

Distr.
GENERAL

CES/SEM.47/22
21 February 2002

ENGLISH ONLY

**STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

**COMMISSION OF THE
EUROPEAN COMMUNITIES**

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

**Joint UNECE/Eurostat Seminar on Integrated Statistical
Information Systems and Related Matters (ISIS 2002)**
(17-19 April 2002, Geneva, Switzerland)

Topic III: Object-oriented technologies, component architecture

GESMES/CB – A UML MODEL

Contributed paper

Submitted by the European Central Bank, Germany¹

Keywords

Statistical data exchange, multidimensional data, time series, metadata, EDIFACT, XML, design pattern.

Summary

GESMES/CB is nowadays the most commonly used data format for the exchange of time series data and descriptive metadata. This paper formulates the core data model of GESMES/CB in the Unified Modelling Language (UML).

I. INTRODUCTION

1. GESMES (Generic Statistical Message) is a data format developed for the exchange of multidimensional arrays. It is an international (UN/EDIFACT) standard. See http://www.unece.org/trade/untdid/d01b/trmd/gesmes_c.htm. Additional material and links concerning GESMES can be found via <http://www.gesmes.org>.
2. GESMES/CB is a GESMES profile developed in 1997 for the exchange of time series and related metadata. It supports all needed functionality while it is simpler to process. See <http://www.ecb.int/stats/gesmes/gesmes.htm> for documentation.
3. GESMES/CB has become a very successful electronic data interchange (EDI) format since then. It is the exclusive message format for exchanging statistical data within the European System of Central Banks;

¹ Prepared by B. Bodenstorfer (Bernhard.Bodenstorfer@ecb.int).

the Bank for International Settlements (BIS) and Eurostat encourage its use for all time series data exchanges. Since several software products supporting GESMES/CB are available, including also free software, interest outside Europe has become increasingly important.

4. The use of GESMES/CB in its EDIFACT expression is already very stable and is also expected to be fully supported in the future. In addition to this, the global interest and the emergence of XML demand further standardisation efforts for GESMES/CB. Besides a couple of clarifications and improvements to satisfy new needs of the user community, refactoring the EDIFACT message format into XML is the major present challenge.

5. To aid the ongoing discussion, the successful data model of GESMES/CB has been formulated using the Unified Modelling Language (UML). The set of UML-diagrams will most likely ease the access to GESMES/CB for IT experts who are not familiar with EDIFACT.

6. This paper presents the core of the GESMES/CB data model expressed in UML. It is based on the GESMES/CB User Guide [1], which is the reference document where the GESMES/CB format was first laid out.

II. BASICS

A. Intuitive Model

7. Before the model is presented in UML, this section tries to give a brief introduction to GESMES/CB for the readers less familiar with it. More comprehensive information can be found in [1].

8. GESMES/CB is about multidimensional arrays of time series. A time series is a vector of observations in time order. Each observation carries a value and a status. The observation value represents a real number or indicates a missing value. The observation status specifies the meaning of the value, e.g. whether it is an actual value or a forecast. An observation can also carry more information; for instance, to support breaks in the series or comments. This is largely configurable in GESMES/CB. A time series can be represented in a table, one line per observation.

period	value	status	confidentiality	pre-break value	comment
January 2000	1.17	A	F		
February 2000	1.66	A	F		
March 2000	1.95	E	C		Estimated on 14 Feb 2000

9. Every time series has a key. The keys are multidimensional. Typically, a key would be written like M.AT.CPI.CH.Z, where the values for all dimensions are separated by a dot. The example above could mean that the designated time series is a monthly series about Austria's consumer price index, percent change per year, while "Z" is added for technical reasons in this particular case.

10. Time series with keys that differ only in the frequency value form sibling groups. A sibling group key could be written as *.AT.CPI.CH.Z, but many existing systems simply omit the frequency dimension value.

11. It is possible to attach textual or coded attributes to observations, time series, sibling groups, or the whole collection of data. In the example above, the comment “Estimated on 14 Feb 2000” is such an attachment as further columns would be. Following the traditional EDIFACT expression of GESMES/CB, the observation value, status, confidentiality, and pre-break value are not regarded as attachments, but this fact is not reflected in the UML-model.

12. The structure of the time series keys, particularly the number of dimensions, is defined in a key family. All time series keys governed by one key family have the same dimensions. The key family also specifies which attributes can be attached to which kind of objects. Key families do not change frequently since they define the structure.

13. The meaning of all dimensions in a key, along with other information and attribute fields is expressed by statistical concepts (like frequency, reference area, and phenomenon). Key families assign exactly one statistical concept to each such field. In a relational database table one could imagine one column per statistical concept, though some columns may happen to be only very sparse (e.g. that for pre-break value).

14. Key families may also restrict the set of valid data values. Typical restrictions are those to a text format and to a code list. Code lists are lists of values together with a description for each of them.

B. Notational Remark

15. Hereafter, the names of classes in the presented core data model will be written with first capitals. The classes used in this model are: Characteristic, Characteristic Value, Code, Code List, Cube, Data Family, Dimension, Dimension Value, Frequency, Key, Key Family, Key Family Component, Key Structure, Observation, Observation Key, Period Set, Sibling Group, Sibling Group Key, Statistical Concept, Structural Definition, Text Format, Time, Time Series, Time Series Key, Value Set.

C. Deviations from [1]

16. The model presented here tries to capture GESMES/CB as it is laid out in [1]. Nevertheless, careful modernization has been applied in the following places in order to simplify the model and discussion:

17. The original notion “key family” from [1] is expressed here by two classes: Key Family and Data Family. The class Key Family covers the structural side, whereas the class Data Family expresses the meaning as a data container.

18. The class Attribute has a slightly different meaning here than “attribute” in [1]. Actually, this terminus is used inconsistently there, once to exclude so-called array cells and once to include them.

19. Time is considered a Dimension and part of the Key Structure. This relationship is not justified by [1], where Time is not treated as part of a Key. But this way the identification of Observations can be treated along the same lines as that of Time Series and Sibling Groups.

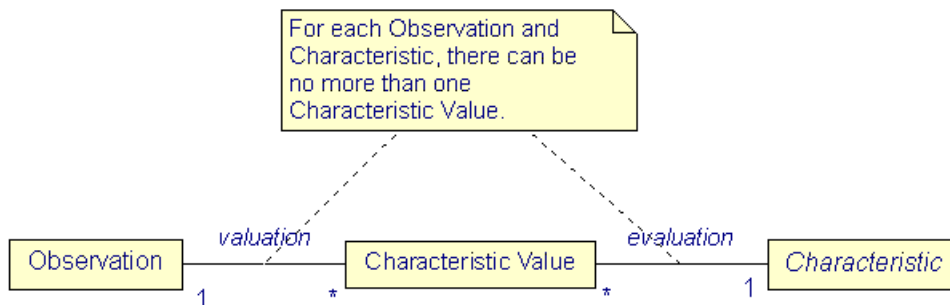
20. Since [1] focuses on the EDIFACT syntax representation, time and time format are presented separately there. Here, they are viewed as a single Dimension, which completely describes a time period.

III. STORING DATA

A. Observations and Characteristic Values

21. It could be tempting to define observation value, status and other information about observations as attributes in UML. However, GESMES/CB allows key families to freely define such fields. To support this flexibility, GESMES/CB uses the Measurement pattern. For a thorough and more comprehensive treatment of the Measurement pattern see [2], Chapter 3, but be aware of the differences in the class naming.

22. Observations are the elementary objects of care in GESMES/CB. Attached to an Observation are the mandatory Characteristic Values for the Characteristics “observation value”, “observation status”, and others. Attachment of values to Observations is expressed by the Measurement pattern:



23. Characteristics describe the phenomenon type. They are part of the knowledge level information, whereas Observations are part of the less fixed operational level. The multiplicities in the above diagram indicate that one Observation in general carries several Characteristic Values, where each of them corresponds to one Characteristic.

24. The Measurement pattern is the central pattern in GESMES/CB. Around it, GESMES/CB arranges additional structure; on the side of the Observation it introduces aggregations and identification by Keys, while on the side of Characteristic it adds semantic description and restrictions for Characteristic Values.

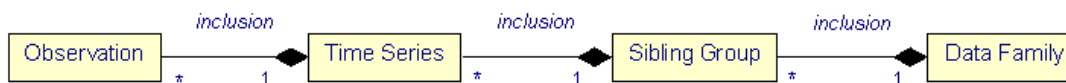
B. Observations and their Aggregates

25. Observations are not the only objects of care in GESMES/CB. Their aggregations (Time Series, Sibling Group, and Data Family) make up for the rest of those.

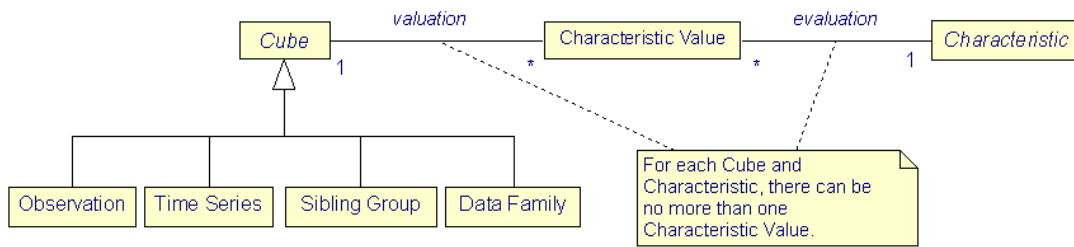
26. A Time Series is a vector of Observations. All of those must have the same period length, e.g. one year, one month, or one day. This common length determines the frequency of the Time Series.

27. A Sibling Group is an aggregate of Time Series.

28. A Data Family is an aggregate of Sibling Groups.



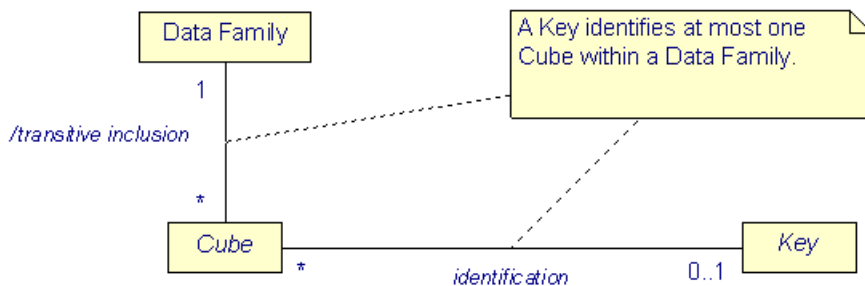
29. Characteristic Values can be attached not only to Observations, but also to their aggregates. This common behaviour is factored into the abstract parent class Cube. Observations and all kinds of aggregates are Cubes. The general Measurement pattern for GESMES/CB is depicted in the diagram below:



IV. IDENTIFICATION AND KEYS

A. Namespace Pattern

30. To identify objects from within a collection, GESMES/CB uses the Namespace pattern. The idea is that an identifier only together with a namespace specifies an object. If the identifier varies, then the object varies, and if the namespace varies, then the object varies, too. Hence, the same identifier can be used to denote different objects in different namespaces. Not just by chance the pattern looks similar to the Measurement pattern, where also two entities together determine a third one. For Cubes the namespaces are Data Families, and the identifiers are Keys.



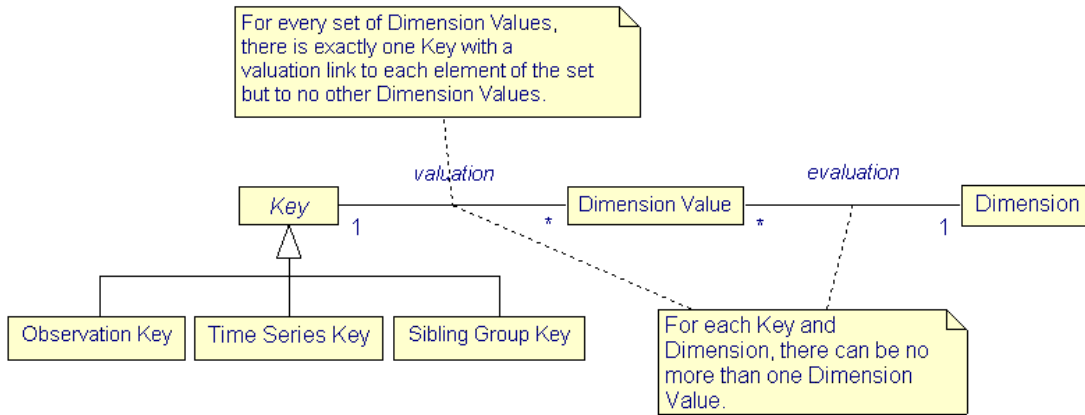
31. The association “transitive inclusion” is derived from the association “inclusion” in order to talk about Time Series and Observations within a Data Family in a natural way. A Sibling Group is transitively included in a Data Family if and only if it is included in it. A Time Series is transitively included in a Data Family if and only if it is included in a Sibling Group, which is in turn included in the Data Family. An Observation is transitively included in a Data Family if it is included in a Time Series that is in turn transitively included in that Data Family.

32. The multiplicity 0..1 for the Key in the above diagram is due to the fact that a Data Family, too, is a Cube. However, it needs no Key to be identified within itself. The multiplicity could be changed to 1 by a convention involving the empty Key for the Data Family itself.

33. The Namespace pattern is also used to identify Code Lists, Statistical Concepts, and Key Families within a set of Structural Definitions, in which case the identifier is a simple string.

B. Multidimensional Keys

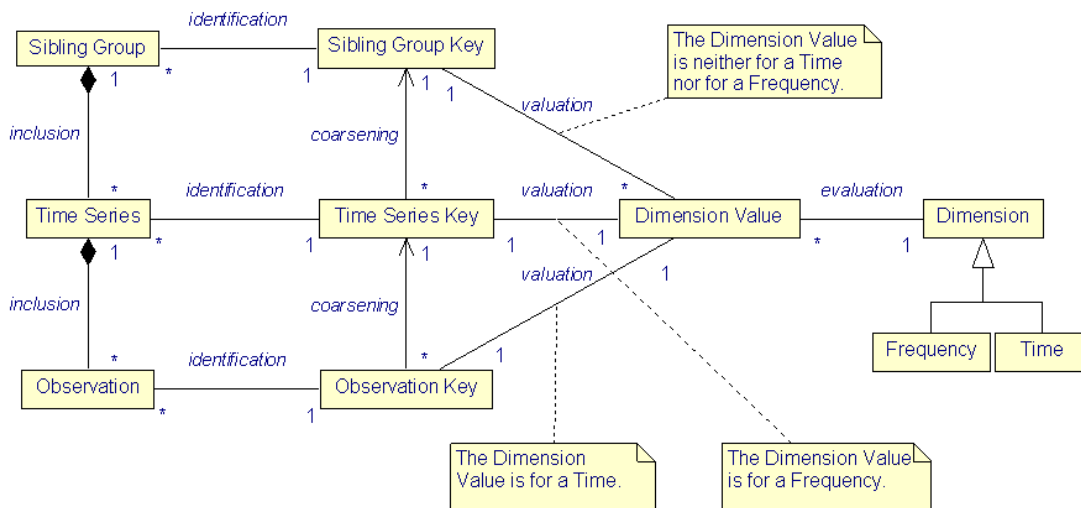
34. GESMES/CB uses Keys to identify Cubes. Keys carry values for various Dimensions. The pattern to express this relationship reminds of the Measurement pattern, once again. In fact, this pattern makes it possible that a Key Family can freely define the Dimensions available to form Keys.



35. The additional uniqueness constraint on the association “valuation” means that Keys are completely determined by the Dimension Values they carry. This is necessary to allow unique identification of a Time Series, or generally a Cube, within a Data Family by giving all relevant Dimension Values.

36. A Time Series Key carries no time period information, because Time Series aggregate Observations for various periods. Similarly, a Sibling Group Key carries no frequency information because Sibling Groups aggregate Time Series of different frequencies. Of course, it carries no time period information either.

37. The subclasses of the abstract class Key are related to each other in parallel to the inclusion relationships between the identified Cubes. Thus, the Observation Key adds only time period information to a Time Series Key, and the Time Series Key adds only frequency information to a Sibling Group Key.



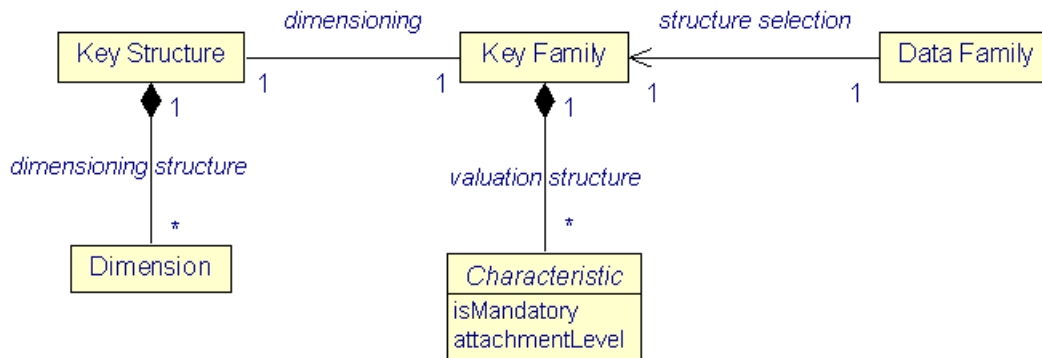
38. If a Sibling Group includes a Time Series, then its Key must be in a coarsening relationship with the Time Series' Key. This essentially guarantees that all Dimension Values but that for Frequency are fixed throughout a Sibling Group. An analogous rule holds for Time Series and Observations and their Keys with the effect that only the Dimension Value for Time can vary for the Observations within a Time Series.

39. As the above diagram indicates, the association “identification” has the constraint that only a Sibling Group Key can identify a Sibling Group, only a Time Series Key can identify a Time Series, and only an Observation Key can identify an Observation.

V. DEFINING THE OVERALL STRUCTURE

A. Dimensions and Characteristics

40. A Key Family defines the structure of a Data Family. It defines all Dimensions and Characteristics that can be used in the Keys or for the Cubes in the context of that Data Family.



41. The attribute “attachmentLevel” of a Characteristic defines whether Characteristic Values for the Characteristic can link to Observations, Time Series, Sibling Groups, or the whole Data Family. For example, the observation status may only attach to an Observation, whereas a title may only attach to a Sibling Group, and a compilation remark only to the whole Data Family.

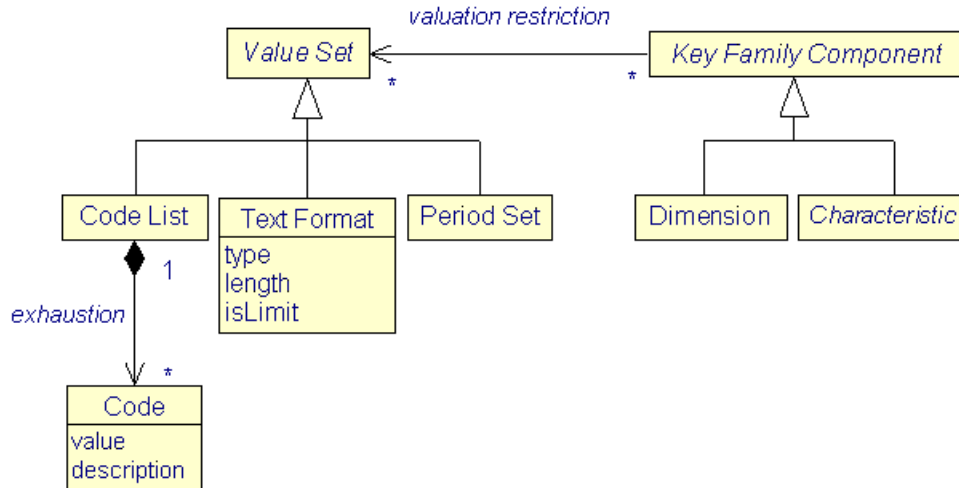
42. The boolean attribute “isMandatory” defines whether there must be a Characteristic Value for this Characteristic for every suitable Cube throughout the Data Family. Here “suitable” means that the type of Cube fits the attachment level of the Characteristic. For instance, observation value and status are mandatory. All Observations must have Characteristic Values for these, while this is not the case for pre-break value.

43. At the moment, there is a discussion within the GESMES/CB user community whether multiple Data Families can select the same Key Family to specify their structure. This is not yet allowed; however, in future the multiplicity of the association “structure selection” may be changed to “many” on the side of Data Family.

44. There must be exactly one Time and one Frequency in the Key Structure of every Key Family. This requirement may also be relaxed in the future in order to support other kinds of data than time series, which are based on periodic observations.

B. Restricting Values

45. Dimension Values and Characteristic Values are linked to Dimensions or Characteristics respectively. That is, they are linked to Key Family Components via the association “evaluation”. Key Family Components can define a set of valid values for them, and this is modelled as a valuation restriction to a Value Set.



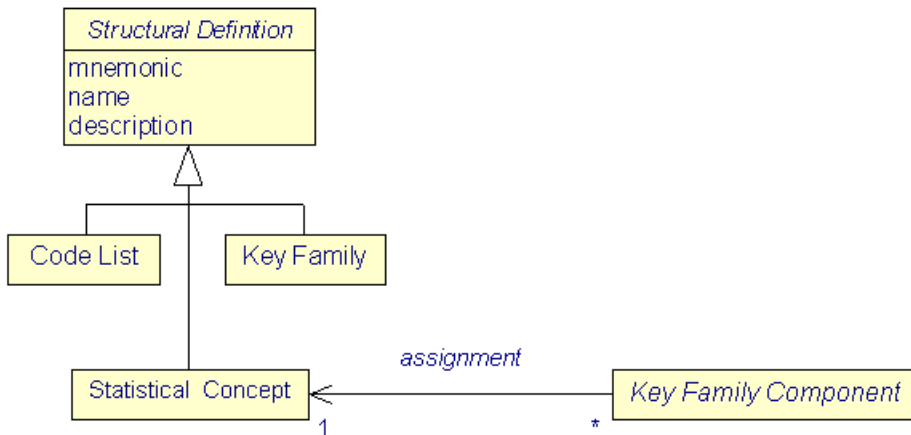
46. If a Key Family Component is a Dimension other than Time, then there must be a valuation restriction to a Code List. If a Key Family Component is a Time, then there must be a valuation restriction to a Period Set. There is a discussion whether to drop the first requirement.

47. A Text Format restricts the length and permitted characters in the textual representation of a value. The length restriction can either be to match the prescribed length or to be less than or equal.

48. The current GESMES/CB in EDIFACT format allows only three combinations of valuation restrictions: Valuation restriction to a Text Format alone; to a Text Format and a Code List; or to a Text Format and a Period Set. The last alternative is chosen if and only if the Key Family Component is a Time.

C. Adding Semantics

49. Key Family Components represent phenomenon types. The meaning of such a phenomenon type is documented by a Statistical Concept, which in turn carries textual information for the human reader.



50. In the context of a single Key Family and its Key Structure, no Statistical Concept may be assigned more than once. However, sharing Statistical Concepts between Key Families is allowed and is considered good practice. For instance, many Key Families may use the same Statistical Concept “reference area” for a Dimension.

D. Division into Structure and Data

51. The core data model of GESMES/CB shows a clear separation between information on the rather stable knowledge level and that on the frequently changing operational level. This division parallels that in [2], Chapter 3.

52. In this model, the classes on the operational level are Characteristic Value, Cube, Dimension Value, Key, and their subclasses.

53. The classes on the knowledge level are Code, Key Family Component, Key Structure, Structural Definition, Value Set, and their subclasses.

E. Possible Simplifications

54. It is not a secret that the presented model allows several simplifications. While these should be considered when deriving an implementation model, they make the analysis model less explicit and therefore convey less information about GESMES/CB. This section nevertheless gives a few ideas for such simplifications.

55. The subclasses of Key could be collapsed into the class Key, and similarly the subclasses of Cube into the class Cube. In this instance, the classes Cube and Key would no longer be abstract.

56. The classes Dimension Value and Characteristic Value could be replaced by a single class Value. The association “evaluation” would then connect one Key Family Component to possibly many Values. Navigation via this association would still allow distinguishing between both original types of Values.

57. Since a Key is completely determined by its Dimension Values, this class and its subclasses are technically not necessary. In this case the Dimension Values would be directly linked to Cubes. However, it may be a bit more difficult to formulate the model constraints, since the class Key facilitates talking about the model constituents.

58. The association “dimensioning” between Key Family and Key Family Structure is one to one. This suggests that these classes could be collapsed into one. In this fashion, even the associations “dimensioning structure” and “valuation structure” could be merged, as long as a distinction can still be made between Dimensions and Characteristics.

VI. BEYOND THE CORE MODEL

A. Additional Classes and Constraints

59. This paper is concerned with the core data model. In particular, it does not describe the available messages, e.g. those specified in [1] in EDIFACT format. Messages represent transactions on data structured according to the GESMES/CB data model. While they were not part of this modelling exercise, they may become subject to future efforts.

60. In focusing on the core model, a number of details and the model periphery have been omitted. This section briefly sheds some light on a few of these omissions.

61. The Data Exchange Context is the uppermost point of reference for successful data exchange. It captures the additional knowledge necessary to successfully exchange and understand data. The class Data Exchange Context encapsulates the outer space of the model; this can be interpreted as an application of the Facade pattern, see [3]. In reality, a Data Exchange Context could be a mixture of human knowledge, legal texts, system hardware and software, along with the environment.
62. Data Exchange Parties have a relationship to the exchanged data and to the Data Exchange Contexts they operate with. Particularly, a Data Exchange Context contains facilities to identify Data Exchange Parties.
63. A Data Exchange Context must also make available the Structural Definitions which are necessary to validate and interpret data. These are published by Maintenance Agencies, which are also identifiable using a Data Exchange Context.
64. Time and Frequency have a close relationship concerning the possible Dimension Values for these Dimensions within a single Key. This relationship ensures that, for instance, a Time Series Key for a Time Series with daily Observations carries the correct Dimension Value for Frequency.
65. The Characteristics observation value and the pre-break value are particularly important since they represent the primary values of analytical interest. To make a purely semantic difference between these essential Characteristics and others like the observation comment, the abstract class Characteristic has a subclass Measure and another, called Attribute. The only Measures in a Key Family are observation value and pre-break value. All further Characteristics are Attributes.
66. A number of constraints are defined in [1] which have been omitted here for brevity. An example of this is the maximum length for Code values in Code Lists.
67. Other constraints are not made explicit in [1], but are obviously in place due to the limitations of the GESMES/CB-EDIFACT message. For instance, it is not possible to mix Structural Definitions coming from different Maintenance Agencies within one Key Family.
68. In [1], the Dimensions in a Key Structure and the Dimension Values in a Key are ordered. This is necessary for GESMES/CB-EDIFACT, but probably not for other syntaxes like XML. Hence, ordering is not required in the core model.

B. Restriction Environments

69. In most cases, additional conventions may be enforced in a Data Exchange Context. Of course, these restrictions must be respected then by all Data Exchange Parties. Often such restrictions will be the result of technical limitations, such as a maximum number of Key Family Components per Key Family.
70. Another typical restriction would be that many systems expect that “FREQ” is the mnemonic of the Statistical Concept assigned to the Frequency in a Key Family. Moreover, fixed Code values for the associated Code List could be assumed.
71. Also an agreement to use further classes and associations can be part of a restriction environment. Equally well, additional features of the core classes may be prescribed.

VII. CONCLUSION

A. Future Prospects

72. The GESMES/CB data exchange format is not only used by central banks, but has also been successfully applied by national statistical institutions, for example. To reflect this broader usefulness, it is envisaged to change the letters CB (reminding of “central bank”) to TS (for “time series”) in due time, giving GESMES/TS.

B. Acknowledgements

73. The presented UML-representation of GESMES/CB was developed in close collaboration between relevant divisions from Banca d’Italia, BIS, ECB, and Eurostat. The author wants to express his gratitude to all participants in the UML modelling exercise.

References

- [1] BIS Data Bank Services, ECB Statistical Information Systems: “GESMES/CB User Guide”; Release 2.00, 2000.
- [2] Martin Fowler: “Analysis Patterns”; 10th printing, Addison Wesley, 2001.
- [3] Erich Gamma et al.: “Design Patterns”; 20th printing, Addison Wesley, 2000.