Distr. GENERAL
03 May 2013
WP.18
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata
(Geneva, Switzerland, 6-8 May 2013)
Topic (ii): Case studies and tools

CLASSIFICATION MANAGEMENT SYSTEM DEVELOPMENT FOR STATISTICS NEW ZEALAND

Working Paper

Prepared by Andrew Hancock, Statistics New Zealand and Arofan Gregory, Metadata Technology North America[12]

I. Introduction

1. Statistics New Zealand (Statistics NZ) is currently undertaking an organisation-wide ten year programme of change (Statistics 2020 Te Kāpehu Whetū) to implement Statistics NZ's Strategic Plan 2010-2020.

2. The Transformation Programme that underpins Statistics 2020 is about addressing risks, realising opportunities, and creating a dynamic, responsive, and sustainable organisation. It involves changing the way the organisation works; who it works with and how; changes to the systems, tools, and processes it uses; and to the skills, attitudes, and behaviours needed for success.

3. A significant part of the programme is to mitigate legacy computer systems and bring about efficiencies to systems and processes by transforming the way statistics are delivered. This has provided the opportunity for Statistics NZ to rethink the way it develops, maintains and implements classifications and standards for official statistics. As a result the Classifications and Related Systems (CARS) that is currently the corporate repository for all classifications, concordances and coding indexes is to be replaced.

[1] **Liability statement:** Statistics New Zealand gives no warranty that the information or data supplied in this paper is error free. All care and diligence has been used, however, in processing, analysing and extracting information. Statistics New Zealand will not be liable for any loss or damage suffered by customers consequent upon the use directly, or indirectly, of the information in this paper.

II. Background/Brief History of CARS

4. As part of the move from a mainframe computer network to a PC based network in 1996, Statistics NZ introduced the Classifications and Related Systems (CARS). The role of CARS was to be the corporate repository for all classifications, concordances and coding indexes used in the production of official statistics. The system manages definitions and versions of classifications, concordances (translations between versions) and codefiles (lists of probable survey responses and classification categories to which they are coded) used for coding survey responses.

5. CARS is Statistics NZ's main classification management tool, although not all classification management occurs within CARS. Over the years, issues with the limitations of CARS have been addressed through the development of work-around systems that compensate for some gaps in CARS functionality. CARS is now coming to the end of its shelf-life as it operates on a Centura/Sybase platform, of which both Centura and Sybase are legacy systems, and so needs replacement. This provides an opportunity to address the limitations of the original CARS, and improve the manner in which other systems connect to CARS.

A. Objectives of CARS

(a) To centrally store all economic, social and geographic classifications data used by Statistics NZ. The data stored includes standard classifications and survey specific classifications which are not standard. Historical classifications are also stored when they are required for the analysis of historical data (including time series).

(b) To provide common ways to update and access classifications data.

(c) To facilitate the use of standard classifications in all Statistics NZ statistical data by having all standard classifications data stored in one central place and readily accessible to all.

B. Benefits of having CARS

(a) Common storage facilities will make it easier to find and access classifications data, and this will benefit the survey development process.

(b) Common ways to store, update and use classifications data will both help reduce the time and resource required to operate surveys, and also eliminate the need to provide separate storage and updating facilities in each survey system.

(c) All concordances being used are accessible thus reducing time-consuming work in producing data for publication sometimes involves finding out whether a concordance exists or manually constructing one because that concordance can't be found (even though it may exist somewhere else in SNZ).

(d) The provision of time series analyses is enhanced because concordances in CARS can precisely document the changes between versions of a classification.

(e) The provision of a centralised computer-assisted-coding facility with CARS considerably reduces the time required to develop survey processing systems, and improves coding accuracy and consistency. The same computer-assisted-coding facility is available in a packaged form to external users. Another benefit of using the computer-assisted-coding facility with CARS is the coding consistency gained when several surveys coding to the same version of a classification all use the same source codefiles.

III. Rationale for Moving to a Classification Management System (CMS)

6. The rationale for moving to a new classification management system is due to:

(a) The need to mitigate a legacy system

(b) The need to move from a classification repository system to a full classification management system

(c) The need to reduce proliferation of like classifications and versions

(d) A desire to introduce a new approach to the management, storage and dissemination of classification related attributes and entities.

7. This has led to a vision of a system which is concepts based, that allows greater relationships to be established between relevant attributes, that is more efficient and automated in the authorisation and dissemination process, and which allows greater search and discovery of classification information. The main aspect is that rather than store copies of information as single standalone classifications and versions, the traditional components of a classification will be stored separately to enable greater reuse and reduce duplication ie store once and use in multiple locations.

IV. The CMS Application Model and its Relationship to Other Standards

8. In considering the model needed to develop such an application, it is useful to look at how such an application model relates to other standards and models which have been important for classification systems more generally.

9. It is important to clarify terminology here, because different standards and models are designed for different purposes. For the CMS model, there were some existing models and standards which served as the foundation upon which the CMS application model was built Other standards and model were ones which would be important for integrating the CMS with other systems, both as input from legacy systems, and as output, so that classifications could be used easily by other applications working with or disseminating data. Here, the terms which are used are "application model" – the model which exists within the software application in memory, which is entirely application-specific; "implementation model" – usually a standard model which is application-agnostic, but which sometimes closely resembles an application model (often standard models fit into this category); "reference model" – a model which is conceptual, and used primarily to facilitate good communication among the user community. Note than some models can perform more than one of these functions.

10. The foundational models for the CMS were several: ISO/IEC 11179, the Neuchatel Classifications model, SKOS (an RDF vocabulary for describing concept systems), and the generic Statistical Information Model (GSIM). Two of these models have been significant for the statistical community for a long time: ISO/IEC 11179 describes the representations of data element, which is obviously an important to classification management. Neuchatel provides a strong model for classification management, and is a good model for this purpose. Both of these models act as reference models and as implementation models, using the definitions above: they can serve as the models for creating application models, and this has been done in several different NSIs in the past.

11. Another important model in developing the CMS application model was GSIM. This is an important step forward in terms of classification models, because it emphasizes the importance of concepts and how they are used in different ways. When focussing on the relationships between classifications, it is important that the uses of concepts are captured in every place in which they are used. GSIM gives such a model in a way which other standards have not done in the past.

12. SKOS is a very widely implemented RDF vocabulary, used to describe "concept systems". This term should be understood in the way the Semantic Web community uses it, however, as it has a much broader meaning than within the statistical community. To the Semantic Web community, a "concept system" is anything which relates a set of concepts or ideas, used to describe, represent, or catalogue other objects. As a technology, RDF is very good at relating different types of objects, and because of this it is possible to end up with many different concept systems. (Anecdotally, claims are made that SKOS is the second most-used vocabulary in RDF, after Dublin Core.) It is interesting to note that the DDI Alliance is soon to release an extension to SKOS, XKOS, which refines it for use in

describing statistical classifications specifically. Both SKOS and XKOS served as models for the CMS application model.

13. Other models which are important to the CMS model, but which are not foundational, are those from the DDI and SDMX standards. These standards provide implementation models (using the terms given above), although neither were designed for classification management. What they are designed to do is to use classifications as the values of dimensions, attributes, and measures in aggregate data sets; and also to use classifications to organize collections of data sets (Category Schemes in SDMX); and to use classifications as the representations of variables and as response domains in questionnaires.

14. Looking at the history of how these various standards and models have been developed, there is an interesting pattern: with the exception of SKOS, many people have been involved in one or more of these developments. The most extreme case of this is GSIM, in which there were participants from the Neuchatel group, SDMX, DDI, ISO/IEC 11179, and XKOS. Thus, it is not surprising that many of the same ideas are being progressively refined as these related standards have been developed.

15. The CMS application model is the result of many years of thinking and implementation, based on the various standards and models which have come before. It is important to understand, however, that while a standard model must support requirements of a broad nature, coming from across a potentially diverse set of organizations and applications, the CMS model does not have this requirement. It is designed only to support classification management within a single organization, and thus can be more strictly defined than a standard implementation model would typically be.

V. Overview of the CMS Classification Model

16. However, the model can be generally described before going into detail in some of the more interesting portions of it. Discussion will focus on two aspects of the model: it's emphasis on the uses of concepts, and the richness of the relationships between different classifications.

17. There are several packages within the model:

(a) Core – This portion of the model focuses on identification, versioning, and describing contexts within which classifications are used.

(b) Classification – This package gives a general model for classifications in their generic sense, and then gives more specific extensions for formal statistical classifications and derived classifications.

(c) Coding – This package describes the relationships needed for integration with the SNZ coding system, and hold constructs such as synonyms, and synonym sets.

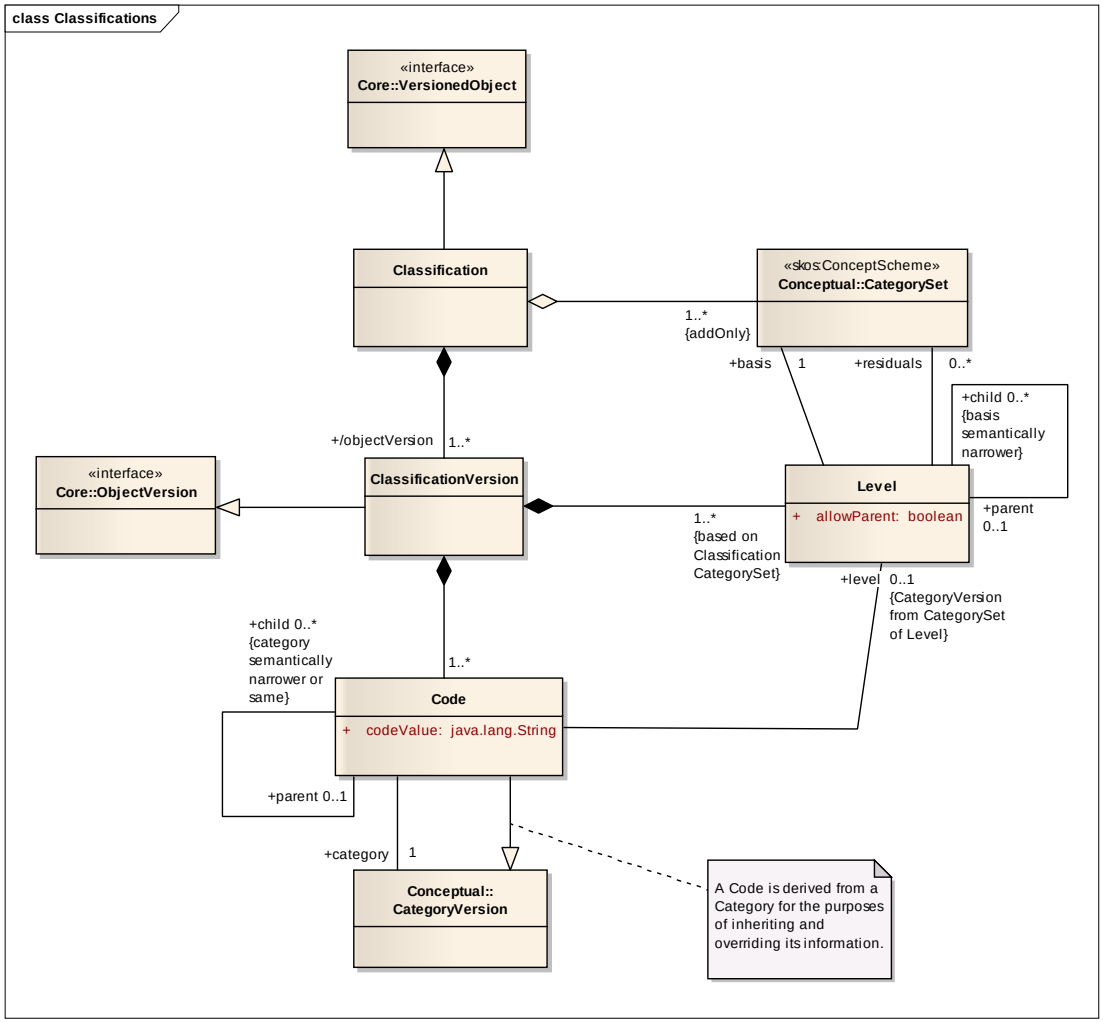
(d) Conceptual – This is the place where the concepts and their uses are modelled, along with the model for categories (that is, units of meaning).

(e) Concordances – This package describes all the relationships which can exist in concordances.

18. The packages presented in more detail here are Classification, Conceptual and Concordance. These packages show how rich use of concept linkages and other types of relationships are expressed in the CMS application model.

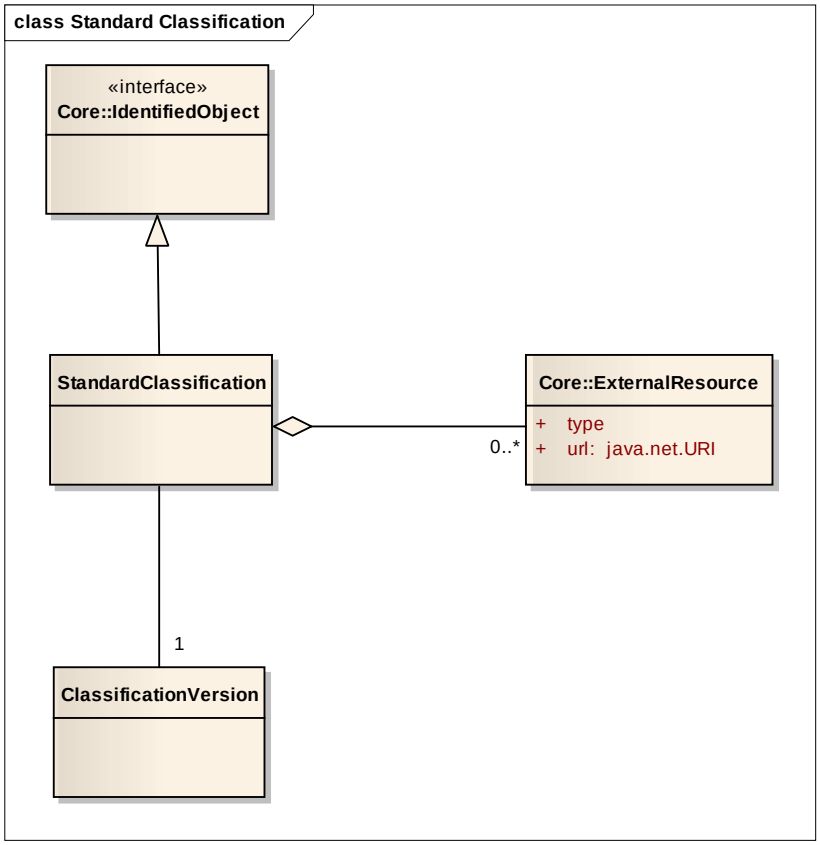
A. Classification

19. There are three basic models. The most general classification model is shown below:

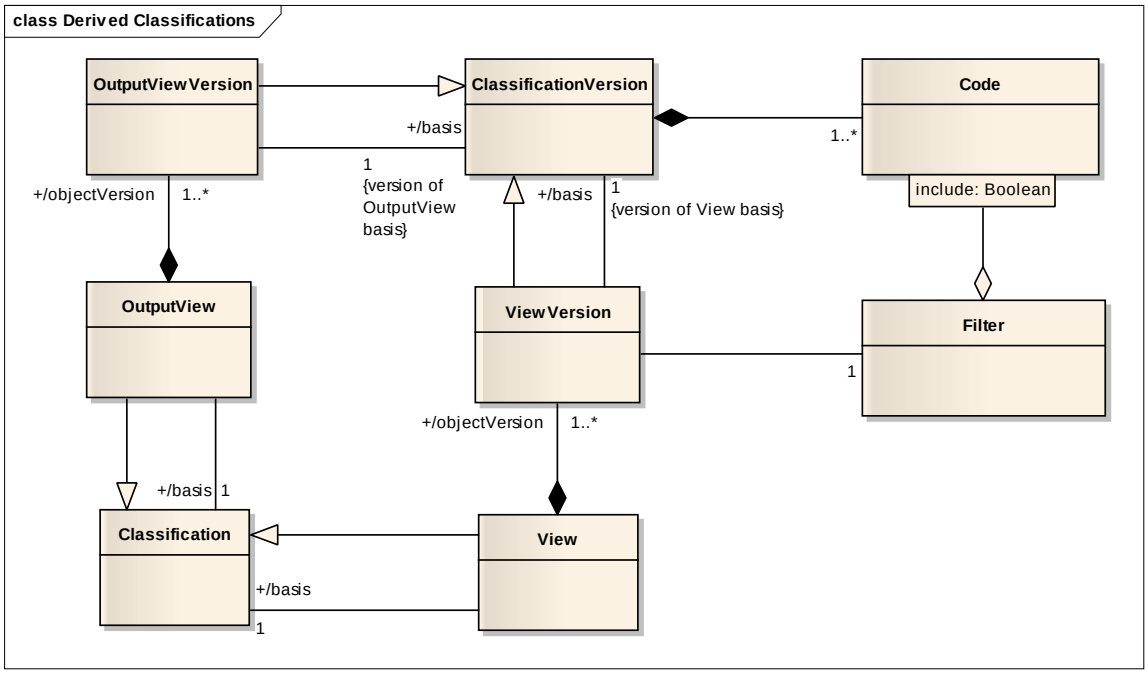


20. This shows that Classifications are potentially composed of several ClassificationVersions, and that these are associated with sets of Codes, and are levelled. This is very similar to Neuchatel. Additionally, Classifications are the basis for Concept Schemes, which are taken from SKOS.

21. The Standard Classification model is quite simple, doing nothing more than providing an object representing the published Classification version, and associating it with one or more URIs, so that it has a physical location.



22. The Derived Classification model is more interesting. The technique used here is to assign "views" to Classification versions, so that the multiple uses of a single Classification can be managed, with a single view of how each use is related to the central Classification.

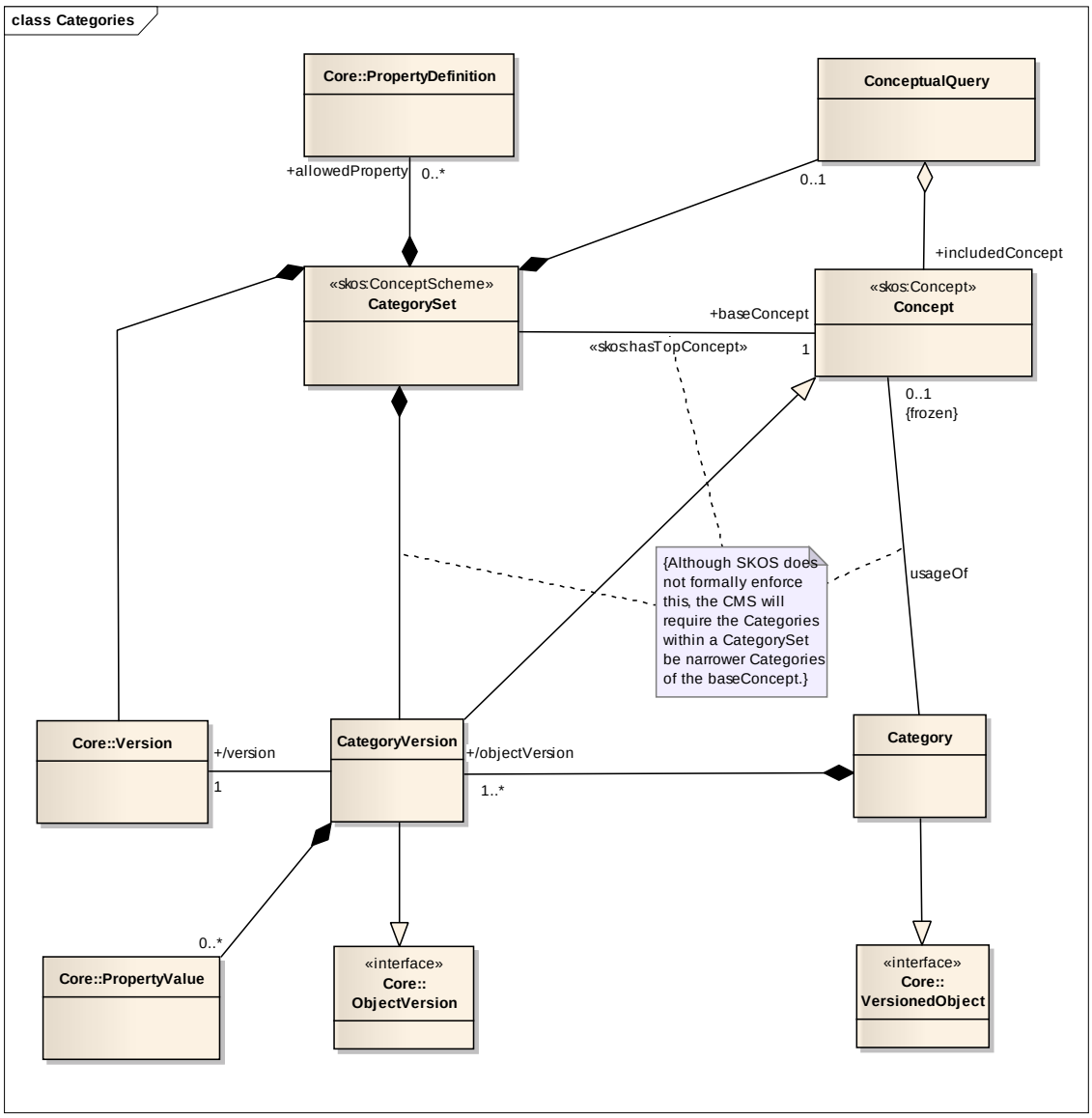


23. Note that the "views" are themselves versioned, and that there is a distinction made between an "output view" and a "view". An Output View is the use of a Classification for dissemination purposes, and may include filtering on the classification itself, a collapsing of nodes, etc. It is always

associated with a concordance, so that the Output View has a known relationship with a Classification.

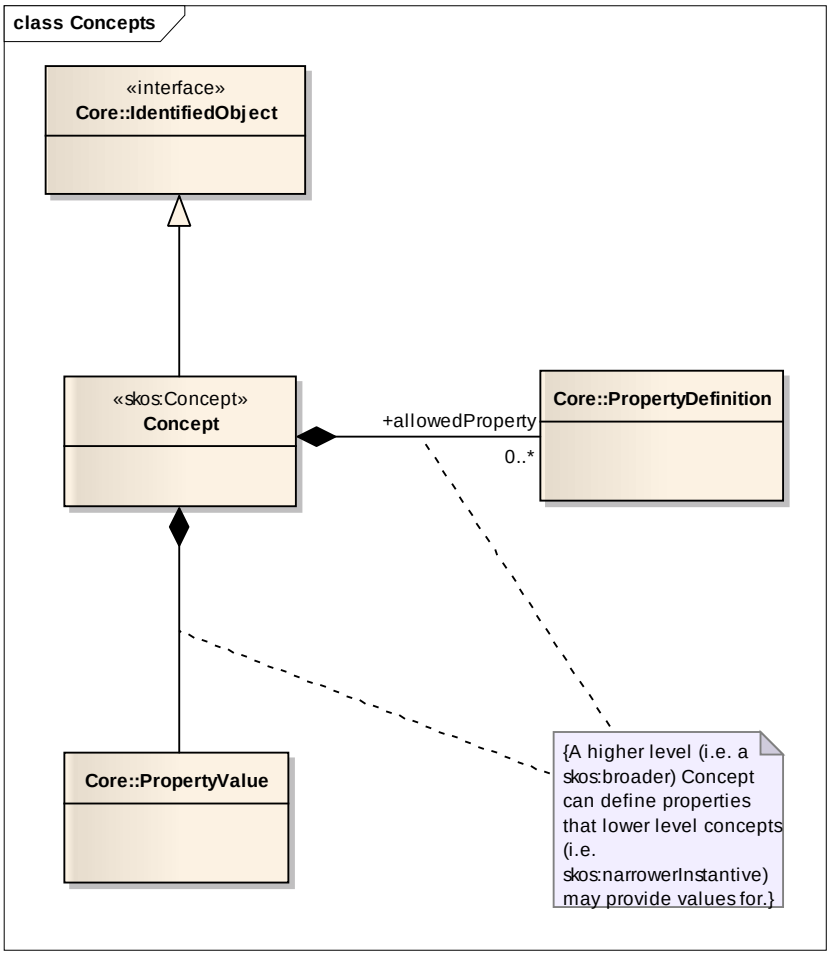
B. Conceptual Package

24. There are two diagrams in the Conceptual Package, one looking at Categories, and the other at Concepts. These two diagrams illustrate the richness of the use of concepts in the CMS model.



25. This shows that sets of categories have a base Concept. All Concepts are narrower in their definition, but are instantiations of the base Concept. Note also that categories – the objects which represent the units of meaning found in the nodes of classifications – are also related to Concepts: they are the use of Concepts. This is very similar to what is found in GSIM, and is a very powerful construct. It provides for the navigation within and across classifications according to the meanings they are composed of, which can be very powerful for disseminating data, and also for understanding the exact relationships between different data sets and structures.

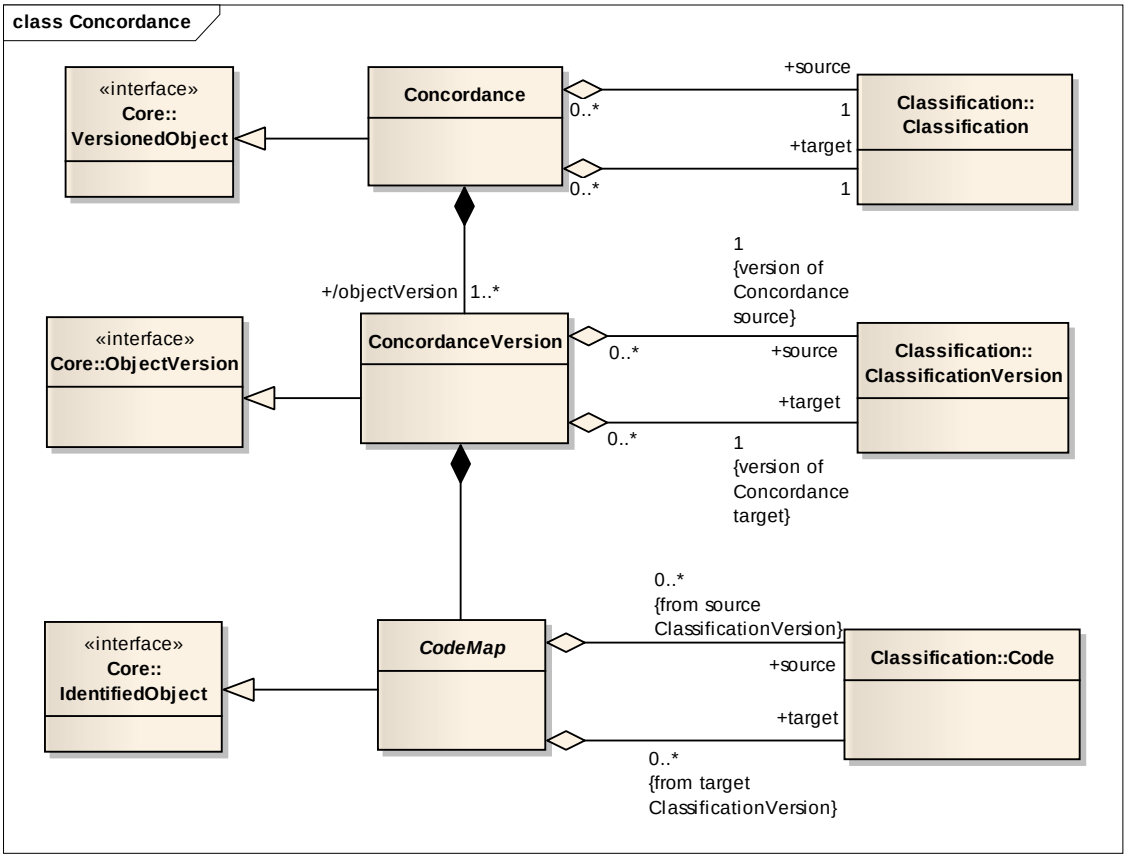
26. The Concept diagram shows more regarding the modelling of Concepts.



27. What is significant here is that Concepts used for different purposes may have different property sets. As Concepts play many different roles here, they need this flexibility within the CMS. It is also important to note that the Concept here is a SKOS Concept.

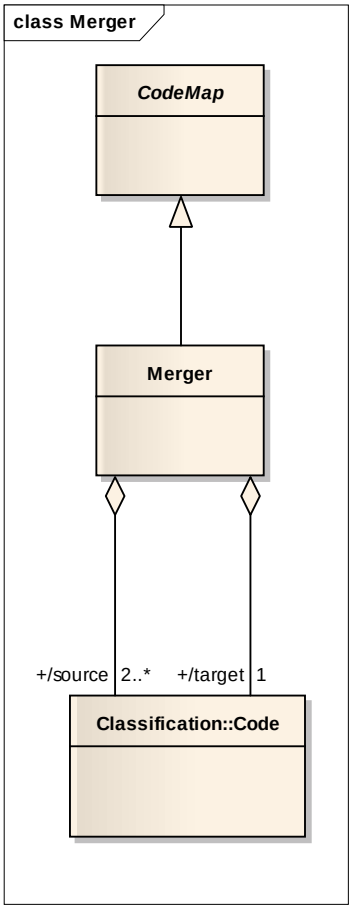
C. Concordance Package

28. There are many different diagrams within the Concordance package, modelling many different types of relationships between classifications. These include Merger, Deletion, Breakdown, SplitOff, and many other functions important to classification management in describing how two classifications are related. The diagram shown here is the central one, however.



29. This is similar to Neuchatel and GSIM (which bases its concordance model on Neuchatel). Here, are versions of Concordances which group sets of CodeMaps (the correlation of specific Codes used by different classifications).

30. To show how different functions are supported, here is the Merger diagram, showing how two or more nodes from the source classification are merged to form a single node in the target classification.



31. There are similar models for splitting a node of a classification into two or more. As stated, there are many different operations of this type. They can be modelled simply, but in every case they are instantiations of the CodeMap class, adding whatever specific information is needed to support that operation, and making the relationship a manageable object. One can see how, should additional functions be required, this model can be easily extended to support whatever additional information is required.

32. It should be noted that the Code in this model functions as the managed combination of a Category (the unit of meaning) a Concept, and the representation. While this is a simplification of the GSIM model, where there is a separate object for the node of the classification, this is in essence an implementation of the GSIM model. Within the controlled confines of a particular application, this type of simplification can be done without any harm, whereas, in a generic model like GSIM, many different ways of achieving the same goal would need to be supported, hence the extra objects.

VI. Conclusions

33. The CMS system is a classification management system, and not merely a repository. It contains a wealth of information about the concepts which are used in classifications, and has a rich and flexible set of information about the relationships and properties of classifications and their component structures. The model is designed to be flexible and extensible, to accommodate further system developments.

34. The CMS model builds on many of the best features of other models and standards, and is capable of supporting these. The richness of the information held in this model will provide a great deal of power when processing and disseminating data, because of the granularity with which it has associated concepts and other types of relationships between classifications, the many derived classifications which are taken from them, and the nodes of which they are composed.