EnviroInfo 2004 (Geneva)
Sh@ring – EnviroInfo 2004
 ISBN: 28 29 30 275-3

Data Quality, Abstraction and Aggregation in the LEDA Traitbase

Michael Stadler[1], Dierk Kunzmann[2], Jürgen Schlegelmilch[1] and Michael Sonnenschein[3]

Abstract

In providing an open European-wide web-based database of plant traits relevant for the conservation and sustainable use of biodiversity in changing European landscapes, the LEDA project aims to support different user groups by providing information about plant traits related to the key features of plant dynamics. The trait database (Traitbase) is currently built from scattered national database initiatives, literature sources and new measurements contributed by persons from the ecological and botanical research community.
Here, we discuss the problems of data processing and output customisation arising from the different information needs of three user groups and the heterogeneous data quality in the Traitbase due to the different types of data sources. To this end, we first identify the user groups and their needs. Then, we introduce the terms aggregation and abstraction for the given context. After this, we discuss issues of data quality and finally we go into details of the aggregation methods preferred in generating output for the different user groups.

1. The LEDA Traitbase

The LEDA project (Life-history traits of the northwest European flora – a database) is funded by the European Union (contract no. EVR1-CT-2002-40022) and aims to support different user groups by providing information about plant traits related to the key features of plant dynamics (Knevel 2003), (LEDA 2002). For this purpose,

[1] OFFIS, Escherweg 2, D-26121 Oldenburg, Germany, Email: {stadler|schlegelmilch}@offis.de, Internet: http://www.offis.de
[2] Carl von Ossietzky University of Oldenburg, Landscape Ecology Group, P.O. Box 2503, D-26111 Oldenburg, Germany, Email: dierk.kunzmann@uni-oldenburg.de, Internet: http://www.uni-oldenburg.de/landeco/
[3] Carl von Ossietzky University of Oldenburg, Dept. for Computer Science, D-26111 Oldenburg, Germany, Email: Michael.Sonnenschein@Informatik.Uni-Oldenburg.DE, Internet: http://www-ui.informatik.uni-oldenburg.de

LEDA implements an open European-wide database (*Traitbase*) of plant traits relevant for the conservation and sustainable use of biodiversity in changing European landscapes. The Traitbase will be accessible through a user-friendly web interface for data retrieval and data mining. Starting with 3000 species of the flora of northwest Europe and about 30 plant traits, the database is being built from scattered national database initiatives, literature sources and new measurements contributed by persons from the ecological and botanical research community.

Today, the database schema of the LEDA Traitbase is fixed, the interfaces for web-based data input and for simple queries can be used by project partners, and a query interface offering different aggregation and abstraction methods is constantly being improved. Upon its completion in November 2005, the Traitbase software will additionally include a specific data input component for public contributions supervised by an editorial board, an exemplary link of the Traitbase to a geographical information system on plant distribution (FloraWeb), and a data mining component for identifying new interrelations in data sets extracted from the Traitbase.

The LEDA consortium consists of 10 partners from six countries co-ordinated by the University of Oldenburg in cooperation with the University of Groningen. The institute OFFIS (Oldenburg) is responsible for all the computer science aspects in the project.

1.1 The Traitbase user groups

We distinguish three user groups in LEDA. A user is assigned a role at time of registration to the Traitbase services. He or she may later on change his role and thereby get different output from the system.

The first user group consists of political decision makers who need quite abstract, easy to understand and at the same time unambiguous answers to rather unspecific queries. They are initially assigned the role *decision maker*. A typical question of such users to the Traitbase might be: “Which of the given habitats is more promising for restoration of a given list of endangered species in a short time?” Although the answer should simply be “this habitat”, it requires an analysis of values of several traits of importance to persistence, regeneration and dispersability. It is most probable that the access of decision makers to the Traitbase is indirect via decision support systems that need a small amount of key-values for yielding understandable output. Thus, this user group is best supported by delivering a single value or a single aggregated value per trait and species (at a given location and a given habitat type) or a single diagram per trait and group of species.

The second user group consists of land-use managers in a general sense and referees in environmental agencies. They are assumed to have at least basic knowledge of relationships between plant species, their traits, growth, and evolvement under given environmental conditions. For their decisions they need rather specific infor-

mation about plant traits of species in given regions or given habitat types. We assign to these the initial role *referee*. An exemplary question of this user-group is: "How often should grazing be allowed on a given meadow in order to favour a specific rare species?". To answer this type of questions, rather specific information about plant traits of species in given regions or given habitat types is needed. Thus, referees should be free to select from multiple predefined aggregation methods.

The third user group consists of researchers from the fields of botany and ecology. They are initially provided with the role *researcher* and have very specific questions as e.g. "Is there a correlation between seed mass and seed number for a given species?". Researchers are able to properly interpret results computed from data of mixed quality and using different aggregation methods. They are also able to assess the ecological and statistical quality of results on their own, given that they know about the quality of the raw data and the aggregation methods used. Thus, this user group should be able to freely configure system output and have access to raw data.

1.2 Data Types of LEDA Traits

According to the type of data used for representation, three kinds of traits are accounted for in the Traitbase. The representation of *nominal traits* is based on nominal values. In the Traitbase, such values denote categories. Nominal values offer the least possibilities for processing. However, there is the possibility of mapping nominal values to ordinal or numeric values before processing them. An example of a nominal trait within the Traitbase is *shoot growth form* with possible values as *lianas, climbers and scramblers, stem erect, stem ascending to prostrate* and *free-floating plants.*

The representation of *ordinal traits* is based on ordinal values. Ordinal values can be ordered by meaning and within the Traitbase context they denote categories. It is not possible to perform mathematical calculations on ordinal values since the distance between adjacent values differs. An example of an ordinal trait is bud bank – vertical distribution, which for each soil layer has the three possible values: *no buds per clonal fragment*, *1 to 10 buds per clonal fragment*, *more than 10 buds per clonal fragment*.

Values of *numeric traits* are represented by numbers. It is possible to perform standard mathematical calculations with these numbers. An example of a numeric trait within the Traitbase is *canopy height* where the height of plants in meters is given in numbers.

It is important to note some properties of traits. First, the value of a trait can sometimes be expressed in different ways (e.g. the trait stem tissue density can either be expressed through nominal values woody, semi-woody and non woody or by a density value in g/m^3 as well). Secondly, a trait record often contains helper values

in addition to the actual value of the trait to be measured. These helper values must be accounted for when performing data aggregation. Third, values of numeric traits may be given as maximum, minimum, mean, or median and they are adjoined with information about the number of samples and the number of replicates used for obtaining the value. Optionally, a standard error or a standard deviation may be given for such traits, as well. Thus, a single trait record can either represent a real single observation (number of samples = 1) or it can represent a single result obtained by following the collection and measurement protocols given by the *LEDA data standards* (LEDA 2004). It is also possible, that nothing is known about the number of samples or number of replicates for values obtained by following precise rules, if they are not given in literature but are still the only values so far available for a given species and a given trait.

1.3 The Concepts of Aggregation and Abstraction in LEDA

We use the term *generalisation* to denote the process of mapping detailed data to more general information, being both easier accessible to interpretation by humans, and of lesser precision. Thus, generalisation makes data easier to interpret at the cost of information loss. A special kind of generalisation maps multiple records of data each containing values for a given set of attributes to a single record of data containing aggregated values. We refer to this kind of generalisation as *aggregation*.

There is another type of generalisation that is relevant within the context of queries to the Traitbase. It is essentially used to attain a common level of detail for data from different sources or to adapt it to the needs of users. We refer to this type of generalisation as *abstraction*. Both, aggregation and abstraction are used to present statements about plant traits in a way easy to comprehend. Another means used is the display of data in diagrams.

2. Abstraction

There are different types of abstraction used within the context of the LEDA Traitbase, namely *classification* and *domain mapping*.

2.1 Classification

Classification serves the purpose of obtaining more manageable results to queries. It essentially consists of mapping values to classes and thereby allows evaluation of data to occur on domains of smaller cardinality. Classification is offered as a method for displaying data about many LEDA traits offered to referees and researchers. To this end, classes are predefined and associated with meaningful names on a per-trait

basis. Other traits are associated with a number k of classes, each of which represent an interval of possible values calculated by dividing the range of allowed trait values into k intervals of equal width. Besides being useful for generating output for users, classification can also be to applied to data before carrying out aggregation operations.

2.2 Domain mapping

Often, data available about traits is of different level of detail. This amounts to data with different domains and is inconvenient both, because for ordinal, non-numeric values, data aggregation operators assume a common level of detail for producing meaningful results and because data of different detail displayed together is confusing to users. For numeric values, common ways of domain mapping are rounding and truncation. However, these are non-problematic cases, since they are readily implemented and sometimes in most programming and query languages and mostly performed automatically. For nominal and ordinal values, however, no such standard-methods are readily available and well-known. In the Traitbase context, we use hierarchical domain mapping that relies on hierarchies that have to be defined for each attribute domain. For example consider EUNIS habitat type keys (EUNIS 2003), a hierarchical code for habitat types used within LEDA. When given keys A, A1, A3 and B, domain mapping to the lowest common level of abstraction would yield the values A, A, A and B that can be handled in a single aggregation operation.

A special case of domain mapping is applied, when values for a trait may be given either as nominal value or as numeric value stored in different attributes. In order to apply aggregation or to homogeneously display data, it is necessary to perform a mapping to one of those attributes.

3. Data quality

When performing data aggregation, it is of importance that data quality be accounted for. Either data to be aggregated is of same quality. In this case the aggregation result can be regarded both as statistically and ecologically sound, if the same holds for the input data. Or, data to be aggregated is of different quality. In this case, special measures have to be taken in order to assure ecological or statistical soundness of the aggregation result. Within the given context we identified two types of data quality: While statistical data quality refers to the availability of all values needed to perform specific aggregation operations in a statistically sound way and the assumption of elementary prerequisites (as underlying normal distribution for mean values), ecological data quality refers to the methods used for obtaining and noting trait values (e.g. estimations noted as min/max intervals).

3.1 Categories of data quality

In our project, we introduced three categories of statistical data quality and three categories of ecological data quality in order to allow for separate aggregation of data with different quality:

(SS) Suited for statistically sound aggregation. Values representing single observations always belong to this category. Values representing multiple observations only belong to this category, if they are mean values with given sample size n and one of standard error or standard deviation, or if they are median, minimum or maximum values with given sample size n.

(HS) Suited for statistically sound aggregation after application of heuristics. An example of such values are pairs of minimum and maximum values from estimations denoting the border of an interval covering an unknown number of observations.

(NS) Not suited for statistically sound aggregation. This category e.g. contains mean values without given n and standard error, as often found in literature.

(SE) Suited for ecologically sound aggregation due to an exactly known collection and measurement protocol. An example of this is the LEDA data standard applicable to new data collected for entry into the Traitbase.

- (SE1) Exact values obtained by measurement, observation or experiments. This might be the exact number of seeds on a shoot or ramet of a given species, at a given location and a specific point in time.
- (SE2) Approximate values obtained by extrapolation.
- (SE3) Approximate values obtained by estimation. An example for this is the canopy height of a plant that is often estimated from drawings found in florae.

(HE) Partially suited for ecologically sound aggregation due to generally known method of measurement and collection methods differing from those described in the LEDA data standard. Most data from literature and imported data from databases fits in this category. Apart from stating that it has been collected and measured following the LEDA data standards, there is no means in the Traitbase database schema to distinguish between data standards. Thereby, most data from literature and other databases will fall into this category.

- (HE1) Exact values obtained by measurement, observation or experiments.
- (HE2) Approximate values obtained by extrapolation.
- (HE3) Approximate values obtained by estimation.

(NE) Not suited for ecologically sound aggregation due to unknown collection and measurement methods. All values from records having "unknown" entered as general measurement method or trait specific measurement method (if available) fall into this category. Normally, one would have to assign records with unknown collection method but known method of measurement to this category as well, but since there is no attribute provided for storing information about collection methods in the Traitbase and often there is a good guess for collection methods when a trait-specific measurement method is given, we assign such records to category HE.

The more characteristics values have in common with respect to heuristics applied to aggregate them and with respect to their ecological data quality, the less ambiguous and the easier to interpret the aggregated values will be. Therefore, categories HS, SE1, SE2, HE1 and HE2 are sometimes further refined on a per-trait basis. Categories involving estimation should not be distinguished since there is no objective means of judging the quality of estimation. When selecting values for aggregation, it is important to consider both, statistical and ecological data quality. In order to denote a value, that has a statistical quality S and an ecological quality E, we use the notation (S,E).

As an example of ecological data quality categories, we might consider the trait *seed number production per single stem*. Following the LEDA data standards the seed number can either be obtained by counting the seeds found on a single stem. The resulting value would then be of data quality (SS, SE1) since it is a single observation and an exact measurement. In many cases however obtaining the value through counting would introduce too big a workload, which is not desirable given that it takes about two hours in average to collect and measure a single trait for a single species. To alleviate this problem, the LEDA data standards for the trait in question allow for a combination of counting the seeds per flower and the number of flowers on a single stem and subsequently extrapolating the total number of seeds from these results by multiplying the obtained values. Naturally, the result of extrapolation won't be as exact as by counting the total number of seed. They would be of data quality (SS, SE2). Another allowed method for obtaining the seed number is to count the seed number per flower and to estimate the number of flowers and then extrapolating the total seed number, which might be appropriate for bushes. Last but not least, it might even make sense to estimate both values and then extrapolate or even to do a direct estimation (educated guess) of the total seed number. The latter methods are of data quality (SS, SE3). In order to ensure the most precise results possible, it makes sense to aggregate only values from a single category.

As an example of refining ecological data quality categories on a per-trait basis, consider the trait *leaf size*, where measurements can either be carried out on leaves with petiole (leaf stalk) and rachis (central axis of a pinnate leaf) or on leaves without those components. The LEDA data standards specify that values for leaf size

should be obtained using a scanner and appropriate software using a number n of leaves. The median of values obtained for a single species is then stored as measurement result which can be assigned to category (SS, SE1). In order to be able to distinguish between different kinds of leaf size measurements, two subcategories of SE1 related to trait leaf size have been introduced. These are SE1L1 and SE1L2, the former only containing leaf size values that have been obtained measuring leaves with petiole and rachis and the latter containing leaf size values obtained by measuring without petiole and rachis.

3.2 Selecting data quality for aggregation operations

By default, and for decision makers aggregation is only carried out on data of quality levels (SS, SE) and (SS, HE). Users from groups referees can additionally choose to extend aggregation to data having a statistical quality level (HS) and an arbitrary level of ecological data quality. Researchers may instruct the system to perform aggregation on all available data regardless of its quality. Note, that the system will always output appropriate warnings with results achieved by such operations. We always assume, that lower data quality levels subsume higher data quality levels.

4. Data Aggregation

Depending on trait type, attribute types and data quality, there are many different types of data aggregation that are useful for generating query results for the different Traitbase user groups. The following sections give an outline of aggregation possibilities offered by the Traitbase query interface and specify the standard methods of aggregation on a trait-type, data quality and user group basis.

4.1 Frequency of values

We distinguish between the absolute frequency of values and the relative frequency of values. The absolute frequency of values is useful for comparing the frequencies of all possible outcomes of a nominal or ordinal trait. A means for comparison can either be a list of values, a bar chart or a pie chart. In order to use this kind of aggregation for numeric traits, a classification should be performed in order to map trait values to classes and then calculating the absolute frequencies of occurrence of those classes. Whenever estimated values are taken into account in an aggregation operation, absolute frequencies are preferred, since they allow users to take into account shifts between neighbouring classes.

4.2 Ratio of Frequencies

Sometimes it is of interest to express the frequency of a subset of possible values with respect to a different subset of such values. When applied to ordinal or nominal values representing categories, this may be interpreted as calculating the frequency of a set of categories to the frequency of another set of categories. Dependent on the trait, different ratios of frequencies might be of interest. A meaningful ratio is a good candidate for single values delivered to users at the management level. The seed longevity index (Thompson 1998) expressing the potential longevity of a species in an environment calculated from the amount of seed in different layers of the soil seed bank is a prominent example of applying this kind of aggregation.

4.3 Modal values

The value with the highest frequency of all values allowed for a trait is called the modal value. The modal value is especially interesting for nominal traits, where each value denotes a category. The category represented by the modal value is then called the modal category. For most ordinal, category-valued traits, output of the modal category has been chosen as default for supporting decision makers and referees.

4.4 Quartiles and median value

The median value and other quartiles are computable for traits whose values are ordered, i.e. for ordinal and numeric traits with arbitrary underlying frequency distributions. Whenever calculating quartiles for numeric attributes, values are weighted according to the number of samples they represent. For decision makers and referees, quartiles or median value have been chosen as standard output for aggregation of values for numeric traits, if available values do not represent interval borders and they are of data quality (SS,SE).

4.5 Minimum and maximum values

In the given context minima and maxima can be interpreted in two different ways. First, minima and maxima calculated for a given set of records can be regarded as separate values. Secondly, they can be interpreted as upper and lower borders of an interval. In the former case, minima and maxima can be calculated for attributes mean, minimum or maximum of numeric traits. In the latter case the minimum of minima and the maximum of maxima are both calculated and then interpreted as borders of a resulting interval. The latter method of aggregation is proposed as default method for responding to decision maker and referee queries whenever available data consists of min/max pairs representing intervals.

4.6 Mean values

If statistical soundness is to be preserved, a mean value can only be calculated for numeric traits with an underlying normal distribution. Since the number of samples from which values in the Traitbase are calculated may vary (i.e. it may contain mean values of multiple observations as well as values resulting from single observations) a weighted mean has to be calculated. In general the values of traits accounted for in the LEDA Traitbase cannot be assumed to be normally distributed. In biology, it is common to calculate means even from such values. Thus, for researchers we offer the possibility of calculating not-statistically sound means, but users are reminded of the lacking statistical support and calculating means is never offered as default aggregation method.

5. Conclusion

Different kinds of data aggregation are needed to support the different user groups of an ecological information system. The results of such operations depend both, on the level of statistical data quality of raw data and on the level of ecological quality of this data. In order to allow a precise interpretation of aggregation results, users have to know about the data quality of raw data used. To produce results of known ecological quality, it is useful to apply aggregation operations only to sets of data with homogeneous ecological data quality. Whenever this is not possible due to data availability issues, it is preferable to use aggregation methods as e.g. the calculation of modal values that have a rather low error propagation. As a basic principle, when delivering information to decision makers only data of highest available ecological

and statistical data quality should be taken into account and preferably a single aggregation method per characteristic of interest should be used.

Bibliography

EUNIS (2003): http://eunis.eea.eu.int/habitats.jsp

FloraWeb : http://www.floraweb.de

Knevel, I.C. (et. al.) (2003): Life-history traits of the Northwest European flora: The LEDA database, Journal of Vegetation Science, Vol. 14, P. 611-614

LEDA (2002): http://www.leda-traitbase.org

LEDA (2004): Knevel, I.C. (editor): Collecting and measuring standards for Life-history traits of the Northwest European flora, to be published

Thompson, K. (et. al.) (1998): Ecological correlates of seed persistence in soil in the Northwest European flora, Journal of Ecology, Vol. 86, P. 163-169