MODELLING LEVELS IN THE STATISTICAL INFORMATION SYSTEM OF THE BANK OF ITALY.

Vincenzo Del Vecchio
Banca d'Italia, SISC

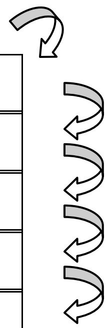
October 2002

The Idea of a Hierarchy of Models

In the electronic and data processing (EDP) literature, the idea of a hierarchy of models emerged a certain time ago. The principle is that an information system can be described by different levels (layers) of modelling in a hierarchy in which the model of some level is described in terms of a model of the hierarchically upper level and it also describes one or more models in the hierarchically lower level, as displayed in Fig. 1.

In brief, starting from the reality that has to be described (call it the "zero" level), in the first level we have the data extensions, that is, models of parts of the reality, followed, in the second level, by the data definitions, that is, models of the data. The third level contains the methods used to produce the data definitions, that is, models of models of data (meta-models). Finally the fourth level contains the methods that produce other methods, that is models of the meta-models (meta-metamodels).[1]

L4	Methods that define methods (meta-metamodels)
L3	Methods for the making of definitions (meta-models)
L2	Definitions (models of data)
L1	Extensions (data)
L0	Reality



= describes

Figure 1. *The Levels Hierarchy*

The purposes of such a structure are on one side the self-documentation of the information system (see also "infological completeness" [Sundgren, 1991: 11]; a model cannot exist if there isn't an upper level model that describes it) and, on the other side, the harmonization of the information system, that is, the existence of a generic model (in the upper level) that describes many others (its instances in the lower level).

The higher we go up in the hierarchy, the more abstract and general becomes the modelling: in practice, however, there is no need to have more than four levels, as the fourth level-model potentially allows us to obtain the whole hierarchy because a fourth level-model can be considered

[1] Sometimes, in the EDP literature, "models" are called "schemas" and consequently "meta-models" are called "models" and "meta-metamodels" are called "meta-models".

as a model that "generically defines models" and, hence, is also able to define itself, so eliminating the need of yet higher levels.

Apart from EDP applications, the principle turns out to be very useful in the statistical information system modelling. At the moment, in the Banca d'Italia, for example, this principle is the fulcrum of the structure of its statistical information system for credit and finance also beyond the data representation horizon. In fact, the hierarchy of levels allows a link between many aspects such as the internal architecture of the information system, languages, roles of involved actors, administration and processing competencies. In what follows, we aim to describe the learning that derives from this experience.

Examples of EDP Standards and Applications Based on the Idea

According to the idea of a hierarchy of models, in 1986 an international standard was proposed by ISO/ANSI for the design and implementation of a generic Information Resource Dictionary System (IRDS)[2]. An IRDS can be considered an information system that describes another information system. The information about information is also known as meta-information [Sundgren, 1991: 11], so an IRDS can be considered as a meta-information system. The proposal became an ISO standard [ISO, 1990] and was enriched later on [ISO, 1993; 1998].

The proposal is based on a multi-layered structure consisting in four levels in which every level has the purpose of defining the immediately lower level, as described in the previous subsection. The first level (the data) is considered external to the IRDS, the second level has the purpose of defining the data and is considered as the content of the IRDS, the third level contains the structure (that is, the model) of the IRDS and is itself defined by the fourth level, fixed by the standard.

An application of the same idea is the structure of the catalogue of some relational data base management systems (RDBMS) available in the market. In this case, the first level is the user data, that is, the content of the user relational tables (or views), the second level contains the definition of the user relational tables (that is, the content of the RDBMS catalogue tables), and the third level is the definition of the catalogue tables. The fourth level can be considered the model used to define the catalogue (this last level is, in principle, usable for the definition of other catalogues).

A more recent and ambitious application of the same principle is the four level-structure proposed by the Object Management Group (OMG), an organisation for the standardization in the object-oriented field of the software development in which many of the leading software production firms in the world participate. In the OMG standard (see [OMG, 2000]), the four levels are called M3, M2, M1, and M0, respectively. M3 is the fourth level (the meta-metamodels level). Only one model at level M3 is necessary to define all the M2 level-models (meta-models). The OMG standard for the M3 model, called MOF (Meta Object Facility), is able to define itself. Examples of the third level (M2) meta-models are the UML meta-model and the relational meta-model. Correspondently, at the second level (M1) there are UML models and relational models relevant to a specific subject. First level (M0) contains data.

[2] See [ISO,1986], [ISO,1987]

Hierarchy of Models in Modelling Statistical Information Systems

All the examples of the previous section of multi-layered structures have been concretely used in implementing statistical EDP applications. Relational databases were introduced in the statistical information system in the second half of the eighties and are by far the main tool used to store statistical data. The IRDS standard was known thanks to the work of a current member of Banca d'Italia's Statistical Department [Limongelli, 1986/87; Limongelli et al., 1988/89]. The idea was applied to build a simple experimental data design and documentation tool, developed in 1988/89 by Del Vecchio and Limongelli, also on the basis of a work of an IBM specialist [Magnani, 1985], and used to design the data structure of a specific statistical EDP application (as an IRDS, the tool was based on the second and third level of the hierarchy). More recent developments rely on the OMG standards.

The four level hierarchy of models was also applied to the conceptual[3] modeling of the statistical information systems that support the activity of the institutional functions of the Bank, with an approach that is almost identical to the one described in [Grossman, 2002].

As in the EDP examples, in the statistics case a fourth level-model has the purpose of defining "structures" suitable to define third level-models. Such structures are not specific of statistics, they are instead more general and usable also in other fields (for example, the operational systems). That is to say that a fourth level-model contains structures able to define any kind of methodology, possibly shared by all of them. An important feature of a fourth level-model is its self-describing property, that is, the ability of its structures to describe themselves and, therefore, to make the existence of levels higher than four superfluous.

The specificity of the statistical field is located at the third level. A "statistical" third level-model, in fact, is considered the formal representation of a methodology for statistical description of the reality (that is, a descriptive statistic methodology). A third level-model contains structures able to give a concrete and possibly formal shape to statistical methodological rules. The existence of a "statistical" third level is a consequence of the recognition of the specificity of the statistical methodology with respect to others such as, for example, the EDP ones used for the implementation[4].

A model of the second level can be considered the definition of a specific statistical information segment[5], that is, the definition of data and processes relevant to a specific subject. Therefore, second level-models are specific subject-matter models produced using a certain statistical methodology (that is, a third level-model). Note that, according to the general four level hierarchy idea, the notion of "data model" (data that is the definition of *other* data) is more specific than the more common notion of "metadata" (data that describes other data in some way). For example, a "quality datum", that is, a datum measuring, or reporting the quality of another datum, can be considered a "metadatum", yet it is not the "model" of the latter. In the four level model approach,

---

[3] The term "conceptual" is used to mean "independent from the implementation", so that many possible practical implementations can be made of a "conceptual" model, each one following its own set of implementation rules (that is, its own implementation model). In principle, there can also be models not implemented in an EDP environment

[4] Note that the EDP implementation of a certain model "M" of the "statistical" hierarchy is made by using a third level-model of the EDP hierarchy (for instance, the relational meta-model): the EDP model of such an implementation (for instance, the relational model), therefore, appears to be in the second level of the EDP hierarchy, independently of the level of "M" in the "statistical" hierarchy.

[5] The term is introduced to indicate a self-consistent part of an information system with an autonomous existence and evolution, like, for example, a survey, a stove pipe, a processing line, etc. and coincides with the notion of "domain" introduced in item **2.5.8** of Chapter 3.

both the original datum and the quality datum are considered level 1-data and have their definition in a level 2-model. So a level 1-model happens to contain also metadata, and the relationship between a datum and its “not definition metadata” is not of ‘type-instance’ type and it takes place within the *same* level, not between different ones.

As displayed in **Fig. 2**, a model of the first level is the extension of a statistical information segment, that is, an occurrence of a second level-model. In simpler words, it is a set of values that correspond to a definition. Different sets of values, therefore, are different first level-models. More than one extension may correspond to the same definition such as in the case that the measurement process generating the data extension is performed more than once or, likewise, owing to the evolution of the data content in time: an update in the data content gives origin to a new extension, different from the previous one but with the same definition.

As usual, the reality to be described is located at the zero level..

L4	Methodologies that define methodologies
L3	Descriptive statistic methodologies
L2	Statistical information segment definitions
L1	Statistical information segment extensions
L0	Reality

= describes

Figure 2. *The Levels Hierarchy in Statistics*

[Del Vecchio, 1997] contains a description of the principles on which the third modelling level is based in Banca d’Italia. Most of the described notions are derived from the descriptive statistical methodology and can be placed in fact at the third level (for example, statistical set, statistical variable, statistical function, …).[6]

The Hierarchy of Models and the Practice of Harmonization

The hierarchy of models can be seen as a conceptual tool to deal with the complexity of the statistical information systems. It appears to be useful in the design and the operation of information systems as well as in the analysis of existing ones and in the effort of harmonizing and standardizing them, independently of how the implementation is done.

The practical application of the idea leads to identify many models on every level and the type-instance relationships between models belonging to consecutive levels (see Fig. 3), roughly:

- a fourth level-model for each general modelling methodology used in practice;
- a third level-model for each descriptive statistical methodology used in practice;
- a second level-model for the definition of each information segment;
- a first level-model for each extension of an information segment.

[6] Nevertheless, there are also examples of subject matter definitions and of extensions.

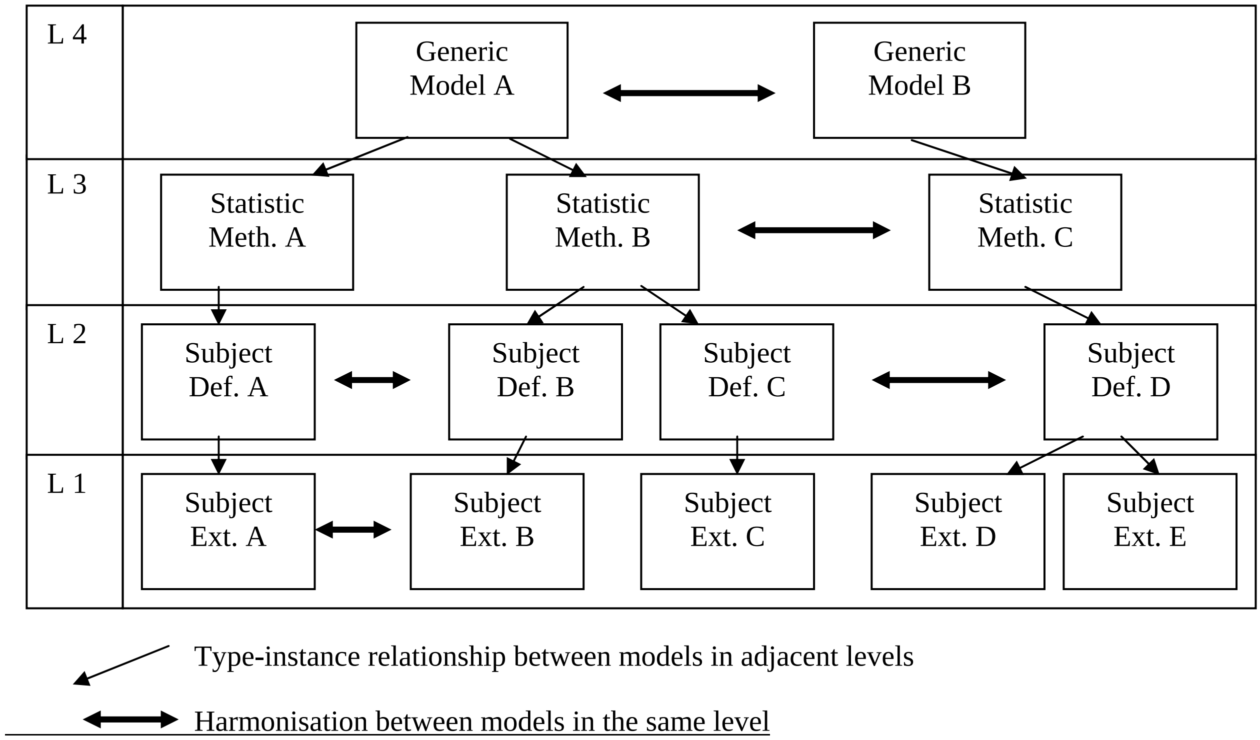


Figure 3. *The Hierarchy of Models*

Such a schema of decomposition and description, applicable to a single information system and different information systems of different organisations alike, could provide a guideline for harmonization efforts.

The harmonization effort takes place within each level (see Fig. 3). Models at different levels, in fact, have different purposes and their objects[7] are different because the goal of a certain level is to describe the lower one. On the contrary, it makes sense to compare and possibly harmonize models at the same level when their objects are also partly the same.

Harmonization happens to be by far simpler if models are defined by means of the same upper level model, that is, using the same modelling method. For this reason, the ideal situation would consist in having only a single fourth level-model, as suggested by the ISO and the OMG standards, able to model any other kind of third level-methodology.

Likewise, the best situation for statistic modelling would consist in having a unique descriptive statistic methodology, able to model every kind of subject-matter statistical information segment.

In practice, however, the existence of many competing modelling methodologies cannot be always avoided. In fact, a unique methodology may be not the best way to satisfy different needs; furthermore, methodologies can evolve, can be inherited from the past, the power to unify them may lack, especially if they are owned by different units or organisations.

The harmonization between different models in levels 1 and 2 can be very important. The mapping between different models enables to convert a model into another, to exchange their contents (data and definitions), to share parts of the model and to ensure some degree of coherence between them.

[7] What the model describes.

Obviously, harmonization in the context above is significant if models are devoted to describe a part of reality that is, to some extent, common, for example, if the discipline (economics, medicine, physics, …) which statistics refer to is the same.

Models and Languages

Every model in the hierarchy gives rise to a language used for defining and naming its structures, and to the possible operations on them. In defining models, terms can be borrowed from the natural language, but they assume a more specific and formal meaning in the model context. Moreover, the same natural term can be used in different models, assuming different meanings in each one. Therefore, for the proper comprehension of the meaning of the terms, knowledge of the context in which they are used (that is, the natural language, or a more formal model) is necessary. That is to say that the meaning of a term belongs to the model in which the term is defined.

This principle, applied to the hierarchy of models, leads to four levels of languages each one corresponding to a modelling level:

- a generic modelling language for each fourth level model;
- a statistic methodology language for each third level model (in which the terms are derived from the descriptive statistic methodology);
- a subject matter definition language for each definition model (in which the terms are basically derived from the discipline to which the model refers, like economics, medicine, physics, …);
- a subject matter extension language for each extension model (in which the terms are the symbols used in the extensional representation).

Because models in two different levels are oriented to describe different aspects, the meaning of a term possibly used in two models, owing to the different levels, cannot be the same. So, it is good practice to avoid the use of the same term in different levels, because such a situation could generate misunderstandings, as the term would have more than one meaning. The same recommendation is valid also if two models of the same level use the same term with different meanings. When it is impossible or not convenient to use different terms, term use should be accompanied by the indication of the context to which it refers (that is, the model). On the contrary, it is possible and desirable that two different model in the same level may share the same term with the same meaning. In conclusion, the structure of the terminology of an information system appears to be composed by a co-ordinated group of terminologies, strictly corresponding to the structure of the models.

Finally the “four level hierarchy of models” is a model of its own and has its proper language that contains terms with a particular meaning according to the models defined in the hierarchy. An example is the term “level” that, in this context, means “modelling level” but in a third level-model may mean “classification level”.

Levels and Roles

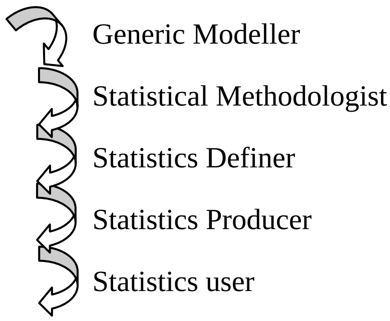
The hierarchy can also be used to distinguish different roles in the information system. Basically, the idea is that a “role” consists in using the model on a certain level in order to produce models in the lower level. Proceeding from up to down (cf. **Fig. 4**):

- the “generic modeller” produces general purpose models (fourth level-models);

- the “statistical methodologist” uses a general purpose model to produce statistical methodologies;
- the “statistics definer” uses a statistical methodology to produce subject matter definitions;
- the “statistics producer” uses a subject matter definition to produce statistics;
- the “statistics user” uses statistics to understand the reality and possibly produce actions on it.

As a parallel with roles defined in [Sundgren, 1991: 11], the “statistic user” appears to be the “client”, while “data definer” and “data processor” can be considered two different kinds of administrator (respectively of the definitions and of the data).

L4	Methodologies that define methodologies
L3	Descriptive statistic methodologies
L2	Statistical information segment definitions
L1	Statistical information segment extensions
L0	Reality




= produces

Figure 4. *Levels and Roles*

The roles defined above are independent of the nature of the “actor” (the role executor) that can be human or software artefacts. Therefore, the same role can be played in principle by people or by machine.[8] In this role-playing, the upper level model supplies specifications to the “actor” that interprets and applies them in order to produce the lower level model. When this behaviour is enforced in practice, the system is “active” because the upper level model drives actor behavior. The schema shows how software artefacts can be made active: they have to be driven by the respective upper level model.

In the approach followed in the Banca d’Italia, the hierarchy of models (or, better, its implementation) is actually used to drive the data processing system. At the moment, all of the major EDP software packages in the statistical system of Banca d’Italia are founded on the idea of “active” models in hierarchy. To process a level, the software is driven by upper level. For example, to produce a level 1-model (for example, a set of statistical data), software is driven by its level 2-model, expressed in a formalized subject matter language and therefore highly independent from the technical aspect of the implementation. This allows subject matter experts to act directly as definers, almost without the necessity of EDP expertise or intervention of EDP experts.

Moreover, to produce a level 2-model, the work of subject matter experts is supported by software tools, driven by the level 3-model in use, with the double goal of helping the definer and of

[8] From an organizational point of view, the responsibility of a certain task is always human, so in every role there must be a human responsible, but here it is the “activity performing” perspective that is being focused rather than the organizational, so it is preferable to define roles independently from the nature of the actor. Another thing to keep in mind is that different roles do not necessarily mean different people because nothing prevents a person from playing more than one role. Nevertheless, it is convenient to keep the distinction of roles in order also to allow different people to play different roles.

enforcing the compliance of the defined level 2-model with the level 3-model. The definition of level 3-models according to the level-4 one is supported in a similar way.

Models and Competencies on Them

The hierarchy of models provides a method to distinguish the competencies between different units, establishing a high-level link between the information system structure and the organisation involved in running and using it. Every model, in fact, can be in charge of a different responsibility (the owner of the model). On the other hand, any number of models, also in different levels, can be in charge of the same responsibility (see Fig. 5). The bi-dimensional schema allows implementing many configurations of competencies, spread between two extreme and ideal situations. The first one is the vertical decomposition, according to the subject of the statistics, in case the whole hierarchy of models (all of the four levels L1 through L4) relevant to a certain subject is left in charge of the same unit. The second one is the horizontal decomposition, according to the roles, when a whole level is left in charge of the same unit. Practically, the tendency to have only one model in the level 3 and 4 drives toward an intermediate situation, like the imaginary one drawn in Fig. 5.

The owner of a model has the duty (and right) to define it (to create and modify the model). To perform such a task, the owner has to use the relevant upper level model, so he must have the right to know that one as well. Whenever someone has the right to know a model, he may need also to know the upper level model used to define it. Defining or understanding a model, therefore, implies the knowledge of the whole upward chain of higher-level models, but does not require the knowledge of models in the levels below. Such criteria help in defining access rights of the various actors.

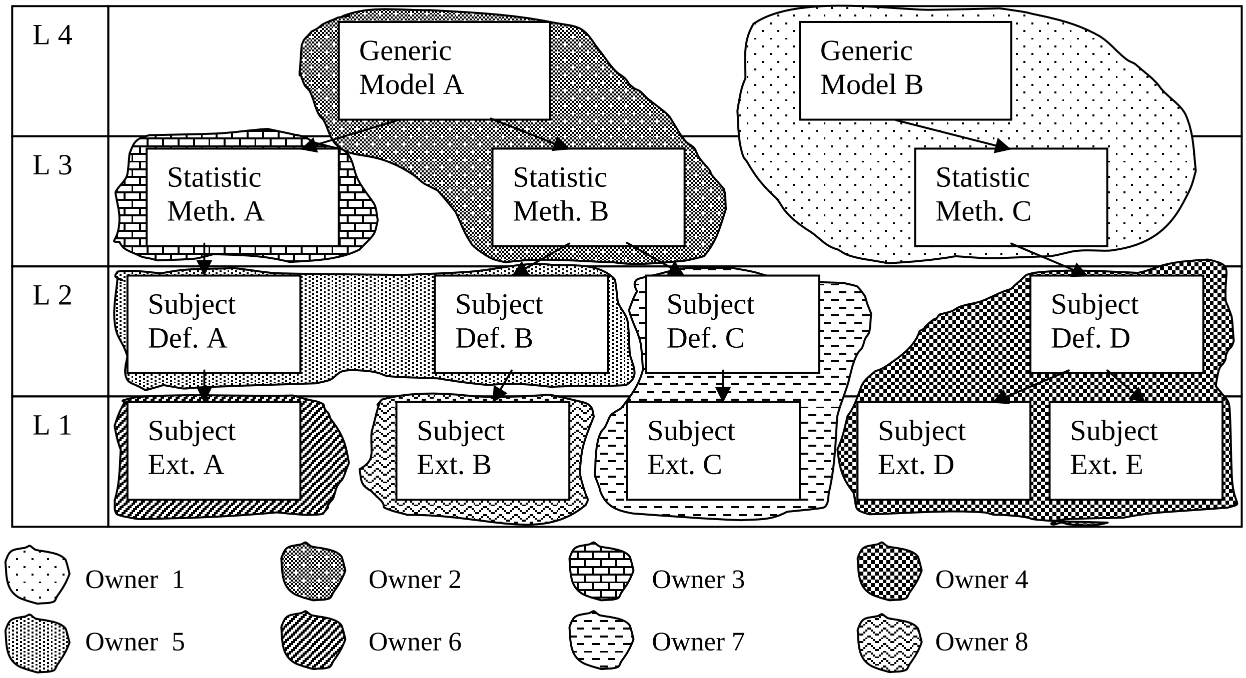


Fig. 5 - *Models and Competencies*

Conclusions

The approach of the multi-layered hierarchy of models to describe information systems meets many objectives at once; in fact it appears to be:

- oriented to obtain “infological completeness”;
- a means to give formal structure to one or more information systems, decomposing them in self-consistent parts;
- a guide for the effort of harmonizing and standardizing, in particular for the definition of terms;
- a method to distinguish roles;
- a suggestion on how to make “active” information systems;
- a first way to define competencies.

As a matter of fact, it appears to offer a synthetic and high-level vision of statistical information systems, connecting different perspectives in which they are usually seen.

References

Del Vecchio V. [1997] La rappresentazione dei dati e dei concetti statistici. Report, Banca d'Italia. English translation [2001] Statistical Data and Concepts Representation.

Grossmann W. [2002] Structures for metadata – Metanet project discussion document. April, 2002.

ISO [1986] ISO/TC97/SC21/WG3 Information Resource Dictionary System (IRDS) – Part 1 – Core – Working Draft, Revision 11, 27/6/86, Number N169.

ISO [1987] ISO/TC97/SC21/WG3 Information Resource Dictionary System (IRDS) – Part 1 – Service Interface – Working Draft, Revision 2, 31/3/87, Number N308.

ISO [1990] ISO/IEC 10027:1990 Information technology – Information Resource Dictionary System (IRDS) Framework.

ISO [1993] ISO/IEC 10728:1993 Information technology – Information Resource Dictionary System (IRDS) Service Interface.

ISO [1998] ISO/IEC 13238-3:1998 Information technology – Data Management – Part 3: IRDS Export/Import Facility.

Limongelli D. [1986/87] Studio e realizzazione di un prototipo di information resource dictionary system per un ambiente di produzione software. Tesi di laurea, Università degli Studi di Bari - Corso di laurea in Scienze dell'Informazione (Anno Accademico 1986/87).

Limongelli D. et al. [1988/89] A Prototype for the Integration of Information Resource Dictionary System and PCTE. In: Computer Standards and Interfaces, Vol.8. North-Holland.

Magnani P. [1985] Progettazione di un Data Base: i dati e i concetti fondamentali. Sistemi e Automazione **256**.

OMG [2000] Meta Object Facility (MOF) Specification. Version 1.3, March, 2000.

Sundgren Bo [1991] Statistical Metainformation and Metainformation Systems. Report, Statistics Sweden.