Project no. 006463
EuMon
EU-wide **mon**itoring methods and systems of surveillance for species and habitats of Community interest

Instrument: Specific Targeted Research Project

Thematic Priority: Biodiversity conservation

D17: Recommendations for the coherence, scientific quality, and cost-effectiveness of species monitoring schemes

Due date of deliverable: Month 24 (October 31, 2006)
Actual submission date: January 26, 2007

Start date of project: 1.11.2004 Duration: 42 months

University of Debrecen, Department of Evolutionary Zoology and Human Biology

Revision: Final draft

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	**PU**
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)	

and time and cost-effectiveness of species monitoring schemes

Deliverable 17 of EuMon's Work Package 2.1

By Szabolcs LENGYEL (UNDEBR)

with contributions by Pierre-Yves HENRY (MNHN), Eva PAPASTERGIADOU (UNPAT), Klaus HENLE (UFZ), Dirk SCHMELLER (UFZ), Piotr NOWICKI (UJAG), Tiiu Kull (ESTAGUN), Kaire Lanno (ESTAGUN) and Eszter Déri (UNDEBR)

Table of contents

0.1. POLICY SUMMARY

The general aim of this deliverable is to help in identifying suitable approaches for the establishment of newly initiated schemes, for the improvement of existing schemes, and to contribute to moving towards higher coherence among monitoring schemes in Europe. The direct objectives are to provide a detailed list of criteria for judging and quantifying the scientific quality of species monitoring schemes, to outline ways for establishing coherence within and among schemes, and to estimate the time- and cost-effectiveness of schemes. All criteria proposed can be qualitatively characterised or quantified from data entered in the DaEuMon database available at http://eumon.ckff.si/monitoring.

Scientific quality of monitoring schemes (chapter 2) is proposed to be characterised by the following measures:

- spatial and statistical representativity of data collected,
- statistical power: ability of monitoring schemes to detect trends,
- measurement precision of monitoring schemes: reliability of data from monitoring,
- degree of refinement of the sampling design,
- scientific knowledge requirements for data collection, and
- use of state-of-the-art field and statistical methods.

Coherence (chapter 3) can be defined in several ways. The DaEuMon database can be used to characterise criteria related to:

- the internal consistency of individual monitoring schemes,
- compatibility among monitoring schemes, and
- coverage of Natura 2000 species and sites and of species for which countries have medium to very high responsibility by the monitoring schemes.

Time and cost-effectiveness (chapter 4) is proposed to be measured as the ratio of (i) the level of information obtained by the scheme and (ii) the effort necessary to conduct the scheme. The level of information encompasses the quantity and quality of information provided by the scheme and can be measured by:

- the areal coverage,
- taxonomical and ecological extent and
- scientific quality of monitoring schemes.

Effort required to run the schemes can be of two kinds.

- Time requirements, measured by manpower, including both professionals and volunteers.
- Financial costs are composed of
    - personnel costs, that can be estimated by manpower plus arbitrary salaries (not in DaEuMon), and
    - costs of materials and equipment.

We do not evaluate schemes or develop overall ranks for all schemes. Rather, we provide these guidelines to filter out schemes that can be recommended as examples of "best practice" schemes given the trade-offs described or schemes that are particularly suitable for integration into broader (geographically or taxonomically) monitoring schemes. This should allow coordinators of monitoring schemes to identify scopes for improvements of their schemes given their specific constraints. A full evaluation should involve the simultaneous evaluation of scientific quality, coherence, and time and cost-effectiveness for monitoring schemes. Because monitoring schemes differ largely in geographic scope, taxonomical extent, time and cost requirements, we argue that full evaluations should be carried out on smaller sets of similar schemes. Guidelines for a synthesis approach using composite measures of quality for such smaller sets are given in chapter 5 of this document.

0.2. BACKGROUND AND GENERAL COMMENTS

The aim of this document is to develop criteria for evaluating the scientific quality, coherence, and time and cost-effectiveness of species monitoring schemes. The document has been developed in parallel with D20 ("Recommendations for coherence, scientific quality, and time and cost-effectiveness of habitat monitoring schemes"), partly because several criteria used for evaluations are similar for monitoring schemes regardless of whether they focus on species or on habitats. Although the structure and wording of the two documents show similarities, the emphases of D17 and D20 are different, containing specific criteria applicable to either species monitoring or habitat monitoring.

The DaEuMon database can provide data to quantify some of the criteria given below, whereas some criteria cannot be evaluated because the database will not have information on the topic. However, it is important to list the latter criteria in addition to the ones that can be tested directly from the database in order to provide a complete set of criteria both for WP5 and for future external reference. When data from DaEuMon can be used for the evaluation, specific references for this possibility are given or will become obvious from the text (e.g. when referring to database question numbers).

Some of the ideas presented may seem at first as overly optimistic and simplistic to be used in a general evaluation of monitoring schemes. However, we considered it important to include all ideas because these recommendations will provide a foundation for further work conducted mainly by WP5 of EuMon and can help generate further ideas for users/coordinators, whose work will focus on testing and selecting the best approaches for their evaluation of monitoring schemes.

1. INTRODUCTION (OBJECTIVES AND DEFINITIONS)

Species monitoring programmes are highly variable. In contrast to habitat monitoring, species monitoring is often more traditional, more established, and more variable in scope. Species monitoring schemes range from one species or population to many populations and taxa, and from local through regional to continental or global scales and from one short period to decades-long monitoring programmes. Due to this high variability in the scope and extent of species monitoring, it is essential that the EuMon project, which collects and analyses information on monitoring schemes from European countries, develop criteria as to the coherence, scientific quality, and time and cost-effectiveness of species monitoring schemes.

The general aim of this deliverable, therefore, is to help identifying suitable approaches for the establishment of newly initiated schemes, for the improvement of existing schemes, and to contribute to moving towards higher coherence among monitoring schemes in Europe. The direct objectives are to outline ways for establishing coherence within and among monitoring schemes, to provide a detailed list of criteria for judging and quantifying the scientific quality, and time- and cost-effectiveness of species monitoring schemes. The outputs from this document may help in the development of the basis for integration of monitoring schemes (input to D16), and to provide practical input to WP5 in the development of indicators and tools.

The interpretation of this document requires that the concepts central to understanding are defined early on. Each chapter, therefore, starts with defining what is meant by scientific quality, coherence, and time or cost-effectiveness of monitoring schemes. Because most

evaluations are made on extractions from the DaEuMon database, the conclusions depend to a large extent on the quantity and quality of the information present in the database. At the preparation of this document, 385 species monitoring schemes were available in the database as a general background. However, an attempt was made to draw general conclusions that can be applied when there will be even more schemes present in the database.

2. CRITERIA PROPOSED FOR SCIENTIFIC QUALITY

2.1. Criteria derived from entries to the DaEuMon database

2.1.1. Reason for launching

Chances are that if a monitoring scheme is launched because of scientific interests, it is probably better designed in a scientific sense and its results are analysed better (or may be easier to analyse). Therefore, if the reason for launching (question 5) is "scientific interests", it can be reasonably assumed that the monitoring scheme has a higher-than-average scientific quality. This assumption needs to be tested, however, with answers given to other questions in the database. Questions that may provide a surrogate measure of scientific quality are outlined below in sections 2.1.2 to 2.1.8 of this document. A 'signal' for scientific quality may be present if there is a difference in some measure of scientific quality among schemes launched for different reasons, and if variables associated with higher scientific quality are more frequently associated with the reason of "launched from scientific interest". Therefore, the approach is of interest also if reversed, as it provides an opportunity to ask whether studies not launched for scientific reason differ in sampling design, representativity, or other relevant criteria.

2.1.2. Spatial and statistical representativity of data from monitoring

Representativity, the property that a smaller set of entities (a sample) can be reliably used to draw inferences on the entire set of entities (statistical population), is relevant in judging scientific quality from several aspects. Here, the spatial representativity is of central importance. A spatially representative monitoring is either conducted in the entire range of the focal species or uses an appropriate sampling design. The sampling design should ensure that the results of species monitoring are representative at the national or regional scale. The scale of representativity always depends on the goal of the monitoring scheme. Thus, the main question of interest here is that on geographical scope (question 6). For achieving coherence with 2010 target goals, it should also be possible to extrapolate the results of monitoring to any geographical scale or political level, if combined with other schemes.

To evaluate whether a sampling design is statistically representative, additional information can be used. For example, if the sampling design is stratified or is based on randomisation, there is a good chance that the sampling design ensures spatial representativity. The criteria for representativity are summarised in **Table 1**. The scoring can be based on a simple nominal scale (**Table 1**) or on the relative position of a scheme compared to the best one according to this criterion (e.g. number of actually monitored sites per maximum number possible).

Table 1. Spatial representativity scores for answers to questions in the DaEuMon database. The scoring presents criteria for coverage of species: all sites (representative) versus some sites (less representative).

Question in database	**If answer is:**	**Representativity is scored as:**
6. Geographical scope	international	representative
	national	representative
	regional	not representative (except for endemic/restricted species) *
	local	not representative (except for endemic/restricted species) *
S4: Sampling design	stratified	representative
	not stratified	may or may not be representative
S6: Choice of sites to be monitored	exhaustive	representative
	systematic	representative
	random	representative
	based on expert knowledge	not representative
	other	not representative

* The EuMon database does not directly contain information on whether a scheme monitors endemic or restricted species.

An optional measure of representativity is the relative area monitored per species per country. This measure can be calculated as the ratio of the number given in answer to S11 and area of the country (question 7; country areas have to be obtained from other sources) and yields a percentage of the country actually monitored. If the value is equal or close to 1, then representativity is high ("fully representative"). If it is lower than 1, classes of representativity can be developed, e.g. high-medium-low representativity. However, it is important to draw attention to two potential problems. First, many coordinators gave the actual area monitored, e.g. sum of plot areas and not the area at which results can be extrapolated (although this was explicit in the comment to this question). Another problem is when a species is endemic or restricted to a small region, in which case even a small percentage value of national area can secure high representativity for the monitoring. However, this problem is relatively small; if the restricted/endemic status is known (as is the case with many such species), it is easy to go back to DaEuMon and assess representativity.

2.1.3. Ability to detect trends: statistical power

A central criterion for the evaluation of a monitoring scheme is whether it can statistically detect a change of a certain size in the population or distribution of species. The probability of detecting a trend or change of certain size is statistical power. Statistical power is a function of measurement precision, annual background variation in the variable of interest, and effect size (the strength of the trend or certain size of the change one wishes to detect).

Measurement precision (or 'error', in analogy with experimental design theory) is a composite measure of several features of the monitoring system. Measurement precision is the most important measure of scientific quality, and can be estimated quite well from data collected in DaEuMon; therefore, we present ways to estimate it in a separate section (2.1.4.).

The background or 'ordinary' variation of the variable of interest (e.g. distribution, population size) from year to year is difficult to obtain for most of the taxa and is certainly not covered by DaEuMon. It is thus advisable to search literature sources to obtain estimates of this

variation for the taxa of interest and to extract these estimates (at least orders of magnitude) from published data on the species of interest (S26) (e.g. US Fish & Wildlife Service, Patuxent Wildlife Research Centre compilation of annual variation in various taxa). A caveat in using data on orders of magnitude is that the yearly variation becomes a decisive factor in the statistical power to detect trends by being far stronger than the properties of the schemes themselves. Two alternatives may be (i) to develop a qualitative assessment (or informed guesses) of temporal variation and/or (ii) to use ranked, ordinal-scale data on variation (large variation > medium variation > small variation) from literature sources.

Effect size is the strength of the trend or extent of change one wishes to detect as significant. Effect size is often an arbitrary value obtained from informed guesses. For monitoring schemes, the effect size is likely to be influenced by what is considered as an alarming change in the population or distribution of certain species. The reporting guidelines prepared by the Scientific Working Group of the Habitats Committee of DG Environment of the European Commission suggest using the following threshold values:

- trend in distribution range: 1% annual decline OR 5 % annual decline in favourable reference range,
- trend in population size: 1% annual decline OR 10% annual decline in favourable reference population OR population structure deviating from normal.

These threshold values are highly ambitious and can be achieved mostly by schemes that have very high measurement precision and low background variability. If the three parameters (measurement precision or "error", background variability, and effect size) can be quantified, then the statistical power to detect a trend can be determined. The precision of the power estimate will depend largely on measurement precision as this is the component most strongly related to the features of the monitoring schemes. Such an estimated statistical power can then be evaluated against a set of criteria developed to judge efficiency over a range of monitoring objects in a range of countries, regions etc. The NATURA 2000 system and the IUCN system provide explicit sets of criteria to evaluate the conservation status of species.

2.1.4. Ability to detect trends: measurement precision

Measurement precision is the most important component of statistical power. The reason it is separately treated here is that several entries to the DaEuMon database can be used to quantify the measurement precision of monitoring schemes in several ways. To estimate precision, the following variables are available from DaEuMon:

S2: nature of data collected: presence/absence data vs. counts of individuals (S2),
S18, S19: number of years monitored = difference in ending year (S19) and starting year (S18)
S7: detection probability can be accounted for
S12: number of sites monitored per year = sampling effort + spatial variation (or precision of annual status estimates)
S13: number of samples per site = replicates per visit
S15: annual frequency of sampling (related to among-year variation)
S16: number of visits per site = temporal replicates to compute detection probability or to adjust for phenology

So, how can we use the information from the above questions to quantify precision of yearly estimates (sampling error, measurement error)? First of all, it has to be noted that based on the database entries, there may be two options:

- when precision can be reliably estimated: sampling and/or measurement error (including detection error) can be estimated,
- when precision cannot be reliably estimated: sampling and/or measurement error (including detection error) cannot be estimated satisfactorily due either to the lack of detailed information or the inapplicability of the declared sampling design.

One option to get around this problem may be to define a minimum number of sites monitored by a scheme (S12) and to include those schemes that fulfil this criterion so that the monitoring can be used to produce estimates that are reliable. For instance:

- if the number of visits per year (S16) equals 1 and the number of samples per site (S13) also equals 1, then sampling error and measurement error cannot be estimated, or the estimate is not reliable,
- if the number of visits per year (S16) is greater than 1 and/or the number of samples per site (S13) is also greater than 1, then sampling error and measurement error can be estimated, i.e., the estimate is reliable (though it still may not be precise).

2.1.5. Statistical power judged by coordinator vs. estimated from database entries

One important measure of the quality of a monitoring scheme is how well it functions according to its designers/implementers and how well it functions in reality according to independent criteria. The match between statistical power supposed by the coordinator and the requirements set by the NATURA 2000 goals needs to be an important factor in making recommendations as to a best practice monitoring scheme. The reliability of the scheme as judged by the coordinator is given in question S9 ("Minimal annual change you think you can statistically detect"). Three scenarios can be deduced from the answer to this question:

- If the data entry is missing, it may mean that the coordinators do not consider the ability to detect trends important or do not know the precision of their scheme or that precision is difficult to estimate.
- If an entry is given and it is lower than the thresholds for conservation status evaluation, then statistical power is supposedly high (sufficient).
- If an entry is given and it is higher than the thresholds for conservation status evaluation, then statistical power is supposedly low (insufficient).

Schemes with an experimental design can be considered as monitoring schemes designed to detect specific changes in time, due to given causes of change. Probably in most of these cases it can be assumed that the experiment was appropriately designed to document the changes properly. In such cases, the trend is no more the slope for a continuous temporal change, but a size effect associated with a specific treatment. If this assumption is correct, the issue of experimental design may be discussed here, in the section "ability to detect a trend". Here, question S5 ("Use of experimental design") is of relevance:

- if the answer is "before/after experimental design" (DaEuMon entry value 1 or 3), the experiment is appropriate for monitoring purposes,
- if the answer is "comparison with control" (value 2), the experiment is ideal for monitoring,
- if the answer is "no experimental design" (value 0), then the statistical power can be used to detect temporal trend.

2.1.6. Sampling design refinement

The refinement of the sampling design of a monitoring scheme may be related to its representativity. If a sampling design is more refined in structure, it is likely that it is capable of accounting for more aspects of spatial and temporal variation in the variables of interest, e.g. species distribution or population size. Therefore, schemes using a more refined sampling design may be more representative. The refinement of a sampling design can be deduced from answers to the following questions:

S2: main data collected (presence/absence, counts, mark-recapture, age/size structure, phenology)
S4: stratification of sampling design (yes/no)
S5: existence of experimental design (no, before/after comparison, controlled treatments, before/after and controlled treatments)
S6: selection of sites for monitoring (exhaustive, random, systematic, personal/expert knowledge, other)
S7: accounting for detection probability (yes/no)

The answers to these questions may be used to develop an index of refinement for the sampling design. The index can be based on arbitrary values (e.g. schemes using stratification receive a score of 1, those not using stratification receive a score of 0), and can be calculated in ways deemed most appropriate by the user, i.e., values can be added, multiplied, etc.

2.1.7. Scientific/biological knowledge requirements for collection of monitoring data

If we assume that the scientific/biological knowledge requirements in a monitoring scheme and the scientific quality of the scheme are positively correlated, the proportion of professionals participating in the monitoring scheme indicates scientific quality. The proportion of professionals can be calculated from numbers given in answer to questions S21 and S22 as: S21/(S21+S22). Furthermore, monitoring schemes that require training or expert knowledge (S23) may have a higher scientific quality than those that do not require such measures. However, before using these criteria to evaluate scientific quality, the above assumptions need to be tested by using data from DaEuMon. Scientific output from the monitoring schemes is difficult to obtain from DaEuMon, although answers given to question 8 (References) may provide hints whether scientific publications are prepared as an output or not. Therefore, as an example on testing the relationship between involvement of professionals and scientific quality, the proportion of professionals can be used as an independent variable and existence of scientific publications (e.g. a binary variable from answers to question 8 References) can be used as response variable in a logistic regression. A possible bias in this logic can be if coordinators did not provide references on their scheme even though they have some or if the reference list given is not exhaustive. One option to estimate this bias is to relate the level of statistical analysis (S8) to occurrence of scientific publications in 8 References, assuming that results from more sophisticated analyses are published more often.

It is necessary to emphasise that all of the above needs a rigorous testing before they can be proposed for the general evaluation of scientific quality. Especially, the relationships between the number of professionals and either the presence (and possibly quality) of references or high scientific quality needs to be tested and judged robust to use in estimating scientific quality.

If the assumptions do not hold, the scientific knowledge criterion is not recommended for direct use in the evaluation. In this case, the information given in questions S21-S23 can be used to build cost-efficiency indicators, assuming that whoever collects monitoring data (S21-S22: professional/volunteers, S23: trained/not trained), their value to quantify state and trends in distribution and population size is the same.

2.1.8. Use of state-of-the-art field and statistical methodologies

The scientific quality of a monitoring scheme is likely to be higher if the data are collected and/or analysed by more up-to-date, more sophisticated methods. For example, accounting for detection probability in the field is viewed as more modern than not doing so (S7), and data analysis is more sophisticated than lack of data analysis and if it involves General Linear Models or other advanced statistics beyond simple graphics and descriptive statistics (S8). A special case is when data are analysed by persons/institutions different from those collecting the data. This is often the case with large-scale national or international projects. In such cases, data quality and analysis usually are both sophisticated.

There are at least two ways to incorporate methodological complexity in a general evaluation of monitoring schemes. Firstly, the questions related to methodological complexity can be considered separately, and each of the more "modern" or "sophisticated" entry value can get a higher weight in the evaluation. Second, an index of methodological complexity can be developed, which could summarise information from several questions. Questions that are relevant for this aspect are S7 (Detection probability), S8 (Type of data analysis) and possibly S23 (Training/expert knowledge required).

It is important to note here that the use of modern methodology generally leads to high scientific quality. However, high scientific quality may also be achieved by relatively simple data collection and analysis. Therefore, one cannot simply declare poor scientific quality for a monitoring scheme that does not use the above modern methods of data collection and analysis. This difference in the relationship between modern methods and scientific quality needs to be considered in the evaluation of monitoring schemes or analysis of data from DaEuMon.

2.2. Other optional criteria for scientific quality

The criteria listed in this part cannot be derived directly from the DaEuMon database, but we list them here for the sake of completeness. The criteria may be used on a subset of monitoring schemes with details of suitable depth or in future attempts to evaluate monitoring schemes once data become available.

- Other refinements of the sampling design:
    - in time: Is the phenology of species monitored considered? Is frequency of sampling adjusted to the phenology of the species monitored?
- Are background variables or drivers of potential changes monitored? (land use, management, causes of change: fragmentation, degradation etc.)
- Are the results published in peer-reviewed scientific journals/books or in other forms, e.g. in reports or in grey literature?

3. COHERENCE

Coherence within and among monitoring schemes can be defined in several ways. Within schemes, coherence can exist between the stated goals and methods used in the monitoring scheme (definition 1). The main meaning of coherence here is that whether monitoring schemes are convergent enough to allow the drawing of common inferences from them. This aspect will especially be important for the suggestion of optimal approaches, which will need to have high internal coherence, i.e., whether the recommendations about quality, time constraints etc. fit each other. Coherence among schemes may indicate the level of compatibility among the schemes (definition 2). For instance, schemes that have a similar scope, goals, and methods may be more coherent with each other than schemes differing in these aspects. Coherence among the schemes, therefore, is directly related to integration, i.e., more coherent schemes may also be better for integration. The examination of coherence from this perspective is the focus of D16 (for species monitoring) and D19 (for habitat monitoring). Therefore, it will not be elaborated in much detail here. Finally, coherence of the monitoring schemes may be interesting from the perspectives of the NATURA 2000 system (definition 3). Questions that are important in this aspect concern the coverage of NATURA 2000 sites by the monitoring activities, as well as the coverage of NATURA 2000 species and habitats by the monitoring schemes. This aspect of coherence is also related to the development of methods determining national responsibilities and conservation priorities, a task of WP4. WP4 further studies the coverage of species and habitats of Community interest by the existing monitoring schemes. Therefore, here we only examine whether the monitoring schemes provide sufficient information to assess the complementarity of NATURA 2000 sites.

3.1. Coherence between goals and data used in monitoring

In this part, we review coherence between the stated goals (objectives) and the main data collected in a monitoring scheme. We use information entered for questions S1 (Aim of monitoring: population trend, distribution trend, community/ecosystem trend) and S2 (Main data collected: presence/absence, counts, mark-recapture, age/size structure, phenology). We investigate all potential combination of goals/data types and evaluate the appropriateness of data types as a function of the goals stated (**Table 2**). Appropriateness is scored in three ways: appropriate (high coherence between goal and data type), poorly appropriate (low coherence) and inappropriate (negligible coherence; criteria extracted from Appendix 2 of D2).

Table 2. Coherence between goals and data types collected in species monitoring schemes.

S1: goals	S2: data type	Derived information	Evaluation
population trend	presence/absence	spatial/temporal representation	poorly appropriate
	counts	population size estimates over time	appropriate
	mark-recapture	local population size estimate	appropriate on local scale, poorly appropriate on larger scales *
	age/size structure	ratio of young/aged individuals or cohorts	poorly appropriate (only under special conditions)
	phenology	temporal representation	inappropriate
distribution trend	presence/absence	spatial representation	appropriate
	counts	relative density in space	appropriate
	mark-recapture	local population size estimate	poorly appropriate (small-scale, time-consuming)
	age/size structure	ratio of young/old individuals or cohorts	inappropriate
	phenology	temporal representation	inappropriate
community / ecosystem trend	presence/absence	spatial representation	appropriate
	counts	relative density in space	appropriate (especially for species of high indicator value)
	mark-recapture	local population size estimate	poorly appropriate (small scale, time-consuming)
	age/size structure	ratio of young/old individuals or cohorts	inappropriate
	phenology	temporal representation	inappropriate

* Mark-recapture can be considered appropriate if e.g. the species has one or a few local populations.

3.2. Coherence with EU directives and 2010 expectations

To determine the coherence of monitoring schemes with the expectations derived from NATURA 2000 system and with the goals of the 2010 target (jointly referred to as objects of Community importance), it is essential to quantify the status and trends in population sizes and distributions. For the following evaluation of coherence, only those schemes can be used that provide an opportunity for such quantification, similar to NATURA 2000 reporting obligations for the member states.

A straightforward way to evaluate coherence here is to quantify the percentage of schemes in DaEuMon that are clearly relevant for the 2010 target. With regard to section 2.1., a further separation is necessary for schemes that monitor distribution trend and those that monitor population trend. The evaluation process involves the following steps:

1. Compilation of lists of species under the Bird and Habitats Directives for each country.
2. Identification of the species that are actually monitored. This needs to be based on a routine to be developed using data entered in DaEuMon.

3. Separation of actually monitored species based on coherence between the goals and data types used in monitoring (section 2.1).
4. Calculation of the proportion (%) of species of Community interest that are actually monitored, followed by separate calculations of the proportion of species whose monitoring is considered appropriate by the criteria proposed in section 2.1.

The results are largely dependent on the exhaustiveness of DaEuMon (Step 2). The evaluation of coherence will be reliable if all monitoring schemes are entered for a country, whereas it will be less reliable if many monitoring schemes operating in a country are missing from DaEuMon. Thus, for countries well represented in DaEuMon (as of October, 2006: e.g. Poland, France, Spain, Hungary), the above approach will provide reliable results and a higher potential for compatibility and integration. In any case, this approach provides a basis for a kind of gap analysis: "Which species of Community importance are NOT monitored adequately?"

Another way to characterise coherence would be to consider if a scheme monitors many or only a few species of Community interest. For instance, if one scheme monitors 100 Directive species, whereas other schemes for the same country monitor only a few (or even a subset of these 100 species), the latter schemes can be considered as less coherent than the first scheme. In such cases, a suggestion of improvement would be to recommend to monitor more species than actually monitored or to better align the two schemes to increase cost-efficiency and coherence with national monitoring goals.

Finally, the qualitative estimation of statistical power, as outlined in sections 2.1.3. and 2.1.4., is also important from a coherence point of view. The qualitative method will enable EuMon to give an indication, which percentage of the monitoring schemes are likely to gain high, medium, and low power to detect trends of the threshold values given in the EU guidelines given the design and crude information (e.g. on temporal variation of taxa). Furthermore, it could enable EuMon to estimate how many years monitoring schemes need to be run before they can detect trends of the magnitude chosen in the guidelines. Likewise, it would be possible to assess how much more precise estimates are needed or how constant a species must remain, to be able to detect the trends foreseen in a given time period. Thus, with DaEuMon we can identify schemes that are coherent with EU guidelines on evaluation criteria for conservation status (cf. threshold values of changes in distribution and population size).

4. CRITERIA FOR TIME AND COST-EFFECTIVENESS

Efficiency or effectiveness is usually determined as the achievement relative to some investment allocated to do the work. For monitoring schemes, an analogy can be to evaluate "potential" of a scheme per "effort". Potential in this respect is defined as the quantity and quality of processed information that can be obtained in the scheme, whereas Effort can be any measure of investment, e.g. time requirements expressed in manpower (personmonths) or money. The Potential per Effort (or PpE) can then be used as a measure of time or cost-effectiveness of monitoring schemes. The present document defines both Potential and Effort indicators but will only mention the possibility of the division to obtain PpE because the exact estimation or calculation of PpE will largely depend on the goals of the individual users. Different answers may be reached, for example, if one divides different Potential indicators with different Effort indicators. Therefore, the users can conduct the interpretation of these PpE ratios independently, by focusing on the questions they address. The aim of this part of the document thus is to provide guidelines for readers to evaluate for themselves how

time/cost-effective their approaches are and how the compromise between Potential and Effort can be optimised by improving the congruence between the goals and designs of monitoring schemes.

4.1. Potential indicators

"Potential" variables are numerators in the ratio of achievement and investment proposed to describe the time and cost-effectiveness of monitoring schemes. There are several measures of Potential available from the DaEuMon database. We propose that the Potential of a monitoring scheme depends on the area involved, the ecological and taxonomical spectrum of monitoring, and the scientific quality of the monitoring. In other words, the main questions are how large the monitored area is, how wide is the range of species monitored, and how well is monitoring conducted. Most of the variables important as Potential variables are thus also relevant for evaluating the scientific potential of monitoring schemes. Therefore, the following overview will refer back to chapter 2 regarding scientific potential and its components (representativity, statistical power/measurement precision etc.).

4.1.1. Areal coverage of species monitoring

In an ideal case, the proportion of area monitored per total area of the distribution range of the target species (coverage or spatial representativity) is an important measure of the potential of the monitoring scheme. The DaEuMon database contains information on the total area to which the results of monitoring can be applied (S11). Thus, although the actual area monitored is not known, the coordinators, ideally, have already provided the area to which their results can be extrapolated. In such cases, only the distribution range of the species monitored is required to estimate the proportion of the species' range actually monitored (areal coverage of monitoring).

4.1.2. Taxonomical and ecological extent of monitoring

Answers to several questions offer an opportunity to quantify the taxonomical coverage of monitoring schemes. Besides the areal extent of monitoring, these basic information are the primary source of estimating the Potential produced by monitoring schemes. For example, the Potential is probably higher of a monitoring scheme if it monitors not two but 20 or 200 of species within a taxonomic group. Therefore, the number of species monitored is a basic information directly related to the Potential of a monitoring scheme. The questions that can be used for such purposes are as follows.

- Number of species involved in monitoring (taxonomical coverage of species monitoring schemes):
    - S26: identity and number of taxonomic group(s) monitored;
    - S27: number of species monitored.

One question can be used to estimate the ecological coverage of species monitoring schemes (S30: Number of habitats where species monitoring is conducted). This measure is a crude estimate how extensive the species monitoring scheme is in an ecological sense and can be secondarily important. For example, if two monitoring schemes focus on the same taxonomical group, the one, which monitors this group in more habitats, is supposed to have a higher Potential according to the conceptualisation proposed.

4.1.3. Scientific quality

Scientific quality (i.e., how monitoring is conducted) is a composite measure of several attributes, that are reviewed in chapter 2 (especially section 2.1.). Of the variables listed there, representativity of data from monitoring (questions 6, S4, S6), statistical power (S2, S11, S12, S15, S18, S19) and measurement precision (S7, S9, S12, S13, S15, S16) are primarily relevant to estimating the scientific quality-related Potential of a monitoring scheme.

In addition to these variables, the availability of information on potential drivers (S31, S32) or background variables can be important as this information enables one to address causes of changes observed during monitoring. The availability of such information greatly adds to the scientific Potential of monitoring, whereas this information is not necessarily and directly related to the scientific quality of schemes (*sensu* chapter 2). A monitoring system not involving background variables may be independently high, medium, or low in scientific quality.

It has to be noted here that two measures of scientific quality (2.1.7. Scientific/biological knowledge requirements for collection of monitoring data and 2.1.8. Use of state-of-the-art field and statistical methodologies) will need to be considered separately. The rationale here is that the involvement of many professionals and/or many state-of-the-art methods may result in higher costs for a monitoring scheme, but also can disproportionately increase the Potential or "extent" of the scheme. Therefore, the involvement of professionals and modern methodology needs to be studied separately for each monitoring scheme or within a group of similar schemes.

4.2. Effort indicators

Effort variables are denominators in the ratio of achievement and investment proposed to describe the time and cost-effectiveness of monitoring schemes. There are two important measures of Effort available from the DaEuMon database, time (estimated by manpower in person-days) and financial costs.

4.2.1. Indicators for time requirement of monitoring schemes

Time or manpower variables can be deduced from the database in two ways. Firstly, the answer to question S24 (Manpower [in person.day] needed per year to run the scheme) gives an estimate of the yearly time effort. To account for the fact that some monitoring schemes are not run every year (question S15: frequency of monitoring), manpower is averaged per year (i.e., divided by S15). The total time requirement to run a monitoring scheme is then given by S24 / S15.

Another group of questions allows the quantification of the time requirement for fieldwork per year (in person-days). The number of sampling sites (S12), the number of sampling occasions per year (S16), and the time requirement necessary for one sampling occasion (S17) can be used to calculate the time requirement for fieldwork in a year. The total time requirement for fieldwork per year is therefore given as S12 * S16 * S17 / S15).

The outcome of the relationship between two indicators (Total time required to run scheme and Fieldwork involved) gives an opportunity for several interesting comparisons. For example:

- if S24 = S12 * S16 * S17 / S15, this means that the monitoring scheme consists entirely of fieldwork and S24 is sufficient for further use,
- if S24 < S12 * S16 * S17 / S15, i.e., fieldwork is more than the total time required per year, at least one measure of manpower given by the coordinators is wrong and the estimates will not be reliable,
- if S24 > S12 * S16 * S17 / S15, we can calculate the manpower necessary for non-fieldwork (e.g. sorting samples, identifying species, data entry and analysis in the lab) as S24 - S12 * S16 * S17 / S15.

4.2.2. Financial resources indicators

Two kinds of basic information on the financial resources of monitoring schemes can be retrieved from the DaEuMon database. Personnel costs and costs for material and equipment given by the coordinators can be used to characterise the extent of monitoring schemes in a financial sense.

4.2.2.1. Personnel costs

Personnel costs are not directly present in the database but may be estimated from time variables. The basic idea is to compute the equivalent of human resources in money. The Total time requirement for running the monitoring scheme (question S24) multiplied by an arbitrary daily salary (Y) indirectly estimates personnel costs. Thus, personnel costs can be estimated as S24 / S15 * Y.

It should be kept in mind that personnel costs strongly vary by country, status of the participants, and employer. Therefore, several categories of arbitrary salary appear warranted. Forgoing such distinctions may result in biases. For example, Eastern European schemes may show more cost-effective merely due to lower average salaries as compared to those in Western Europe. Similarly, not taking into account that volunteers or non-trained personnel may earn less than professionals involved in the schemes may overestimate personnel costs.

It is also important to keep in mind that the ratio of professionals and volunteers can be used to compute how much money is saved in the monitoring schemes thanks to the involvement of volunteers. To make such a distinction to account for training necessity, we may define two arbitrary salaries, Y1 for trained and professional participants and Y2 for untrained and volunteer participants (and include Y1 and Y2 in formulae hereafter). For example, if we assume that volunteers participate in fieldwork and professionals conduct lab work, the total amount equivalent to total human resources per year is then given by S24 / S15 * (Y1[S24 - S12 * S16 * S17] + Y2[S12 * S16 * S17]).

The percent of the total manpower that is saved thanks to volunteer involvement is given as (S24 / S15) * (S22/[S21+S22]). If we want to calculate the amount saved thanks to volunteer involvement per year, it can be expressed as (S24 / S15) * Y1 * (S21/[S21+S22]).

4.2.2.2. Material/equipment costs

The annual costs of materials and/or equipment used in the monitoring scheme is directly given in answers to question S25 ("How much do you spend on material and equipment per year (in €)?"). This information, therefore, can be readily be incorporated into "Total costs of monitoring per year" and can be given per year as S24 / S15 * Y + S25.

4.3. Indicators for time and cost-effectiveness

Once the quantification of indicators for Potential and Effort has been made, composite indicators for time and cost-effectiveness can be devised. The general idea, again, is to establish a ratio of "achievement/investment" or Potential per Effort (or PpE). Considering that three types of Potential indicators and two types of Effort indicators are deduced from the DaEuMon database, six combinations can be envisioned (illustrated in Table 3).

Table 3. Examples of potential pairing of Potential and Effort indicators to develop composite measures of time and cost-effectiveness of monitoring schemes.

	Effort indicators	
Potential indicators	*Time requirement*	*Financial resources*
Coverage of habitat monitoring	Area monitored per year	Area monitored per unit manpower/money
Taxonomical (and ecological) extent of monitoring	Number of species / taxa monitored per year	Number of species / taxa monitored per unit manpower/money
Scientific quality	High or low scientific quality per year	High or low scientific quality per unit manpower/money

The simplified composite measures shown above are only recommendations. Both the Potential and Effort indicators as well as PpE-calculations may provide guidelines that may help users/coordinators to evaluate their own schemes or as many schemes as they wish. The users (coordinators) may want to define their own indicators rather than strictly following the guidelines proposed here. Such indicators devised by the users may be more relevant to the specific questions asked by them and to the specific monitoring schemes under study.

5. SYNTHESIS: RECOMMENDED LOGIC FOR EVALUATIONS

In this section, we suggest a synthesis-approach for the evaluation of monitoring schemes. Obviously, this approach is not the only one and users of DaEuMon should be able to compare whatever they want, including even very different monitoring schemes, e.g. distribution of lichens and population trends of birds. The aim of this section, therefore, is to provide some guidelines on how users can do such an evaluation for themselves. We do not evaluate schemes or develop overall ranks for all schemes. Rather, we provide the following guidelines to filter out schemes that can be recommended as examples of "best practice" schemes given the trade-offs described above or schemes that are particularly suitable for integration into broader (geographically or taxonomically) monitoring schemes.

For the technical part of evaluation, we suggest using a filter-like approach. As an example, one should imagine a table containing monitoring schemes as rows and different features of the schemes as columns. Then for each relevant entry for a given scheme, qualifications on scientific quality, time/cost efficiency, and coherence as suggested above can be added. For example, for representativity (section 2.1.2.), one should add "representative" in a separate column for schemes showing "international" or "national" entries in question 6 (Geographical scope) and "not representative" for schemes answering "regional" or "local". Moving on to e.g. S4 (Sampling design), one then could select "representative" for schemes with entries "stratified". Not stratified = may be or may not be representative, it depends on properties of the rest of the sampling design. Therefore points given by „stratified" should be small compared to declaring random / exhaustive or systematic.
Repeating these steps for each of the relevant columns (database fields), one should be able to use filtering to search for schemes that are marked as "representative" for each relevant question. In the example above, filtering for "representative" would yield the schemes that are international or national in scope and that use a stratified sampling design.

This approach would enable the users/coordinators to focus on the questions they feel relevant and/or to use their own ranking system, if they wish. For EuMon, such an approach would provide an opportunity to calculate average potentials (for quantifiable criteria) or ideal values (for qualitative criteria) and to quantify the proportion of schemes reaching an above-average rank (for quantitative criteria) or a satisfactory rank (for qualitative ones). Furthermore, there is an opportunity to reveal trade-offs (or synergies) between criteria in case we find negative (or positive) correlations. Finally, the relationships found can be repeatedly tested for all the relevant taxonomic groups.

For the quantifiable indicators (scientific quality, Potential and Effort indicators), numeric values can be calculated that may characterise the feature one is interested in. To calculate these indicators, one could follow the formulae given for the respective feature (chapters 2 and 4) or could develop one's own formulae.

A synthesised approach for the evaluation of monitoring schemes could be a hierarchical approach that starts with the questions most relevant to the end-users and studies questions of increasingly smaller importance and larger detail in subsequent steps. Therefore, it is recommended that the following logic be applied on a smaller relevant set of similar monitoring schemes.

1. The coherence between stated goals and data types should be established. This step is the first filter, as appropriate schemes are selected for further assessment, and poorly appropriate ones receive a lower score or are treated separately.
2. The coherence between EU goals (NATURA 2000, 2010 target) and monitoring should be established. Here the ability of the schemes to quantify status and trend in distribution and population size of species of Community importance needs to be evaluated.
3. Build a composite indicator for scientific quality based on the following primary indicators: representativity, statistical power, and precision. Secondary indicators, such as causes of change, requirement of scientific knowledge, and state-of-the-art methodology, should also be considered. However, it may be useful to treat primary and secondary indicators separately.
4. Calculate or estimate Potential indicators.
5. Calculate or estimate Effort indicators.
6. Build appropriate composite PpE measures, contrast Potential or one of its components with time and/or financial costs (total costs and real costs).

Several issues may arise with a unified evaluation process. First and most importantly, monitoring schemes differ greatly in geographic scope, taxonomical extent, time and cost requirements, etc. This variation results in coherence, scientific quality, and Potential and Effort indicators varying greatly across different schemes. It follows that a simplified composite measure (e.g. an index of general quality) cannot be reasonably calculated and/or meaningfully compared across many monitoring schemes. A joint evaluation applying a detailed analysis is likely to lead to high variation in the indicator values and hard-to-explain relationships among schemes. Therefore, it is important to stress that such composite measures and indices should be calculated only for sets of schemes that are similar to some degree in their stated goals and methods. For any concrete evaluation, the above-listed logic is advised for sets of schemes that have similar goals and/or focus on similar taxonomic groups. If such sets of schemes are identified, schemes could be compared according to their achievement of the goal and their costs. This way many of the points along the above logic would be the same (or at least similar) for the schemes under study, which would reduce the list of relevant indicators to a set of 3-4 indicators per scheme. An evaluation of the similar schemes based on 3 or 4 indicators should then be straightforward.

A related issue is that it is advisable to carry out the complete evaluation all at the same time. For instance, some schemes may stand out as high in scientific quality but may perform low on time and cost-effectiveness. Therefore, to have a full picture, we propose that effectiveness is simultaneously evaluated with coherence and scientific quality of monitoring schemes. A full evaluation, therefore, should involve the evaluation of coherence, scientific quality, and time and cost-effectiveness for monitoring schemes.

6. CASE STUDIES: EXAMPLES OF THE APPLICATION OF CRITERIA DEVELOPED

6.1. Plants

This case study analyses plant monitoring schemes (schemes monitoring several different taxonomic groups, plants among them, are not included). There are 41 such schemes in DaEuMon (*database date 06.11.2006*).

6.1.1. Criteria proposed for scientific quality

If monitoring scheme is **launched** from scientific interests, it may be better designed in a scientific sense and results may be easier analysable. From plant monitoring schemes 17% are launched from scientific interests, whereas more than half of the schemes (54%) are launched because of national law.

Schemes **representativity rate** was calculated based on three aspects (answers to questions about geographical scope, sampling design, and choice of sites to be monitored). One scheme was representative in all three aspects (scheme nr 577, common plants survey). 22% of the schemes were representative in two aspects and 61% in one aspect, 15% of schemes were not representative based on these questions.

The measurement precision can be estimated based on detection probability, number of sampling sites, number of samples per site, annual frequency of sampling, and number of visits per site. But the result depends on how much weight (points) to give to each of those questions and which formula to use exactly for calculation. For the present case the answers to each previously named question were divided into classes and certain amount of points was given to each class and points summed. Answers and points given were as follows:

- detection probability: 1 (1 point), 0 (0 p);
- nr. of sampling sites:
- 1 (0.5p), 2-10 (1p), 11-100 (1.5p), 101-500 (2p) and > 500 (2.5p);
- nr. of samples per site: 1 (0.5p), 2-5 (1p), > 5 (1.5p);
- annual frequency of sampling: 1 (0.5p), ≥ 2 (1p);
- nr. of visits per site: 1 (0.5p), ≥ 2 (1p).

With this approach the scheme nr 40 was among the best regarding precision rate. However, answers (and hence precision) depend on monitoring purpose and species under investigation. For different purposes also different sampling methods are appropriate.

One variable for estimating measurement precision is Number of years monitored. The longest schemes have run already for decades (the oldest one has run since year 1800, scheme nr 40).

The statistical power judged by coordinator is given as minimal annual change that can be statistically detected. For 63% of schemes the change was 10% or more. For 17% the data entry was missing (so the precision is difficult to estimate or it is not considered important).

Degree of refinement of the sampling design can be estimated based on questions S2, S4, S5, S6, and S7. However, the result depends on how much value to give to each of those questions and which formula to use. Here to answers of each named question certain amount

of points was given and the points summed. Using such method for example the scheme nr 297 was with quite a good degree of refinement. However, within WP5 future work is needed to put up a formula equally applicable in all cases.

Proportion of professionals in different schemes is between 1-100%. The majority of schemes (68%) includes only professionals and requires also training, what may indicate a more refined sampling design and statistical analysis, but as volunteers can do very good work, the amount of professionals engaged is not a good indicator of sampling design refinement.

The scientific quality of scheme can be higher if field and statistical methods are more up-to-date. Of 41 plant monitoring schemes 3 allowed accounting for detection probability, analysed data, and used advanced statistics. For the majority of schemes (56%), only one of those points was filled-in in the questionnaire/was fulfilled.

6.1.2. Criteria proposed for coherence

Coherence between goals and data used in monitoring was calculated based on table 2. Where there were multiple goals, only one was considered (the one what was first listed). For 66% of schemes the data type was appropriate, 22% poorly appropriate and 12% inappropriate.

To estimate **coherence with EU directives and 2010 expectations** for all plant monitoring schemes was beyond the scope of this case study. Besides, for several schemes the species names monitored are not provided making an assessment difficult.

6.1.3. Criteria proposed for time and cost-effectiveness

The important measure of the monitoring value is the **proportion of area monitored** per total species distribution area. Since it requires the distribution areas of all monitored species (which are not in the database), it was not possible to calculate this measure for this case study.

When examining **the taxonomical and ecological extent of monitoring**, the number of species monitored for different schemes is between 1 (32% of schemes) and 1000. For 85% of schemes the number of species monitored is below 100. The three schemes that monitor the highest number of species are schemes nr. 297 (500 species), 40 (570 species) and 837 (1000 species).

If information about **potential drivers** is available, it adds to the scientific value of monitoring. For plant monitoring schemes such information is known for the majority of them (88%).

There are several points available to measure the **effort indicators** of schemes (such as time requirement and financial resources indicators). Total time requirement to run the monitoring scheme (calculated as S24/S15) ranges from 1 to 325 (manpower in person days per year) and total time requirement for fieldwork per year (S12*S16*S17/S15) from 1 to 10000.

Personnel costs as well as material/equipment costs are highly variable between different monitoring schemes. If we take for example 40 € for an arbitrary daily salary, the schemes personnel costs (S24/S15*salary) range between 100 and 13000 €. To estimate the real costs

spent for plant monitoring for particular countries it would be better to use average national salaries for that and split it into the proportion of volunteers and professionals contributing to the monitoring for getting more accurate comparison of schemes among countries with different economical level.
In conclusion, ten monitoring schemes are listed in random order in table 4, which got higher point scores in all aspects considered in this case study. However, this should not be misunderstood as a general indicator of "goodness". Rather, these schemes perform comparably well in terms of efficiency. One scheme can be excellent according to some criteria (e.g. efficiency) but perform less when other criteria are used for the assessment. Goodness of the scheme (methods used etc.) depends strongly on the monitoring purpose, monitored species, and other criteria (even when a scheme does not score high for a particular criteria, it can still be appropriate for its purpose).

Table 4. Monitoring schemes receiving higher point scores in present case study.

Scheme nr	Scheme name
773	Relict, temperate-like, conifer forests from S Spain as early-warning indicators of climate change
795	Plant species monitoring
796	Moss species monitoring
812	Invasive alien plant species
577	Common plants survey
484	Monitoring of vascular plants
40	National bryophyte mapping
426	Monitoring the spread of *Phillyrea angustifolia*
297	Monitoring of vascular plants in The Warta Mouth National Park
17	Rare plants - plot monitoring

6.2. Insects other than butterflies and dragonflies

This case study analyses insect monitoring schemes (schemes monitoring several different taxonomic groups: BEETLES, ORTHOPTERA, OTHER INSECTS excluding: BUTTERFLIES, DRAGONFLIES). There were 16 such schemes in DaEuMon (*database date 12.12.2006*).

6.2.1. Criteria proposed for scientific quality

If monitoring scheme is **launched** from scientific interests, it may be better designed in a scientific sense and results may be easier analysable. Of the invertebrate monitoring schemes 38% (6 out of the 16) are launched for scientific interests, whereas half of the schemes (50%) are launched because of different legal obligations, only 2 schemes were conducted for management/restoration (12%) purposes.

Schemes **representativity rate** was calculated based on three aspects (answers to questions about geographical scope, sampling design, and choice of sites to be monitored). None of the schemes were representative regarding all three aspects. 12.5% of the schemes were representative for two aspects, 68.5% for one aspect, and 19% of schemes were not representative based on these questions.

The measurement precision can be estimated based on 7 questions: nature of data collected (S2), detection probability (S7), number of sampling sites (S12), number of samples per site (S13), annual frequency of sampling (S15), number of visits per site (S16), and number of

years monitored till 2006 (2006-S18). We assigned scores to each category of each of these questions scores as shown in 6.1.1. We emphasize that these points should not just be summed, because there should be different weights for each question depending on their importance for the assessment goal. There are schemes where the answer of S16 is 0 visits per site. Three schemes account for detection probability. They collect presence/absence data which has less power for detecting trends than abundance data. Most of the studies that are implemented on more than 15 sites are running more than 3 years (9 out of 11 schemes), so they seem to be the most precise schemes (a result that is corroborated by assessing the answers to the other questions as well).

The statistical power judged by coordinator is given as minimal annual change that can be statistically detected. For 37.5% of the schemes, the statistical power is supposedly low (insufficient), for 50% of the schemes the change was 10% or less, so the statistical power is high (sufficient), and for 12.5% the data entry was missing. In the latter case, the issue was may be under-considered by the scheme, and thus it may be similar to achieving low statistical power.

Statistical power based on the experimental design used (S5). Assuming that the experiment was appropriately designed to document the changes properly, a scheme can be regarded as ideal if there is a before/after comparison and a control area as well, whereas schemes are appropriate if there is some experimental design (no matter whether control are or before/after comparison), and no answer or no experimental design is not appropriate. Only one scheme is rated ideal (0.06%), 31% are appropriate, and most of the schemes (68%) are not appropriate for assessing causes of change.

Degree of refinement of the sampling design can be estimated based on questions S2, S4, S5, S6, and S7. Here to answers of each named question certain amount of points was given and the points summed. If an answer was ‘representative’, we gave 1 point, if it was not, it received a 0. If the score of S2+S4+S5+S6+S7 > 1, i.e., the scheme was representative in any respect, then we rated the scheme as representative. This counting resulted in a 50%:50% distribution of representative and not representative schemes.
The proportion of professionals can be calculated from numbers given in answers to questions S21 and S22 as: S21/(S21+S22). Furthermore, monitoring schemes that require training or expert knowledge (S23) may have a better design than those that do not require such measures but they may be more difficult to implement on large scales. There is one scheme, which needs only volunteers (No. 214), and does not need any expert knowledge. 44% of the schemes include only professionals and most of them (except 1 scheme) require also training.

The scientific quality of a scheme can be higher if field and statistical methods are more up-to-date (relevant questions here are S7, S8, and S23). Only 3 schemes are rated as having high scientific quality based on statistics. 63% (10 out of 16) use only basic statistical methods or no statistics. Three schemes do not allow a categorization, because the answer is “other”. It means that the issue was probably under-considered.
Of the 3 high schemes using more up-to-date methods, none allowed to account for detection probability, a surprising result.

6.2.2. Criteria proposed for coherence

Coherence between goals and data used in monitoring was calculated based on table 2 in D17. Most of the schemes had more than one answer for the S1 question and because of this all of the schemes (100%) fell in the appropriate category.

To estimate **coherence with EU directives and 2010 expectations** for all insect monitoring schemes is too labour-intensive for this case study. Besides, for several schemes the species names monitored (S28) are not mentioned.

6.2.3. Criteria proposed for time and cost-effectiveness

The important measure of the monitoring value is the **proportion of area monitored** per total species distribution area, but because of the same reasons named in the case study on plant monitoring an assessment could not be made.

Assessing **the taxonomical and ecological extent of monitoring**, the number of species monitored (S27) for different schemes is up to 500. For 56% of the schemes, the number of species monitored is below 100.

There are several indicators available to measure the **effort** of schemes (such as time requirement and financial resources indicators). Total time requirement to run the monitoring scheme (calculated as S24/S15) ranges approximately from 1 to 800 person days per year and total time requirement for fieldwork per year (S12*S16*S17/S15) from 1 to 684 person days per year. There was one scheme (No. 630) for which these data were not available. Comparing the time requirements for field work and total time requirement, 38% of the schemes can be regarded as complex, because the total time needed for the monitoring is more than the time needed for the fieldwork, so these schemes include other work than fieldwork. However, the majority (44%) of the schemes were guessed incorrectly by the coordinators because the total time needed is less then the time needed for fieldwork. We assume that coordinators more guessed than adequately measured the time requirements or accidentally confused the two questions. Two schemes (13%) comprise only fieldwork.

Personnel costs as well as material/equipment costs are highly variable among monitoring schemes. If we take for example 40 € for an arbitrary daily salary, the schemes personnel costs (S24/S15*salary) range approximately between 133 and 32,000 €. There was again one scheme (No. 630) for which these data were not available.

6.3. Scientific quality of butterfly monitoring schemes

The aim of the analysis was to evaluate scientific quality of the 31 butterfly monitoring schemes gathered in the EuMon database by 30. November 2006. The scientific quality was evaluated according to the criteria of data representativity and measurement precision as suggested in sections 2.1.2 – 2.1.4 of the present deliverable. In addition, we tested the performance of potential alternative criteria proposed, such as scientific interest being the reason of launching the scheme (cf. section 2.1.1) and the proportion of professionals among the monitoring participants (cf. section 2.1.7).

6.3.1. Criteria of representativity

We adopted the following three criteria of data representativity:
(i) stratified sampling design (S4);
(ii) exhaustive, systematic, or random site selection (S6);
(iii) coverage (S11) broader than 10,000 km^2.

It should be noted that the last criterion replaced that of geographic scope of the scheme (cf. Table 2), because the answers concerning geographical scope appeared clearly subjective and almost totally inconsistent with the coverage or the number of sites, at least in the case of butterfly monitoring schemes. It seems that the co-ordinators treat their schemes as national whenever they survey sites located in more than one region regardless whether their results are really representative for the whole country.

A scheme was given one point for each of the three representativity criteria fulfilled and subsequently its total score was calculated as the sum of points. A single criteria has been fulfilled by approximately one third of the butterfly monitoring schemes (for each criteria n = 11 out of 31), but there are only two schemes meeting all three representativity criteria: schemes nr. 405 and 925.

6.3.2. Criteria of precision

We adopted the following three criteria of measurement precision:
(i) possibility to account for detection probability (S7) or mark-recapture data type (S2), because mark-recapture methods by definition make it possible to account for detection probability – see Deliverable 12. (Any negative answer in S7 in the case of schemes using mark-recapture is thus assumed to be either due to misunderstanding of the S7 question or nomenclatural discrepancy – in mark-recapture nomenclature capture probability is used instead of detection probability);
(ii) number of sites (S12) larger than 15, based on the rule of thumb proposed by Joseph et al. (2006);
(iii) number of samples per site (S13) larger than 3, based on the standard of the Pollard walk method (200-m transect divided into four sections, Pollard 1977) that has proven effective in numerous butterfly studies.

The number of samples per year (S16) was not used as criterion as done for multiple butterfly surveys within a season are done to account for phenology rather than to provide temporal replicates. The number of monitoring years, which can be computed using the information on starting (S18) and ending year (S19) of monitoring as well as its annual frequency (S16) as (S19 − S18 + 1) / S16 was not used either, because the majority of schemes had undefined ending year. However, it should be underlined that virtually all butterfly schemes have been conducted for long enough to provide at least 5 annual estimates and consequently their temporal coverage can be considered adequate.

As in the case of representativity, for each of the three precision criteria fulfilled a scheme was given one point, and subsequently its total score was calculated as the sum of points. The majority of butterfly schemes have an adequate number of sites monitored (n = 18), but fewer collect enough samples per site (n = 13) or account for detection probability (n = 15). Promisingly, among the two schemes that met all the three precision adequacy criteria there was the Swiss BDM Butterfly Monitoring that performed the best also along the representativity criteria. The other one was the Butterfly Monitoring Ukraine (scheme no. 39).

In addition, under of the assumption that annual variation of butterfly populations is roughly similar among most species (see review of their coefficients of variation by Thomas et al. 1989), the index of potential statistical power of each scheme was estimated as: number of sites × number of samples per site × number of years (until present, i.e. with 2006 assumed as the ending year) × detection probability factor (5 if detection probability can be accounted for, and 1 if not; based on typical variation in butterfly detection probability, Nowicki unpubl. data). It turned out that there was absolutely no correlation (Kendall's $\tau = -0.07$, $P = 0.5973$) between the potential statistical power index and statistical power assessed by scheme co-ordinators (S9), which implies that the latter can be regarded, at most, as a measure of co-ordinator confidence in their scheme.

6.3.3. Performance of alternative criteria

Schemes launched for scientific reasons were in general characterised by slightly better precision adequacy, but on the other hand they had lower representativity. However, the differences were not significant in both cases (Kolmogorov-Smirnov test: $D = 0.1625$ for precision adequacy, and $D = 0.2792$ for representativity; $P > 0.10$ in both cases). Proportion of professionals correlated neither with representativity (Kendall's $\tau = 0.04$, $P = 0.7285$) nor precision adequacy (Kendall's $\tau = 0.06$, $P = 0.6493$). This is in fact not surprising, because it is the total number of participants rather than the proportion of professionals that matters more for achievable precision and representativity of a monitoring scheme, and a high number of participants can normally be achieved only through extensive involvement of volunteers. The above argumentation is supported by the clearly negative relationship between the proportion of professionals and the potential statistical power index (Kendall's $\tau = -0.35$, $P = 0.0061$). To sum up, none of the alternative criteria proved useful for the evaluation of the scientific quality of butterfly monitoring schemes.

6.3.4. References

Joseph LN, Field SA, Wilcox C, Possingham HP (2006). Presence-absence versus abundance data for monitoring threatened species. Conservation Biology 20:1679–1687

Pollard E (1977). A method for assessing changes in the abundance of butterflies. Biological Conservation 12:115–134

Thomas JA, Clarke RT, Elmes GW, Hochberg ME (1998). Population dynamics in the genus *Maculinea* (Lepidoptera: Lycaenidae). In: Dempster JP, McLean IFG (eds) Insect Population Dynamics in Theory and Practice. Symposia of the Royal Entomological Society 19. Chapman and Hall, London, pp 261–290