackage scientists; preferably a large number of metrics should be calculated.
- *Exclusion of numerically unsuitable metrics*: In order to reduce the long lists of metrics that are quickly and easily processed by software packages, filter procedures have to be applied. These procedures include the identification and exclusion of numerically unsuitable measures, for example, measures with a narrow range of values or with many outliers and extreme values, which can be simply revealed by box-whisker plots (Hering et al. 2006). However, where there are a high number of candidate metrics it may be more convenient to apply statistical tests.
- *Definition of a stressor gradient*: It is mandatory that the data set used for development includes data on a gradient of sites, ideally including unimpacted (reference) sites and heavily degraded (poor and bad) sites.

  An environmental stressor gradient is ideally represented by a set of sites of one water type covering the whole range (high, good, moderate, poor, and bad sites) of the environmental stressor that is to be targeted by the assessment system. The classification may be a continuous measure or on the division into five classes or even into the two classes "unstressed" and "stressed", only. A continuous gradient is preferred, while the simple division into "unstressed" and "stressed" sites should only be used if the number of sites is limited or few environmental data are available. Analysis of the gradient may be restricted to a single stressor or may include the impact of multiple stressors if stressors can not be separated (i.e., if sites are affected by more than one stressor simultaneously). For description of the impact of a single stressor, physical, chemical, or hydromorphological data on the individual sites can be used. Suggestions for environmental parameters representing stressor gradients are:

- Data on BOD_5, oxygen content or redox potential in sediments to describe the impact of organic pollution;
- Data on BOD_5, $N\text{-}NO_3$, *Escherichia coli*, eventually combined, for an index addressing water pollution in general terms;
- Data describing the trophic status of sites such as concentrations of phosphorus and nitrogen compounds;
- Data on priority substances;
- Data that characterise the morphological situation of a site such as the percentage of micro- or mesohabitats;
- Data on catchment land use for describing general stress gradients (Böhmer et al., 2004);
- Several of the above mentioned data, or other data (e.g. invasive species), to describe more general types of stress.

A statistical analysis such as PCA (Principal Component Analysis) can be used to reduce the number of variables by i) calculating hypothetical main gradients of the environmental dataset and ii) identifying redundant (co-correlating) variables. The direct analysis of metrics and abiotic environmental data is possible with Redundancy Analysis (RDA). The advantage of direct ordination procedures is their aim to fit the main abiotic and biotic gradients.

- *Correlation of stressor gradients and metrics*: Correlating the results of a metric to the stressor gradient is a central part of the procedure, which can be processed either by looking for significant differences (t-Test, U-Test) or by running rank correlation analysis (e.g., Spearman, Kendall). It is also possible to use Pearson' product moment correlation in cases of large data sets. Thus, a simple scatter plot may be used to aid the judgement on the strength and quality of metric-stressor correlations. Non-parametric regression models can also be applied for explorative analyses of the pressure–response relationships to identify non-linear patterns (e.g. Schartau et al. 2007).
- *Selection of candidate metrics*: An ideal metric should be responsive to stressors, have a low natural variability, provide a response that can be distinguished from natural variation, and be interpretable (Hering et al. 2006). A candidate metric's results must show a significant correlation to the stressor gradient. This correlation can be positive or negative, either across the whole stressor gradient or measured for a part thereof (e. g. only moderate to high quality sites). Metrics fulfilling this criterion are, in principal, suited to assessing the degradation of the ecosystem type and can be selected as candidate metrics.

  Numerous papers describe the possible approaches to metric selection (e.g. Holland 1990; Barbour et al. 1992, 1999; Karr and Kerans 1992; Karr and Chu 1999; Buffagni et al. 2004; Hering et al. 2004; Ofenböck et al. 2004; Vlek et al. 2004; Pont et al. 2006). Based on existing knowledge and literature information, the candidate metrics are selected on the basis of knowledge of the aquatic biota within a geographical entity.

  After having selected the candidate metrics, they need to be evaluated for efficacy and validity. This means that inappropriate metrics have to be eliminated from the process. Metrics have to be considered as inappropriate if they: (1) are less than robust and have a temporal and/or spatial variability exceeding variability caused by anthropogenic influences; (2) do not reflect human impairment and have little relationship to the impacts; and (3) are not well founded on ecological principles and understanding; for example, a correlation of land use with the invertebrate feeding-type “parasites”.

Only those metrics that show a quantitative impact-response change across a stressor gradient that is reliable, interpretable and not diffused or obscured by natural variation, must be selected. Moreover, different types of metric should be considered.

If none of the calculated metrics fulfils the criteria it should be considered generating “new” metrics, e.g. based on species predominantly occurring in stressed or unstressed sites following an indicator species analysis.

- *Selection of core metrics*: Candidate metrics, which can be identified as robust and most informative are scrutinized further for consideration for inclusion in the assessment system. To be selected as a core metric two major aspects have to be considered: (1) the metrics should cover the different metric types and “indicative parameters” (Table 1) and (2) redundant metrics need to be excluded. Metrics that show strong inter-correlations (e.g., Spearman’s r > 0.8) with one another are defined as redundant. The identification of redundant metrics is aided by triangular cross-correlation matrices and, in the case of redundancy, the correlation of each pair of metrics with the other metrics is compared in order to finally omit the one that shows the highest overall mean correlation. For the selection of appropriate core metrics, statistical analysis aimed at identifying those variables, which show the strongest relationship to certain environmental stressors, are recommended.
- *Distribution of metrics within the metric types*: In case a multimetric index is targeted, it should preferably contain at least one metric from each type (Table 1) and, therefore, reflect multiple dimensions of biological systems (Karr and Chu 1999, Hering et al. 2006). The possible combinations of metrics resulting from the selection of candidate metrics must be correlated to the stressor gradient used to select the candidate metrics. For this purpose, all metric results are first scaled by transformation into a score ranging from 0 to 1 (100 %). This enables the calculation of means for all candidate metrics (see next step). Those metrics whose combination results in the strongest significant correlation to the stressor gradient should be selected as core metrics.
- *Metric normalisation*: The upper and lower anchors mark the indicative range of a metric, i.e. the values that are empirically set and defined as “1” (upper anchor) and “0” (lower anchor), respectively, to normalize a metric’s result. The upper anchor corresponds to the upper limit of the metric’s value under reference conditions. The upper anchor is derived by the approaches outlined in the previous section on the normalisation of the common metrics (Step 7).

  The lower anchor corresponds to the lower limit of the metric’s value under the worst attainable conditions. If data on sites of bad ecological quality are available, the lower anchor should be set as a percentile (e. g., 5% or 10%) of all metric values of the bad ecological quality sites, or at the lowest value obtained or obtainable. If there are no data on bad ecological quality sites but data on sites representing different degrees of stress are available, the lower anchor can be obtained by extrapolation. In practice, each metric result must be translated into a value between 0 and 1 (Ecological Quality Ratio), using the following formula:

  $$\text{Value} = \frac{\text{Metric\_result} - \text{Lower\_Anchor}}{\text{Upper\_Anchor} - \text{Lower\_Anchor}}$$

  for metrics decreasing with increasing impairment, and

$$\text{Value} = 1 + \frac{\text{Metric\_result} - \text{Lower\_Anchor}}{\text{Upper\_Anchor} - \text{Lower\_Anchor}}$$

for metrics increasing with increasing impairment. Values > 1 are set to 1.

The resulting metric value for a given site is finally expressed as an EQR. The EQR represents the relationship between the values of the biological parameters observed for a given body of surface water and the values for these parameters under the reference conditions applicable to that water body. The ratio is expressed as a numerical value between zero and one: high ecological status is represented by values close to one and bad ecological status by values close to zero.

Generation of a Multimetric Index

The aggregation of metrics into a Multimetric Index should ensure that each metric type is represented by a similar number of metrics (e.g. Karr and Chu 1999). Nevertheless, the final selection of metrics for a Multimetric Index should produce the strongest multimetric view of biological condition in relation to the pressure(s) of interest. Therefore, we do not recommend a fixed number of metric types or measures per metric type. Different combinations of metrics (always including the relevant metric types) should be correlated against the stress gradients as done earlier for the selection of candidate metrics. The finally selected multimetric index should be among those metric combinations best correlating to the stress gradient.

Setting class boundaries

The final output of a multimetric index provides a score that represents the overall relationship between the combined values of the biological parameters observed for a given site and the expected value under reference conditions. This score is - as for single metrics - expressed as a numerical value between zero and one. This range can be subdivided into any number of categories corresponding to various levels of impairment (compare section “metric normalization”).

Boundary setting is examined in the intercalibration process, thus it has to be well founded. Boundaries can be set using discontinuities in the relationship between anthropogenic pressure and the biological response. Furthermore, the use of paired metrics that respond in different ways to the influence of the pressure allow for defensible boundary placement (e.g. % sensitive taxa compared to % of impact taxa for benthic invertebrates in rivers and lakes). The high-good boundary may be derived from metric variability at reference sites (e.g. 5th percentile value). The quality classification can also be calibrated against pre-classified sampling sites (e.g. pre-classification based on expert judgment). In case of a nearly linear dose-response relationships we propose quality classes with equal ranges to provide five ordinal rating categories for assessment of impairment in accordance with the demands of the WFD (boundaries placed at 0.8, 0.6, 0.4 and 0.2).

Uncertainty estimation

The estimation of uncertainty for the assessment system will be based on the field dataset collected in WISER. Separate guidelines will be provided by workpackage 6.1.

Product

The product will be a methodological description to be documented in a standard form (to be published on the project's website) covering the following:

- Water-body types to which the method can be applied
- Stressor addressed
- Underlying field and laboratory procedure
- Indices used and how they are calculated
- Threshold values between ecological status classes
- Existing calculation tools, if applicable
- Estimation of uncertainty
- Interpretation of the results

This standardised description will also be part of the relevant deliverables, which are mainly scheduled at the end of the project (Table 2).

Table 2: Relevant deliverables describing the assessment systems to be developed in the individual WISER workpackages.

WP	Deliverable no	Deliverable name	Due at
3.1	3.1-5	Report on integrated phytoplankton tools for use in ecological status assessment	Month 36
3.2	3.2-3	Report on the most suitable lake macrophyte based assessment methods for impacts of eutrophication and water level fluctuations	Month 24
3.3	3.3-3	Report on assessment of European lakes using benthic invertebrates	Month 24
3.4	3.4-4	Report on fish indicators for ecological status assessment of lakes affected by eutrophication and hydromorphological pressures	Month 30
4.1	4.1-4	Manuscript on the review of multi-species indicators synthesised with WP results	Month 36
4.2	4.2-3	Report/manuscript on benthic macroflora indicators for coastal waters	Month 36
	4.2-4	Report/manuscript on benthic macroflora indicators for transitional waters	Month 36
4.3	4.3-3	Manuscript on the responses of existing indicators to hydromorphological changes, including modelling of the ecological potential	Month 30
	4.3-4	Manuscript on indicators for hard bottom substrates	Month 30
4.4	4.4-5	Final report indicating the potential for modelling approaches for fishes in transitional waters and the conclusions regarding harmonising suitable metrics and approaches for wider use	Month 36

References

Barbour M, Gerritsen J, Snyder B, Stribling J (1999) Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish. US Environmental Protection Agency Office of Water. Washington DC:339pp.

Barbour M, Plafkin J, Bradley B (1992) Evaluation of EPA's rapid bioassessment benthic metrics: metric redundancy and variability among reference stream sites. Environmental Toxicology and Chemistry 11:437-449.

Birk S (2010a) Overview report of biological assessment methods used in national WFD monitoring programmes - Methods for coastal and transitional waters. First draft. University of Duisburg-Essen, Essen:192 pp.

Birk S (2010b) Overview report of biological assessment methods used in national WFD monitoring programmes - Methods for lakes. First draft. University of Duisburg-Essen, Essen:184 pp.

Birk S, Hering D (2009) A new procedure for comparing class boundaries of biological assessment methods: A case study from the Danube Basin. Ecological Indicators 9:528-539.

Böhmer J, Rawer-Jost C, Zenker A, Meier C, Feld CK, Biss R, Hering D (2004) Assessing streams in Germany with benthic invertebrates: Development of a multimetric invertebrate based assessment system. Limnologica 34:416-432.

Borja A, Dauer DM (2008) Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. Ecological Indicators 8:331-337.

Borja A, Franco J, Valencia V, Bald J, Muxika I, Belzunce MJ, Solaun O (2004) Implementation of the European water framework directive from the Basque country (northern Spain): a methodological approach. Marine Pollution Bulletin 48:209-218.

Breine JJ, Maes J, Quataert P, Van Den Bergh E, Simoens I, Van Thuyne G, Belpaire C (2007) A fish-based assessment tool for the ecological quality of the brackish Schelde estuary in Flanders (Belgium). Hydrobiologia 575:141-159.

Buffagni A, Erba S, Cazzola M, Kemp JL (2004) The AQEM multimetric system for the southern Italian Apennines: assessing the impact of water quality and habitat degradation on pool macroinvertebrates in Mediterranean rivers. Hydrobiologia 516:313-329.

Buffagni A, Erba S, Birk S, Cazzola M, Feld C (2005) Towards European Inter-calibration for the Water Framework Directive: procedures and examples for different river types from the E.C. project STAR. Quad. Ist. Ric. Acque 123:1-467.

Carstensen J (2007) Statistical principles for ecological status classification of Water Framework Directive monitoring data. Marine Pollution Bulletin 55:3-15.

European Commission (2008) Commission Decision of 30 October 2008 establishing, pursuant to Directive 2000/60/EC of the European Parliament and of the Council, the values of the

Member State monitoring system classifications as a result of the intercalibration exercise. Official Journal of the European Union L332:20-44.

Hering D, Feld CK, Moog O, Ofenböck T (2006) Cook book for the development of a Multimetric Index for biological condition of aquatic ecosystems: Experiences from the European AQEM and STAR projects and related initiatives. Hydrobiologia 566:311-324.

Hering D, Meier C, Rawer-Jost C, Feld CK, Biss R, Zenker A, Sundermann A, Lohse S, Böhmer J (2004) Assessing streams in Germany with benthic invertebrates: selection of candidate metrics. Limnologica 34:398-415.

Herlihy AT, Paulsen SG, Van Sickle J, Stoddard JL, Hawkins CP, Yuan LL (2008) Striving for consistency in a national assessment: the challenges of applying a reference-condition approach at a continental scale. Journal of the North American Benthological Society 27:860-877.

Holland AF (1990) Near Coastal Program Plan for 1990: Estuaries. U.S. Environmental Protection Agency Environmental Research Laboratory, Office of Research and Development, Washington, DC:259 pp.

Karr J, Chu E (1999) Restoring life in running waters: better biological monitoring. Island Press, Washington.

Karr JR, Kerans BL (1992) Components of biological integrity – Their definition and use in development of an invertebrate IBI. In: Midwest Pollution Control Biologists Meeting, Chicago, Ill., 1991. U.S. Environmental Protection Agency, Region V, EPA-905/R-92-003, pp. 1-16.

Kelly MG, Bennett C, Coste M, Delgado C, Delmas F, Denys L, Ector L, Fauville C, Ferreol M, Golub M, Jarlman A, Kahlert M, Lucey J, Ni Chathain B, Pardo I, Pfister P, Picinska-Faltynowicz J, Rosebery J, Schranz C, Schaumburg J, van Dam H, Vilbaste S (2009) A comparison of national approaches to setting ecological status boundaries in phytobenthos assessment for the European Water Framework Directive: results of an intercalibration exercise. Hydrobiologia 621:169-182.

Ofenböck T, Moog O, Gerritsen J, Barbour M (2004) A stressor specific multimetric approach for monitoring running waters in Austria using benthic macro-invertebrates. Hydrobiologia 516:251-268.

Pardo I, Uszko W, van de Bund W, Owen R, Poikane S, Bonne W, Kelly MG, Pont D, Birk S, Bennett C (2010) Revision of the consistency in Reference Criteria application in the phase I of the European Intercalibration Exercise. University of Vigo, Vigo:38pp.

Pont D, Hughes RM, Whittier TR, Schmutz S (2009) A Predictive Index of Biotic Integrity Model for Aquatic-Vertebrate Assemblages of Western US Streams. Transactions of the American Fisheries Society 138:292-305.

Pont D, Hugueny B, Beier U, Goffaux D, Melcher A, Noble R, Rogers C, Roset N, Schmutz S (2006) Assessing river biotic condition at a continental scale: a European approach using functional metrics and fish assemblages. Journal of Applied Ecology 43:70-80.

Schartau A, Moe S, Sandin L, McFarland B, Raddum, G (2008) Macroinvertebrate indicators of lake acidification: analysis of monitoring data from UK, Norway and Sweden. Aquatic Ecology 42(2):293-305.

Schmedtje U, Birk S, Poikane S, van De Bund W, Bonne W (2009) Guidance document on the intercalibration process 2008-2011. Guidance Document No. 14. Implementation Strategy for the Water Framework Directive (2000/60/EC):55 pp.

Stoddard JL, Larsen DP, Hawkins CP, Johnson RK, Norris RH (2006) Setting expectations for the ecological condition of streams: the concept of reference condition. Ecological Applications 16:1267-1276.

Stoddard JL, Herlihy AT, Peck DV, Hughes RM, Whittier TR, Tarquinio E (2008) A process for creating multimetric indices for large-scale aquatic surveys. Journal of the North American Benthological Society 27:878-891.

van de Bund W (2009) Water Framework Directive intercalibration technical report. Part 1: Rivers. Joint Research Centre, Ispra:140pp.

Vlek HE, Verdonschot PF, Nijboer RC (2004) Towards a multimetric index for the assessment of Dutch streams using benthic macroinvertebrates. Hydrobiologia 516:173-189.