

Brief Report – HILT III Project

UNESCO Thesaurus to DDC mappings: Third Summary – Thousand Sections

George Macgregor - Sept. 2006

This brief report explores the feasibility and cost of performing mappings between two knowledge organization systems (KOS) for the purposes of the High-level Thesaurus (HILT) project (phase III), which is exploring suitable architectures for machine-to-machine (M2M) terminology services. KOS-to-KOS mappings between the UNESCO Thesaurus and the 'thousand sections' of the Dewey Decimal Classification (DDC) are considered.

Comment on approach

The process of estimating the feasibility and cost of mapping terminologies to the 'Thousand Sections' of DDC entails an approach not normally used in intellectual mapping. Traditional approaches entail mapping the chosen term from the satellite terminology to the best and most relevant match in the target terminology. However, in order to investigate the possibility of offering mappings to the DDC Thousand Sections, it is first necessary to examine the DDC schedules, then seek out potential terms to be mapped from the satellite terminology to the chosen DDC number, and then implement a mapping from the satellite to the target terminology. This aspect should (probably) be noted if this work be documented in JISC reports. It is also worth noting that – conceptually speaking – this approach undermines the model of HILT as a 'spine-based' terminology server.

Issues and methodology

After some preliminary investigations with mapping UNESCO, it is clear that there exist several independent conditions which make accurate calculations problematic regarding cost and time.

Each satellite terminology will cover knowledge at different levels of specificity, with some (e.g. AAT, GCMD, MeSH, etc.) disregarding most other areas of knowledge in order to provide a discipline specific terminology. Thus, in some cases (such as UNESCO below) the terms are too broad (or too narrow) to adequately map to the ten, hundred and thousand sections. In practice this means that the DDC number of, say, 900 (History & geography) has mapped to it the UNESCO terms of 'History' and 'Geography' separately. In some of the cases below, one DDC number might have four mapped terms in order to reflect the subject coverage of the DDC class. Within particular disciplines and DDC numbers, it is possible to expect that there may be as many as six such mappings to a single DDC number (multiple-to-single mappings), depending on the specificity of the satellite terminology. This also occasionally leads to inaccurate or confusing mappings.

Since one-to-one mappings are unlikely, and because we are unable to predict how many mappings within a given terminology might be mapped to a single number, it is extremely difficult to provide accurate figures of how many mappings would be required, at what cost, and at how long it might take to implement.

The set below includes mappings to 38 separate DDC numbers; however, the set includes a further 12 UNESCO to DDC mappings (50 mappings in total), thus reflecting the need to have multiple satellite terms mapped to a single DDC number (see, for example, 070; 390; 900; 090; etc.). These 12 mappings account for a large proportion of the total set (24%).

The set below was selected purposively and cannot be said to be fully representative of either the satellite or target terminology; however, it might be possible to ascertain a **vague approximation** of number and cost of mappings by adding 24% of the mappings required to the total number of

DDC ten, hundred and thousand sections. Consideration of vacant DDC classes also has to be undertaken. These vacant DDC numbers are either unassigned or are left vacant for the use of auxiliary tables. These currently total 95 numbers, 94 of which reside within the thousand sections. 1 (i.e. 040) is the only unassigned hundred division.

The number of mappings required is thus:

$$10 + 100 + 1000 - X + Y = 1259$$

Where X is the number of vacant classes, which is 95. This value will remain constant across all mapped terminologies.

Where Y is the possible number of additional mappings based on an occurrence rate of 24%. In our case this is 244. This value is **not** constant across all terminologies and will vary depending on the terminology being mapped.

Figures for cost and time per mapping are based on Leonard Will's evaluation of HILT II, in which he proposed revised figures following inconsistencies found in the HILT II Final Report calculations. Leonard Will proposed a cost of £5.25 per mapped term, at 7 minutes per term. Two financial years have passed since the evaluation was published. Inflation has consequently been added at the current rate of 2.4% (over two years). This provides a cost per mapping of **£5.50**.

Therefore:

$$1259 * 5.50 = 6924.5$$

$$1259 * (7 / 60) = 146.89$$




According to the above calculations, the cost of mapping UNESCO to the ten, hundred and thousand sections of DDC could be said to be **£6924.50**. This would take **146.98** hours to accomplish.
















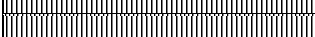

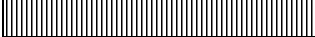
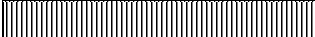
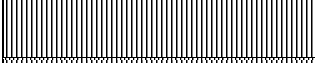
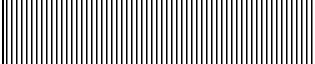
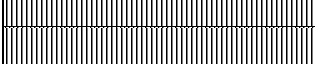
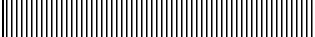


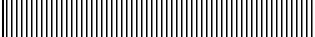
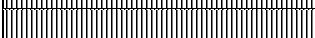




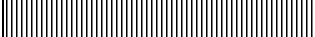
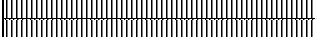


Conclusion

It is manageable to map the current HILT terminologies to the ten, hundred and thousand sections of DDC; however, it is simply too difficult to provide accurate figures across all terminologies since the Y value will vary considerably between all terminologies. This will be particularly the case in discipline specific terminologies (e.g. MeSH, AAT, etc.) and low specificity terminologies (e.g. IPSV, UNESCO, etc.). Thus, the rough-and-ready investigation documented above will have to be replicated across all the HILT terminologies in order to ascertain the Y value. Even after collecting this data, the possibility of providing accurate figures will be difficult since it is impossible to predict how many multiple terms in the satellite terminology will, in reality, arise.

No match in the satellite terminology is denoted by a grey box in column 2.

Column 1 (Class/Division/Section) are denoted by shaded boxes as follows:

Ten main classes	
Hundred divisions	
Thousand section	

Class/Division/Section	UNESCO term	DDC no.	DDC caption
	Computer science	000	Computer science, information & general works
	Information	000	Computer science, information & general works
	Philosophy	100	Philosophy & psychology
	Psychology	100	Philosophy & psychology
	Religion	200	Religion
	Social sciences	300	Social sciences
	Linguistics	400	Language
	Science	500	Science
	Technology	600	Technology
	Arts	700	Arts and recreation
	Literature	800	Literature
	History	900	History & geography
	Geography	900	History & geography
		010	Bibliography
	Information sciences	020	Library & information sciences
	Information	030	General encyclopedic works
	UNASSIGNED	040	UNASSIGNED
	Information	050	General serial publications
	Information	060	General organizations and museology
	Journalism	070	Documentary media, educational media, news media; journalism; publishing
	Publishing	070	Documentary media, educational media, news media; journalism; publishing
	Educational media	070	Documentary media, educational media, news media; journalism; publishing
	Information	080	General collections
	Manuscripts	090	Manuscripts, rare books, other rare printed materials
	Rare books	090	Manuscripts, rare books, other rare printed materials
	Metaphysics	110	Metaphysics
	Epistemology	120	Epistemology, causation, humankind
	Parapsychology	130	Parapsychology and occultism
	Spiritualism	130	Parapsychology and occultism
	Philosophical schools	140	Specific philosophical schools and viewpoints
	Psychology	150	Psychology
	Logic	160	Logic
	Ethics	170	Ethics (Moral philosophy)
	Philosophy	180	Ancient, medieval, eastern philosophy
	Philosophy	190	Modern western and other non-eastern philosophy

	Trade	380	Commerce, communications, transportation
	Communication and development	380	Commerce, communications, transportation
	Transport	380	Commerce, communications, transportation
	Trade	381	Commerce (Trade)
	International trade	382	International commerce (Foreign trade)
	Postal services	383	Postal communication
	Telecommunications	384	Communications Telecommunication
	Railway transport	385	Railroad transportation
	Inland water transport	386	Inland waterway and ferry transportation
	Maritime transport	386	Inland waterway and ferry transportation
	Maritime transport	387	Water, air, space transportation
	Air transport	387	Water, air, space transportation
	Transport	388	Transportation Ground transportation
	Metrology	389	Metrology and standardization
	Customs and traditions	390	Customs, etiquette, folklore
	Etiquette	390	Customs, etiquette, folklore
	Folklore	390	Customs, etiquette, folklore