Duplication of concepts in UDC

Andrew Buxton
UDC Editorial Team

Abstract: The paper describes a problem particular to universal knowledge classifications with a disciplinary structure. These types of classification present concepts subsumed to the disciplines in which they are studied and thus have to resolve the problem of concepts being repeated in different fields of knowledge. The author looks into how the impact the repetition of concepts in the UDC disciplinary structure may have on information retrieval. He considers advantages and disadvantages of different approaches in presenting re-used concepts in the scheme.

1. Computer searching of UDC numbers

Probably the most common method of "information retrieval" is using keywords (or often key phrases). They are used in a post-coordinated fashion, i.e. they are assigned as a list of separate terms at the indexing stage but may be combined together at the search stage. A document may be indexed with the terms:

mining, gold, Cornwall, history

and can then be retrieved by a search on one or more of them:

gold AND Cornwall

The Universal Decimal Classification is basically a pre-coordinated system, i.e. the various concepts are combined into a single number at the indexing stage. For example, "history of gold mining in Cornwall":

622.342(410.197)(091)

Computer searching of UDC numbers has been discussed by Buxton (1990). Given a fairly sophisticated retrieval system, it might be possible to retrieve all documents on Cornwall using (410.197) or all documents on history using 93. However, it would not be possible to retrieve all documents on gold solely from the number because there is no consistent notation for gold. The concept "gold" is always coloured by the discipline in which it occurs:

336.743.22	Gold currency
546.59	Gold (inorganic chemistry)
549.283	Gold (mineralogy)
553.41	Gold deposits (economic geology)
622.342	Gold ores (mining)
669.21	Gold (metallurgy)
671.1	Gold and silver articles (Industries: precious metals)
671.412.1	Gold coins (coins)

In a retrieval system based on UDC numbers, users would be dependent on the subject index to find all the different places at which documents on gold might appear. This could be done through an intermediate display (rather like the "disambiguation page" in Wikipedia) asking which of the above

they wanted to search on. In practice it might not be very easy to decide: where would one look to answer the question "How does gold occur naturally?" Also there is increasing interest in studies which transcend old disciplinary boundaries, sometimes creating new disciplines. "Women's studies/gender studies" is an obvious example.

As far as constructing schedules goes, the occurrence of concepts in more than one discipline ("co-occurrence") makes the schedules bigger and more complex. It also requires additional effort in creating alphabetical index, making sure that frequently sought concepts are always expressed consistently in words and are indexed irrespective whether they are a subject of study or an object to which the actual subject is related in some way - for instance an agent related to a process/operation/action or the result/product of such an operation.

Obviously, in terms of managing a universal scheme in which concepts are bound to occur in more than one field, the best solution would be to have a unique notation representation of concepts whenever possible. This would mean that one would have just one number for gold and use it in synthesis wherever required. Chemistry has the most detailed schedules of substances and so might be the best candidate for providing the number. Then we could have:

336.743:546.59	Gold currency
549:546.59	Mineralogy of gold
553:546.59	Mineral deposits of gold
622.3:546.59	Mining of gold
669:546.59	Metallurgy of gold

This approach has already been adopted in UDC for history. Instead of numbers for each country and each period of that country's history, the number is synthesised from 94, the place auxiliary and the time auxiliary, e.g. 94(44) "1940/1944". There are some problems with this: (a) historical countries do not always match modern countries and (b) the time periods are specific to the document so one would need quite a sophisticated retrieval system to find documents where the period covered overlapped with the period of interest.

There is in fact a schedule of "common auxiliaries of materials" at –03:

-032	Naturally occurring mineral materials
-032.1	Air
-032.2	Water
-032.3	Carbonaceous and hydrocarbon minerals
-032.4	Ores
-032.42	Gold ores. Silver ores
-033	Manufactured mineral-based materials
-034	Metals

This mixes together pure chemicals and mixtures and is clearly not based on chemical composition. It raises questions about when –032.42 is to be used in combination and when one should opt for, e.g., 622.342. Also if more detail is required (e.g. for a particular silver ore) should it be applied to –032.42 or do we take recourse to the existing detail in 54?

2. Disadvantages

1. This leads to longer numbers, e.g. nine figures for mining of gold instead of six, and see also the simple numbers for the most common animals in farming below. This is a wider issue in the revision of UDC and will not be considered further here.

2. It has implications for where the numbers file. It would mean that e.g. mining of gold would file rather early in mining before the various methods of mining if the number was just 622:546.59. This is always going to be an issue if synthesis is used rather than enumerating main numbers. However, one approach is to have a specified main number for breakdown by that facet, e.g.

622.3 Mining of specific minerals, ores, rocks

3. It would mean that the entities in that facet would always be grouped according to the discipline that the schedule was taken from. So if animals were taken from the zoology schedule, they would be grouped according to zoological taxonomy in animal husbandry rather than the way that would be more obvious in farming:

636.1 Domestic equines. Horses
636.2 Large ruminants. Cattle. Oxen
636.3 Small ruminants. Sheep. Goats
636.4 Pigs. Swine
636.5 Poultry

Biologically cattle, sheep and goats are even-toed ungulates along with deer, camels and giraffes. Horses are odd-toed ungulates along with tapirs, zebras and rhinoceroses. Poultry are birds which would also include game. (However, a taxonomic breakdown would perhaps remove the bias towards Western farm animals!) Pets, hunting, animal products, etc. would be similarly affected. Also the schedule of plants in gardening, 635, is related to their interest to gardeners rather than their botanical taxonomy:

635 Gardening
635.1/.8 Vegetables. Kitchen gardening
635.1 Root vegetables
635.2 Edible tubers and bulbs
635.3 Plants with edible stalks, leaves or flowers
635.4 Other green vegetables
635.5 Salad vegetables
etc.
635.9 Ornamentals. Decorative plants

In inorganic chemistry, ionic compounds are grouped under the cation (generally metal), e.g.

546.722 Ferrous oxide
546.722'226 Ferrous sulphate
546.722'284 Ferrous silicate
546.762 Chromium(II) oxide

If this schedule were used in mineralogy it would go against the usual grouping by anion, e.g.

549	Mineralogy
549.6	Silicates
549.611.11	Sapphirin

This happens in other places too:

553	Economic geology
553.67	Deposits of magnesium, aluminium, iron, etc. silicates
631.8	Fertilisers
631.842	Nitrate fertilisers. Potassium and sodium nitrates
66	Chemical technology
661.68	Production of silicon and its compounds
661.683	Sodium and potassium silicates
691.27	Fibrous and foliate silicates
691.31	Calcium silicate stone (The division here is on the basis of natural/artificial stones, not chemical composition).

In biochemistry and pharmacology, there are several common groupings which cut across chemical composition, e.g. vitamins, hormones. (Issues in the classification of chemicals are discussed in Buxton (2011).)

4. Another problem with using the chemistry schedules wherever there is a substance facet is that many common substances are not pure chemicals but mixtures. Some common mixtures have been included in the current chemical schedules, e.g.

546.176	Aqua regia. Mixture of hydrochloric and nitric acids
546.217	Air

but steel appears only in the common auxiliaries and in technology.

It would be clumsy to have to synthesize numbers for mixtures from the individual numbers, e.g. 546.17+546.21 for air (not entirely accurate) or 549.261+546.26 for steel (this does not cover the large range of different steels).

The schedule of common auxiliaries includes many common substances which are mixtures, but as remarked above it is incompatible with a chemical taxonomy.

5. Many of the substances required are quite complicated chemically. For example, minerals often have variable constituents. The feldspars, currently at 549.651, are a group of aluminium silicates with varying amounts of potassium, sodium and calcium. A synthesized number to express this would be quite complicated.

6. In some contexts it is necessary to distinguish the physical state of a substance (the "allotrope") although it is the same chemically. Solid, liquid and gaseous forms are quite easily dealt with by an auxiliary. But diamond, graphite, soot, charcoal and buckminsterfullerene are all forms of carbon and do not follow a pattern that applies to other elements.

3. Conclusion

At first sight there appears to be a lot of unnecessary duplication in the UDC where the same concepts are listed several times in different places. This is in contrast to thesauri consisting of keywords which can be selected in whatever combinations are required both at the indexing and retrieval stages. However, a classification is more complicated than a thesaurus. It needs to combine concepts together at the indexing stage and have a citation order to say in what order they should be combined. In its role as a way of ordering books in a library, it needs to have a reasonably simple and concise notation. If it aims to cover all areas of knowledge, it needs to have a sensible arrangement of concepts in each discipline, even where the same entity is studied in more than one discipline. Concepts which are distinct and and needed in one discipline may not work so well in another one. Sadly it appears that these requirements are incompatible with synthesising compound numbers from simple numbers for each concept and treating them like keywords.

References

Buxton, A. Computer searching of UDC numbers. *Journal of Documentation*, 46, 3 (1990), pp. 193-217.

Buxton, A. Ontologies and classification of chemicals: can they help each other? In: *Classification & Ontology: Formal Approaches and Access to Knowledge: Proceedings of the International UDC Seminar 2011*. Eds. A. Slavic; E. Civallero. Wurzburg: Ergon Verlag, 2011, pp. 109-128.