Classical and Symbolic metadata setting for biological datasets

Haralambos Papageorgiou*, Maria Vardaki*

*Department of Mathematics University of Athens Panepistemiopolis, 15784, Athens Greece
hpapageo@uoa.gr, mvardaki@uoa.gr

Abstract. We consider an extension of classical data analysis into symbolic data analysis to describe the management process of biological datasets produced by multi-source clinical studies. Such extension leads to more complex data types and tables and the metadata under consideration hold information both on classical (original) and the symbolic data. In this paper we model these metadata items in a structured object-oriented schema for symbolic data revealing their relations. A number of transformations are also discussed both for classical and symbolic classes of our model in order to illustrate how the applied transformations on symbolic data depend on the related classical data setting.

1 Introduction

Symbolic data serve not only to summarize large datasets, but they also lead to more complex data tables, thus enabling the manipulation of huge datasets (Bock and Diday, 2000; Billard and Diday, 2006). Using the symbolic data techniques, data are aggregated into macrodata, forming Symbolic Objects (SO) and Symbolic Data Tables (SDT) (Bock and Diday, 2000; Diday and Noirhomme-Fraiture, 2008; Noirhomme-Fraiture, 1997).

A symbolic data table constitutes the main input for symbolic data analysis (Diday, 2002). It looks like a classical data table where each cell represents symbolic data, since each row corresponds to a symbolic description of a group of individuals and each column corresponds to a symbolic-valued variable (Noirhomme-Fraiture and Brito, 2011).

Consider a modern, state-of-the-art information system. As expected, it stores a considerable amount of microdata, macrodata and related metadata for each piece of information imported. In the case of an information system that manages biological datasets collected from multi-source clinical studies, due to confidentiality reasons (as emphasized by UNESCO, the Nuremberg Code, the Helsinki Declaration, etc.) all data are imported, randomized and further used in the form of macrodata. Symbolic analysis techniques are especially useful for managing large datasets from multiple sources; therefore they can adequately manage, among others, biological macrodata resulting from various clinical studies.

Since, even in the classical data setting, aggregate data can be of little value to any data consumer if explanatory information (metadata) does not accompany them (definitions, patients' eligibility criteria, the study parameters, the risk factors, how data were collected and

manipulated, etc.), in the symbolic data setting the need of structured metadata is more demanding. For example, whenever a SDT is used by information systems, its construction and handling should be automatically accompanied by the appropriate documentation (metadata) which, in turn, greatly improves the quality of the produced results by reducing the dangers of data and metadata mismatches (Papageorgiou and Vardaki, 2008). Importing such amounts of information does increase the burden of investigators' work but the advantages of doing so are substantial, especially when, for instance, someone attempts to combine the findings from multiple related studies to increase the size of the population or the number of variables studied.

In this paper, we attempt to extend classical data into a symbolic data setting to describe the process of clinical studies' data management. We take as a guideline the statistical, process-oriented metadata model introduced by Vardaki et al. (2009) to describe the process of medical research data collection, management, results analysis and dissemination. Our approach does not interfere with the design of the clinical study, the sample selection and the data collection stages of the above model, but we extend the management of the datasets in order to hold metadata both for the classical (original) and the symbolic data, enabling the use of symbolic analysis techniques. For this purpose, we introduce an abstract object-oriented metadata model designed in Unified Modeling Language (UML) which can hold metainformation for the classical (original) clinical data and also the necessary metadata for the symbolic data setting. A set of operators/transformations is applied for further symbolic data analysis.

2 From classical to symbolic data/metadata setting in clinical studies

In classical data analysis, the statistical population, the sample derived through a sampling method, as well as the sampling units examined (called individuals thereafter) and the related (classical) variables, are the key issues to be evaluated when conducting a survey. In symbolic data analysis, the symbolic objects are the central items. Generally, symbolic objects (u) are defined as triplets $(\alpha,\, R,\, d)$ where d is a description, α is a membership function which defines the extension of the SO and R is a comparison relation between descriptions. Some basic references on the topic are: Bock and Diday (2000), Billard and Diday (2006), Diday and Noirhomme-Fraiture (2008).

The symbolic setting used in this paper is explained by the following (simple) example:

Let us consider data collected in two nearly identical randomized clinical trials one undertaken in Australasia and one in Canada, with the aim to study the effect of starting chemotherapy immediately in asymptomatic patients with metastatic colorectal cancer (Ackland et al., 2005). Numerous variables for each individual participating in the trial are registered (geographical, demographical, medical status, etc.) as required by the protocol of the study, some of them considered as risk factors, such as: age, weight, gender, diagnosis stage of cancer, etc. The collection of all classical data values for both trials' patients is presented in Table 1 (representing the survey), where each row refers to the classical variables' values recorded for a single patient (individual) denoted by i: $i = 1, \ldots, 168$.

i	Area of study	Treatment schedule	Gender	Age	Prior chemotherapy	Survival (months)	...
1	Australasia	Immediate	M	79	No	16.2	
2	Australasia	Delayed	F	50	No	11.0	
3	Canada	Immediate	F	76	Yes	13.2	
4	Canada	Delayed	M	45	No	9.3	
5	Australasia	Immediate	F	46	Yes	15.4	
6	Australasia	Delayed	F	38	No	9.2	
...	...	...	...	...	...	...	
168	Canada	Immediate	M	46	Yes	11.7	

Table 1. Sample Dataset : Classical Data

In symbolic analysis, the objects denoted by u_i are classes of the initial patients satisfying a set of properties. As illustrated in Table 2 u_1 for example, represents the group of patients of the Austalasian trial which received immediate treatment, 76% were men (M) and 24% women (F), with age ranging between 46 and 80 years, the 28% of them had received prior chemotherapy and the median of survival of the group was 15.5 months (time measured until the end of trial).

u	Area of study/Treatment schedule	Gender (%)	Age	% received prior chemotherapy	Survival Median (in months)
u_1	Australasia/Immediate	$\{M\,(76), F(24)\}$	[46,80]	28	15.5
u_2	Australasia/Delayed	$\{M\,(73), F(27)\}$	[36,77]	24	11.9
u_3	Canada/Immediate	$\{M\,(76), F(24)\}$	[56,80]	26	11.9
u_4	Canada/Delayed	$\{M\,(73), F(27)\}$	[50,78]	24	10.2

Table 2. Sample Dataset : Symbolic Data (mixed variables)

The variables in Table 1 are classical variables whereas the variables in Table 2 are symbolic-valued variables (SVars).

If we select:

$d = \{$(Area of study = Australasia) $\wedge$ (Treatment schedule $=$ Immediate)$\}$ and denote by $\alpha(i)$ the membership function where, for a patient participating in the above clinical trials

$\alpha(i) =$ True iff [(Area of study $=$ Australasia) $\wedge$ (Treatment schedule $=$ Immediate)], then this condition is satisfied only by the first row of Table 2.

In our case, the R relation is the simple "belongs $(\in)$" linking the SVar "Area of study/-Treatment schedule" with the particular description (d) of interest.

In the case of the above symbolic data setting, a prerequisite to enable new knowledge extraction and the manipulation of the corresponding SOs, SDT as well as their underlying concepts is the development of an appropriate set of statistical metadata items that would hold ***additional*** information for the construction of a SO and a SDT, the SVars, as well as the relation of the symbolic data setting to classical data (Papageorgiou and Vardaki, 2007).

More specifically, metadata considered for the extension of classical data to symbolic data should describe at least the following:

- the original data used for the creation of a SO and a SDT

- the SO, SDT and each SVar as well as the components for their creation (individuals, groups, variables, conditions, the membership function, etc.)
- the process of the SVar, SO and SDT creation (keep the processing history) from the corresponding classical data.

3 Metadata modeling of biological datasets for symbolic analysis

Consider the case of the clinical studies described in the previous section. The entire process of such a study follows several stages which, in very broad terms, can be divided into: (i) design and development, (ii) patient accrual and data collection and (iii) follow-up and analysis (quality control for completeness and accuracy, study monitoring, analysis and end results).

Metadata are part of every representation of data in each of the above steps of a clinical study. Column headers, titles, descriptions and footnotes of tables or introductory text are only simple examples of metadata that are useful for the thorough understanding of the data. In the case that a SDT is used as input, metadata are particularly useful especially if the user is not very familiar with symbolic analysis and the associated presentations of symbolic data.

Consider now an individual (patient) with specific eligibility criteria and risk factors which enters a clinical study performed in different research centers as the studies presented by Ackland et al. (2005). Then, metadata for the classical (original) data should be transformed using the metadata that correspond to the class membership variables into ones describing the new set of objects. This will include documenting the new SVars, the SOs, SDT, the relations denoting the operators and the descriptions.

The metadata model discussed in this section extends the model for multisource biological data/metadata introduced by Vardaki et al. (2009) to its symbolic setting in order to allow for combined results (Figure 1). We illustrate how symbolic data depend on the individuals of a sample, the statistical population and other elements of the original/classical data (see also Papageorgiou and Vardaki 2008), thus extending the classical clinical datasets setting in order to be used by the symbolic analysis techniques. The upper-right part of the model indicates the classical setting, while the resulting symbolic part is deployed in the lower-left area. The *survey* (denoted by the class *clinical trial/study* in the model) and the *symbolic data table* are the two central classes of the model representing the main components for classical and symbolic analysis techniques respectively.

For better understanding of the relations of classes of the model presented in Figure 1, the classical part is discussed in Section 3.1 and the symbolic part of the model is further examined in Section 3.2.

3.1 Modeling the classical data/metadata setting

As already mentioned, the central class of the classical setting part of the model is the survey. Indicative attributes of the *clinical trial/study (survey)* before the initiation of any process mainly concern its general framework like for example, the hypothesis testing, the name of the trial, its type (field trial, retrospective survey, case study, drug trial, etc., according

to the categorizations applied), purpose, phase code (I,II,III,IV), disease condition examined (diagnostic or not), clinical syndrome, as well as specifications of the time coverage anticipated (start date from patients accrual until the end of observations), the geographical coverage, etc.

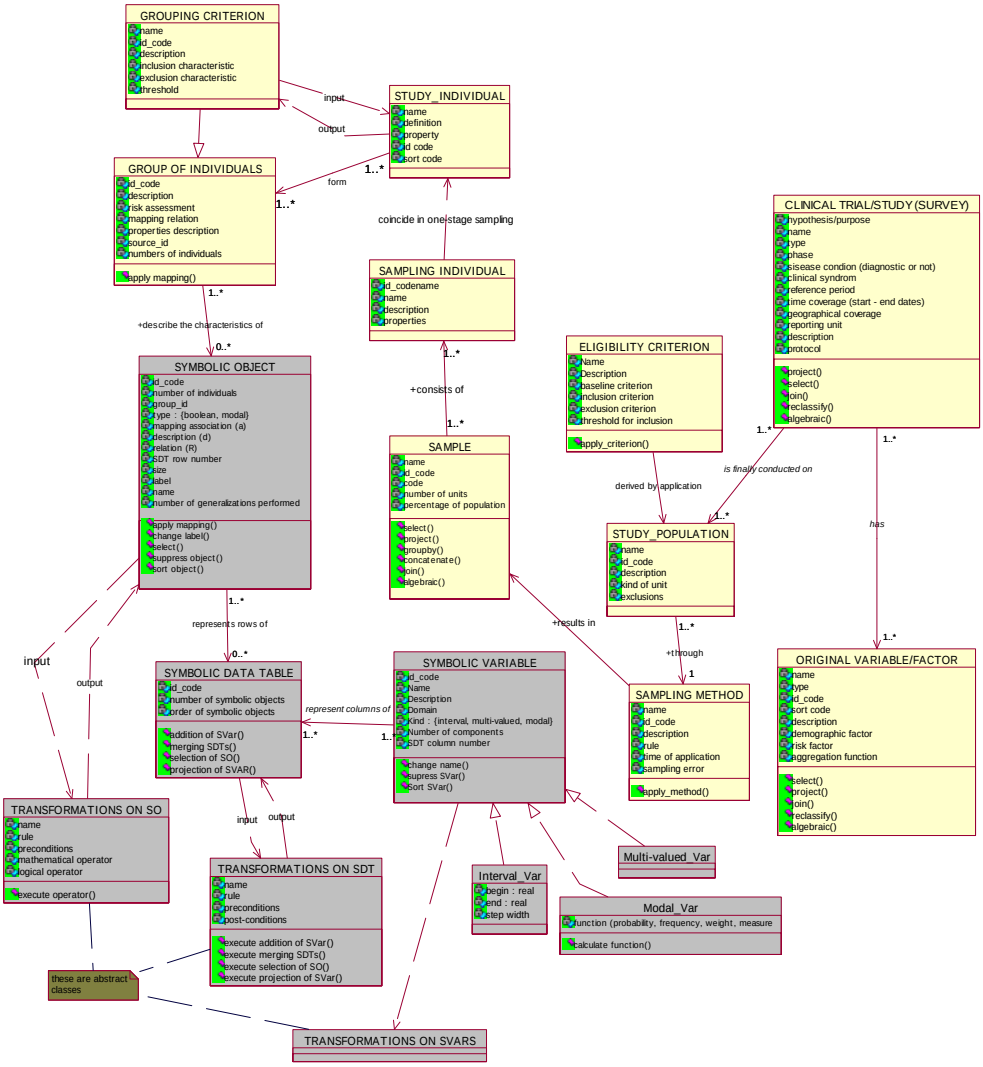

FIG. 1 - *Metadata model for the extension of Classical Data into Symbolic Data to describe the process of medical research data management.*

Initially, we examine a *population of interest* (consisting of all eligible patients) which is examined according to one or more *eligibility criteria*. These are inclusion or exclusion factors applied on every previously considered eligible patient to verify its condition. Finally, one or

more *study populations* are derived which include individuals satisfying the eligibility criteria. For example, in cases of non-diagnostic disease conditions, a "low threshold for admission" is required (examined as attribute in eligibility criterion). We must properly define the study population in use (i.e., the analytical definition and description, exclusions, as well as a database identification code (denoted by id-code) since a center's database can handle different study populations simultaneously or may need to be considered as a retrospective survey in a future experiment.

Usually it is not desirable or feasible to examine every unit included in the study population, so a *sampling method* is applied to derive a smaller, but representative enough, set of units (*sampling individuals*), the *sample*. The selected sample consists of sampling individuals having specific properties. The sampling units, in the case of one-stage sampling, form the individuals of each study (*study individuals*) which finally participate in the symbolic data setting. Since biological variation is such that patients with the same medical disease will almost certainly show varied responses to a given treatment, clinical trials inevitably require groups of patients (Gardner et al., 2001). Therefore, we should apply one or more grouping criteria to patients and form groups (*groups of individuals*). Each formed group of patients has different characteristics and properties as well a unique database id-code for further database processing.

Finally we describe each *original variable/factor* which defines and documents the characteristics of the study subjects and of the entire clinical study.

It should be noted that the consideration of samples in the model serves mainly in selecting the appropriate Individuals to form Groups and as a pre-condition for the relative transformations of Section 4 to be valid. Other tasks like i.e. performing inferential analysis, is not in the scope of this paper.

A more extended model of classical data/metadata, their relationships, dependencies and operators have been broadly discussed in Vardaki et al. (2009) and Vardaki and Papageorgiou (2010).

3.2 Modeling Symbolic data/metadata

Whenever composing groups of individuals and symbolic objects, there is a need to describe their process of synthesis and also provide essential metadata both for the interpretation of the results and for the handling of the output for further processing.

In our case, the main link class with the classical setting described in Section 3.1 is the formulation of the *symbolic object* from a *group of individuals* derived through the application of a *grouping criterion*. For example, in the two randomized trials described in Section 2 one grouping criterion applied to both studies has been the "treatment schedule" defined by the related classical variable/factor.

Although the classical definition of a SO given in Section 1 does not directly consider the groups of individuals as part of the SO description, in clinical trials, the information on associated groups of individuals is essential. Therefore, in order to further use a derived SO, we should hold metadata on the groups description, grouping factors applied, number of individuals included in each group, the source of the clinical trial, denoted by source-id (i.e. Australasian or Canadian center in our example), etc.

The *Symbolic Data Table* is the central class for symbolic data analysis and it is related with the SO and SVars classes by the way it is constructed, where each row represents the

description of a SO and each column is a SVar. Therefore, any operator applied on a SO or a SVar affects the corresponding SDT they represent.

Concerning a *SVar*, its description, domain and kind are kept as attributes; a symbolic valued variable can be of a kind either as an interval variable (its values are intervals or ordered categorical values), a multi-valued variable (its values are sets of values), or a modal variable, which is more complex than the others. A value of a modal variable is a set of pairs, where each pair consists of a value observed in the specific group of individuals and its relative function that can be calculated using a frequency, probability or weight distribution (for more information see Billard and Diday, 2003). The SDT column number is also considered by the model for further use in any operator process.

Since clinical study datasets (original data) but also their symbolic settings are aggregated data (macrodata), a user of such biological information would benefit if he/she could be able to trace back quickly the processing history of specific datasets creation with minimum response-time. A way of achieving this goal is by using such a highly structured model as the one discussed in this section ***embedded with operators and transformations*** for the manipulation of data. Similar transformations as the ones performed for classical data can be applied for symbolic data having a number of pre-conditions as discussed in the following section.

4 Transformations on symbolic data

A transformation is the result of simple or complex processing steps on clinical data, either classical or symbolic. It models any post-processing steps applied to datsets, describing their outcome on both the data and the metadata. Examples of transformations range from the simple merging of two clinical studies, to the evaluation of a complex clustering algorithm.

In accordance to a classical data setting, a number of transformations can be executed for symbolic data since our model keeps information about the series of processes that have been applied on the data of a survey or a SDT. The model's structure specifies which data and metadata items we will capture; the operators permit the execution of processing of metadata items included in the model, while the transformations ensure the validity and automation of data/metadata manipulations.

For symbolic data setting, the operators permit the execution of a specific process required for the extension of classical data to symbolic data and each of these processes is denoted in the operators part of each class by an "*apply process()*" operator. For example, the "*apply mapping()*" in classes *group of individuals* and *symbolic object* is required in order to denote the application process of the mapping relation and the mapping association accordingly. In classical data, similarly, the "*apply method()*" included in the *sampling method* class is an operator necessary for the derivation of a sample.

Each transformation identifies a set of rules (preconditions) to be valid. The pre-conditions are used by an automated system to decide whether a transformation can be applied, thus minimizing possible errors. Furthermore, each transformation defines post-conditions which are used when a chain (workflow) of two or more transformations must be optimized, describing the properties of its results. For instance, if the post conditions deduce that two transformations are interchangeable, then the system can arrange the order of their evaluation in such a way that resources and processing time is minimized.

As illustrated in the model of Figure 1 specific transformations can apply on SOs, SVars and SDTs. Transformations vary from simple ones, like 'Sorting of a SO', 'Suppression of a SO', 'Change label', 'Sorting of a SVar' and 'Suppression of a SVar' for symbolic objects and symbolic-valued variables accordingly, to more complex transformations which need a number of pre-conditions to be valid.

The application of transformations is denoted by the relative operator for its execution in each associated class of the model. Nevertheless, for demonstration purposes we illustrate the input and output classes of each transformation with the use of abstract classes like the "Transformations on SDT", "Transformations on SO" and "Transformations on SVars" which are not logical classes but have been included only for demonstration purposes.

In this paper, we do not intend to include all possible transformations since we expect users to add, or modify transformations, simply by writing and publishing the SQL code required and the pre- and post-conditions. We should note however, that all transformations, by their definition, have the closure property, meaning that, the application of a transformation on a SDT produces a new SDT. The closure property is important as it allows for the chaining of two or more transformations to describe more complex processing flows. This feature permits our model to track the history of processing steps applied on clinical trials, effectively allowing users to confirm their correctness and, most importantly, to assert the quality of the information produced. Indicative transformations are described as follows:

Addition of a symbolic-valued variable (SVar) in a SDT

We can add a symbolic-valued variable to an existing SDT only if the study population of the classical data for the particular clinical trial has such a measurable characteristic. This transformation can be applied when we manage datasets *from the same clinical trial*, since the main precondition of this transformation is that the SVars under consideration refer to the same study population and sample. In the database and the resulting SDT, existing data tables will be extended with an additional column for the new SVar.

Merging SDT

This transformation is very useful when different institutes perform clinical studies in various countries (or in general, for clinical datasets having different study populations), as discussed in Ackland et al. (2005). In such cases, a new SDT is produced and further processed in the database containing data from the previous two SDTs. In order for this transformation to be valid, it is required that:

a) the SVars of both SDTs are equivalent (see Vardaki et al., 2009, for more information on equivalence relations),

b) the intersection of the two original study populations is empty (void). This pre-condition is needed to ensure that the resulting groups of individuals forming the SOs of the SDT do not incorrectly contain duplicates that may lead to biased results.

Merging of SDT can be performed by joining, one under the other, two symbolic data tables with equivalent SVars having the same order in each table. Therefore, we produce a new table (SDT) having the same columns (SVars) as the two initial tables and its rows will be the rows of both SDTs under consideration.

Selection of SOs

It is common in a clinical trial, like in every survey, that an investigator may need only subsets of the data collected. When we require only a part of the study individuals for further anal-

ysis, in fact we apply an exclusion grouping criterion on the individuals, thus producing a new group of individuals describing a new SO. The result of applying this transformation is a new SDT holding only a subset of the initial SOs satisfying the selection criterion. For example, in our example of the clinical trials examined by Ackland et al. (2005), we can apply a selection transformation to remove from the SDT the groups of patients with risk factor: "Weight ≥ 90 kgs". The excluded patients may be represented by one or more SOs in the SDT depending on the partitioning of the classification we use for the "Weight" variable. In fact, if we consider the SVar "Weight" as an interval variable with step width: 10, then, in our example, we can have the following partitioning into 4 groups: $\{[60, 70), [70, 80), [80, 90), [90, 100), [100, 110)\}$, (initial inclusion criterion for trial entry of an individual (i) regarding the weight risk factor was: $60 \leq i < 110$). In this case, two SOs will be removed from the SDT, the ones that are characterized by the weight group $[90, 100)$ and $[100, 110)$.

Projection of SVARs

This transformation represents a symbolic-valued variable removal from the SDT. It is a transformation that removes an entire column from a SDT and implicitly defines a SDT holding only a subset of SVars. For example, in clinical trials, we can apply projection for removing the symbolic-valued variable that reveals the risk factor "weight" of the group studied.

The abstract classes "Transformations on SDT", "Transformations on SO" and "Transformations on SVars" are provided for the execution of each of the above transformations. Regarding the discussed "Transformations on SDT", the input can be one or more SDTs and the output of each transformation is one new SDT. Series of transformations can be applied, each one producing a new table which is uniquely stored in the database and the history of the various SDTs produced with the application of various transformations is maintained.

5 Conclusions and suggestions

In this paper, we do neither consider new terminology nor intend to include all metadata required for symbolic data analysis of biological datasets. Our contribution is that we structure metadata in such a way that the processing of classical into symbolic data/metadata setting can be partially automated. For example, if someone attempts to add a new individual (i') who has description (d') to a group of individuals of a SO (u) with description (d), the model will warn the user in the case that the description (d') does not match the description (d) of the target SO. This is achieved by the system's support of transformations which formally describes the result of simple or complex processing steps on symbolic clinical data.

Further steps should include the definition of a quality framework in the form of metadata for all stages followed in extending the classical setting into the symbolic data setting for the clinical study data collected and managed using the structured statistical model described in this paper. Such quality assurance framework would allow for automatic quality assessment in all processing steps of further Symbolic Data Analysis techniques.

Acknowledgement

The authors would like to thank the referees for their valuable suggestions.

References

Ackland S.P., M. Jones, D. Tu, J. Simes, J. Yuen, A.M. Sargeant, et al. (2005). A meta-analysis of two randomised trials of early chemotherapy in asymptomatic metastatic colorectal cancer, *British Journal of Cancer 93 (11)*, 1236-1243.

Billard L. and E. Diday (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis, *Journal of the American Statistical Association* 98 (462), 470-487.

Billard L. and E. Diday (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley, ISBN: 978-0-470-09016-9.

Bock. H. and E Diday (2000). *Analysis of Symbolic Data*, Springer-Verlang, Berlin, ISBN 3-540-66619-2.

Gardner D., K. Knuth, M. Abato, S. Erde, T. White, et al. (2001). Common data model for neuroscience data and data model exchange, *Journal of the American Medical Association 8*, 17-33.

Diday. E. (2002). *An introduction to Symbolic Data Analysis and the Sodas software*, JSDA Electronic Journal of Symbolic Data Analysis, 0 (0).

Diday. E. and M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley, ISBN 9780-470-01883-5.

Noirhomme-Fraiture M. and P. Brito (2011). Far beyond the classical data models: symbolic data analysis, *Statistical Analysis and Data Mining*, 4(2), 157-170, Wiley.

Noirhomme-Fraiture M. (1997). *Zoom-Star, a solution to complex statistical objects representation*. St. Howard, J. Hammond and G. Lindgaard (Eds.), Proc. INTERACT '97, Sydney

Papageorgiou, H. and M. Vardaki (2007). *Quality Issues in Symbolic Data Analysis*, Selected Contributions in Data Analysis and Classification. P. Brito, P. Bertrand, G. Cucumel, and F. De Carvalho (Eds.), 113-122. Springer.

Papageorgiou, H. and M. Vardaki (2008). *A Statistical Metadata Model for Symbolic Objects*, Symbolic Data Analysis and the SODAS Software, E.Diday & M. Noirhomme (Eds.), 67-80. Wiley.

Vardaki, M. (2004) *Metadata for Symbolic Objects*, JSDA Electronic Journal of Symbolic Data Analysis, 2 (1).

Vardaki M. and H. Papageorgiou (2010). *On Quality Assurance and Assessment of Biological Datasets and Related Statistics*, Advances in Computational Biology, H. Arabnia, (Ed.), 89-97. Springer.

Vardaki M., H. Papageorgiou, and F. Pentaris (2009). A Statistical Metadata Model for Clinical Trials' Data Management, *Computer Methods and Programs in Biomedicine* 95, 129-145.