

# Understanding Classification Schemes

Introduces the concepts involved in building and managing Classification Schemes and the patterns for categorising and organising them

**Adrian Miley**

[Adrian.Miley@mileywatts.com](mailto:Adrian.Miley@mileywatts.com)

# What is a Classification Scheme?

- A *Classification Scheme* is a controlled vocabulary of terms (*Classifiers*) that together form a complete and exhaustive group of values used to categorise and describe the characteristics of other things. Classification Schemes only have meaning when applied to other data.
- Classification Schemes are defined with different levels of rigour as to the Classifiers in the scheme but follow common underlying patterns in how they are organised.
- Examples of Classification Schemes are many and varied and are used over the entire area of human knowledge.
- Different types of Classification Schemes are known by many alternative names such as:- *Enumerated Domain, Taxonomy, Ontology, Dictionary, Folksonomy*.
- It is the degree of rigour in defining terms that characterises the various types of Classification Scheme.

# Defining Classification Schemes

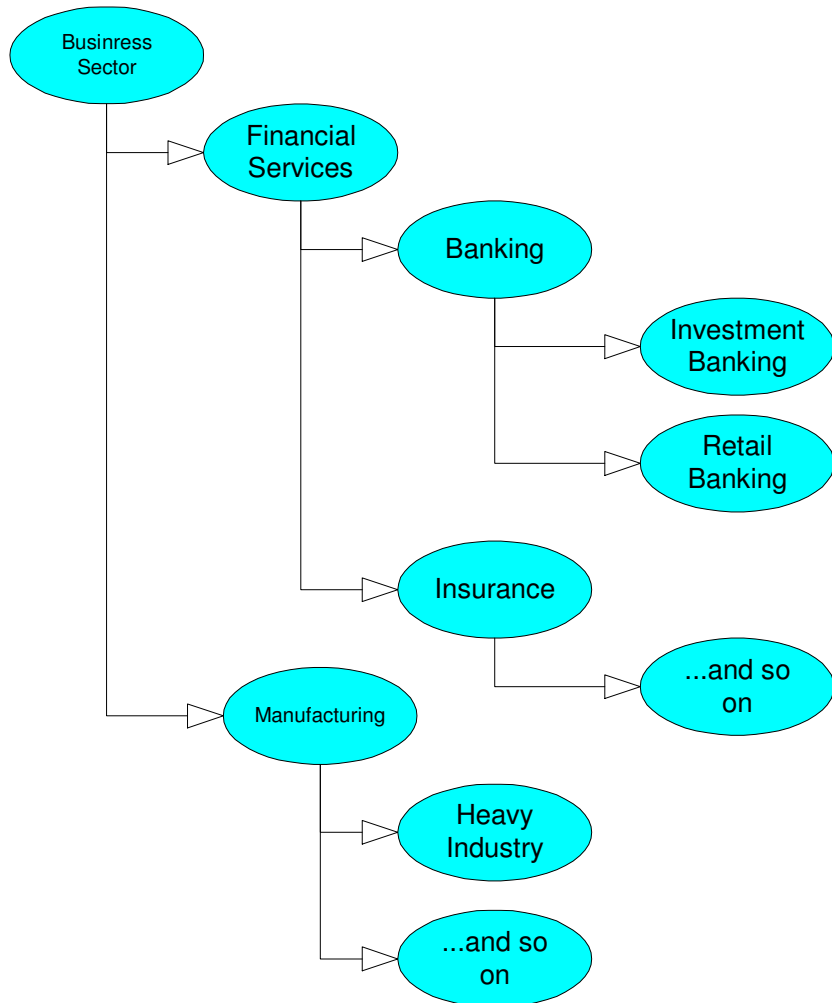
- Basic Classification Scheme Organisation
  - **Disparate Classification Scheme** or **Classifier Group** – a discrete set of Classifiers without any internal organisation.
  - **Hierarchical Classification Scheme** – a set of Classifiers organised into a “top-down” organisation where each Classifier has one parent and many children. Each level further refines the previous level.
  - **Network Classification Scheme** – A set of Classifiers where the Classifiers can be linked to any number of other Classifiers in an n-dimensional network.
- Derived Classification Schemes
  - **Collective Classification Scheme** – a Classification Scheme that is a sub-set (intersection) of another Classification Scheme.
  - **Aggregated Classification Scheme** – a Classification Scheme that is a union of two or more other Classification Schemes.

# Classifier Groups

- The simplest of Classification Schemes where the classifiers in the group have no internal organisation. Otherwise known as **Enumerated Domain** or **Code-list**.
- Each Classifier Code (the value that identifies the Classifier) must be unique within the Scheme. This is necessary because Classifier use is generally by reference to the Code.
- May be ordered or unordered but sequence is not significant.
- For example:
  - [Day Of Week] := {Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}
  - [Gender] := {Male, Female, Sequential Hermaphrodite, Synchronous Hermaphrodite, Neuter, Chimera, Transsexual }

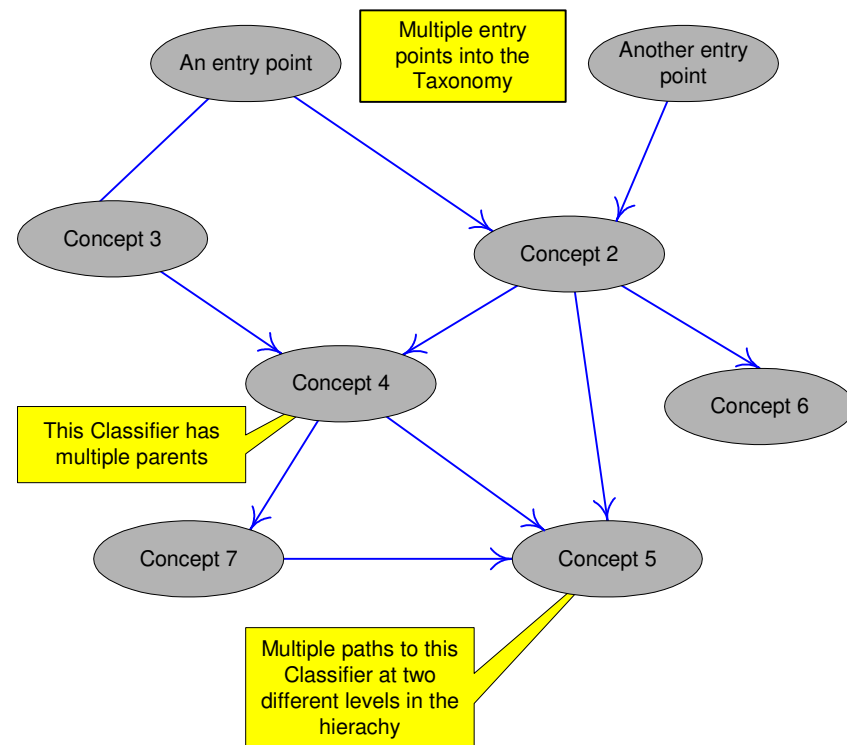
# Hierarchical Classification Schemes

- A Hierarchical Classification Scheme is a tree-structure where each element of the system (except for the top element) is subordinate to a single other element.
- Generally used for Classification Schemes that contain many (100+) Classifiers.
- The use of each term also implies that all parent terms in the hierarchical path back to the root term are also applicable.
- Generally only leaf-element is usable and other nodes are for internal organisation and navigation.



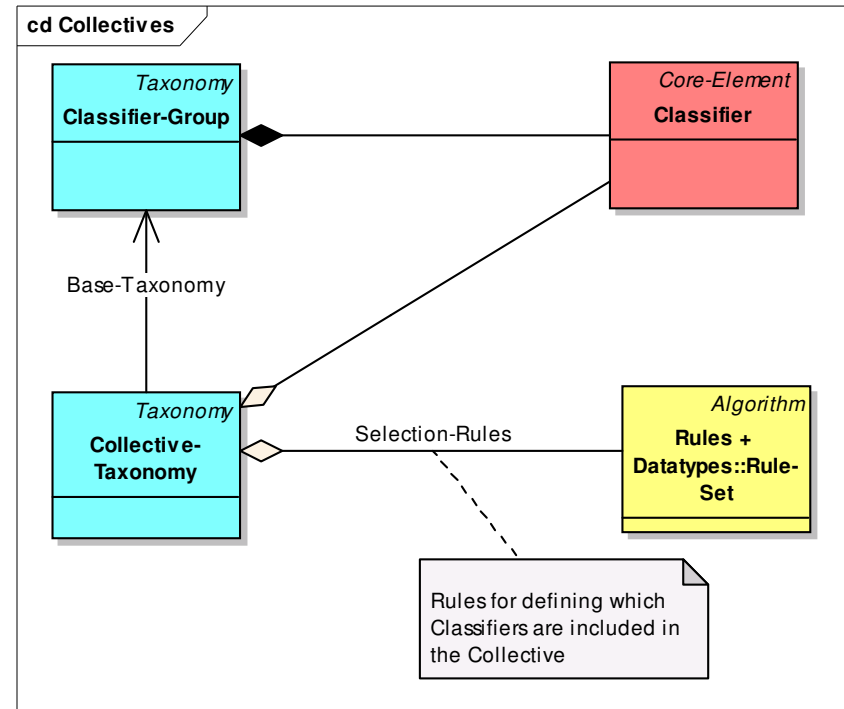
# Networked Classification Schemes

- A Networked Classification Scheme where the Classifiers can be linked to any number of other Classifiers in an n-dimensional network.
- The network may also contain a hierarchy as well.
- Enables multiple views of the same Classifiers.
- Each set of links may identify different relationships between Classifiers e.g. “Similar To” as well as “Generalises” and “Refines”.
- Restriction is that relationships should not be recursive (forms an infinite loop).



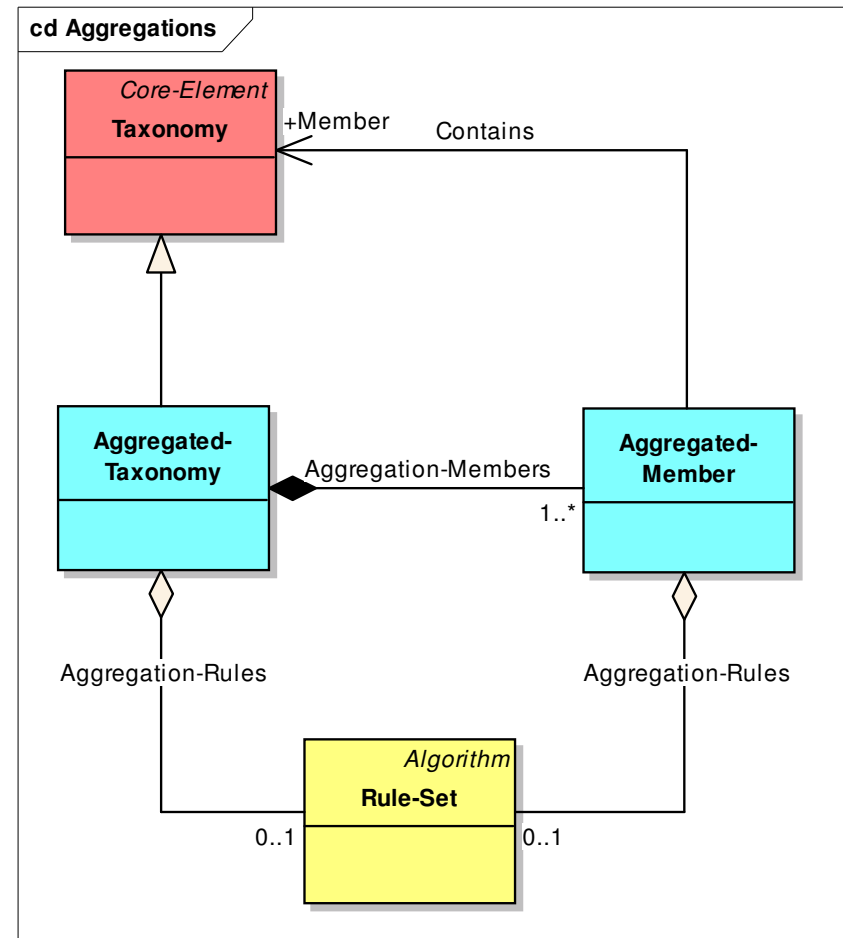
# Collective Classification Schemes

- A sub-set of a Classification Scheme based on defined selection criteria.
- May inherit the organisation of the underlying Scheme or may be disparate but cannot define an alternative organisation.
- Included Classifiers may be derived statically (at the point the Selection Criteria are defined) or at publication time (at the point the Collective is released for use).



# Aggregated Classification Schemes

- Formed by aggregating the contents of two or more other Classification Schemes.
- May not change internal organisation of underlying Classification Scheme but may create links between the underlying Classification Schemes.
- May have selection rules for defining which members of underlying Classification Scheme to include.

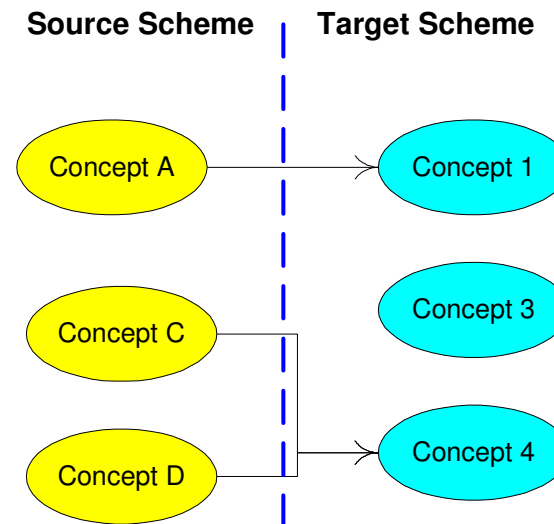
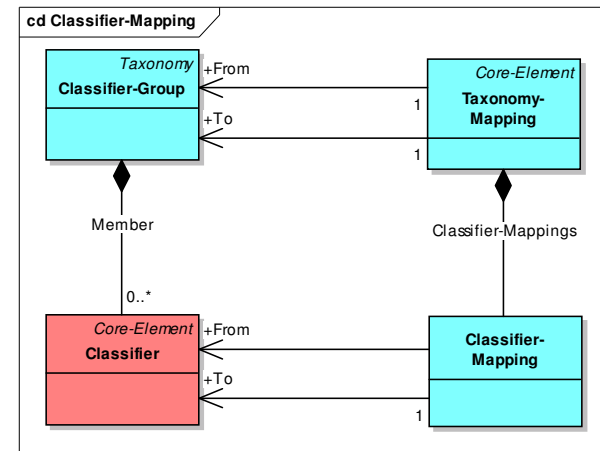


# Extending Classification Schemes

- Linking Classification Schemes
  - Mapping Classification Schemes – defined transformations between Schemes
  - Code Combinations
- Changes over time
  - Classifier is historic equivalent of Classifier
  - Validity Date Range – the date range that the Classifier is valid for.
  - Usability Date Range – the date range that the Classifier can be used to categorise other data.
  - Membership Date Range – the date range that the Classifier was a member of a Classification Scheme.

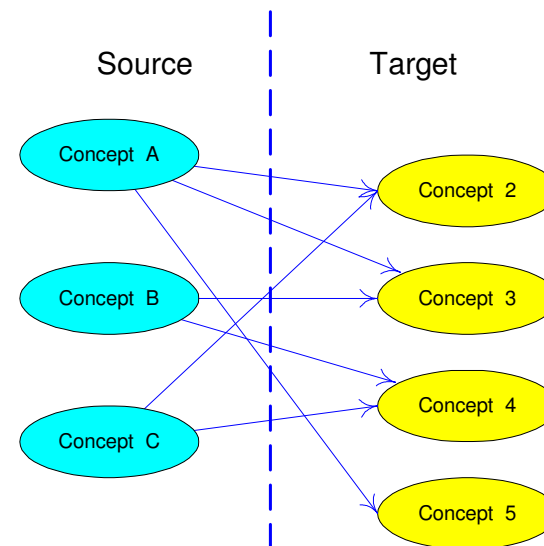
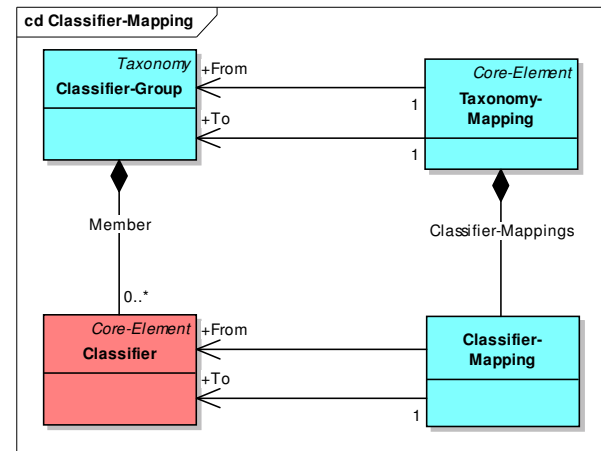
# Mapping Classification Schemes

- Two Classification Schemes covering similar subject area are mapped to each other to produce a transformation from one scheme to the other.
- Useful when external data sources are being mapped to an internal common data repository.
- Transformation usually only from Source to Target but may be bi-directional.
- Every Classifier in Source should map to Target but not every Target must have a Source.



# Code Combinations

- Selection of a Classifier in one Scheme defines the allowed Classifiers from another Scheme.
- Useful for “cross-field” validation.
- Each Classifier in Source can be linked to any number of Classifiers in Target but not every Target needs a Source.
- Links are bi-directional because the order of navigation (validation) should not be significant.



# Valid Date Ranges

- A Classification Scheme may evolve over time and not all Classifiers are valid for the entire life of the Classification Scheme. There are three sub-sets:
  - The sub-set of classifiers that may be present in existing data.
  - The sub-set of Classifiers that may be used to classify new data
  - The sub-set of Classifiers that will become valid in the future.
- Hence each Classifier have their own Validity Date Ranges within the Validity Range of the Classification Scheme.
- Codes may also be reused but valid dates of reused codes cannot overlap.

