Project no. GOCE-CT-2003-505298
ALTER-Net

A Long-Term Biodiversity, Ecosystem and Awareness Research Network

SERONTO

A *S*ocio-*E*cological *R*esearch and *O*bservation *Onto*logy: the core ontology

Bert van der Werf[1], Mihai Adamescu[2], Minu Ayromlou[3], Nicolas Bertrand[4], Jakub Borovec[5], Hugues Boussard[6], Constatin Cazacu[2], Toon van Daele[7], Sabina Datcu[2], Mark Frenzel[8], Volker Hammen[8], Helena Karasti[9], Miklos Kertesz[10], Pirjo Kuitunen[11], Mandy Lane[4], Juraj Lieskovsky[12], Barbara Magagna[3], Johannes Peterseil[3], Sue Rennie[4], Herbert Schentz[3], Katharina Schleidt[3], Liisa Tuominen[13]

[1] Alterra – Wageningen University and Research Centre; The Netherlands
[2] Bucharest University – Department of Systems Ecology; Rumania
[3] Umweltbundesamt – Federal Environment Agency; Austria
[4] Centre for Ecology and Hydrology; United Kingdom
[5] Centre of Academic Sciences; Czech Republic
[6] French National Institute for Agricultural Research – INRA; France
[7] Research Institute for Nature and Forest – INBO; Belgium
[8] Helmholtz Centre For Environmental Research – UFZ; Germany
[9] University of Oulu; Finland
[10] Institute of Ecology and Botany; Hungary
[11] University of Jyväskylä; Finland
[12] Institute of Landscape Ecology – Slovak Academy of Sciences; Slovakia
[13] Finish Environmental Institute – SYKE; Finland

Deliverable type: Report

Deliverable reference num.: 4.I6.D2

Instrument: Network of Excellence
Thematic Priority: Global Change and Ecosystems (Sub-priority 1.1.6.3, Topic 6.3.III.1.1)
Due date of deliverable: 28.02.2008
Submission date: 09.04.2009
Start date of project: 1st April 2004
Duration: 5 years
Deliverable lead contractor: Umweltbundesamt GmbH (Austria)
Revision: 1.0
Work Package: I6
Document ref number: WPI6-2009-10

SERONTO

A *S*ocio-*E*cological *R*esearch and *O*bservation *Onto*logy: the core ontology

CONTENTS

Short summary

SERONTO is an ontology developed within ALTER-Net, a Long Term Biodiversity, Ecosystem, and Awareness Research Network funded by the European Union. ALTER-Net addresses major biodiversity issues at a European scale. Within this framework SERONTO has been developed as a new approach to deal with the problem of integrating data from distributed datasources stored and collected at different locations within the European Union. SERONTO is a work in progress product of a group of people with diverse scientific backgrounds.

The ontology is a formal description of the concepts and relationships for the most important aspects of biodiversity data derived from monitoring, experiments and investigations. SERONTO is an ontology that enables seamless presentation of data from different sources in a similar conceptual manner.

SERONTO aims to facilitate meta-analysis, data mining, and data presentation across a wide variety of datasets collected for different purposes. SERONTO consists of a core ontology accompanied by (research) domain specific ontologies. The SERONTO core ontology describes the fundamental concepts, relationships and structure. The domain specific ontologies (e.g. species, geography, water, vegetation) extend these concepts, relationships for their specific needs and requirements. The development of the core is based on statistical methodological concepts.

Important considerations in designing SERONTO were:

- *Repeatability*: The ontology should be capable of holding enough meta-data that another person can repeat the experiment or observation at another place and time. It is not obligatory, however, to provide all information for all datasets; for instance, some information may be missing for old datasets.
- *Transparency*: It must be possible to record and retrieve meta-data describing what actually happened. SERONTO includes concepts for sampling designs and for things going wrong and documenting data collection under less than ideal conditions. If data and meta-data are available in this way, it's clear what assumptions must be made to combine data and correctly interpret analyses.

Important concepts in the SERONTO core are:

- *Physical things* - the research object or experimental unit
- *Parameters* – the measurement, classification and treatment of the investigation item
- *Value sets* – placeholders for time series and other complex data
- *Reference lists* – nominal values, such as species lists
- *Methods* – used for each parameter, including units, scale, and dimensions
- *Sampling structure* – the origin of the research object or population, and the way it was chosen
- *Groupings of objects*, such as experimental blocks, on which observer, time or other aspects are assigned or related to
- *Additional information*, such as actors (observer, observer groups and institutions), project information, etc., can be attached to several different concepts.

1 Introduction

1.1 Problem

During the last decennia there's an increasing threat to the European biodiversity due to human activities. For countering biodiversity loss by means of management measures it is essential to have a good understanding of biodiversity and the pressures it faces. Due to different languages and schools the biodiversity research has been fragmented. New field research, European wide, is very costly and takes much time and effort. Many questions, such as climate change have however an urgent character. Re-use of combined biodiversity data and their pressures measured in different contexts with data-mining and other meta-analysis techniques seems appropriate.

1.2 Data

Biodiversity data found in Europe is in general complex, often originating from long term monitoring schemes. Monitoring schemes in which biological along with physical-environmental data are monitored for such a long period that sampling and measurement methods often have been changed at some point. Depending on the research question the spatial scale of the data can range from individual to landscape and with the inclusion of satellite images even larger. Time scales can be short or span several decades and the purpose of the data gathering varies between datasets resulting in different types of experimental units. For a good understanding of the pressures on biodiversity socio-economic data should also be considered as well as data coming from experiments. Experimental data can have observations which lie outside the natural distribution and thus being able to describe situation which does not (yet) exist. Thereby diminishing the need for extrapolations. With socio-economic data language problems are evident. To add to this complexity, not all data are stored in files, spreadsheets or databases. There are still huge amounts of useful data on paper. Metadata about the data such as sampling and measurement methods are often missing and not always recoverable. In order to use this incomplete data in a meta analysis setting assumptions has to be made explicit or derivable from the data.

1.3 Context

In order to halt the loss of biodiversity by 2010 the European Commission established in 2004 the ALTER-Net project within the Framework VI research program. ALTER-Net, A Long Term Biodiversity, Ecosystem and Awareness Research Network, addresses major biodiversity issues at a European scale. The work presented here is done within a workgroup of ALTER-Net which had the objective to construct a framework within which can be built a system to manage biodiversity data, information and knowledge from the Network and to make them available to scientists, policy makers and the public. The first step in this is making observations available for scientists as this should be at the most detailed level. Information derived from that can at a later stage be made available for other audiences.

1.4 Solution

An ideal system to share and combine such diversity of data should be able to use a consistent view on data whatever the origin and domain of it. It should be possible to make new combined datasets for further scientifically sound analysis. It should be flexible enough to add new scientific domains and extend existing ones and to hold incomplete data. Queries on the data should be able to use concepts (e.g. give me all experimental units where a chemical parameter has been measured) across different domains. In addition to that, data should be distributed, but for the user being viewed as coming from one database.

1.5 Semantics

Umweltbundesamt, Federal Environment Agency (UBA) already had long time experience in storing and retrieving long term monitoring data. They developed the software application MORIS which uses class concepts and semantics in a successful way. As this software is in German and current ontology programs such as protégé makes it more easy to use semantics in a class-instance environment, it was decided to re-engineer their solution using the latest concepts. The experiences they had with MORIS proved valuable. For a exelent and concise introduction into ontologies see http://en.wikipedia.org/wiki/Ontology_(information_science).

1.6 Design concepts ontology

For defining SERONTO we used the free software package protégé ™. The owl file type is owl dl, in which reasoning is still possible. The conceptual design of SERONTO is a core ontology with several separate domain ontologies derived from it. The core contains basic classes and relationships for observations, but no domain knowledge for ecology, soil science, etc.. This information is in separate owl files, which uses the classes from the core to derive their concepts and relationships from. In this way e.g. querying the data for further study needs only to use the abstract concepts from the core. In this paper we will focus only on the core-ontology.

1.7 Standards

In future SERONTO should provide information required by different standards. As a lot of these standards are being developed right now (INSPIRE, OGC, ISO…). These emerging standards are not available in the form of owl files yet. Hence it was not feasible to integrate some of them at this point. However as ontologies are quite flexible, it can be expected that adjusting SERONTO to standards or mapping to standards will not be a great challenge.

1.8 Core ontology

As pointed out by several authors (Valente and J. 1996) states that a core ontology contains elements that are as generic and method-independent as possible, emphasizing good engineering practice and pragmatism in building core ontologies, meaning modular and minimum number of inclusions. (Valente and J. 1996; Valente 1997) taking this further and define four principles by which a core ontology should constructed: *Parsimony, Clear theoretical basis, Categories versus terms and Coherence*. With parsimony they mean no redundancy, "...core ontologies, being the reflection of a theory, require the application of Occam's razor. The distinctions it makes must be pivotal ones, which have no spill or redundancy with other terms in the ontology. The advantage, of course, is simplicity and elegance". They further pointed out that the reuse of the categories increases this way. Although this is a very good principle, it is however unavoidable that there is some redundancy. This arises when things can be explicit stated or can be derived from attributes at the same time. (e.g. defining certain experimental blockings or deriving it from observer fields, see below). If there is a good interface on the core, such ambiguity can be prevented. The second principle is about the knowledge domain of the core "A core ontology should have a clear theoretical basis. A core ontology will embed in one way or another some basic view of what this domain is about, what its components area and how they interrelate.". As SERONTO is an ontology about observations, it is obvious that the theoretical basis lies in statistical methodology. This methodology contains terms and concepts for describing and analyzing 'raw' data from diverse origin. For the third principle of using categories instead of terms they state "Core ontologies should attempt to define basic categories of domain knowledge, where categories are not top level terms in an abstraction hierarchy, but rather knowledge types". And for the fourth principle coherence they state "By coherent we mean more than that the basic categories should be consistent

and complete for every level of detail. They also should be part of a framework that by itself makes sense. ".(Doerr, Hunter et al. 2003) states further that "The goal of a core ontology is to provide a global and extensible model into which data originating from distinct sources can be mapped and integrated". He further adds to this: "*...a core ontology provides a underlying formal model for tools that integrate source data and perform a variety of extended functions. As such, higher levels of complexity are tolerable and the design should be motivated more by completeness and logical correctness than human comprehension*".

Those four principles are relevant for designing SERONTO. However it is not possible to be very strict on all principles due to the diversity of the group developing SERONTO. The parsimony means that if there is already a global solution which covers all, no specific solutions for specific situations should be created (as that specific solution is already part of the global one). This is not maintained, there was a two way split within the group, some preferring easiness of input in protégé (which leads to several additional solutions for special cases) while others prefer more rigid and abstract (and thus less human comprehensible) concepts. SERONTO must be seen as 'work in progress'.

1.9 Naming convention

An ontology should use a consistent naming convention to make it more readable. SERONTO uses the naming conventions from (Schober, Kusnierczyk et al. 2007), Especially important is "Each name should be intuitively meaningful to human readers and linguistically correct. Well-known terms by domain experts should be used". In compliance to this, SERONTO has class names in lowercase and instances starting with uppercase with the underscore "_" as word separator. Classes should be in singular form. The properties starts with lowercase an preferable containing a verb. For properties no word separator has been defined, every noun should start with uppercase instead. The preferred spelling and language is British English.

For the properties we decided on top of that, that each one has an inverse. Usually written as 'hasClassname' versus 'isClassnameOf' e.g. hasProject and isProjectOf.

1.10 Versioning

With an eye to the future SERONTO is already prepared for the use of LSID's (Life Sicence Identifier, see http://www.ibm.com/developerworks/opensource/library/os-lsidbp/). For each name (class and property) a LSID can be constructed using the filename, which consists of name, year, month, day combined with the class or property name with a version number added e.g. SERONTOCore20090205 with class projectV1. In order to keep the ontology working, while being able to change it, the old versions for classes and properties are never removed, they can only become obsolete. Unfortunately there's no LSID server yet, but SERONTO is already partly prepared for that.

1.11 Domain of the core

Having concluded that ontologies are a promising way to proceed, next thing is finding a consistent view on data whatever the origin and domain. Fortunately there is already such a conceptual consistent view on data available. Many database structures are build from the perspective of the data gathering, observation and measurements. As a concept for the SERONTO core a statistical perspective was chosen which is build from the perspective of the data processing and analysis. Almost every scientific field uses statistical methodology to obtain and analyze their data. Therefore it's natural approach to use concepts developed in this field. Statistical methodology handles about methods of observation as well as analyzing their results. With the data and the information about the data, further called meta-data, it must be possible to do a scientific responsible analysis or meta-analysis of the data. Another important scientific concept related to this is the notion of repeatability. In

the ideal case, the data and meta-data contained in the ontology must be sufficient for other people to repeat the observations or experiments with other people and at another place in exactly the same manner.

1.12 Components of analysis

Viewing from the analysis point of view, we can find out what is needed. Then translating these findings to the observation ontology. In an analysis of a single dataset there are five important aspects to consider: *dependent variables, independent variables, random variables, the scale of the variables and assumptions*. The *dependent variable* is also called the response variable. Unless we only want to have a simple estimate such as mean value or variance an analysis contains at least also an independent variable or a random variable. The *independent variable* contains values which are believed to determine the response in a deterministic kind of way as opposed to the random variable which has random deviates as result. The *random variables* originates from all sources of variation which are often not under control of the experimenter/observer, from design aspects or from aspects which cannot be repeated (e.g. a particular individual as experimental unit) or randomized (e.g. time). Some common sources of variation are sampling design, experimental blocking, treatment and measurement design, experimenter, observer, measurement devices and time of measurement/treatment. However this list cannot be made complete.

Each random deviate is based on a discrete or continuous probability distribution function. The usual one is the normal distribution with zero mean and having an estimated variance, but other probability density functions are possible as well, univariate as well as multivariate. This distributions which can be assumed for the dependent variable depends for a great extend on the *scale of the variable* (measurement scale), for which we develop an extended description see below. Furthermore, almost every analysis can only be done under certain *assumptions*. Next to distributional and model assumptions often assumptions of non-biasedness and representability has to be made. Assumptions also arise from errors or things going wrong in the experiment/observation. Furthermore the sampling structure is of utmost important in generalizing the findings of an analysis.

When combining multiple datasets, we furthermore should be aware of different methods used for similar variables, different units of observation and very important different experimental units. Care should be exhibited when combining mean values with individual values because a particular value can end up in multiple ways in the same analysis. Care should also be taken in combining data from different grouping-levels of experimental units (e.g. observations on forest plots combining with observation on single trees in a forest).

In summary:

Aspects important for a ‘single’ dataset

- Dependent variables
    - Measurements/observations
- Independent variables
    - Treatments
    - Classifications in stratified sampling
    - Measurements (e.g. co-variables or in meta-analysis)
    - Often time
- Sources of variation
    - Sampling structure
    - Observer/Experimenter
    - Method
    - Design (blockings)
    - Place
    - Etc.

- Measurement scale
    - Distributional assumptions
    - Preferred analysis
- Assumptions

Extra aspects important for combining data from different sources

- Methods used
    - Relationship between methods
- Units and dimensions
    - Conversion between units
- The entity of observation/treatment
    - Experimental units
    - How to group them
    - How are they selected

An even shorter summary can be written down as Who, What, Where, When and How (and Why). In the ideal case all information about the data should be available. This however will seldom be the case. Historical data are often incomplete in meta-data. Using such data in combination with other data needs additional assumptions e.g. about method, representivity. The lack of information guides the assumptions needed. If meta-data is available however, there must be fields available to store it.

If we view data from an administrative point, the 'statistical' description of the data is not enough. Project information and purpose of research should be added. Owner and intellectual property rights should be considered too. In SERONTO however, we do add project information but ignore for the moment the property rights.

Questions has arisen if the ontology should contain the real observed values. A common approach is defining datasets (e.g. the excel sheet) and describing the meta-data for the dataset as in the Ecological Markup Language (EML)[1] or OBOE (Madin et al. 2007). In SERONTO we try to describe the single observations itself. The reason for this is that in monitoring data, experimental units (e.g. a particular tree) can end up in multiple datasets. It is according to us easier to describe data at the level of the single observation in this case then to try to relate observations within different datasets by means of an ontology. Thus avoiding addressing particular records in different datasets.

As the problem of combining distributed databases is a more technical one, it will not be dealt with in this paper.

1.13 Design Principles domain core

To sum it up, the concepts of the core are derived from scientific principles and lean heavily on statistical methodology. Important considerations in designing SERONTO were

- Repeatability: The ontology should be capable of holding enough meta-data that another person can repeat the experiment or observation at another place and time. It is not obligatory, however, to provide all information for all datasets; for instance, some information may be missing for old datasets.
- Transparency: It must be possible to record and retrieve meta-data describing what actually happened. SERONTO includes concepts for sampling and experimental designs and for things going wrong and documenting data collection under less than

[1] See http://knb.ecoinformatics.org/software/eml/

ideal conditions. If data and meta-data are available in this way, it is clear what assumptions must be made to combine data and correctly interpret analyses.

- Machine readability: All information needed to 'understand' the data must be contained in the ontology. Either by definitions or derivable from the relationships. It must be possible to set up algorithms to arrive at conclusions without additional information from textual reporting or verbal information.
- Mapability to other standards: it should be possible to map as much meta-data as possible to other existing standards, but not at the expense of the other requirements. Standards could be EML, ISO 19115, QML, ABCD, INSPIRE, etc…).

2 Ontology classes

For readability purposes in the following description of the core-ontology, the class, property and instance names are written without the version information. In the pictures the oval shaped non-filled text boxes are classes, while the filled ones are instances of classes. Rectangular boxes, none filled, are used for the values of a data type property such strings, date and numeric values (see Figure 1).

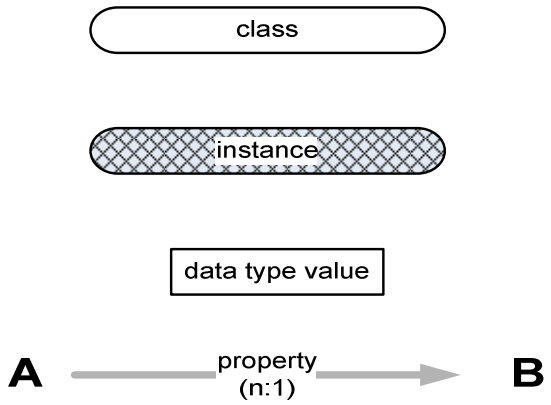

Figure 1. Basic elements in figures in document.

The property is shown by an arrow, when A and B are classes then there will be some numbers between brackets (e.g. n:1). This should be read as follows (in Figure 1) A can have 1 B attached by means of the property and B can have more than 1 different A's attached by means of the inverse property. Those numbers are not shown when depicting properties of instances.

The description of the classes below is not exhaustive. Some extra classes can be found in the owl file. As the ontology is still in progress, the actual current content can deviate.

2.1 The basic 'observation'

An example of a set basic sentences describing an observation is as follows: the height of a dominant tree in a forest plot is h1 cm at time t1, h2 cm at time t2, etc.. The method of measurement is a triangulation method. Those two sentences can be written more common: The parameter V with scale S and unit U obtained by the method M on the experimental unit E from the population P has the values h1 at time t1. A simplified overview is pictured in Figure 2.

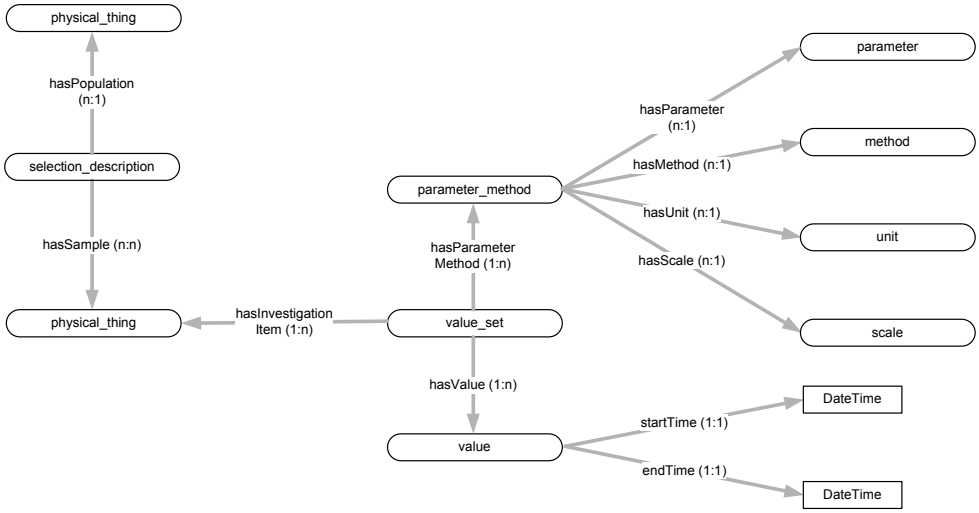

Figure 2 - Base classes and properties from the core to describe an observation. For explanation see text.

Translating the example described above into the schema of figure 2, it becomes

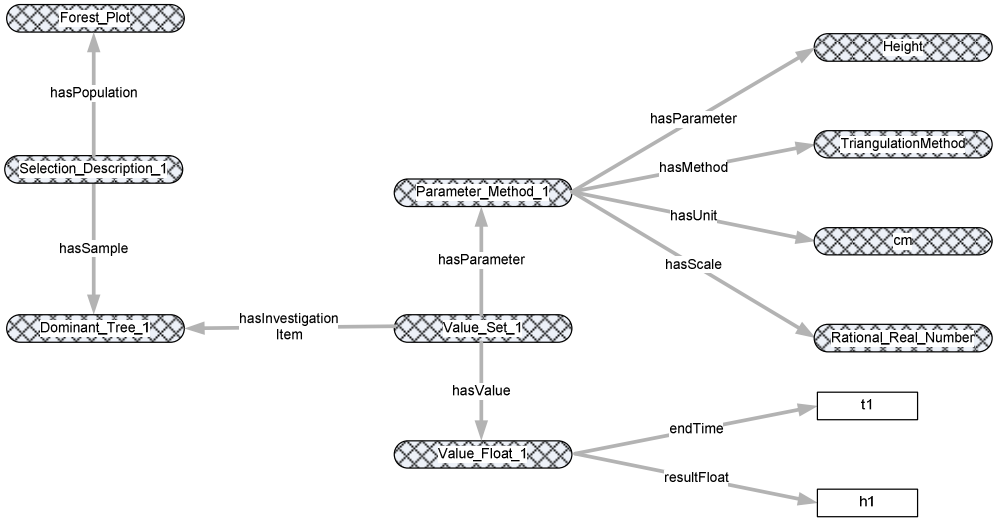

Figure 3. Example of an height measurement from a dominant tree in a forest plot.

We will first describe and explain the classes depicted in Figure 2, adding additional classes when needed for the description.

2.2 Parameter

A parameter in SERONTO is basically the same as a variable in statistics. The confusion between the term parameter and variable arises from the viewpoint of the problem. From a statistical point of view the variable represents the observed (or measured) values, which can vary. A parameter in statistics is often a model parameter which has to be estimated from the data and is considered fixed for a certain population. This last is more like ontologies, the parameter is a more or less fixed item (which is a class) in the ontology. The ontology can be seen as comparable to the model in statistics. The parameter is a class from which all parameter/variables measured or applied can be instantiated. The idea is

either by using subclasses (e.g. chemical parameter) or by using the grouping_description to be able to query the instances in the ontology at higher aggregated levels. For example give me all chemical parameters on soils.

2.3 Methods

A Parameter can represent a measurement, treatment or a classification (including random variables). To interpret the values of a parameter a method must be attached to it. But as one method can be used for different parameters (e.g. using a ruler to measure the length or width of something) and a parameter can be measured by different methods (e.g. ruler or using laser for length), a class parameter_method has been created as a placeholder. This parameter_method has the properties hasParameter, hasMethod, hasScale and hasUnit among others. In principle the parameter_method can be seen as a relationship with attributes. As this is not possible to express in owl, a separate class had to be defined.

Methods can be very diverse. In the example above the triangulation method is mentioned for estimating the height of a tree. This method can be subdivided into a sequence of sub methods, some of which necessarily have to be done before another. The whole composition of sub methods encompasses the triangulation method Figure 4.

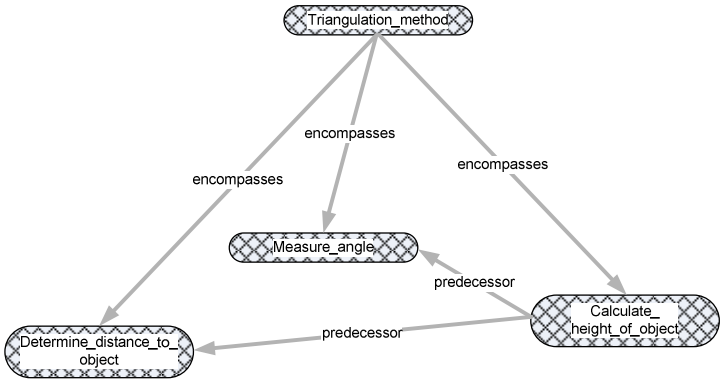

Figure 4. Triangulation method for measuring e.g. the height of a tree.

As can be seen in Figure 4, the triangulation method consists of three sub methods: 'determine distance to object', 'measure angle' and 'calculate height of object'. The sub methods 'determine distance to object' and 'measure angle' must be carried out before the calculation of the height can be done.

We distinguish different types of methods. The most important ones are:

- *Classification_method:* assigns nominal values and is very often used in stratified sampling
- *Measurement_method:* used for measuring things other than classifications
- *Treatment_methods:* describes how treatments are carried out

2.4 Units and dimensions

Units and dimensions are defined in the SERONTO core, but stored in a separate owl file. Basically we distinguish three classes of units: singular_unit, power_function_of_unit and compound unit. The relevant base classes are shown in Figure 5.

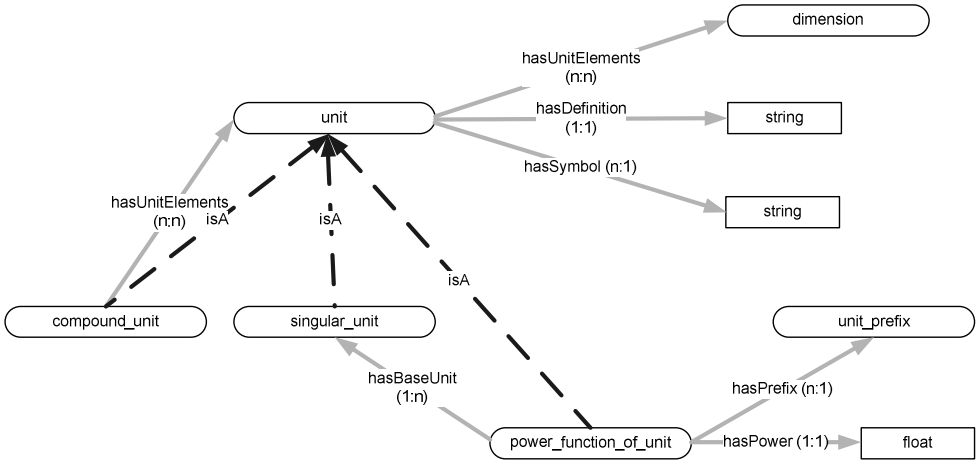

Figure 5. Base classes and properties of the units and dimension sub-ontology.

As an example the unit Newton (kg m s-2) is given in Figure 6.

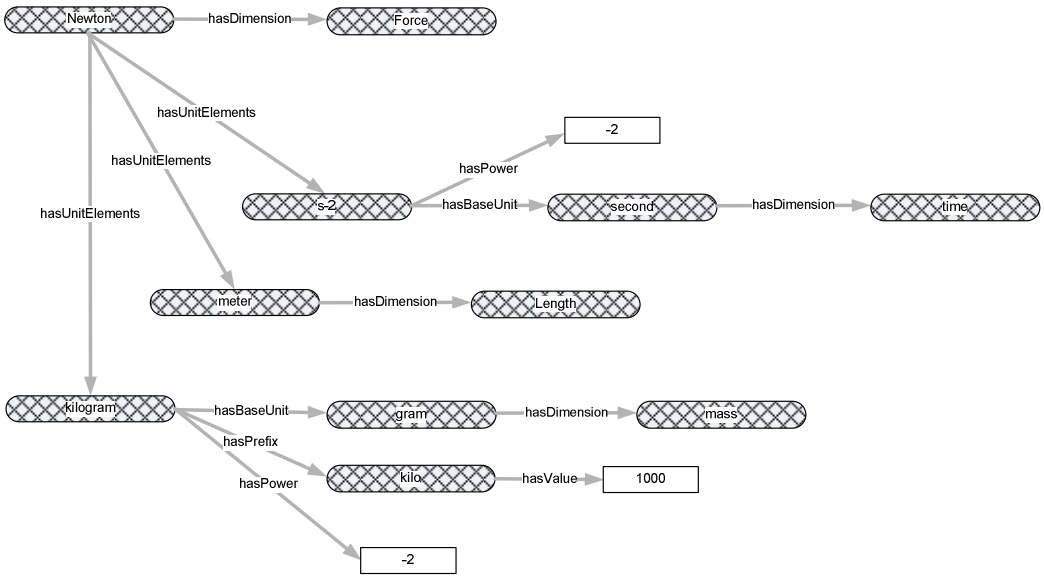

Figure 6. Description of the unit Newton N=kg m s-2.

Next to the definitions of the units and dimensions, classes for unit conversion are included. The basic conversion formula can be seen as NewUnit = Offset + Multiplier * (Unit – Location) ^ Power, where Offset, Multiplier, Location and Power are constants. As an example Kelvin = 273.15 + 0.5555*(Fahrenheit + 32)^1. If this is filled in consistently, conversions to standards (e.g. SI units) can in principle be done automatically.

2.5 Measurement scale and domain

The measurement scale is an important aspect of a variable, because from the scale the possible statistical models can be inferred or what assumptions has to be made to do a particular analysis. For example, if the scale is nominal then usually the number of observations within the single nominal class are analyzed (other option could be fractions or percentages within that class). If the nominal variable has two levels then a binomial error distribution can be assumed and with more then two levels a multinomial distribution can be assumed or an appropriate log-linear model and link function can be used (McCullag and

Nelder 1989). In literature there is a discussion on how valid the measurement scales are for determining statistical models and tests. The original definition of the measurement scales originates from Stevens (Stevens 1946; Stevens 1951). His ideas are summarized as (after (Khurshid and Sahai 1993))

- *Nominal*: Assigns numbers or text to observations to identify or classify
- *Ordinal*: assigns numbers or text to observations in sequence, from lesser to greater amounts of the measured characteristic,
- *Interval*: assigns numbers to observations that reflect a constant unit length between units of measurement
- *Ratio*: assigns numbers to observations to reflect quantity with reference to an absolute zero point.

Going from Nominal to ratio the scale becomes 'stronger'. A stronger scale can always be presented as a weaker one and analyzed accordingly, this can be useful when combining data from different sources. Originally, knowing the scale of the measurement implied the analyses to be done. However since the time of Stevens original division a lot of new statistical techniques and insights have been developed. And as (Velleman and Wilkinson 1993) among others pointed out, not all variables can be caught into this division of the original four groups. Notable exceptions are

- *Percentages*, which looks like a ratio scale are bounded at two ends and cannot tolerate even arbitrary scale shifts.
- *Scores*: bad, good, unknown could be considered nominal, but within certain analysis the order bad, unknown and good could be assumed, which makes it an ordinal
- *Observations lying within a varying range of values*: e.g. mortality data where for first weeks the day number is known, then precision changes to week number, month number and finally a 'truncated' value of mortality of survivors after 1 year.
- *Circular measurement* e.g. angles where degree 361 is same as degree 1 (e.g. could do some circular statistics or use polar coordinates for analysis)
- *Cyclic measurement* e.g. month where January of year 1 is not the same as January of year 2 but is however similar in certain aspects

(Mosteller and Tukey 1977)presented for this purpose an alternative list but this doesn't cover all mentioned exceptions either.

- *Names*
- *Grades*: Ordered labels such as Freshman, Sophomore, Junior, Senior
- *Ranks*: Starting from 1, which may represent either the largest or smallest
- *Counted* fractions: bounded by zero and one, these include percentages
- *Counts*: non-negative real integers
- *Amounts*: non-negative real numbers

In addition (Velleman and Wilkinson 1993) stress that such divisions do not restrict the choice of analysis. In our opinion every statistical model or test is based on certain assumptions. As a consequence of this, given some measurement scale, no type of analysis should be forbidden beforehand. If the researcher is willing to make the assumptions for his dataset it will be primary his or her responsibility.

In designing the ontology we considered the 'scale' aspects which are needed for analyses as well as needed for combining data (e.g. unified approach of nominal and numerical scales). Also the scale must be applicable not only to measurements but also treatment and classification variables. For the scale we devised the following classes Figure 7.

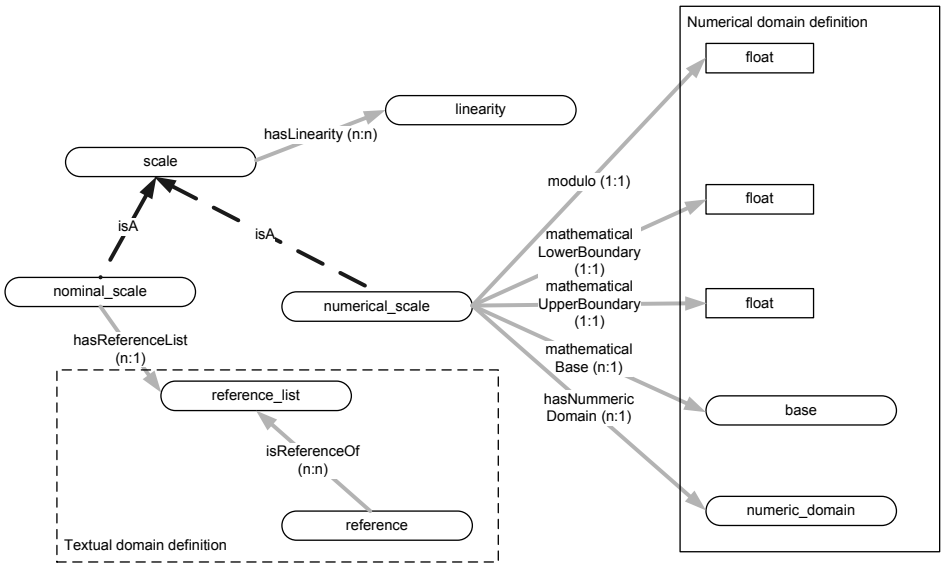

Figure 7. Classes for describing the scale of a parameter.

The scale can be divided having string values (nominal_scale) or having numerical values (nominal_scale). The basic characteristic for the scale is the domain definition (which values can be chosen) and the interpretation of those values, which we call the linearity. The linearity has as possible instances (nominal, ordinal, interval circular and cyclic). The domain of a numerical scale can be defined by the lower and upper mathematical boundary (this is not the boundary which can be measured by the device, but are the real domain boundaries), the mathematical base (decimal, binary, hexadecimal, etc), the numerical domain (ℕ: Natural numbers (including 0) , ℤ: Integer numbers, ℝ: Real numbers, ℚ: Rational numbers (integer division)) and the modulo for cyclic or circular measurements. The domain for the nominal scale is a reference_list (a list of string values: reference). Examples

Example *Fraction*: Continuous; domain: ℝ; bounds: [0, 1]; linear; base 10 see Figure 8

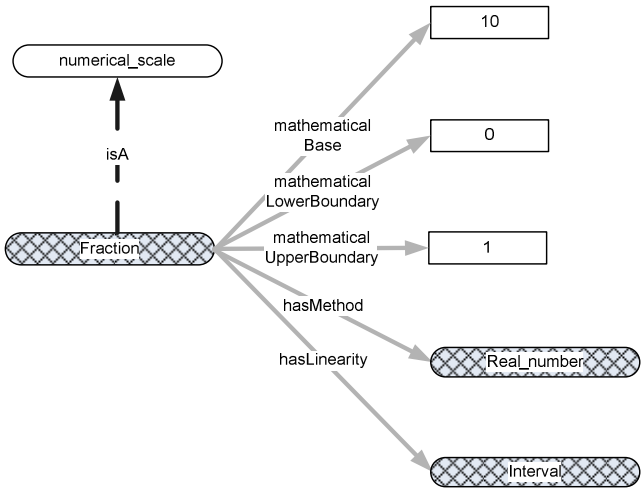

Figure 8. *Fraction*: Continuous; domain: ℝ; bounds: [0, 1]; linear; base 10

Example *Degrees*: Continuous; domain ℝ; bounds (-∞, ∞); Circular: mod(2π); (0: north -> to parameter): base 10 see Figure 9.

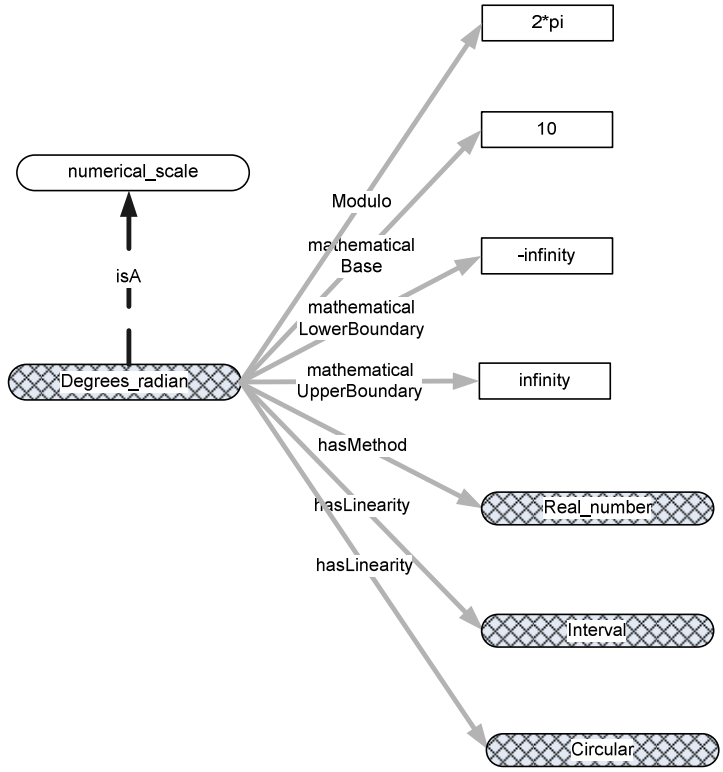

Figure 9. *Degrees*: Continuous; domain ℝ; bounds (-∞, ∞); Circular: mod(2π); base 10

Example for a nominal variables with the ordinal classes Young, Middle_aged and Old see Figure 10.

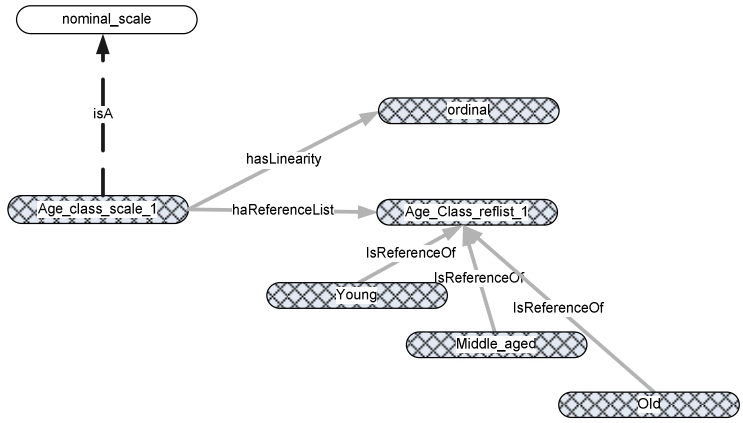

Figure 10. Definition of ordinal-nominal scale for an Age class reference list with the classes Young, Middle aged and Old.

NB: If the zero is absolute or not can almost always be derived from the units/dimension. Analyzing data without knowledge of units and dimensions is usually not a good idea. Cyclic needs a bit more explanation. When for instance someone measures the number of months after some event, then month 13 has the same month name, it is different but also somewhat equal. In analyzing the data with automatic methods, the computer should be able to derive such things from the data and give a warning or using it accordingly. In the case of month nr, the lower boundary would be 0 (or 1), upper boundary infinity, modulus 12 and linearity should be interval. But you can also argue that the linearity should be ordinal, because not all months have equal length in days. This is often determined by the

viewpoint of the one who analysis the data (or when storing the data from the observer). There can be conflicting viewpoints, but that shouldn't be prevented. It is not very likely that the list of linearity's is exhaustive yet, therefore the list of linearity's can be extended.

It should be stressed that the scale boundaries as defined in the ontology are the theoretical boundaries. In addition to the mathematical restrictions, there can be restrictions on the devices for measurement, which can limit the outcome for theoretical range. This kind of information should be attached to the methods of measurement, just like the precision and accuracy.

2.6 Values

Values are coupled to the object and parameter by means of the value_set. Values are divided according to the scale into nominal and numerical values (value_nominal and value_float resp). The value can relate to a time period: startTime and endTime. If the value is relevant at a point in time rather than a period, the endTime should be filled in only. In Figure 11 the relevant classes are shown:

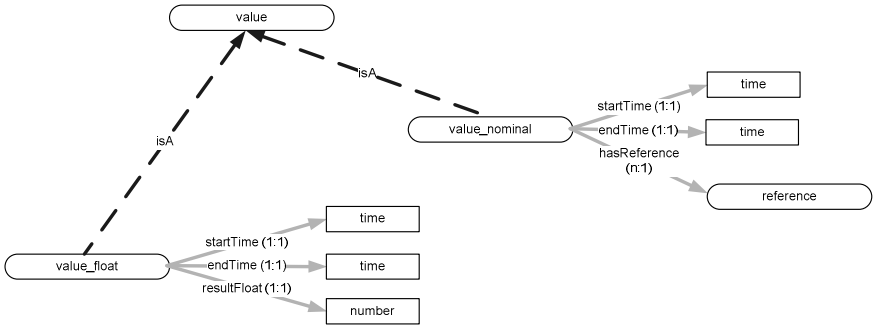

Figure 11. Basic classes for values.

The reference in the nominal value should be attached to a reference list as well by means of the scale see Figure 10.

During the development of the domain ontologies the need was expressed to be able to attach a reference to any object while the reference should be kept attached to a reference_list. Although not everyone in the group agreed on this solution, the reference_element was created with properties hasReference and hasRefernceList (Figure 12)

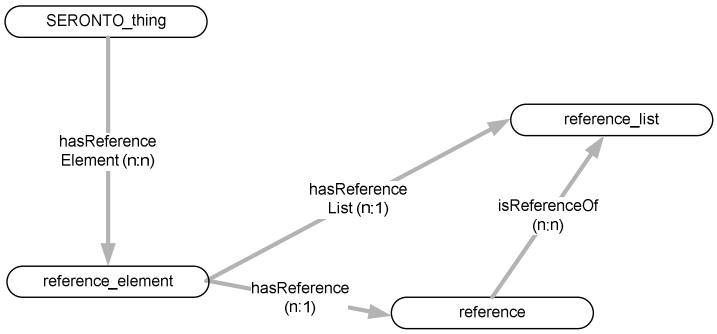

Figure 12. The reference_element class coupling reference, reference list and any object to each other.

This leads to another way of assigning (nominal) values to an object, in fact to any object. This leads to a problem, no time of observation is attached anymore. In order to provide for that, the relationship hasHelpObject has been defined which has the value_set as domain (Figure 13).

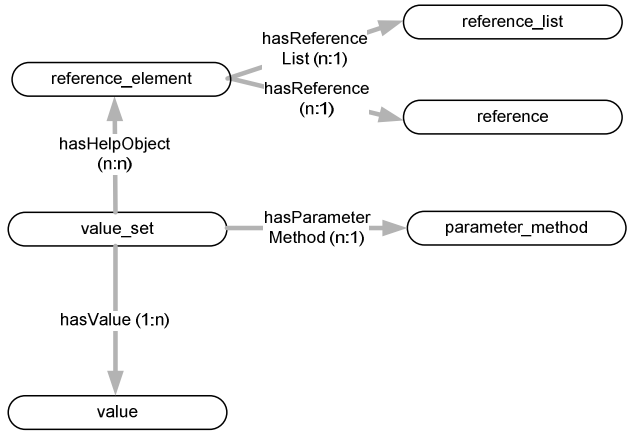

Figure 13. The hasHelpObject relationship, this relationship attaches additional nominal values to a value_set.

The use of this is shown in (Figure 14) where an example measurement of a vegetation relevee is depicted. The hasSpecies and the hasLayer are derived properties (!) from hasHelpObject and situated in the domain ontology. The time of measurement and methods can be derived by means of value_set (inherited upward).

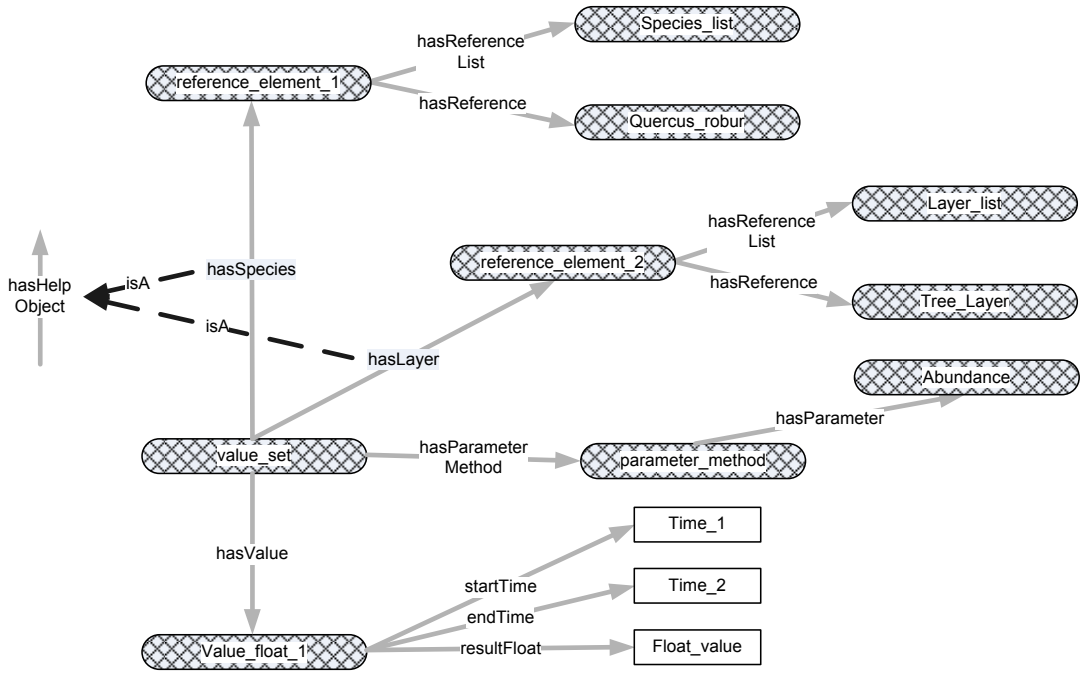

Figure 14. Example of vegetation relevee which uses hasSpecies and hasLayer derived from hasHelpObject.

2.7 Time series

Time is different from other parameters, although it can be treated the same way (for instance when measuring survival time (or date)). Time in SERONTO is seen as a pivotal point to group observations together. If we have two different types of observation on the same physical object e.g. weight and height of a particular organism, then by inspecting the time of measurement, we can decide if it can be coupled. Another way to group observations (e.g. samples) is by the use of a grouping description (see 'sampling' below) or by e.g. the project number. Automatic data devices often generate huge numbers of data, all with their time-stamp. To limit the data storage and processing time, it was decided to store the results of a time series with one single value_set in stead of giving each value its own value_set. An example is shown in Figure 15.

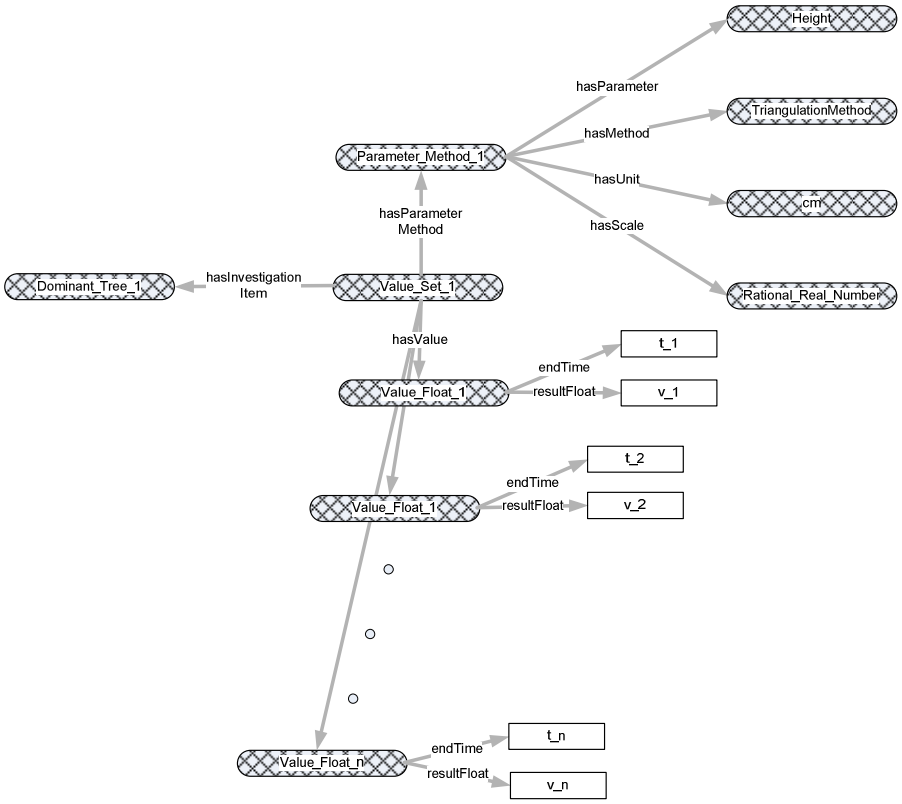

Figure 15. Example of a time series in SERONTO.

2.8 Sampling

The sampling structure is of very importance in the analysis of data. For a large part it determines the random factors as in a variance component model (ref), but also it leads the researcher while interpreting their result back into the real world. For example if a sampling is only conducted to females, then strictly the results of the analysis can only be applied to females from that population sampled. Often however with additional assumptions it will be applied to other populations as well or even to males. There is a close relation between experimental designs and sampling structures. All sampling schema's can be regarded as repeats of at least two steps. First define the population(s) followed by selecting individuals out of which the population exists. In SERONTO we defined the class selection_description for this (Figure 16).

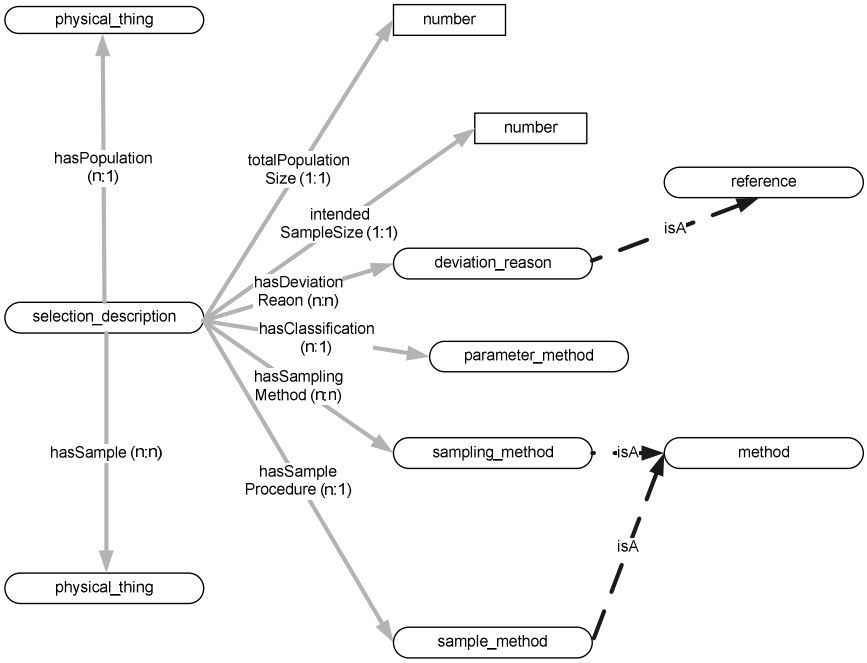

Figure 16. The basic class for describing a selection step in sampling.

The selection description connects the sample with the population from which it is drawn (properties hasSample and hasPopulation). The sampling_method describes the type of method used e.g. random sampling, convenient sampling, stratified sampling, etc.. The actual way it has been done is described in the sampling_procedure e.g. the actual taking a of soil sample, taking a leaf sample, taking a water sample. If a stratified sampling has been used, the population is divided into the groups indicated by the parameter_method's attached. From each of the so created subpopulations the actual sample has to be drawn and this doesn't have to be necessarily with the same method (e.g. we divide the population first into females and males, and then taking all females, but using a random sample for males). It is throughout SERONTO assumed that all characteristics defined for the population (by means of parameter_method or higher level selection_description) are inherited by the samples. In this way we can define characteristics which are 'inherited' by the subsamples. It should be stressed that the characteristics of the top level populations should be described as extensive as possible. In Figure 17. only the habitat as parameter is given, but country, size, age, etc., if known should also be given. The reason for putting in an 'exhaustive' list is that it makes it much easier to combine appropriate other forests from other observations to derive knew knowledge.

In Figure 17. a simple random selection of forests having a particular habitat is given. It is intended to take 10 forests in total.

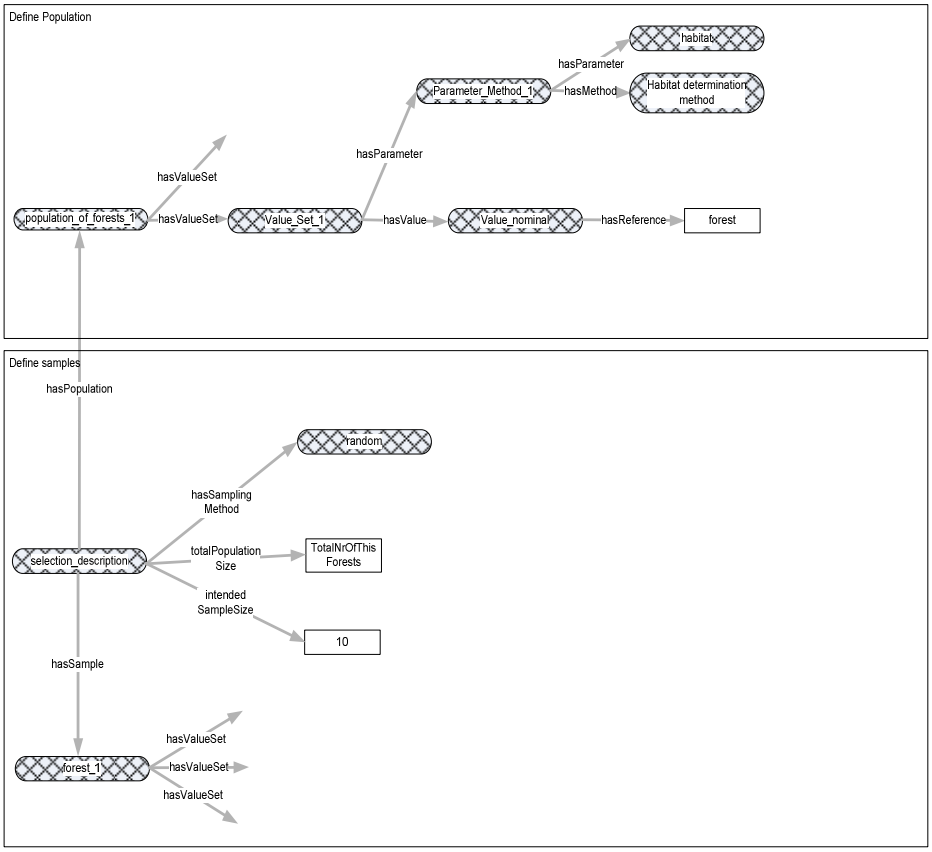

Figure 17. Example of a simple random selection of forests having a particular habitat.

In Figure 18 a more complicated stratified sampling is shown. Within the forest of a particular habitat, subpopulations of deciduous or broad-leaved forest are defined. From those subpopulations 10 forests are intended to be selected. A further selection could be carried out by defining subpopulation of plots (e.g. old or young) from which single trees are selected. Within the trees we further select leaves (e.g. with method convenient i.e. only leaves that are reachable). This can go on to even more detailed levels.

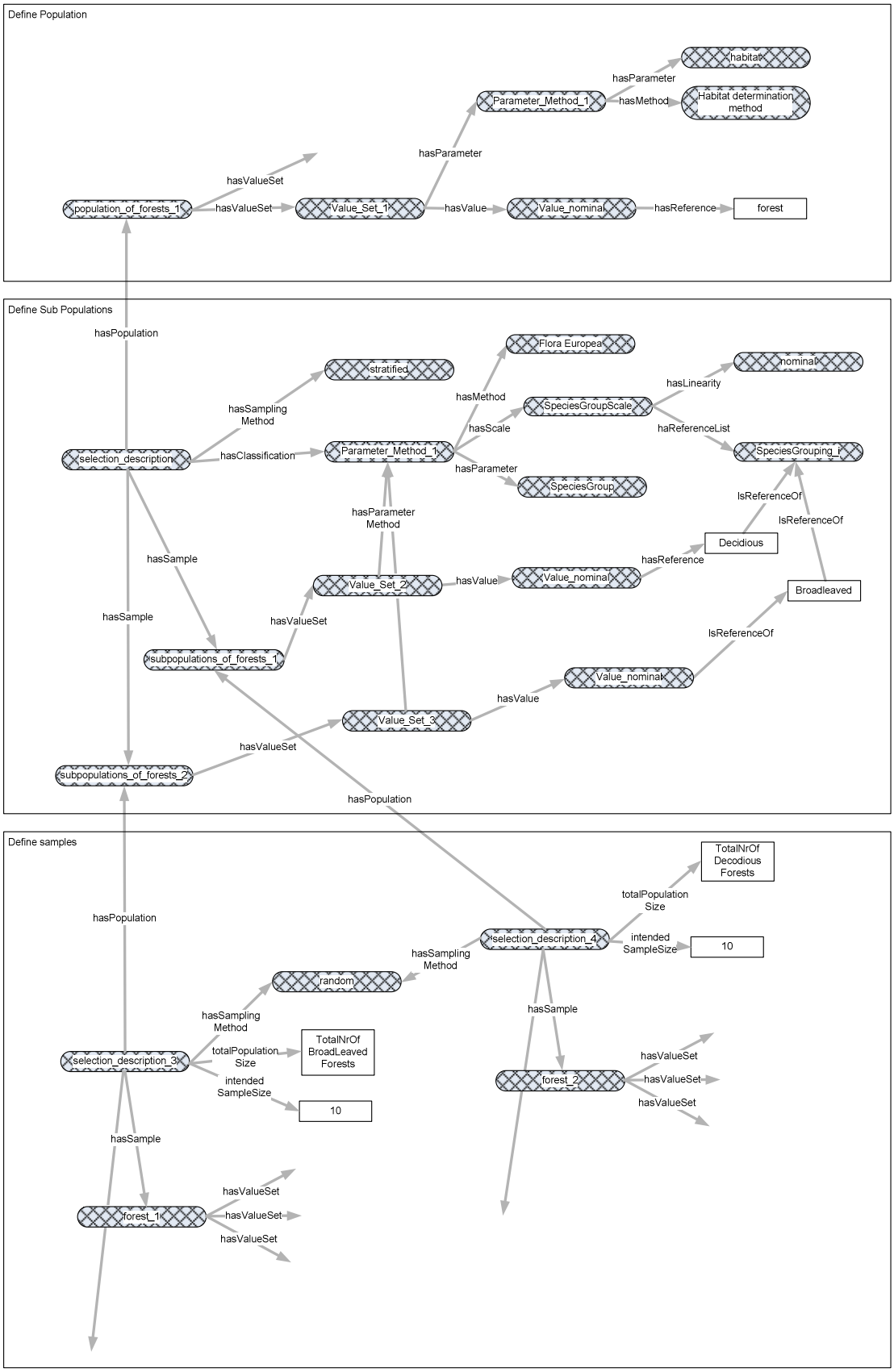

Figure 18. Example of a stratified sample: deciduous and broad-leaved forest populations.

2.9 Counts

Counts can result from three ways. One way is direct counting and assigning it to a parameter, the second way is to count the number of observations belonging to a certain class value (nominal value, reference) and the third way is deriving it from the number of samples selected out of a population. The last case has something special. Usually this last number is known in advance. If not then this can be regarded as an observation too and can be modeled with the usual statistical methods (e.g. log-linear models). When the number is know in advance, e.g. 20 trees from a forest plot then making a model predicting this number is senseless, but if there are no 20, then the actual trees sampled is a result (measurement). This is why the relationships intendedPopulationSize, totalPopulationSize and hasDeviationReason have been added. If the reason is 'can't find any more' then modeling this number is appropriate. In addition to that, the total population size can be necessary for estimation (mean, variances, etc.), especially when there are low numbers or when the whole population has been sampled.

2.10 Groupings

To be able to correctly describe a large variety of experimental designs, the grouping_description was added. The grouping_description is in principle the inverse of the selection_description. A useful example is the Latin square design, where the columns are defined by the layout in the field and the 'rows' by the observer group e.g. as in Figure 19. where I, II, III and IV are the 'columns' (strips of land), A, B, C and D are different treatments randomized within a column and the pattern denotes the observer group (note that each pattern contains all four treatments; this design is useful when there is a gradient perpendicular to the strips and we can expect some differences per observer group).

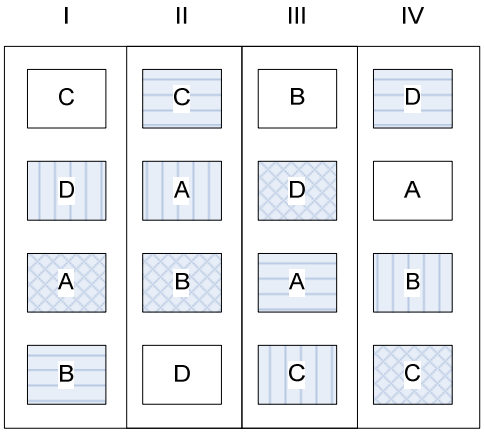

Figure 19. An example of a Latin square design, see text for explanation.

The design of the experiment is put in SERONTO as in Figure 20, where only the grouping of one observer group has been worked out.

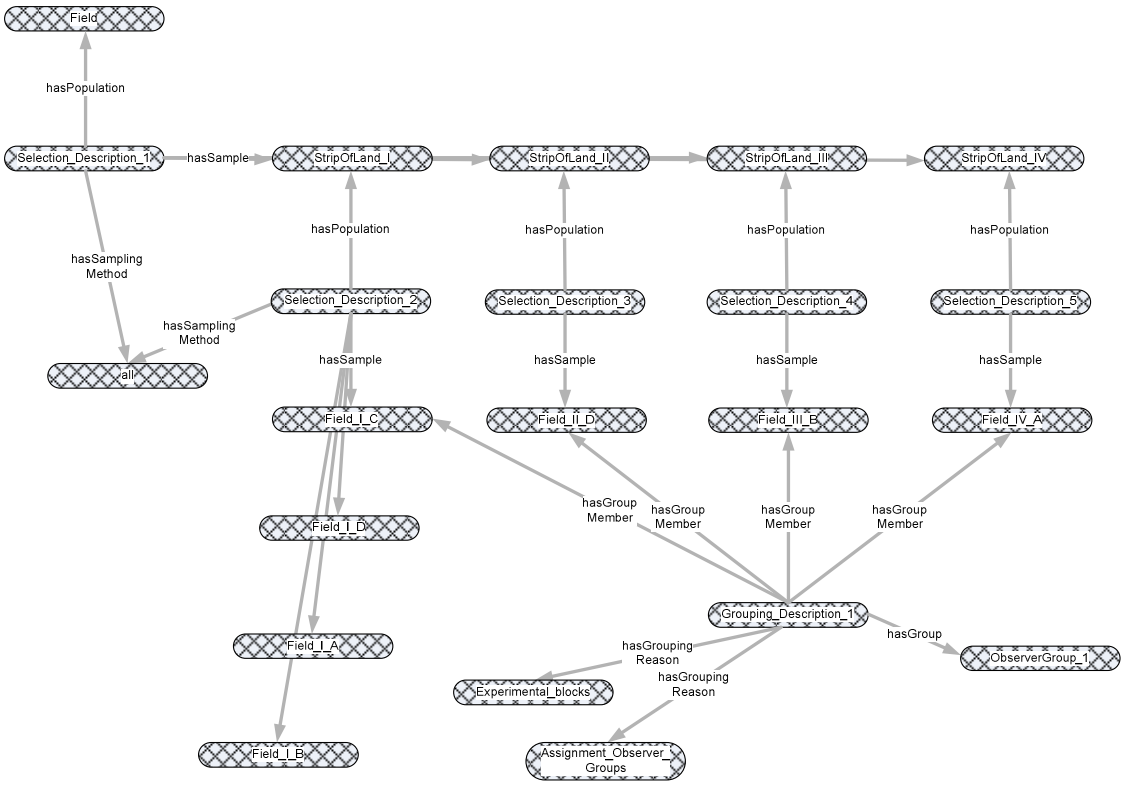

Figure 20. SERONTO representation of part of design of figure 19.

It can be argued that this information is already in the information of the values. Although not mentioned earlier yet each individual value should have an observer attached (or observer group). Analyzing the design by means of studying this could lead to the conclusion of a Latin square design. However, the same design could arise from random assignment of four observer groups. Strictly speaking it would be wrong to analyze that as a Latin square, the design would be the same, but the method of arriving at the design not. Hence the expected values are different and hence we need the additional grouping class in SERONTO. In addition to that there are very complicated designs which are difficult to derive at without knowing.

The grouping description can also be used for grouping parameters together e.g. grouping all parameter measuring chemicals. The purpose of this is being able to query at aggregation levels. The result of the query should show the parameters together with the other information (e.g. experimental unit). The person who does the analysis can then decide which to keep and which not. This use of the grouping description can be distinguished by the grouping_reason. Another property for doing grouping of instances is the lower (and its inverse upper). As this has the SERONTO_thing as domain and as range, all kinds of hierarchical lists can be created.

2.11 Common class properties

The base of all classes is the SERONTO_thing. This class has several administrative properties, which are inherited by all other classes. Administration can be defined for the administration of the classes itself and the administration of the instances (meta-data). For the administration of the classes we have isCreatedBy, isDefinedBy, isEnteredBy and date (date of creation). For reasoning purposes, the properties includes and intersects are added but not actually used yet.

The administrative properties for the data itself are hasOwner (who owns the data), and hasProject.

2.12 The basic class structure

The topmost class in every owl file is the owl Thing. All other classes are derived from that. In comparison to this, all SERONTO core classes are derived from SERONTO_thing. From this we divide the SERONTO world into abstract_thing's, physical_thing's and reference_catalogue. From the last all reference's, reference_list's and reference_element's are derived. From abstract_thing all descriptions are derived, while investigation_item's and device's are derived from the physical_thing's (Figure 21).

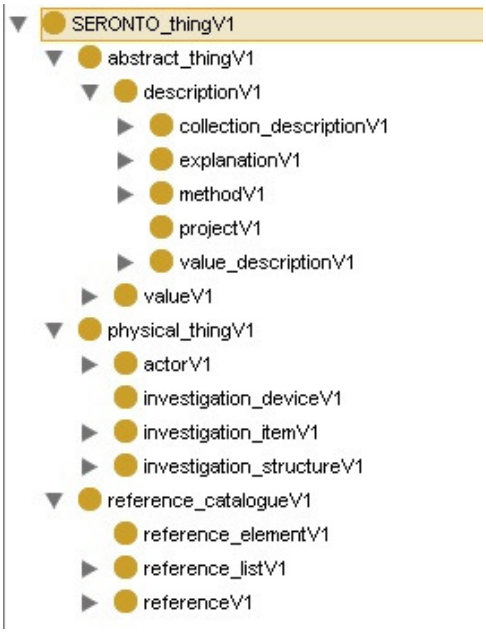

Figure 21. Basic class hierarchy of SERONTO. See text and owl file for explanation.

2.13 The owl file

The ontology is used and developed with the program protégé version 3.3.1 (http://protege.stanford.edu/). The latest version of the core-ontology and additional information can be found at http://www5.umweltbundesamt.at/ALTERNet/index.php?title=Ont:Ontology_Creation_Portal. Worked out examples and a manual for SERONTO can be found here.

3 References

Doerr, M., J. Hunter, et al. (2003). "Towards a Core Ontology for Information Integration." Journal of Digital Information integration.

Khurshid, A. and H. Sahai (1993). "Scales of measurements: an introduction and a selected biography." Quality & Quantity **27**: 303-324.

Madin, J. , Bowers, S., Schildhauer, M, Krivov, S., Penningtond, P. Villa, F. (2007)

An ontology for describing and synthesizing ecological observation data. Ecological Informatics (2) 279-296.

McCullag, P. and J. A. Nelder (1989). Generalized Linear Models, Chapman & Hall.

Mosteller, F. and J. W. Tukey (1977). Data Analysis and Regression. Boston, Addison-Wesley.

Schober, D., W. Kusnierczyk, et al. (2007). Towards naming conventions for use in controlled vocabulary and ontology engineering. 10th Annual Bio-Ontologies Meeting, Vienna, Austria.

Stevens, S. S. (1946). "On the Theory of Scales of Measurement." Science **103**: 677-680.

Stevens, S. S. (1951). Mathematics, Measurement and Psychophysics. New York, John Wiley.

Valente, A. (1997). "Guidelines for Principled Core Ontologies Content areas: ontologies, knowledge acquisition." http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.51.7863.

Valente, A. and B. J. (1996). "Towards Principled Core Ontologies." http://ksi.cpsc.ucalgary.ca/KAW/KAW96/valente/doc.html.

Velleman, P. F. and L. Wilkinson (1993). "Nominal, Ordinal, Interval and Ratio Typologies are Misleading." American Statistical Association **47**: 65-72.

Acknowledgement

This work was supported by ALTER-Net (A Long-term Biodiversity, Ecosystem and Awareness Research Network). ALTER-Net (http://www.alter-net.info/www.alter-net.info) is a Network of Excellence funded by the 6th Framework Programme of the European Commission.

4 Annex A – Handbook SERONTO

FileName: Annex_A_Handbook_SERONTO.pdf