WP. 29
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (iii): Metadata models and terminology

CRISTAL, A MODEL FOR DATA AND METADATA

Contributed Paper

Submitted by Statistics Netherlands, the Netherlands[1]

I. INTRODUCTION

National statistical institutes do not hold a monopoly on the production of statistical information. The competition with other information producing institutes is likely to increase in a near future. Therefore it is strategically important to focus on the special added value that national statistical offices can offer. The *quality* and the *coherence* of the statistical information they supply are probably their trump cards. Hence national statistical institutes should further invest in the quality and coherence of their information to keep ahead of the competition.

For several years Statistics Netherlands has had the policy to store all published statistical information in their output database. As such it has succeeded collecting all statistical output information in one central database. Moreover, all the information in this central output database is freely accessible for everyone on the internet. However, the average *quality* and *coherence* of the information in the output database is still moderate.

When seen from a logical perspective the output database is a heterogeneous collection of independent statistical *cubes* from which clients can select subsets in statistical *tables*. The statisticians who enter the cubes in the output database are completely free to introduce their own definitions, statistical units, classifications and target variables, independent of the other cubes.

Even though the current system with a heterogeneous collection of cubes in the output database is functioning rather satisfactory, the clients of statistical information become more demanding. Clients tend to ask for more background information when they ask for more information about:

1. how the statistical *terms* are defined,
2. how the statistical information *relates* to neighbouring information in the same environment, and sometimes even
3. how the statistical information was *manufactured*.

[1] Prepared by Erik van Bracht. – ebct@cbs.nl

The most acute problem for the current output database is that this kind of background information is insufficiently available. This problem does not only arise in the output database of Statistics Netherlands. This is probably one of the most challenging problems in practically all national statistical institutes.

The solution to this problem is not straightforward. The problem has many organisational aspects, but not all aspects are organisational. Although it is not always recognized, many essential modelling problems have not yet been solved either. Even if the national statistical institutes try many technically advanced solutions, a satisfactory model for the coherence between heterogeneous kinds of statistical information does not seem to exist until now. The Cristal model presented in this paper is an attempt to provide a solution.

II. ON STATISTICAL INFORMATION MODELS

Statistical information can be modelled by means of many different kinds of data models. Basically practically all data models consist of infinitely extensible sets of *mutually related texts, words and numbers*. As a consequence any logical system that is capable to manage any extensible set of texts, words and numbers with any kind of mutual relation generalizes all these statistical data models.

However, statistical information model designers will probably have difficulties with mutually related texts, words and numbers only. Instead they want to work with more tangible and identifiable *entities taken from the real world*. Therefore an interesting *ontological*[2] step takes place in statistical information modelling: certain combinations of interrelated texts, words and numbers are being *interpreted as existing entities*.

The right *interpretations*, i.e. how certain related texts, words and numbers correspond to really existing entities can of course be clarified, again by means of *other* related texts, words and numbers often referred to as the *meta information*. Even if this procedure is sufficient in many practical cases, from a more formal point of view this gives the impression of postponing a fundamental problem: even if the relations between texts, words and numbers can, their *interpretations cannot be formalized*. This means that statistical information models can only relate texts, words and numbers, but they cannot give them a *completely unambiguous meaning*.

In spite of this fact information model designers still strive to do the impossible. As contemporary Don Quixotes they strive to bridge the gap between the formally related texts, words and numbers on one side and the structure of the really 'existing' entities on the other.

The most common of the information models used is the Entity Relationship (ER) model usually implemented in a Relational Database Management System (RDBMS); see [BatiniEtAl1991] and [Thalheim2000]. The ER model distinguishes a finite number of interrelated entity types, each of which has a finite number of attributes. The RDBMS systems implement these entity types by means separate rectangular *tables* that have the attributes implemented as horizontal *fields* in columns. They implement the 'recorded' entities of these types in a list of vertical rows called *records* that have unique *keys* for their identification. A relational algebra determines how information in such tables can be related, joined, intersected and so on.

During the last decade even more efficient specializations of such models, called data warehouses have been developed; see [Kimball96] and [Kimball98]. Also a relatively new modelling technique called Object Oriented (OO) modelling, introducing the behaviour, inheritance and encapsulation of entity types, has gained much popularity in software design; see [Martin&Odell1997].

III. WHY A CRISTAL INFORMATION MODEL?

The question is, if there are already ER models, optimised RDBMS systems, data warehouses, object oriented modelling tools and many other information models available, why should statisticians still need a Cristal information model? The answer is as follows.

---

[2] Ontology is the branch of philosophy that studies the nature of being and the kinds of existents.

The common information models are *insufficiently well suited to maintain the coherence between many heterogeneous collections of statistical information*. They provide insufficient *assistance to maintain coherence* between:

- heterogeneous interrelated collections of changing statistical classifications on different levels of detail.
- heterogeneous interrelated statistical datasets based on the changing statistical classifications.

The Cristal model was especially designed to maintain coherence between many heterogeneous collections of statistical information. The concise description of the Cristal model follows in the rest of this article.

IV. CRISTAL DESCRIPTION

A Cristal is an information structure for statistical information that can be split in two main parts: a formal structure for statistical information (mainly variables, classifications and values) that is already known *before* collecting any statistical observations and a formal structure for statistical information (mainly observation types and observations) that is introduced *after* collecting new observations. These two forms of information may also be called *a priori* and *a posteriori* statistical information. The model for a priori statistical information will be described in the following subsection A. The model for a posteriori statistical information model will be described in subsection B thereafter.

A. A Priori Statistical Information

The most basic elements of a priori statistical information are the a priori *categories,* and the a priori *part-whole relations* between them. However, more is needed to describe statistical information: the extra notions of sets of categories and of category trees are indispensable. Therefore two extra notions are introduced in the model: *levels* for sets and *hierarchies* for trees.

This subsection A is divided into five smaller subsections. The first following subsection 1 describes the basic categories. The second subsection 2 formally describes their part-whole relations. The third subsection 3 describes the levels and the fourth subsection 4 describes the hierarchies. The last subsection 5 describes how all the previous notions are wrapped up in variables.

1. Categories

The word *category* comes from the ancient Greek word *katêgoria*, which is derived from kata ("against") and agoreuein ("to assert"). The Greek Aristotle was the first to use the term *katêgoria* in philosophy. He adapted the term from the legal language, in which it meant "*accusation*", and used it first to mean the same as "*predication"*, i.e. "*the logical affirmation of something about another"*. Since then, partly due to Aristotle's own fundamental categories[3], the meaning of the word category has narrowed to mean either "*any of several fundamental and distinct classes to which entities or concepts belong"* or "*a division within a system or classification*" However, in the Cristal model the term "category" is meant in the original broader sense, which is synonym with "*predicate*", meaning "*something that is asserted about a putative object*"[4].

Identification and Attributes of Categories

Every category in the Cristal model has at least a *name*, a *local key*, a *description* and a *globally unique identification (GUID)* number. The attributes can be extended in two ways. They can be extended with any other kind of *user defined attributes* and they can be extended for their translations in *any other language*. Technically speaking a category uses only its GUID number for its identification, and the name and description of a category are not meant for identification purposes.

---

[3] Aristotle distinguished the following ten 'fundamental' *categories*: substance (what it is), quantity (how much), quality (what kind), relation (with respect to what), place (where), time (when), position (to be put, to lie), possession (to have), action (to do) and passion (to undergo). Later philosophers like Kant, Hegel, Pierce, Whitehead and Ryle have explored different schemes of 'fundamental' categories.

[4] Putative objects are objects that are only supposed to exist by intuition. There is probably no hope for indisputable and logical criteria for existence, which means that the existence of every object will always have a putative status. However, the putative status of existing objects will not be emphasised any further in this document.

2. Category Relations

The basic a priori relations between the categories in Cristal are the *part-whole* relations. The part-whole relations in Cristal are governed by a formal system similar to mereology[5] [Simons87], which organizes the categories in multiple hierarchical structures.

In the literature one can find various kinds of formal part-whole systems in different flavours. Almost all of them use the general structure of a *partial ordering*, which is the basic structure for category relations in the Cristal model. The formal partial ordering relations between categories assumed in the Cristal model will be elucidated in the next subsection.

Partial Ordering

In a part-whole relation, one category plays the *role of the part* and is called *sub-category* while the other plays *the role of the whole* and is called *super-category*.

In Cristal the part-whole relations have a partial ordering structure governed by the following rules:

1) A category is *always* a sub-category of itself. (Reflexivity)
2) If a certain category has both a super-category and a sub-category, then its sub-category is always a sub-category of its super-category. (Transitivity)
3) If two categories are sub-category of each other then these two categories must be identical (Anti-symmetry)

Interpretation of Part-Whole Relations between Categories

If category c_1 is a part of category c_2, then c_1 *logically implies* c_2. In other words: if c_1 can be asserted about an object, then c_2 can also be asserted about it. This can be illustrated by the following examples.

The predicates “is in France” and “is in Europe” can both be asserted about the same object and the first assertion logically implies the second. This is obviously because ‘France’ is considered to be a part of ‘Europe’. However, other situations are also possible: “is a person” logically implies “is a creature”, but ‘person’ is not considered to be a part of ‘creature’. Instead ‘person’ is a specialization of ‘creature’. “is in France” and “is a person” are sub-categories of “is in Europe” and “is a creature” respectively.

3. Levels

In the previous section about category relations, the *transitivity* of category relations was accepted without further comment. However, predicates that correspond to non-transitive relations may exist. These non-transitive relations are the element-set relations. These relations do not fit in a formal mereological system. That is why the Cristal model introduces the notion of *levels* for the special non-transitive predicates.

A level is basically a *set of categories*. However, there are two important differences between levels and sets of categories:

- Two levels can be different, even if they have exactly the same categories[6].
- Two different categories in a level cannot overlap[7].

These two differences will be clarified in the following subsections.

No Overlap between Categories in Levels

Any two parts described in mereology can have a mutual overlap. The same holds for two sets of sets. This is because for instance the set $S = \{\{a, b\}, \{b, c\}\}$ is fully legitimate in set theory even though $\{a, b\}$ and $\{b, c\}$ share the same element b. However, in contrast, two distinct categories in a level may not overlap. This will be elucidated further in the next paragraph.

---

[5] Mereology is a branch of mathematics that is based on the part-whole relation instead of the element-set relation.
[6] In set theory two sets are equal if they have exactly the same elements
[7] The set $S = \{\{a, b\}, \{b, c\}\}$ contains two sets that contain the same element b. Such a set is legitimate in set theory but an equivalent situation is forbidden for a level in the Cristal model.

Suppose that, in analogy with the previous set *S*, categories *a, b, c, x, y* and *z* are such that *a* and *b* are a part of *x, b* and *c* are a part of *y* and *x* and *y* are a part of *z*. This situation is depicted in Figure 1, where the arrows are shown from wholes to parts. In this case the *set* $S = \{x, y\}$ is legitimate, but the *level* $L = \{x, y\}$ is not allowed because *x* and *y* overlap, i.e. they share the same subcategory *b*.

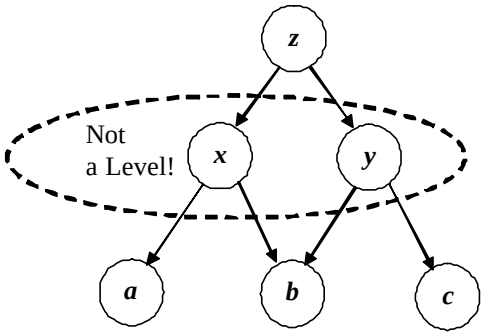

Figure 1.Parts x and y cannot be in the same level because they share part b

The use of this extra restriction in the levels is important for statistics because it guarantees that *double-counting is always avoided* when working with levels.

Identification and Attributes of Levels

There is also another important feature that distinguishes levels and sets of categories. Sets are identified by means of their elements. A set $\{a, b\}$ is always identical with the set $\{b, a\}$, simply because they share exactly the same elements.

However, levels in the Cristal model do not only have elements. Like the categories, levels also have a *name*, a *local key*, a *description* and a *globally unique identification (GUID)* number. The attributes can be extended in two ways. They can be extended with any other kind of *user defined attributes* and they can be extended for their translations in *any other language.* A level uses only its GUID number for its identification, which means for instance that a level with the categories $\{a, b\}$ is not necessarily the same level as a level with the categories $\{b, a\}$.

Ordering of Levels: Refinement Relations

A level can be the super- or sublevel of another level. But for any two levels, one level is not necessarily a sub- or super level of the other. If a level is a sublevel of another level then this level is called a *refinement* of the other level. This indicates that the *refinement relations* between levels are partially ordered as well.

The refinement relations between the levels can be completely *deduced* from the partial ordering of the underlying categories. The following simple rule will be used to deduce the refinement of levels from the ordering of categories: *a level is a refinement of another level if every category in that level has a corresponding super-category in the other level*. Note, however, that a super-level may have categories without any corresponding sub-categories in its sub-level.

4. Hierarchies

The Cristal model offers special support for the use of trees, a structure in which every statistical classification and codification system can be modelled as well. The model for a tree is called a *hierarchy* in Cristal.

A *hierarchy* is a sequence of levels in which each next level is a refinement of the previous level. Moreover, each next sub-level is different from its previous level. The order for the levels in a hierarchy must comply with the order of the refinement relations between the levels.

Identification and Attributes for Hierarchies

Like the categories and levels, hierarchies also have a *name*, a *local key*, a *description* and a *globally unique identification (GUID)* number. The attributes can be extended in two ways. They can be extended with any other kind of *user defined attributes* and they can be extended for their translations in *any other language*. A

hierarchy uses only its GUID number for its identification. Thus, the sequence of levels as well as the name and the description of a hierarchy are not meant for identification purposes.

5. Variables

The Cristal model has two main kinds of variables: *simple* variables and *classification* variables. A simple variable is simply based on a single *set* of values, while a classification variable is based on a mereological system of categories with extra support for multiple levels and hierarchies.

Simple Variables

The simple variables in the Cristal model are introduced to support *sets of values*. These can be numerical values possibly with measurement units, such as '1 meter', '2 meter' and '3 meter', but can also be textual values, such as 'low', 'medium' and 'high'. Theoretically such values should also be treated as categories. However, in practice, statisticians will never give these categories a more specific description, which makes their meaning implicit rather than explicit. Even the relations between values are hardly ever specified explicitly, primarily because they are considered as obvious.

A simple variable is a variable that represents a *set of coordinate*[8] *values*. The simple variables support working with simple sets of values. Two simple variables can be different even if they have exactly the same set of values.

Identification and Attributes of Simple Variables

Like the categories, levels and hierarchies also the simple variables have a *name*, a *local key*, a *description* and a *globally unique identification (GUID)* number. The attributes can be extended in two ways. They can be extended with any other kind of *user defined attributes* and they can be extended for their translations in *any other language*. A simple variable uses only its GUID number for its identification. Thus, neither the set of categories, nor the name and the description of a simple variable are meant for identification purposes.

Classification Variables

A classification variable is a complete mereological system of categories, extended with levels and hierarchies for extra support for sets and trees. It has:

- A collection of partially ordered categories. If a category is in a classification variable, then all its sub-categories are in it as well.
- A collection of partially ordered levels defined on the partially ordered categories. The ordering of levels is in concordance with the ordering of the categories,
- A collection of hierarchies defined on the partially ordered levels. Each hierarchy in the collection is in concordance with the ordering of levels.

Note that each of these three collections can be empty. Next to the three collections a classification variable has also the more common identification features described in the next subsection.

Identification and Attributes of Classification Variables

Like the categories, the levels and the hierarchies, all classification variables have also a *name*, a *local key*, a *description* and a *globally unique identification (GUID)* number. The attributes can be extended in two ways. They can be extended with any other kind of *user defined attributes* and they can be extended for their translations in *any other language*. A variable uses only its GUID number for its identification. Thus, the collection of categories, the collection of levels, the collection of hierarchies, the local key, the name and the description of a classification variable are not meant for identification purposes.

B. A Posteriori Statistical Information

The idea behind the description of a posteriori information in Cristal is similar to the theory of *combinatorial ontology* [Jacquette2002]. The common idea is that only *combinations of known predicates* will be used to describe any *unknown phenomena*. This means that the description of a new phenomenon is in fact nothing more than the creation of a new combination of pre-defined predicates. However, in the Cristal model the term

[8] In this context, coordinate means equal in rank, quality, or significance.

category is used instead of the term predicate. In terms of the Cristal model that means that all the new observations of statisticians can be described by making '*new' combinations* of *'old' categories*.

In the conceptual model of Cristal something may already 'exist' if it can affirm or deny *a subset of* known predicates. That means that as soon as a statistical observer has the conviction that something exists and is worth describing, then he can describe his observation by making a corresponding combination of registered predicates. For his description he does not need to affirm or deny *every* known predicate.

As a consequence the a posteriori information is structured by means of combinations of a priori categories in Cristal. Because of this viewpoint it is the aim in Cristal to enable the construction of *every possible combination of categories*, independent of the variables these categories may relate to. A constructed category combination is called a *datapoint*.

Even if they play only an auxiliary role, the variables can be combined as well. The combinations of variables define *statistical observation types.* A statistical observation type is an ordered sequence of variable references. Potentially every ordered sequence of variable references can define another statistical observation type. Just like the variables, the observation types play also an auxiliary role in the description of the a posteriori information. That is why they must always be accompanied by the actual observations corresponding to these types.

As a result the observation types are always linked to a collection of corresponding observations in the form of datapoints. A single observation type supplemented with a collection of observations of that type is called a *dataset*. The datapoints and datasets will be described in more detail in the next two subsections.

1. DataPoints (Observations)

A datapoint is an ordered sequence of references to categories. Datapoints represent *statistical observations*[9]. Ideally datapoints are totally independent of any kind of variables or types. In Cristal there is *no restriction* for the combinations of categories that can be made. Every logically possible ordered sequence of references to categories is a valid datapoint, irrespective of the kind of variables these categories may relate to.

Identification for DataPoints

Although the preceding definition may seem rather straightforward, there are some important *identification issues* with datapoints. The identification issues for datapoints are more complex than for other elements in the Cristal model, mainly because datapoints do not have a Globally Unique Identification (GUID) number:

- The first, but rather simple, identification issue is that datapoints are *ordered* sequences, which implies that two datapoints are considered different if they refer to the same categories in a different order.

- The second identification issue is more fundamental. Two datapoints may represent two *different* objects in reality even if they have exactly *the same ordered sequence of category references*. In reality there is always something that differs between the two existing objects. However, sometimes the category that expresses that difference is not incorporated explicitly in the sequence of datapoint references. The unpleasant consequence is that exactly the same ordered sequence of references can describe different 'statistical objects' in reality. Because of this it can be useful to permit the existence of *different*[10] datapoints with exactly *the same ordered sequence of category references*.

Footnotes for DataPoints

Every datapoint has an extra optional footnote that gives the opportunity to relate some extra textual information to the corresponding observation. This footnote can be used for many different purposes. It is often used to distinguish the status of observations, for instance to distinguish *preliminary* observations, *provisional*

[9] The term statistical observation does not necessarily mean that it must be observed with the human senses. The statistical observations can also be *deduced* from other observations.
[10] This implies that there is something that can discern these datapoints but is not a category.

observations and *final* observations, et cetera. Obviously, it can be used to attach any other kind of textual information as well.

2. DataSets

The unpleasant identification problem for datapoints can be evaded by the introduction of *local environments* for datapoints. These local environments are called *datasets*. Apart from serving as a *local environment for datapoints*, the dataset has two other purposes:

1. It enumerates the *observations* of corresponding *statistical objects* by means of a set of datapoints.
2. It defines an abstract *observation type* by means of an ordered sequence of references to variables. All datapoints in the dataset comply with this observation type.

These two purposes will be elaborated briefly in the following two subsections.

Set of DataPoints (Set of Observations)

As the term dataset indicates it encompasses a *set of datapoints*. Moreover, each separate datapoint is associated with a *unique* dataset. That means that datapoints cannot be shared between datasets.

Even if it is not *globally* the case for all the datapoints in a Cristal, within the *local environment* of a single dataset the datapoints can be uniquely identified. That means that there may not reside two datapoints having exactly the same ordered sequences of category references *within the scope of the same dataset*. In the end, with the help of datasets, this enables a unique identification for datapoints.

Type of DataPoints (Observation Type)

Apart from being a set of datapoints, a dataset encompasses also the description of a *datapoint type*. Such a *datapoint type* is determined by the datapoint *type description* with an *ordered sequence of references to variables*. Inside a dataset the datapoints must always correspond to their datapoint type in the type description. More particularly this means that the ordered sequence of referenced categories in a datapoint must correspond to the ordered sequence of referred variables in the type description. In other words, if a certain datapoint resides in a certain dataset then the i-th referred category of the datapoint must be a category in the i-th referred variable of the dataset.

Identification and Attributes of DataSets

Every dataset in the Cristal model has at least a *name*, a *local key*, a *description* and a *globally unique identification (GUID)* number. The attributes can be extended in two ways. They can be extended with any other kind of *user defined attributes* and they can be extended for their translations in *any other language*. Technically speaking a dataset uses only its GUID number for its identification, and the name and description of a category are not meant for identification purposes.

V. CONCLUSION

The *coherence* and the *quality* of statistical information are of strategic importance in national statistical institutes to keep ahead of the competition with other institutes. However, the coherence and the quality of statistical information offered by national statistical offices are often still unsatisfactory.

Apart from many organisational problems, there are also problems in information modelling that block the road to more coherence and quality. This is because the most common information models in the information industry are insufficiently well suited to maintain the coherence between many heterogeneous collections of statistical information.

Statistics Netherlands has developed the Cristal model to maintain the coherence of its own statistical information. The Cristal model introduces new concepts taken from mereology (part-whole relations) and combinatorial ontology (datapoints) in statistical information modelling. These concepts – next to other ones (hierarchies for the modeling of trees, levels for the modeling of sets) – make it possible to model how statistical information relates to neighbouring information in the same environment. Further, thanks to the fact

that almost all model elements have attributes to store their description (in possibly several languages), it is always clear how the statistical terms are defined.

Although Cristal does not solve all metadata-related problems (it does not explicitly provide information on how statistical information was manufactured), it does maintain the coherence between many heterogeneous collections of statistical information in a satisfactory way.

VI. REFERENCES

[BatiniEtAl1991], Carlo Batini, Stefano Ceri, Shamkant B. Navathe, Carol Batini, '*Conceptual Database Design: An Entity-Relationship Approach*', Addison-Wesley Pub Co, 1991, ISBN: 0805302441.

[Jacquette2002], Dale Jacquette, '*Ontology'*, Acumen Publishing Limited, 2002, ISBN: 1902683560.

[Kimball1996], Ralph Kimball and W.H. Inmon, '*The Data Warehouse Toolkit, Practical Techniques for Building Data Warehouses'*, John Wiley & Sons, 1996, ISBN: 0471153370.

[Kimball1998], Ralph Kimball, Laura Reeves, Margy Ross and Warren Thornthwaite, '*The Data Warehouse Lifecycle Toolkit, Expert Methods for Designing, Developing, and Deploying Data Warehouses*', John Wiley & Sons, ISBN: 0471255475.

[Martin&Odell1997], James Martin and James J. Odell, '*Object Oriented Methods – A Foundation*', *UML Edition (2nd Edition)*', Prentice Hall PTR , 1997, ISBN: 0139055975.

[Simons1987], Peter Simons, '*Parts: A Study in Ontology*', Oxford University Press, 2000, ISBN: 0199241465.

[Thalheim2000], Bernhard Thalheim, '*Entity-Relationship Modelling: Foundations of Database Technology*', Springer Verlag, 2000, ISBN: 3540654704