1_Introduction

1.1 Background

3. In June 1999, a meeting on terminology was held in Neuchâtel, Switzerland, with participants from the statistical offices of Denmark, Norway, Sweden an Switzerland and the software developers in Run Software-Werkstatt. This was the start of the "Neuchâtel group". The aim of the group was to clarify some basic concepts and to arrive at a common terminology for classifications. The terminology defined the key concepts that were relevant for how to structure classification metadata and provided the conceptual framework for the development of a classification database. The work listed and described the typical object types of a classification database, and the attributes connected with each object type.

4. The development of the model had a practical focus as all of the participating National Statistical Organisations (NSOs) planned to use it in their own imp lementation of a classification database. The most important purposes for developing a classification database were:

1. to make accessibility and maintenance of classifications easier, and
2. to ensure common use of classifications across different fields of statistics.

A central database was the preferred solution because it realised one of the important principles of metadata - document and update once (centrally), and r euse wherever it is relevant. The Neuchâtel terminology model: Classification database object types and their attributes (version 2.0) was released in 2002.

5. Later, Statistics Netherlands joined the Neuchâtel group, and a new version of the terminology, version 2.1 (with one new object and one new attribute), was released in 2004.

6. It was essential for the Neuchâtel group that the terminology should be flexible and independent of IT software and platforms. This resulted in different cl assification database implementations for the participating NSOs, according to specific needs and policies. Also, it was always an important premise for the group that the work should be public and available to anyone free of charge.

7. Many countries have at least partially implemented the model

[1]

. After years of practical experience, several of the implementing countries expressed a desire to see some revisions to the model. As the Neuchâtel group no longer existed, a possible revision was discussed at the 2011 METIS Workshop [2] . Subsequent to the workshop, the METIS Steering Group contacted the UN Expert Group on International Statistical Classifications to work on the revision of the Neuchâtel model. As a result, a joint working group was created, bringing together classification and statistical metadata experts.

8. At the same time, a project sponsored by the High Level Group for the Modernization of Statistical Production and Services was reviewing the Generic Statistical Information Model (GSIM)

[3]

. GSIM provides the information object framework supporting all statistical production processes such as those described in the Generic Statistical Business Process Model (GSBPM)

[4]

, giving the information objects agreed names, defining them, specifying their essential properties, and indicating their relationships with other information objects. In the development of GSIM, the objects related to classifications were mostly drawn from the Neuchâtel Terminology Model.

9. During the revision work it was discussed and decided that for the future the Neuchâtel model for classifications will be part of GSIM. Several objects and attributes have been changed during the revision process, and the revised model will in practice be an annex to GSIM.

1.2 Context and scope

10. Classifications are generally regarded as a special kind of metadata for statistics. They are definitional, content-oriented metadata, ordering and describ ing the meaning of statistical data. A classification database can be described as a register of meta-objects (classifications and related object types), which in turn have their own set of metadata. It forms a more or less technically integrated part of the overall metadata information system of a statistical office.

11. The GSIM Statistical Classifications Model orders the concepts in a two-level structure of object types and attributes. On the first level, it specifies the b asic object types of a classification database (e.g. Classification Family, Classification Series, Statistical Classification, Correspondence Table, Classification Index) and, on the second level, it lists the attributes connected with each object type. It is both a terminology and a conceptual model. It prov ides the conceptual framework for the development of a classification database. This immediate practical purpose has obviously limited its scope. It is not c oncerned with recording all the terms used in this area, nor does it deal with methods or best practices in the development and management of classificatio ns. What it does do, is define the key concepts that are relevant for how to structure classification metadata and, indirectly, how to present information on cl assifications to different kinds of users. Since the GSIM Statistical Classifications Model belongs to the semantic and conceptual sphere of metadata it does not include object types and attributes that are related solely to the technical aspects of a classification database.

1.3 Classification and related concepts

12. According to ISO 704: 1987 (E) Principles and methods of terminology, a term is a word or phrase, which designates a concept. This section sets out s ome central concepts related to classifications and the relationships between them and the terms that will be used to refer to these concepts.

13. In the field of statistics, the term **classification** is normally used to denote one of the following concepts:

1. The general idea of assigning statistical units to categories representing the values of a certain variable.
2. The general concept of a structured list of mutually exclusive categories, each of which describes a possible value of the classification variable. Su ch a structured list may be linear or hierarchically structured. A linear classification is a list of categories, which are all at one and the same level (e .g. the ISO 3166 country code list, or a classification of marital status). In a hierarchical classification the categories are arranged in a tree-structure with two or more levels, where each level contains a set of mutually exclusive categories. The items of each level but the highest (most a ggregated) are aggregated to the nearest higher level. In common usage the term classification often implies a hierarchical classification.
3. One particular structured list of mutually exclusive categories, which is named, has a certain stability and normative status, and is valid for a given period of time (e.g. ISIC Rev.1).
4. One particular named set of several structured lists of mutually exclusive categories, which are consecutive over time and describe the possible va lues of the same variable (e.g. ISIC).

14. The distinction between concepts c. and d. is seldom made explicit. Here as well, the term rather implies a hierarchical classification, and especially one of the group of "large", traditional, well-established and normative standard classifications.

15. **Nomenclature** is a term, which is closely related to classification. Nomenclature has to do with naming. Basically it denotes a list of named entities. Ad ding system and structure to the list turns it into something that resembles a classification. Although they do not have exactly the same meaning, the terms classification and nomenclature are often regarded as synonyms and used interchangeably. Nomenclature is not a term used in GSIM terminology.

1.3.1 Classification Series and Statistical Classifications

16. The conceptual framework of the classification database includes an object type roughly equivalent to concept d above. In the GSIM Statistical Classifications Model, this concept has been named **Classification Series.** The concept of each "structured list of mutually exclusive categories" has been named **Statistical** *Classification.

5

17. A Statistical Classification is a set of categories (Classification Items) which may be assigned to one or more variables registered in statistical surveys or administrative files, and used in the production and dissemination of statistics. The Classification Items are defined with reference to one or more characteristics of a particular population of units of observation. A Statistical Classification may have a flat, linear structure or may be hierarchically structured, such that all Classification Items at lower Levels are sub-categories of a Classification Item at the next Level up. The Classification Items at each Level of the classification structure must be mutually exclusive and jointly exhaustive of all objects in the population of interest.

1.3.2 Some attributes related to statistical classifications

18. Statistical Classifications vary in their relationship to other Statistical Classifications. The following paragraphs discuss the terms that relate to such vari ation.

Classification version

19. A Statistical Classification is a version if it has a certain normative status and is valid from a particular date for a period that may or may not be specified A new version is created when such a Statistical Classification is superseded by the introduction of a new Statistical Classification that differs in essential w ays from the previous version. Essential changes are changes that alter the borders between categories, i.e. a statistical object/unit may belong to different categories in the new and old versions. Border changes may be caused by creating or deleting categories, or moving part of a category to another. These c hanges can occur at any Level of the classification. The addition of case law, changes in explanatory notes or in the names of Classification Items do not le ad to a new version.

20. It should be noted that if a Statistical Classification is superseded by a new version, the two versions will likely serve the same objective or purpose.

21. Statistical Classifications that are related to each other as versions belong to the same Classification Series.

Classification variant

22. A particular Statistical Classification may not meet all the needs of its users. If it is for dissemination or other uses, the classification structure may be ill suited for the purpose at hand (for example, Levels or categories are too general or too narrow, too detailed in one area, and too broad in another). To meet these needs, a number of alternatives may be created, in which the original categories are split or regrouped to provide context-specific additions or alterna tives to the standard aggregation structure. These are called classification variants.

23. A Statistical Classification is a variant of another Statistical Classification if it is built from the Classification Items of that base Statistical Classification. These Classification Items do not need to be all at the same Level in the base statistical classification. To these Classification Items, one or more new Levels may be added. This can include extending the base Statistical Classification with one or several new Levels at the bottom of its base, creating a new lowest Level. It should be noted that variants are typically developed to serve a specific purpose.

24. Variants are commonly of three kinds. These have been named **extension variants, aggregate variants** or **regrouping variants**. There could exist other types of variants. A particular variant could include elements from more than one of these variant types.

25. **Extension variant:** An extension variant is a Statistical Classification that extends the base Statistical Classification with one or several new Levels at the bottom, creating a new lowest Level. An extension variant thus adds new lower Levels to the base Statistical Classification but does not otherwise alter its original structure.

26. **Aggregate variant:** An aggregate variant is a Statistical Classification that groups the categories of a linear Statistical Classification to create one or several aggregate level(s), thus creating a hierarchy.

27. **Regrouping variant:** A regrouping variant is a Statistical Classification that introduces additional or alternative aggregate levels by regrouping categories of the base statistical classification. Two types of regrouping variants have been identified:

1. Regrouping variants which do not violate the structure of the base Statistical Classification: This type of regrouping variant introduces a new level or new levels on top of, or in between existing Levels of a hierarchical Statistical Classification without otherwise altering the original structure of the hierarchy. This regrouping variant consists of all classification Levels of the base Statistical Classification plus the new variant Level(s). The parent Level (if any) of the new variant Level can be either another variant Level or a Level from the base Statistical Classification.
2. Regrouping variants which violate the structure of the base Statistical Classification: This type of regrouping variant introduces a new Level or new Levels on top of any but the topmost Level of a hierarchical Statistical Classification by regrouping categories of the base Statistical Classification in a way which violates its original order and structure. This regrouping variant consists of all classification Levels of the base Statistical Classification below the new variant Level(s) plus the new variant Level(s). In such a regrouping variant, a new variant Level cannot have a base Statistical Classification Level as parent Level.

28. In all variants except regrouping variants which violate the structure of the base Statistical Classification, all Levels of the base Statistical Classification are retained and one or more new Levels are inserted. In regrouping variants which violate the structure of the base Statistical Classification, one or more new Levels are inserted and only the base Statistical Classification Levels below the new variant Levels are retained.

29. It is sometimes debated whether a classification database should be descriptive or prescriptive, the idea being that a prescriptive database will contain only standard classifications, whereas a descriptive database will also contain non-standard variants. In reality, the demarcation between standard and non-standard classifications or between these and more loosely structured groupings is not very clear. It seems, therefore, that the criterion for inclusion in the database cannot be formal status only, but just as much the usefulness and commonality of the information provided. Most of the time the departures from the norm are legitimate, made to meet specific producer requirements or user needs. In any case alternative groupings exist and have to be documented. Indeed, listing the non-standard variants used in a statistical office may be a first and necessary step towards reducing their numbers.

Classification update

30. A Statistical Classification is an update of another Statistical Classification if it supersedes that Statistical Classification but does not differ from it in essential ways. Updates to specific elements of a Statistical Classification may be permissible within the life of a version. They may simply be noted in the context of the element affected or, if the changes are sufficiently numerous or significant, a new Statistical Classification can be issued that supersedes the previous Statistical Classification.

Floating classification

31. A Statistical Classification is said to be floating if it permits updates and essential changes without requiring their recognition through the issuing of a new Statistical Classification. Such Statistical Classifications may be used, for example, in contexts where change in the variable is expected to occur, but irregularly, and such change must be incorporated into the Statistical Classification in a timely fashion. Dates of validity on all elements of these Statistical Classifications allow the reconstruction of the Statistical Classification as it was on any particular date.

1.4 Other terminologies

32. There exist a number of terminologies and glossaries dealing with classification terms. These are either concerned with metadata in general or more specifically focused on classifications. The UN Glossary of Classification Terms is a multi-purpose general glossary of concepts, which also contains information on actual classifications and best practices in the development of classifications. It is much broader in scope than the GSIM terminology.

33. The draft Glossary of Statistical Terms attached to the joint OECD and Eurostat SDMX paper Developing a Common Understanding of Standard Metad ata Components draws heavily on the UN glossary for its classification related terms.

34. There is also the UNECE "METIS" Terminology on Statistical Metadata. This has the term classification scheme instead of Classification Series but the concept is the same.

35. Concepts and terms related to classifications are also found in more general papers, for example, Best Practice Guidelines for Developing International Statistical Classifications, a paper developed by the UN Expert Group on International Statistical Classifications. This paper describes best practices for the development, use, maintenance and revision of classifications and there is close alignment with the GSIM Statistical Classification Model. The usage and scope of the best practice document are, however, different from those of the GSIM Statistical Classification Model.

36. The GSIM Statistical Classification Model terminology should be regarded as a complement rather than a rival to other terminologies in the field. Natura lly there is a certain overlap of terms with the glossaries and papers mentioned above. In most cases there is a general agreement between the concepts a nd the terms used, although the wording of the definitions may vary. Not surprisingly, the one instance in which the terminology is at variance with other ter minologies is in using the term "Statistical Classification" for one particular and well defined concept, and for making a clear distinction between "Classificati on Series" and "Statistical Classification" as explained in previous paragraphs. This and a few other instances of inconsistency are due to the particular foc us and purpose of the GSIM terminology, which calls for quite specific and narrowly, defined concepts.

1.5 Implementation

37. Although the original Neuchâtel terminology was initially developed in the context of the classification database application of Bridge^{NA}, both the termin ology and the conceptual model are generally applicable and not dependent on IT software and platforms. The conceptual model can be used in any conte xt where structured information on classifications is needed.

38. In the context of the Bridge^{NA} system the conceptual framework has been used to develop a general semantic interface for metadata (ComeIn). It has a lso served as a specification for a concept-guided and user-oriented dialogue application, which functions as a browser and a tool for the management of cl assifications. This application is used in Statistics Sweden with the aim of setting up, developing and managing their national classification database.

39. In 2013, a questionnaire investigated the use of standards relevant to classifications and the need for a revised Neuchâtel Model for Classifications. Responses were received from eighteen countries or international organisations: Australia, Austria, Canada, Croatia, Estonia, France, Germany, Ireland, the Netherlands, New Zealand, Norway, Portugal, Slovenia, Sweden, Switzerland, the United States, Eurostat and the ILO.

40.Table 1 contains a summary of the results regarding the use of the Neuchâtel Model for Classifications.

Table 1: Summary of results

Standards and models	**Considering Use**	**Currently in use**
Neuchâtel Model for Classifications	4	11
Neuchâtel terminology		
Classification family	3	13
Classification	2	14
Classification version	2	14
Classification variant	2	9
Classification index	3	10
Correspondence table	2	14
Classification level	2	14
Classification item	2	14
Item change	1	7
Case law	3	4
Classification index entry	3	10
Correspondence item	2	14

1.6 Layout of the GSIM Statistical Classification Model

41. Section 2 gives an overview of the GSIM Statistical Classification Model object types, including a short description. The list is ordered according to an obvious and simple logic.

42. A simplified object graph in Section 3 gives an overview of the main object types and relationships in the conceptual model.

43. Section 4 contains the list of all object types and their attributes. The object types are listed in the same order as in the overview. Each object type is defined by a textual description, followed by a list of the attributes associated with the object type. Each attribute is also described. There has been an attempt also to order the attributes according to some sort of logic and to list them in a consistent way across the object types. Attributes or terms used in the descriptions which are underlined, refer to an object type listed and described elsewhere in the model. While object type terms are unique, the name of an attribute may differ in meaning when the attribute is associated with different object types. Some of the central object types of the model, e.g. Statistical Classification, Classification Item, have quite a number of attributes attached to them. For certain applications some of the attributes will be superfluous. They need not all be used. Time has not allowed a thorough review of the descriptions. We are aware that they are not consistently of one kind, but waver between subject matter oriented and IT oriented language, sometimes genuine definitions, sometimes indicating how the information will appear in the technical application. In spite of good intentions, it has been difficult to keep the conceptual and the implementation levels separate.

44. A worked example for all object types and most attributes, based mainly on the Standard Industrial Classification (SIC 2007), has been added in Appendix 1 to facilitate understanding. In Appendix 2, a checklist of possible content for the introduction to a classification version can be found.

1.7 References

1. ISO 704: 2000. *Terminology Work - Principles and Methods.*
2. *Terminology on Statistical Metadata.* Prepared by the UNECE Work Session on Statistical Metadata (METIS). Conference of European Statisticians, Statistical Standards and Studies No 53. Geneva 2000.
3. *UN Glossary of Classification Terms.* Working document. United Nations. http://unstats.un.org/unsd/class/family/glossary_short
4. Ward, D. and Pellegrino, M. *Developing a Common Understanding of Standard Metadata Components: A Statistical Glossary (draft).* Joint OECD and Eurostat paper for the Workshop on Statistical Data and Metadata Exchange, Washington , D.C., September 2001.
5. *Generic Statistical Information Model*, UNECE, http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model

---

1. Countries that have implemented the model include Austria, Belgium, Bulgaria, Canada, Croatia, Czech Republic, Denmark, Estonia, Germany, Greece, Ireland, Norway, Portugal, Slovak Republic, Slovenia, Sweden, Switzerland and the Netherlands
2. “METIS” was the joint UNECE / Eurostat / OECD group on statistical metadata
3. See: Generic Statistical Information Model
4. See: http://www.unece.org/stats/gsbpm
5. Classification Series corresponds to Classification in the Neuchâtel terminology model. Statistical Classification includes Classification Version and Classification Variant from the Neuchâtel model.