8th International Conference on Environmental Science and Technology
Lemnos island, Greece, 8 – 10 September 2003

A COMMON DATA STANDARD FOR LIFE-HISTORY TRAITS OF THE NORTHWEST EUROPEAN FLORA

MICHAEL STADLER[1], RENÉE M. BEKKER[2], IRMA C. KNEVEL[2,3], DIERK KUNZMANN[3] and JÜRGEN SCHLEGELMILCH[1]

[1]OFFIS, Escherweg 2, 26121 Oldenburg, Germany; [2]Community and Conservation Ecology Group, University of Groningen, P.O. Box 14, 9750AA Haren, The Netherlands; [3]Landscape Ecology Group, Carl von Ossietzky University of Oldenburg, P.O. Box 2503, D-26111 Oldenburg, Germany
E-mail: stadler@offis.de

EXTENDED ABSTRACT

To provide an open European-wide database of plant traits relevant for the conservation and sustainable use of biodiversity in changing European landscapes, the LEDA Traitbase project was recently started. Its aim is to support different users including land use managers (from farmers to landscape planners), environmental agencies, policy makers, and researchers with a tool in their respective work. The central part of the tool will be a trait database (the Traitbase) that can be coupled to consumer-tailored spatial sites and habitat information and will be accessible through a user-friendly WWW interface for data retrieval and data mining. Starting with 3000 species of the flora of Northwest Europe, the database will be built from scattered national database initiatives, literature sources and measurements, focussing on over 20 plant traits that describe three key features of plant dynamics: *persistence*, *regeneration* and *dispersability*.
An important design step in the LEDA Traitbase is the development of a data standard that allows transferring information from existing databases and literature into the new database. It must also effectively support analyses for planning, nature conservation and restoration instruments.
To achieve these goals, the data standard provides definitions for attributes for plant traits, habitat types, geographical references, measurement methods, bibliographic references, and further specifications. The trait attributes were chosen to provide information for a broad spectrum of uses while at the same time ensuring that trait data is not too sparse to present a sound and reliable basis for statistical evaluation.
Other aspects taken into account include the definition of mappings between attribute domains to accommodate for data of different level of detail, laying down permitted types of aggregation, fixing requirements upon data quality – both from the database application domain and computational point of view – and devising procedures for coping with data of different quality.
A precise definition of mappings between domains of attributes present in source databases and domains of attributes in the new database are necessary in order to allow a correct automatic conversion of existing data into a format conforming to the new data standard.
All traits and their possible values will be precisely defined for users or contributors to the LEDA Traitbase. In order to maximise usability of Traitbase data the standard takes into account different aggregation types which are not limited to statistical aggregation but include heuristic types of aggregation defined by experts on a per-trait basis as well.

Key words: Biodiversity, clonal growth, dispersability, persistence, ecological database, regeneration, seed, nature conservation, data quality, plant traits, aggregation methods

1. THE LEDA TRAITBASE PROJECT

Changing land use, pollution, eutrophication and fragmentation, and dereliction of traditional landscapes are major challenges to the preservation and sustainable use of European biodiversity. Recently, an international group of scientists started the LEDA project to build up a Europe-wide database of life-history plant traits. This LEDA Traitbase will provide a species-trait matrix with referenced information under the control of an editorial board.
The major challenges of the project are to (1) to pool trans-national expertise on the functional significance of traits, their classification and measurement, while avoiding unnecessary duplication of national initiatives, and (2) to predict plant biodiversity in a changing landscape. For the latter, we need to know if plants can persist and regenerate in their existing habitats and/or can colonise new habitats. Both abilities depend on their biological traits, i.e. clonal growth, reproduction, seed bank longevity, and dispersability. To be accepted by the public, the LEDA Traitbase needs to be as complete as possible, easily accessible and interfaceable.
The LEDA Traitbase uses a reviewing process supervised by an editorial board and supported by the database to protect quality of stored data.

2. LEDA TRAITBASE DATA STANDARDS

Three months after the start of the project, the LEDA consortium developed LEDA Data Standards and Operational Guidelines, including precise definitions of the trait attributes for plants, a specification of a format for bibliographic references, and a protocol for additional measurements. These standards can be divided into general standards, covering specifications for descriptive data that applies to almost all trait data, and trait-specific standards.

2.1. General Standards

The general LEDA Traitbase standards define the information needed to assess the provenance of the trait data. Each entry in the database must contain the species name, the geographical location, the habitat type, the information source (publication, database, measurement), the general method of measurement, the date of sampling, and similar information.

Habitat type
The information on habitat types includes vegetation type, soil properties, and management policy. For instance the vegetation at the sampling site can be specified using the hierarchical classes defined by EUNIS [9] while for soil properties coarse classes are used or when the information is available the detailed FAO soil classification system is adopted.

Method of measurement
The method employed to get trait data for a plant species is an important information to assess data (cf. sections 3 and 4.3). LEDA distinguishes the following seven coarse categories: (1) estimated, (2) derived from morphology or photos, (3) field or laboratory observations, (4) field or laboratory experiments, (5) experiments according to protocols using field material with known (preferred) or unknown origin, (6) modelling, (7) other.

References
Each entry in the Traitbase has an attribute specifying the source of its data. LEDA uses three types of information sources, namely publications, databases, and personal measurements. For each of these types, the general data standard defines a set of attributes. Publications, for example can be described by giving author names, editor names, title of the publication, publication year, publisher, journal name, and location

specification. Location specifications for journal or book chapter publications include the volume and page numbers, and for 'grey' literature (e.g. theses, reports) the actual location (where to find it) is given. For personal measurements, the name and email address of the person can be specified. In addition to the direct source of the data, each entry has an optional second reference to specify the original source, e.g. for data imported from a database that was originally taken from literature. For each source each separate experiment in a data source is regarded as a trial. The number of each trial ensures the uniqueness of the inserted data and avoids duplication.

Geographical references

To be able to trace back geographical variation within data from different sites, each entry in the Traitbase has a reference to the sampling site. The geographical location of this site can be described with the following attributes: country, study area (whether trait measured in Northwest Europe or not, for data about European plants from other countries), altitude, co-ordinates (longitude/latitude according to the UTM grid) and range. The range specifies the approximate size of the sampling site (Table 1).

country	study area	altitude	latitude			longitude			range
		m	degrees	minutes	direction	degrees	minutes	direction	m
UK	1	95	51	32	N	4	10	W	20
US	0	100	40	57	N	81	47	E	2

Table 1: Example of geographical references in the LEDA Traitbase.

2.2. Trait-specific standards

The focus of the LEDA project will be on plant traits that describe three key features of plant dynamics: *persistence*, *regeneration* and *dispersability* (Table 2).

Functional type	*Traitbase traits*
Persistence	Canopy height, leaf size, leaf distribution along the stem, shoot growth form, specific leaf area, tissue density, clonal growth strategy, bud bank, clonal growth organs, role in plant growth, life-span of a shoot, persistence mother-daughter connection, number of daughter shoot/ mother shoot/year, lateral spread
Regeneration	Plant life span, age of first flowering, seed number/shoot, seed weight, seed size, seed shape, seed bank longevity
Dispersability	Morphology of dispersal unit, releasing height, terminal velocity, external and internal animal dispersal, buoyancy

Table 2: Plant traits included in the LEDA Traitbase given per functional type

For each of these traits, there is a specific data standard describing the standardised protocol of measurements and the type of data, accepted by the LEDA editorial board. The trait standard contains information on, for instance, the trait definition and how to collect treat and measure the samples (see also [3]). The traits fall into two categories: *numerical* and *nominal*. Nominal traits need only one attribute for the actual value together with a list or hierarchy of valid class descriptors and precise class definitions. Numerical traits have attributes for mean, median, minimum, maximum, standard deviation, standard error, and number of observations and replicates together with a range of possible values and the maximum required precision. Single raw values are entered as mean values with one observation as shown in table 3 for the traits *specific leaf area* (numerical) and *leaf distribution along the stem* (nominal). The latter is a sub-

trait of the super trait *growth form* [5]. All known sub-traits of *growth form* use nominal classification systems with supposedly clear definition (or description) of each class [1]. In contrast, *specific leaf area* (SLA) is a more precisely defined trait, being the *one-sided area of a fresh leaf divided by its oven-dry mass, expressed in mm² /mg* ([3], p. 21). So, the numerical traits seem to be more suitable for statistical analysis. However, both traits can be used to predict plant diversity under changing environmental conditions with high quality results (ecological soundness).

	unit	validity range	taxon	nominal value	minimum	maximum	median	mean	std.dev.	std.err.	reference	original reference	method of measurement	n-number	replicates of sampling	date of sampling/collecting	geo reference	soil properties
specific leaf area (SLA)	mm²/mg	0-100	x		x	x	x		x	x	x	x	x	20	10	x	x	x
leaf distribution along the stem	nominal		x	x							x	x	x			x	x	x

Table 3: Two examples (persistence traits) of required (x) data structure.

Plasticity and Variation

In the introduction, we talked about a species-trait matrix. However, this is only a simplified view, since we expect to have several values in a cell, i.e. for a given combination of trait and species. Each value is represented by one entry in a trait-specific table, with all the attributes outlined above. The whole set of values describes the plasticity of the species in this trait, while the contents of the whole table shows the variation of the trait in general. To actually get a matrix with (at most) one value per trait per species, the set of values can be aggregated per cell, for example into a mean. See section 4 for more information about aggregation.

3. DATA STANDARD REQUIREMENTS ON DATA QUALITY

High data quality is essential to get reliable analysis results. The LEDA data standards therefore put emphasis on high data quality, from both a computational and an ecological point of view. Computational data quality is often of syntactical nature, e.g. precision or existence of attribute values, while ecological data quality is mostly concerned with how to properly measure new data. In this section, we discuss both kinds of data quality requirements.

3.1. Computational data quality requirements

Data quality requirements from the computational point of view fall into two categories, namely database schema design rules and statistical soundness criteria.

Database schema design rules require for example that attribute values are atomic, i.e. have no internal structure as far as the database is concerned; this is called the first normal form. Thus, if you want to search for publications by author name, each author name should be an individual attribute value, in contrast to having the list of all authors in one attribute of a publication. Other normal forms are concerned with functional dependencies, uniqueness constraints, or required subset relationships between attributes in different relations (tables).

Statistical soundness criteria ensure that aggregation operations yield meaningful and reliable results. For example, a mean value without the number of observations that it represents cannot be further aggregated and should therefore be disregarded in such analyses (see section 4.3).

3.2. Ecological data quality requirements

Data quality from the ecological point of view addresses, among others, the comparability and reproducibility of measurements using standardised protocols and descriptive data about the sampling sites, such as geographical location and soil properties. For example: How many replicates should be collected to cover a reasonable range of variation among individuals? Depending on the trait in question, 3 up to 25 replicates could be required. Often, only few individuals per species can be found within a sampling site, especially for rare species. The requirements on data quality involve the whole range from required data for a standard case (e.g. 10 replicates) to the absolute minimum of 3 replicates or sometimes even no replicates. In the context of statistical analyses, the results have different data qualities (see section 4.3).
It is indispensable to gather new data following a standard protocol, which encompasses the entire procedure of measurements starting from *plant collection* up to *measuring* and *use of published data*.
An important problem encountered when defining data quality within the LEDA data standards is difference in the point of view that ecologists take compared to computer scientists. Ecologists often have to adapt standards for each species to account for the different anatomy, morphology, and life-history of the species. For example, to measure the *canopy height* many different methods are used. Therefore, the results are often obtained by using different standardised methods albeit representing the same standard. Comparing or aggregating such results is natural and pragmatic to ecologists and it is sound to do so.
Another topic is the ecological soundness of using the mean of maximums for traits such as *seed number production*, even if the corresponding minimum is missing. In these cases, the researcher is more interested in getting the highest potential of a trait, like the upper percentile or quartile from a range of values than the real mean.
What is a low, what is a high data quality or what are different data qualities? From an ecological point of view, the data quality of nominal traits like *leaf distribution along the stem*, primarily depends on the clearness of definitions for each class. For computer scientists, the quality of nominal traits in addition depends on the structure of the taxonomic tree describing which class is a specialisation of which other class. In the case of numerical traits, data quality from both ecology and computer science points of view are equivalent: the more replicate values there are available the higher the data quality is.
In the LEDA Traitbase, high data quality is ensured by a review process performed by the LEDA editorial board consisting of 6 external experts and 10 LEDA consortium partners. A requirement for data to be accepted is that it can be located (e.g. in publications, reports), and a description of the methods used should be provided to the editorial board to allow to check the validity of data.

3.3. Level of detail of stored data

In order to maximise the number of possible queries and to minimize loss of data, it was decided to generally enter data as detailed as possible into the LEDA Traitbase. Modern database systems on current hardware will have no problem to handle the estimated amount of data. The additional work for entering many raw values rather than few pre-aggregated values will be minimized by the user interface of the data input program.
The following rules express the preference of detailed data over pre-aggregated data:

***Raw values vs. aggregated values*:** Values of individual measurements should always be entered separately, if they are separately available. This is in contrast to first aggregating the raw values into, for example, a mean value and then entering that mean together with the number of raw values, the standard deviation and the standard error.

***Raw values vs. value classes*:** Values should always be entered directly instead of first mapping them to classes and then entering class descriptors into the database.

- When a single numerical, ordinal value is given, this value should be entered directly instead of first mapping it into a class and entering that (e.g. “2.5” instead of “low pH”).
- When an interval of numerical values is given, the interval’s upper and lower bounds should be entered (into two suitable attributes) instead of a descriptive value (e.g. lower bound “1” and upper bound “2.5”, instead of “1 < pH <= 2.5”).
- If a class descriptor is given, the bounds of the associated interval should be entered (e.g. if “low pH” is given, and that class descriptor represents the interval from 1 to 2.5, then lower bound “1” and upper bound “2.5” should be entered).
- When a non-ordinal qualitative value as e.g. *triangle* or *rectangle* is given, this value should be entered directly instead of entering *polygon*.

Classes or aggregated values can be calculated from the raw data, if needed as answers to queries. This allows to adapt the classification (class boundaries, number of classes, granularity etc) to changing requirements. However, the described approach of post-aggregation requires appropriate algorithms and mappings to be defined.

4. DATA AGGREGATION CONSIDERATIONS

The LEDA Traitbase encourages users to enter raw rather than aggregated data. However, aggregation is still useful to condense the amount of raw data into meaningful information. Storing raw data allows the LEDA Traitbase to apply different aggregation methods depending on user needs and several considerations. These stem from differing trait characteristics as well as differing soundness concepts in statistics and ecology.
Aggregation requires that values to be aggregated must be stored in attributes on their own, i.e. not mixed up with other values. For example, to aggregate upper bounds of plant height into a mean value, this bound must not be combined with a lower bound into one attribute value like “15m – 25m”; otherwise, database aggregation functions cannot be applied. It follows that non-atomic values such as sets or lists cannot be aggregated.
The LEDA Traitbase will support several types of aggregation: (1) aggregation of numerical values, (2) aggregation of nominal values, (3) non-statistical aggregation and (4) subset aggregation. Aggregation of numerical values into e.g. mean or median values is well understood. We discuss the other types of aggregation in the sections below.

4.1. Hierarchies for aggregating nominal values

Nominal values can be aggregated using a taxonomic tree with the possible nominal values as nodes. These nodes represent classes of values, with those higher up subsuming all classes below them in the hierarchy. The aggregate value for a given set of input values is then their lowest upper bound in the hierarchy. In the example tree in figure 4, an aggregation operation with input values *F4* and *F3.1* will result in value *F* while an aggregation of values *E1* and *F* will result in the special value “*” meaning “*no information*” – there is no common abstraction for *E1* and *F*.

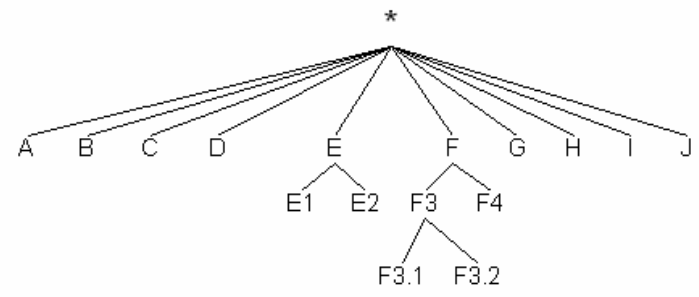

Figure 4: Part of the EUNIS habitat classification tree (as realised in [9])

This type of aggregation lets the LEDA Traitbase cope with data from different taxonomic levels, which is important for merging other data sets into the Traitbase even if their domains for nominal attributes differ from the corresponding LEDA trait standard. However, this requires that a joint taxonomy can be defined prior to aggregation.
Another benefit of this kind of aggregation is that it allows to enter data keeping the domains originally used in the data sources. In principle, this kind of aggregation operation also works for nominal values with an underlying ordinal scale as long as the ordering is preserved by the taxonomic tree.

4.2. Non-statistical aggregation

Besides statistical aggregations, the LEDA Traitbase will support some aggregations of non-statistical nature, which are often needed e.g. for extracting ecologically meaningful statements about a species' dependency on environmental factors. These aggregations use a non-standard algorithm to condense sets of input values, often from different attributes, into a single value. Both nominal and numerical values are allowed as input and output of such algorithms which are often given as decision trees in the literature [8].
An example of a non-statistical aggregation is the calculation of *seed longevity*: The trait *Soil Seed Bank Type* gives information about whether a species seed bank is *transient*, *short-term persistent*, *long-term persistent*, or *present* within a given habitat. The species' seed longevity (a value between 0 and 1 where higher values denote higher persistence and longevity) within a given set of locations can be calculated with the formula [2]

$$seed\_longevity(spec, Locs) = \frac{\#short\_term\_persistent(spec, Locs) + \#long\_term\_persistent(spec, Locs)}{\#transient(spec, Locs) + \#short\_term\_persistent(spec, Locs) + \#long\_term\_persistent(spec, Locs)}$$

where # denotes the number of records in the database for a given species *spec* and set of locations *Locs*. Note that the formula does not take into account the SeedBankType value *present*, because this value does not allow judgement about longevity.

4.3. Subset aggregation

The LEDA Traitbase is intended to support potentially costly decisions and therefore should provide results of high quality. However, the quality of an aggregation result depends critically on the quality of the input values. The LEDA data standards allow input of data of different quality, both statistically and ecologically. However, aggregation over data of mixed quality does not always deliver meaningful results.
Therefore, users can restrict aggregation operations to data sharing common quality indicators only. This will be done by selecting rules to constrain aggregation, with some rules active by default and some that cannot be deactivated. A simple example of a rule that is active by default is the *rule for statistically sound aggregation into mean values*.
In older literature, trait data given as mean values may be lacking any indication about the number N of observations contributing to these mean values, or the standard deviation and standard error. In contrast, recently conducted measurements for the same trait are likely to include both N and the standard deviation. It is thus not sound to aggregate both data into a single mean value for the trait in question.
The *rule for statistically sound aggregation into mean values* forbids statistical aggregation of the former kind of information. Thus, only the subset of high-quality data will be aggregated, giving reliable results. However, sometimes any data is better than no data at all; therefore, users may define statistically unsound aggregations but will be warned about the unknown quality of the result.
In addition to statistical soundness, there are ecological soundness criteria that forbid certain aggregations. These often result from different interpretations of imprecise

definitions. There are not rules for coping with these problems yet, but they are intended to be part of the constraints of the LEDA Traitbase analysis system.

5. LEDA TRAITBASE: PRESENT AND FUTURE

The LEDA Data Standard and the database schema are completed, and rules for ecologically sound aggregation will be devised. In November 2005, the Traitbase will be open and accessible on the world-wide web for all to use. Intensive discussions with stakeholders (land use managers, policy makers, researchers) are on the way to find out how they want to use it. Future plans include the development of expert systems based on the Traitbase.
For future data use and exchange we will work towards the geographical information system (GIS) grid of Northwest Europe that is attached to vegetation relevés information from the syntaxonomic biological system (SynBioSys Europe [7]). Full automatic taxonomic update of European species names from the Euro-Mediterranean plant diversity database (EURO+Med [4]) will be possible in the future.
The LEDA Traitbase will be most useful for everyone if it is as complete as possible, but at present, for only approximately 35% of the species-trait matrix data is known [6]. Therefore, the LEDA consortium is grateful for the help of the scientific community. So, we would like to ask you to contribute to the LEDA Traitbase by contacting the LEDA Traitbase secretariat (see http://www.leda-traitbase.org).

ACKNOWLEDGEMENTS

The LEDA Traitbase Project (contract no. EVR1-CT-2002-40022) is funded by the fifth Framework programme of EESD programme.

REFERENCES

1. Barkman, J. (1988), New Systems of Plant growth forms and phenological plant types. In: Werger, M.J.A, Van der Aart, P.J.M., During, H.J., Verhoeven, J.T.A. (eds.): *Plant form and vegetation structure*, pp. 9-44. SPD Academic Publishers, The Hague.
2. Bekker, R.M., Schaminée, J.H.J., Bakker, J.P. and Thompson, K. (1998) Seed bank characteristics of Dutch plant communities, *Acta Botanica Neerlandica* 47, 15-26.
3. Cornelissen, J.H.C., Lavorel, S., Garnier, E., Díaz, S., Buchmann, N., Gurvich, D.E., Reich, P.B., Ter Steege, H., Morgan, H.D., Van der Heijden, M.G.A. and Pausas, J.G. (xxxx), *Handbook of protocols for standardised and easy measurement of plant functional traits world wide*. In prep.
4. EURO+Med (2003), URL: http://www.euromed.org.uk
5. Klimes, L. and J. Klimesova (1999), CLO-PLA2 - a database of clonal plants in Europe. *Plant Ecology* **141**, 9-19.
6. Knevel, I.C., Bekker, R.M., Bakker, J.P. and Kleyer, M. (2003), Life-history traits of the Northwest European flora: A data-base (LEDA). *Journal of Vegetation Science* (accepted).
7. SynBioSys Europe (2001), URL: http://www.synbiosys.alterra.nel/eu/
8. Thompson, K., Bakker, J. P., and Bekker, R. M. (1997) *Soil Seed Banks of Northwest Europe: Methodology, Density and Longevity.* Cambridge University Press, Cambridge.
9. EUNIS (2003), URL: http://nature.eionet.eu.int/activities/products/eunishaben.html