Principles of Classification

Suggestions for a procedure to be used by ICIS in developing international classification tables for the construction industry

by **John Cann, NBS Services** (UK)

1. Nature of classification

Classification is the grouping together of like objects and their separation from unlike objects (1). The term "objects" as used in this paper includes both material objects (e.g. building, road) and immaterial objects (e.g. magnetism, tensile strength). Classification is achieved by arranging objects into **classes** - a class being a group of objects which share a particular set of properties, no other objects having this particular set of properties.

2. Division of classes

A class may be divided into a number of subclasses, where each subclass is a subset of the original class. The members of a subclass share a set of properties that are a specialised version of the set of properties shared by the original class. The subclass is also said to be **subordinate** to the original class. This process may be repeated and the subclasses divided into a lower level of subclasses. Equally classes may be grouped together into higher level classes which are **superordinate** to the original classes. Classes at the same level of division are described as **coordinate**. The structure: superordinate classes - coordinate classes - subordinate classes is known as a **hierarchy**.

2.1. Principles of division

A class is divided into a particular set of subclasses according to a particular **principle of division** (also called a **characteristic**). For example if we divide the class *Literature* by the *Language* principle of division we obtain the subclasses English Literature, French Literature, German Literature, etc.; if we divide the class *Literature* by the *Form* principal of division we obtain the subclasses Prose and Poetry

In the modelling world the division of classes of objects into subclasses is known as **specialisation**. An example of the specialisation of a class of real world material objects is dividing *Fruit* into Apples, Pears, Oranges, Bananas, etc. However *Fruit* could also be specialised into Red Fruit, Yellow Fruit, Green Fruit, etc.

The modelling world also recognises another way in which a series of classes can be generated from a single class: **decomposition**. This is the dividing of an object into its parts, for example dividing *Fruit* into Fruit Skin, Fruit Flesh, Fruit Seeds. In this case the series of classes, although subordinate to the original class, are not subclasses of fruit because they are not types of fruit.

2.2 Simple classes / Compound classes

A **compound class** is one which reflects more than one principle of division within a conventional class. For example, within the class *Chemistry* the subclass *Vacuum distillation of Ethyl Alcohol* is a compound class since it reflects both a type of chemical substance, and a chemical operation.

A **simple class** is one which reflects only one principle of division

3. Kinds of classification (2)

3.1 Special classifications contrasted with General classifications

Classifications may be **special** - concentrating on, or biased towards a particular subject (e.g. CI/SfB); or **general** - covering the universe of information (e.g. UDC).

3.2 Analytical classifications contrasted with Documentary classifications

Classifications may be **analytical** - systematising physical phenomena (often of the natural world) and thus providing the basis for explanation, prediction and understanding (also known as **scientific classifications** or **taxonomies**); or **documentary** - an aid to the management of documents or other kinds of information, with the aim of making information locatable. UDC (Universal Decimal Classification) and DDC (Dewey Decimal Classification) are documentary classifications; the classification of the animal kingdom into different phyla, genuses and species is an analytical classification. Documentary classifications are generally based on an arrangement by discipline/subject, whereas analytical classifications isolate actual objects/phenomena. For example in an analytical classification we might have a class "Minerals" and divide this into a number of subclasses, one of which will be "Coal". In contrast, in UDC (a documentary classification), coal has no one place - aspects of it appear under the subjects petrology, economic geology, mining, fuel, etc. However, the above categories are not mutually exclusive - UDC incorporates small sections of analytical classifications. Also, CI/SfB, which is in essence a analytical system isolating types of buildings and their parts, has certain sections, mainly in Table 4, which enable it to be used to classify subjects, and hence arrange documents in a library.

3.3 Enumerative classifications contrasted with Faceted classifications

Classifications may also be **enumerative** - an attempt is made to list exhaustively all possible subclasses (including compound subclasses) of interest in a particular class; or **faceted** - only simple subclasses generated by a single principle of division are listed, compound subclasses may then be synthesised from these simple subclasses. In a faceted classification the simple subclasses are grouped into facets, each facet being the total of subclasses obtained by applying a particular principle of division. To be useful and consistent it is necessary to specify an order in which facets are combined when synthesising a compound subclass - this is the citation order. A faceted classification will also have general facets which are applicable to any main class such as the facets of Time and Place.

To illustrate the differences between faceted and enumerative classifications, consider the class **Literature**.(3)

In an enumerative scheme we might have the following subclasses:
English
Prose
Poetry
Sonnet
Ballad
German
Prose
Poetry
Sonnet
Ballad
French
Prose
Poetry
Sonnet
Ballad

In a faceted scheme we might have the following instructions and subclasses:
Citation order:
Language facet is cited before form facet
Language facet
English
German
French
Form facet
Prose
Poetry
Sonnet
Ballad
To create a compound subject such as English Sonnets, we combine, in the correct order, the English subclass from the language facet with the Sonnet subclass from the form facet.

Enumerative classifications are sometimes known as Hierarchical classifications because the successive division by different principles results in a complex tree-like structure with many different hierarchical levels. However, this terminology is misleading since individual facets may have a hierarchical structure, (as can be seen above when the form Poetry is subdivided into forms Sonnet and Ballad) and specifying a citation order also implies a hierarchical relationship between the various facets.

Hunter (4) lists the following advantages and disadvantages of enumerative and faceted classifications.

<u>Faceted</u>

Advantages
C Because complex subjects are not listed, such schemes are easier to compile.
C The schedules are shorter for the same reason but, despite their brevity, they permit the classification of both very simple and very complex subjects.
C New subjects can often be catered for by the combination of already existing concepts.

Disadvantages

C The notation can become long and complex and may be unsuitable for the arrangement of documents on shelves in a library.

C The problem of citation order can cause difficulty.

Enumerative

Advantages

C Such schemes have been generally accepted and widely used throughout the world for a long period of time.

C A fairly short and uncomplicated notation can be used.

C “Notationally” it is easier to display the structure of the scheme.

Disadvantages

C It is impossible to list every conceivable subject.

C There can be a lack of accommodation for even simple subjects.

C New subjects cannot be accommodated and regular revision may be required.

However, to the advantages of faceted classifications may be added the now considerable advantage that they are well suited to computerised systems, including computerised subject searching.

It is possible, indeed common, to have classifications which include a mixture of an enumerative approach and a faceted approach. For example UDC is based on an enumerative classification (DDC), but has many facets which compliment the main enumerative schedules. These include the common auxiliary tables, which are generally applicable facets, and the special auxiliaries which are facets only applicable to particular classes. Even DDC nowadays has a limited number of generally applicable facets.

4. Notation

The primary function of notation is to maintain the order decided upon for the classes. It also provides a reference marker and shorthand label so that a particular class can be rapidly located when, for example cross-referring from an index to a classified sequence.

While notation is vital in a classification of any complexity, it is important to note that “notation is entirely subsidiary to the order of classes: it is its servant. Thinking should always be in terms of the order or grouping (of classes) required, not in terms of notation”(5).

A number of features are desirable in a notation, and these are listed below. However, it should be noted that some of these features conflict - it is not possible to maximise all positive features at the same time.

(a) The notation should have an **ordinal value** that is **intuitive**. This tends to mean that only roman letters and/or Arabic numerals should be used for the main structure of the table. Indeed one of the criticisms of CI/SfB is that the notation does not intuitively indicate the order of the different tables - the significance of devices such as brackets and lower case versus upper case letters in setting order can easily be overlooked by users. Arabic numerals may be treated as decimal fractions (i.e. place an imaginary decimal point before the number when determining the order). This will result in the following sequence: 1...11...12...13...131...2...21...22...23...231...3...etc. Alternatively they may be treated as integers, which would give: 1...2...3...11...12...13...21...22...23...131...231...etc. Both approaches have advantages - treatment as integers is more intuitive for users; treatment as decimals allows for the creation of subordinate classes by simply adding a further digit to an existing number.

(b) It is desirable that the notation is **expressive** - that is it reveals the structure of the classification scheme. For example, in the Literature class in DDC the hierarchy is reflected in the notation, a further digit being added at each stage of division:

8	Literature
82	English literature
821	English poetry
821.3	Sixteenth-century English poetry

One benefit of a simple expressive notation is that it can easily be shortened, if a great deal of detail is not needed, by simply deleting the characters not required, and classifying at the resulting class number. For example, if we decide that Sixteenth-century English poetry is too detailed we can remove the “1.3” and classify at 82, which is, in an expressive scheme, always superordinate to 821.3, and in this case corresponds to English Literature.

However, the obvious question arises “ what if we have more than ten coordinate classes (in a numerical system)?”. Raganathan has suggested two ways this may be accommodated. The first of these, the **octave device**, reserves the final digit 9 for a further set of coordinate divisions. With such a device 9 is never used alone but always introduces a further series of coordinate divisions: 91, 92, 93... 991, 992...and so on indefinitely. However, the resulting numbers do not *look* coordinate and are evidently so only to the initiated. The other device is the **centesimal notation** which provides two numbers at each division thus providing up to a hundred possible coordinate classes in a numerical notation.

A problem with expressive notation is that it will lead to longer class numbers than a non-expressive notation (see below).

(c) It is desirable that the notation should be as **short** as possible. The main way of generally achieving this is to ensure that large subjects are allocated more space than smaller subjects. The decision on what are to be the main classes in the classification is likely to be critical in this. For example DDC allocates the same space to Philosophy and Religion as to the whole of Science and Technology. This tends to mean that important classes in technology have long class numbers. However, a short notation may also be achieved by the rejection of expressiveness (see above), or the use of enumeration rather than faceted synthesis (see section 3)

(d) The notation should be **hospitable** - it should allow the accommodation of new objects. Some class numbers at each coordinate level can be left unused to allow for this, or Raganathan’s octave device can be employed (see (b) above). However, these methods do not allow the object to be placed at the exact point in the classification that it should be, i.e. between its two nearest relatives. An alternative is to extend the notation by adding an extra character at the point in the schedule where the object is to be accommodated. For example suppose “French Windows” had just been invented, and it was decided that they formed a distinct class, separate from both the class "Windows" which is numbered "14" and the class "Doors" which is numbered "15". Clearly it should appear between Windows and Doors in the classification, and so we could add a character to 14, giving 145 (assuming that numerals are treated as decimal fractions). The problem with this approach is that 141-149 might have already been used to refer to subclasses of windows. Also this method would not maintain the expressiveness of a classification.

4.1 Signs/facet indicators

Any classification which employs synthesis to create class numbers must indicate which facets have been used in the synthesis, and what the nature of the relationship is between the facets. There are many ways that this can be done and a full discussion is beyond the scope of this paper. However, UDC uses a colon, **:**, as a powerful device to link class numbers from any part of the main schedules.

5. Other considerations for classification schemes

5.1 Order of citation of facets; order of division

It is important to bear in mind the effects of the chosen citation order for facets (in a faceted scheme) or order of division of classes (in an enumerative scheme). Consider the Literature example in 3.3 - in this example the language facet was cited first and then the form facet. What this means in practice is that all works on English literature will be collected together, as will all works on French literature, but there will be no one place for all works on poetry - English poetry will be separated from French poetry. It is therefore necessary when creating a classification scheme to decide which is likely to be the most useful grouping for users of the scheme.

5.2 Filing order is the reverse of citation order

The filing order is the order in which the facets appear on the shelves (or in a computer database sorted in classified order). For example if we had a book on French literature and a general book on Poetry without bias towards a particular language, which would appear first on the shelf? It can be shown that in order to follow the important classification rule that the general files before the specific it is best for the filing order to be the reverse of the citation order. If this rule is followed and the citation order is "Language facet is cited before form facet" then the general book on poetry will appear before the book on French literature on the shelf. However, this rule is not followed by all classification systems, mainly because it is more complicated for the users to have to remember a different filing order and citation order. Indeed for relatively short, simple schemes the reduction in complexity is likely to outweigh the advantage of sticking strictly to the "general files before the specific" rule. For example in CI/SfB the filing order is the same as the citation order.

6. Indexes, Thesauri

6.1 Indexes to classification schemes, Pre-coordinate Indexes

An index to a classification scheme is necessary:

(a) to indicate the location in the schedules of a sought term;
(b) to collocate distributed relatives.
By "distributive relatives" we mean related subjects which are scattered by the classification because of the citation order of facets (in a faceted classification) or order in which the principles of division have been applied to create an enumerative classification. For example, if the Literature class is divided first by Language and then by Form, then works on Poetry will be scattered throughout the literature section, depending on which language the poetry is written in. An index should enable people who are interested in poetry as a subject, regardless of language to identify the various points in the classification where works on poetry will appear.

An index to an existing classification scheme is a **Pre-coordinate** index - the relationships between the terms in compound subjects has been fixed before the index was created, in this case by the classification scheme.

6.2 Post-coordinate indexes, natural language subject retrieval

In some cases, particularly when retrieving information from a computerised database, it is common to attempt to gather information on a particular subject without the use of a classification scheme.

Suppose we have a computer database consisting of a series of records, each record being a reference to a document. One of the fields for each record is the "Abstract" field, which is a summary of the information that the document presents. A natural language index for this database may be produced automatically using a computer programme. The programme creates an alphabetical list of all words which appear in any abstract (excluding some unimportant "stop words"), together with a cross reference to the record(s) which mention each word. Such a cross-referenced list is a natural language, post-coordinate index. It is **natural language** because words have been used exactly as they appear in the abstract; it is **post-coordinate** because users create their own compound subjects by using whatever combination of words they feel is appropriate.

6.3 Controlled language subject retrieval

If we require more accuracy and consistency in indexing a database, we may assign subject terms to a record using a controlled language. The controlled language consists of a list of preferred terms - that is terms which have been chosen by an expert as the best way of referring to particular subjects. Cross references are provided from natural language terms to the preferred terms which should be used in their place. Preferred terms will be defined where necessary to indicate their scope. An example of a controlled language is the set of "Descriptors" used by the Dialog online database. Some Descriptors may denote compound subjects, but generally they are short terms denoting simple subjects. Indexers analyse the subject content of a document and assign relevant descriptors to that document until all the subjects in the document are covered. Users find compound subjects by combining descriptors or using a descriptor in combination with natural language terms.

6.4 Thesauri

A thesaurus is a common way of presenting a controlled language index so that it may be effectively and efficiently used by indexers and retrievers of information. A thesaurus, however, goes beyond a simple controlled language index by giving broader and narrower terms for each entry, thus introducing hierarchy into the relationship between terms. Therefore a thesaurus, like a classification, imposes a structure on a subject - indeed the two are often developed in tandem (see section 6.5 Thesaurofacets).

An example is the Thesaurus of Engineering and Scientific Terms. Examples of entries from it are given below:

Excavating machinery
    Use Excavating equipment

Evaporative cooling
BT   Cooling
NT   Film cooling
RT   Cooling systems
      Cooling towers

Fixed investment
UF   Capital investment

"*Use*" indicates a cross reference from a natural language term to a preferred term
" UF" indicates a cross reference from a preferred term to a natural language term
"BT" indicates a broader term
" NT" indicates a narrower term
"RT" indicates related terms

6.5 Thesaurofacets, thesauri incorporating a classified sequence

Some thesauri also have a classified sequence of terms which complements the controlled language index, and is linked to it by means of class numbers. Such a thesaurus is the Construction Industry Thesaurus, published by the Department of the Environment. While production of such a thesaurus certainly requires classification, it does not constitute a fully developed classification scheme because it does not give a method or order for combining class numbers from different facets; also the notation is not appropriate for the physical arrangement of documents on shelves.

Another well-known example of the inclusion of a classified sequence and a controlled language index in the same scheme is the *Thesaurofacet* which grew out of the English Electric Company's faceted classification for engineering. This attempts to provide both for subject retrieval and for shelf arrangement, and thus combine the traditional uses of thesauri and classification schemes.

7. Implications for the development of ICIS Classification schedules

First we may ask the question “what kind of classification should we try to produce for ICIS?” Of the types of classification described in section 3, it seems that we are attempting principally a special analytical classification - that is one which systematises knowledge in the field of Construction and thus provides the basis for explanation, prediction and understanding. However, it is important to ask whether we wish that the classification is also good for the arrangement of documents on shelves. If so then a pure analytical classification is unlikely to be sufficient. It also appears that a faceted classification would be more useful than a enumerative classification, mainly because a faceted classification would be better as a tool for the subject indexing and retrieval of information in a computerised database.

Therefore assuming we are producing an analytic, faceted classification what are the stages we need to go through in drawing up the schedules? The following steps are based on the recommendations of Needham (6):

1. Define the main classes for the scheme.

2. Create the facets for each main class

This may be done using a “bottom-up” approach or a “top-down” approach, or often, one followed by the other.

> **(a) Bottom-up approach**
>
> Identify all the objects (i.e. things or concepts) which belong in each class by reference to actual construction works, projects and information. These are known as the isolates. Then group the isolates together, so that in each group the isolates are distinguished by a particular characteristic. For example, if our main class is *Education* we group the concepts *Primary*, *Secondary*, *Further* together because these isolates are all distinguished by the *Ag*e characteristic. Similarly *Geography*, *Mathematics* and *Physics* are all distinguished by the *Subject taught* characteristic and thus belong in the same group. (N.B. *characteristic* here is equivalent to *principle of division* when using a top-down approach). The sum total of isolates distinguished by a particular characteristic is called a facet. The isolates within a facet are called *foci* to distinguish them from the unorganised concepts called isolates.
>
> **(b) Top-down approach**
>
> First identify all the principles of division which are applicable to a main class. Then consider one principle of division at a time, and list all the isolates which result when that principle of division is applied to that main class. When complete this process will yield the facets along with their foci.

The bottom-up approach has the advantage that the isolates have what is called in documentary classification *Literary warrant*. In an analytical classification we may use the analogous term *Practical warrant*. This means that the isolates we have identified are used in practice by people involved in the construction industry, rather than being theoretical terms. Therefore it is recommended to use the bottom-up approach first, and only use the top-down approach to check that there are no omissions.

Sometimes similar facets will be produced for each of two or more main classes, for example a materials facet may be found to be useful in a number of main classes and hence may be removed to form a main class of its own. Another example of this in construction classification is the principle of division “User activity” which may be used to form facets in all of the following classes: Construction Complexes, Construction Entities and Spaces.

3. Set the order of the foci within each facet
The foci should be ordered so that the most similar foci are close together and the most different foci are far apart.

4. Set the citation order for use when combining facets
See sections 3.3 and 5.1.

5. Set the filing order of facets
See section 5.2.

6. Add the notation
See section 4.

7. Make the index
See section 6.

8. The implications for ICIS work of draft standard ISO/CD 12006 - 2 - Part 2 (based on the work of ISO/TC 59/SC 13/Working Group 2)

This standard recommends the tables which should be present in a classification scheme for the construction industry and also recommends the primary principle of division for each table. For example it is recommended that an Elements table should have 'characteristic function' as the primary principle of division (i.e. the one that is cited first) in a similar way to Literature having language as its primary principle of division in the example in section 3.3 of this paper.

It is not yet clear what the fate of the standard will be in this committee draft phase, but it does nevertheless seem sensible to base the ICIS tables on the recommendations of the draft standard.

9. References

1. BS 1000C : 1963 - Guide to the Universal Decimal Classification (UDC). British Standards Institution, 1963.
2. BS 1000M : Part 1 : 1993 - Universal Decimal Classification - international medium edition - English text, edition 2. British Standards Institution, 1993.
3. Hunter, Eric J. Classification made simple. Gower, 1988.
4. Ibid.
5. BS 1000C : 1963
6. Needham, C.D. Organizing knowledge in libraries: an introduction to information retrieval. Andre Deutsch, 1971.

10. Bibliography

C BS 1000C : 1963 - Guide to the Universal Decimal Classification (UDC). British Standards Institution, 1963.

C BS 1000M : Part 1 : 1993 - Universal Decimal Classification - international medium edition - English text, edition 2. British Standards Institution, 1993.

C Hunter, Eric J. Classification made simple. Gower, 1988.

C ISO/CD 12006 - 2 Building construction - Organization of information about construction works - Part 2: Framework for classification of information. 1997.

C Langridge, D.W. Classification: its kinds, systems, elements and applications. Bowker Saur, 1992.

C Marcella, Rita & Robert Newton. A new manual of classification. Gower, 1994.

C Needham, C.D. Organizing knowledge in libraries: an introduction to information retrieval. Andre Deutsch, 1971.

C Ray-Jones, Alan & David Clegg. CI/SfB construction indexing manual. RIBA, 1976.

C Roberts, M.J. et al. Construction industry thesaurus: development edition. Department of the Environment, 1972.