A Multilingual Geographic Feature Classification Index for China and Japan

Merrick Lex Berman
China Historical GIS
Harvard Yenching Institute
September 2002

Abstract:

For an experimental integration of unrelated gazetteer standards, a basic index of Geographic Feature Classifications has been compiled from four Chinese National Standard Tables published in Beijing. The harmonized Beijing list was then crosswalked with the Topographic Materials Codes list (Taipei), and the Japanese Cartographic Feature Codes list, using Alexandria Digital Library Gazetteer Feature Type Thesaurus as the control vocabulary. The feature classifications in the index were also analyzed to determine the degree of semantic equivalence between feature types in the different lists, including indirect matching based on a the relationship to a single control vocabulary type, as well as direct matching between individual types.

1.1 Defining Feature Classification

When developing datasets of geographically-oriented information it is critically important to clearly identify the types and subtypes of data that are being recorded. This is not type classification in the sense of defining a data type as "text" or "numeric" but rather developing classifications to distinguish between *temples*, *mountains*, *burial grounds*, *piers*, *libraries*, and so forth. Many attempts have been made to create domain ontologies to handle such classifications, but the general consensus is that no single standard is capable of defining in advance all of the possible classifications and scales of definition that we will need to use.

Take for example a classification used in one such system: *religious facilities*, which is defined as the preferred term for a number of possible narrower terms: *cathedrals, chapels, churches, convents, monasteries, mosques, novitiates, retreats, sanctuaries, synagogues, tabernacles, temples*. [ADL-Feature] The narrower terms provide a wide enough range for most purposes, and also provide us with a sufficient array of concepts for *religious facilities* so that we would know to use this term if we had to classify *lamasery*. But what if our own work necessitated a clear and unambiguous differentiation between *lamasery, nunnery, lama's dormitory, nun's dormitory, lama's lecture hall, nun's lecture hall, co-ed lecture hall*. If we bundled these all under the term *religious facilities* the purpose of our categorization would be defeated.

One way to think of this is in terms of scale and scope. Whereas, on a national or regional scale (say greater than 1:200,000), and in the scope of general features, it would be enough to distinguish *religious facilities* from *bridges* and *hospitals*. But within the narrower scope of *religious facilities* we might want to distinguish between a *lamasery* and a *nunnery* and so on, regardless of scale. By contrast, if our interest was in the arrangements of shrine objects inside of temples, we might set an applicable scale definition of less than 1:100, and then create classifications such as *incense brazier, candle-holder, gong, lotus throne statue*, and so on.

In both examples, the narrower classifications could be considered as "sub-types" of a broader term, *religious facilities*. This reflects the typical hierachical nature of many feature classification systems. Using one such system to classify the term *pagoda*, we are directed to use the preferred term *tower*. But what if we wanted to classify a more specific type, say a *pagoda of sacred relics* (*bao ta*) ? In this case the *pagoda* unquestionably falls in the *religious facilities* type, and yet by definition must be classified as a *tower*. Here is a case where we must allow our *pagoda* to be associated with multiple types.

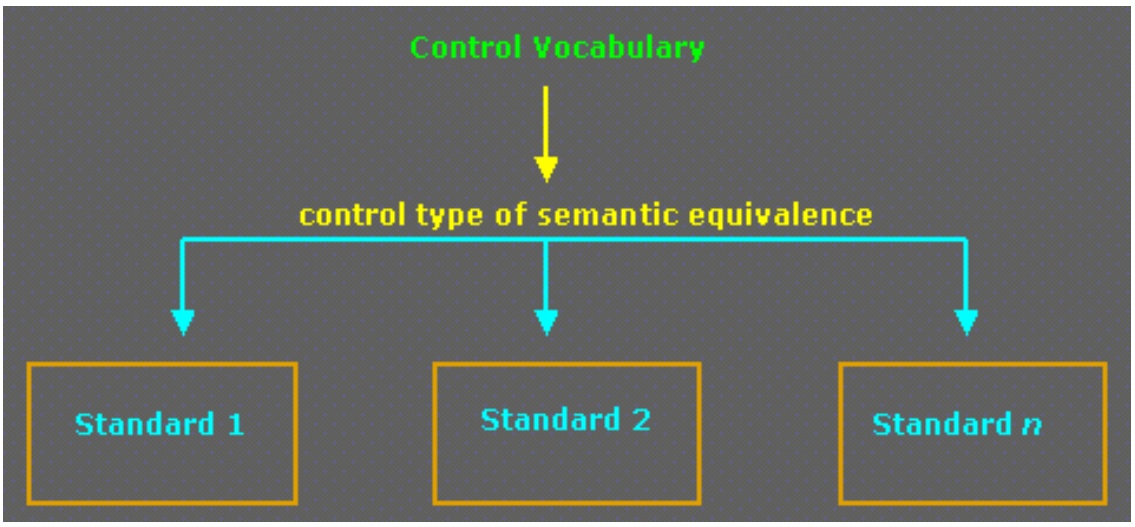
Similar to the problem of relating one feature to multiple types in a single system, there is the problem of trying to associate fixed types from one system to those in another system. The process of associating types from multiple classifications systems to one another, known as "crosswalking," is presented below, together with the results of an experimental Chinese, Japanese, and English crosswalk.

1.2 Feature Classification Naming Conflicts

Let's say that we wish to classify the "dock" feature shown below. On the one hand we might just use our own classification *dock*, or then again, we might want to use a controlled vocabulary of terms, such as those found in a national standard list of cartographic features. Shown on the right, are the terms that we would find if we were to look up the term for *dock* in the government-issued cartographic standards lists published in China, Japan, and Taiwan. The vernacular scripts, the words themselves, are quite different, even though they all are used to classify the same feature. These terms have semantic equivalence, convey the same meaning, but if we do not create a direct association

between these types, we would find it next to impossible to encounter the terms *sanbashi* or *chuanwu*, if we were searching for the term *matou*.

This particular type of semantic heterogenity falls into the category of naming conflicts, which is particularly accute when dealing with data sources in different languages. How can we solve the naming conflicts? The basic idea is to work with each standard separately and to associate all of the terms within each standard to one (or more) of the types found in a control vocabulary. Using this simplified hybrid ontology approach, [Wache, 2001] terms from *standard 1* that are matched to control vocabulary type X, will be indirectly associated with any terms from *standard 2* and *standard n* that also have been matched to type X.



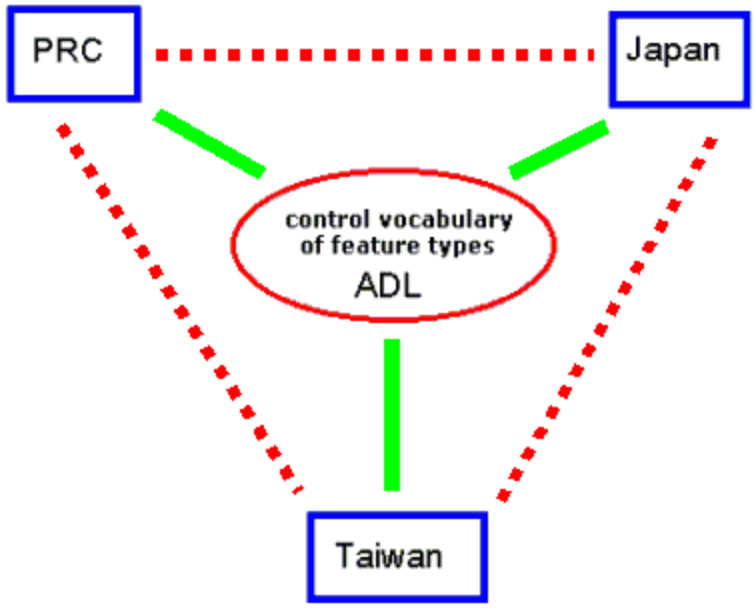
Assuming that we have already completed the matching of types from each standard to a list of types in the control vocabulary, we can now examine the correlations between the standards for any given type in the control list. So for the hypothetical control type X we can see the related types in *standard 1*: 1.A, 1.B, 1.C; in *standard 2*: 2.J, 2.K; and in *standard n*: n.X, n.Y, n.Z. Furthermore, we can now analyze these to see if there are any direct matches, such as *1.B = 2.K = n.Z*, or (*matou = chuanwu = sanbashi*).

In the next section, this model is tested using actual national standards from China, Japan, and Taiwan.

2.1 Crosswalking Existing Standards

For the experimental crosswalk, the control vocabulary used was the Alexandria Digital Library Feature Type Thesaurus (070302 Version) . [ADL-Feature] For the national standards, first a harmonized list was made using four, largely overlapping Specifications for Cartographic Symbols and Chart Symbols, published by the China Bureau of Scientific Standards in Beijing. [GBT-5791] [GBT-7929] [GBT-12319] [GBT-13923]

This harmonized list was crosswalked to the ADL Feature Types, after which the Taiwan Topographic Feature Classification Codes [TW-bianma] and the Japan Base Map Cartographic Symbols [JP-zushiki] were crosswalked separately.



The figure on the left summarizes the process—first, each type found in the separate lists are matched to one or more types in the control vocabulary list (solid green lines). The next step is to focus on a single type in the control vocabulary and compare the subsets of indirectly assocated types in the crosswalked lists. Finally, when indirectly associated types in **all three lists** have semantic equivalence, they are given a direct match (dashed red lines).

Let's look at a concrete example from the crosswalked lists—the control vocabulary term: "piers."

PRC	Japan	Taiwan
码头		
固定顺岸式码头	桟橋（鉄・コンクリート）	船塢
固定堤坝式码头	桟橋（木製・浮桟橋）	渡船碼頭
浮码头	渡舟発着所	湖濱碼頭
趸船式码头		海濱碼頭
栈桥式码头		
引桥式码头		
栈桥式码头		

We can see that 8 types in the PRC list were matched to the ADL control type "piers," and therefore were indirectly associated with 3 terms in the Japan list, and 4 terms in the Taiwan list. Of these terms, we can see a direct match on the concept of *dock*, or *pier*, or a place to tie up a boat that projects from the shore. However, the remaining terms all have shades of meaning that make them dissimilar enough to prevent us from creating any more three-way matches. We could make several two-way matches, (such as *PRC.floating dock = JP.wood or floating pier*), but for the purposes of this study only three-way matches were established as direct matches.

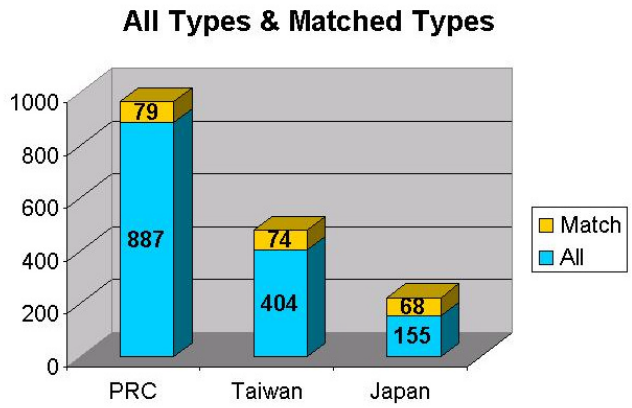
It should be noted that there were some cases in which a type was used in **more than one** direct match.

vernacular	transliteration	translation	semantic category	ADL type
石油井	shi you jing	oil well	oil well	wells
油井・ガス井	yusei gasusei	oil well, gas well	oil well	wells
油井	you jing	oil well	oil well	wells
天然气井	tian ran qi jing	natural gas well	natural gas well	wells
油井・ガス井	yusei gasusei	oil well, gas well	natural gas well	wells
瓦斯井	wa si jing	gas well	natural gas well	wells

An example is shown above, where the type *yusei – gasusei* was clearly a composite of two types that were given separate classifications in the two other lists. Therefore the number of direct matches varies from one list to another.

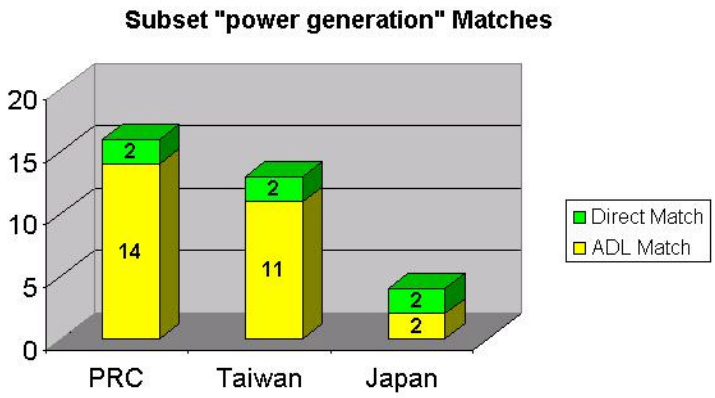
All in all, the amount of work involved in processing the original matches of classifications within a single list to the control vocabulary was much lighter than the second step in which direct matches had to be determined. A description of the results is provided in the following section.

2.2 Results of the Crosswalk



The ratio of direct matches to the total number of feature types in each source list is shown in the chart on the left. Naturally, if a list of terms is more extensive and more specific, it will have a smaller percentage of direct matches compared to the total number of items in the list. And a shorter list of more general terms will have a higher percentage of direct matches. Note that the number of direct matches varies from list to list, owing to the cases in which a single type in one list was used for more than one direct match, as described previously.

Let us drill down to a single item in the control vocabulary, *power generation sites*, and examine the ratio of indirect matches to direct matches.



Here we can see that there were two feature types in the Japan list that were matched to *power generation sites* in the control vocabulary, and that both of these had direct semantic equivalents in the China and Taiwan lists. However, the China and Taiwan lists had a considerable number of additional *power generation sites* features, (some 80 to 85% of the total), which could not be matched due to the limitations of possible direct equivalents in the Japan list. We can surmise that this case would hold true whenever we begin to match highly specific classifications to a limited set of terms in a control vocabulary. No matter how well planned the domain ontology of the control vocabulary, the discrepancy between an existing set of defined terms and newly devised classifications for a specific purpose will always exist.

3.1 Utilizing Crosswalked Feature Classifications

Now that we have a multilingual index of terms crosswalked to a control vocabulary, what is the best way to utilize this information? One possibility is to dispense with the direct matches altogether. Because the number of semantically equivalent types varies so much from one source list to another, and because the complete list of direct matches is merely a distorted abridgement of any source, the additional labor of producing the semantically equivalent types can be dispensed with. Another argument against working on direct matches is that our value judgement will always be questionable at best. The question, 'is a *sacred relic pagoda* = *pagoda*?' reminds us of 'is a *white horse* = *horse*?' , and should probably be left well enough alone.

Why not just match our source types to as many terms in the control vocabulary as necessary? There is no need to comb through the source lists and wonder whether *dock* = *metal or concrete pier*. Associate both source types with the control type *piers*. Then if we searched for all indirect associations for the control vocabulary type, *piers*, the resulting subsets from our source lists would contain *dock* and *metal or concrete pier*. In other words we wouldn't miss anything owing to the lack of a direct semantic equivalence. Below is an example showing the result of a search for all types associated with the control vocabulary term *power generation sites*.

PRC	Taiwan	Japan
发电厂	水力發電廠	変電所
发电站	火力發電廠	送電線
变电室	核能發電廠	
变电所	變電所	
变电站	變壓箱座	
风磨房	公共事業網路	
风车	線路	
通信线	輸送線(高壓線)	
电力线	配電線(電力線)	
杆上的电力线	高壓線塔	
塔上的电力线	電線桿	
地面下的电力线	单线铁路	
电线杆上的变压器		
电线入地口		

In this example we see the actual list of feature types from three source lists. In the future, we may wish to incorporate numerous source lists into the search process, and it may be desirable to add an intermediary step showing the numerical results for each source list, rather than the actual list of terms. For instance our search may begin with a known term from any of the source lists that have been matched to the control vocabulary, such as the Japan term: *hendensho* . The prelimary result of a search for *hendensho* should include:

(a) the control vocabulary terms that *hendensho* has been associated with

(b) the number of indirectly associated types found in the other source lists.

For example, a search for *hendensho* results in:

Control vocabulary type = *power generation sites*

Crosswalked source lists results:

PRC National Standard List – 14 *power generation sites* (out of 887 total types)

Taiwan NGIS Standard List – 11 *power generation sites* (out of 404 total types)

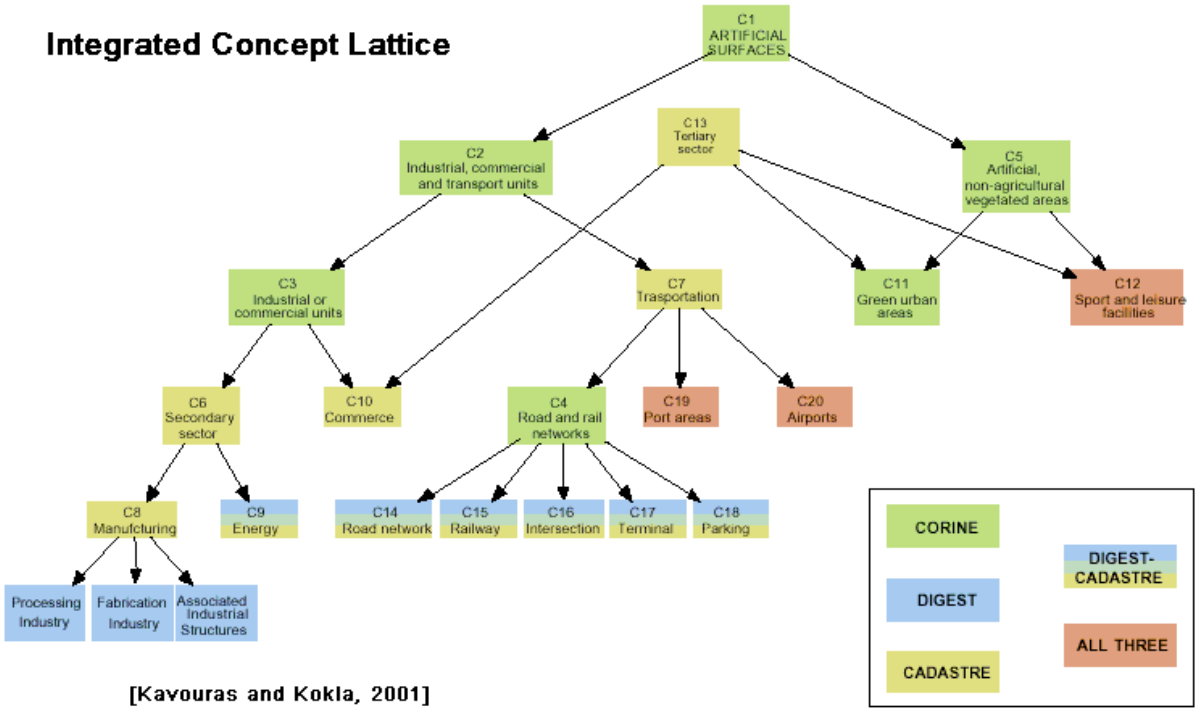
In this way the user is presented with the quantitative results of possible matches from each list. An additional advantage, is that it doesn't matter what type of source lists are added to the crosswalk. All that is needed is a many-to-many linktable between the control vocabulary list, and a cumulative list of all other types. Searches only return relevant hits, therefore an unlimited number of highly specialized feature classifications can be added to the cumulative list of crosswalked terms,enriching the total spectrum of available feature types for the users.

Fortunately, the implementation of the Alexandria Digital Library Gazetteer Content Standard [ADL-Gaz] allows for the entry of multiple classifications for any given feature. Each feature classification being

entered must also identify the source list from which it is taken. Therefore, if at least two type entries were made for each feature, one from the original source list, and one from a control vocabulary (such as ADL FTT) [ADL-Feature] , the result would effectively duplicate the crosswalk technique outlined above. The development of a cumulative crosswalk of multiple feature classification lists would be a valuable tool for researchers compiling new datasets. Searching the cumulative list will facilitate discovery of specialized terms and those projects where they are already in use. In addition the resulting cumulative list would be an excellent basis for a formal concept analysis, outlined in the next section.

3.2 From Crosswalk to Formal Concept Analysis

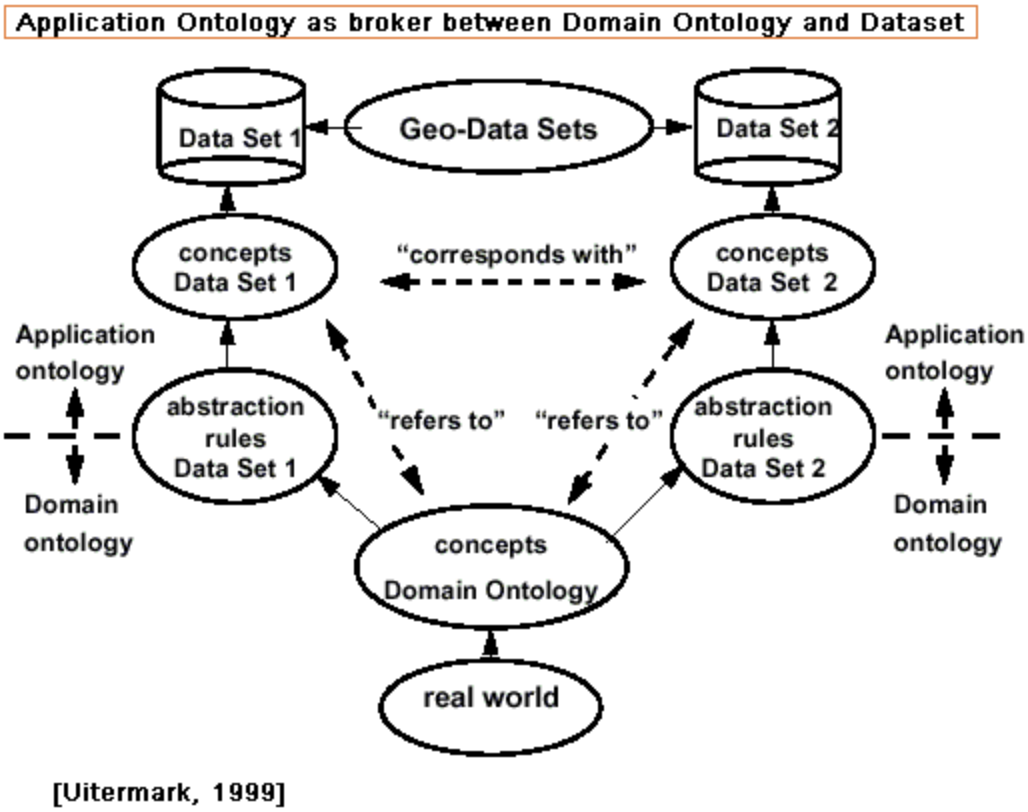
One way in which the crosswalk model discussed here can be used as the basis for future research would be to map semantically equivalent types between source lists (admittedly, an idea rejected as too time-consuming in the previous sections). The idea would be to examine the cumulative list of crosswalked terms in order to find the minimum subset of basic semantic elements they all share. This process, called formal concept analysis, involves semantic factoring of overlapping classifications and the construction of a concept lattice.



The arrangement of terms in the concept lattice is then developed into a single integrated model, for which all of the geospatial concepts and their relations have been defined. Kavouras and Kokla have completed an experimental model of this type, using the CORINE Land Cover Nomenclature, [CORINE]

DIGEST nomenclature for geographic objects, [DIGEST] and the Greek CADASTRE land use terms [HEMCO] as the source lists. [Kavouras -Kokla, 2001]

Formal concept analysis, when used to integrate source lists of geospatial feature types, will produce a new domain ontology for geographic feature classification. The domain ontology becomes, in effect, the new control vocabulary for integration of geographic information using functional logic. To perform this kind of integration, the basic set of concepts in the domain ontology are used to build application ontologies, each of which serves as the functional intermediary for semantic mapping between a specific source list and the domain ontology.



This has been put into practice by Uitermark, van Oosterom, Mars, and Molenaar using two topographic datasets of the Netherlands. [Uitermark, 1999]

Formal concept analysis and development of application ontologies are quite beyond the scope of the crosswalk presented here. Even so, the examples cited above point to interesting future directions for making use of our cumulative multilingual, historical index of feature classifications.

ONLINE RESOURCES FOR THIS PAPER:

PDF of Paper:
http://www.fas.harvard.edu/~chgis/work/docs/papers/lex_pnc_osaka_081602.pdf

Index and data dictionary:
http://www.people.fas.harvard.edu/~chgis/work/downloads/faqs/ADL_xwalk_faq.html

Search Engine:
http://chgis.fas.harvard.edu/tools/xwalk/xwalk_dual.php

Sources:

[ADL-Feature] Alexandria Digital Library Gazetteer Feature Type Thesaurus, Version 070302 http://www.alexandria.ucsb.edu/~lhill/FeatureTypes/ver070302/index.htm

[ADL-Gaz] Alexandria Digital Library Gazetteer Content Standard http://alexandria.sdc.ucsb.edu/~lhill/adlgaz/

[CORINE] Co-ordination on Information of the Environment, Nomenclature. http://reports.eea.eu.int/COR0-landcover/en

[DIGEST] Digital Geographic Information Exchange Standard, Feature Attribute Coding Catalog. Version 2.1 Part 4. http://www.digest.org/DownloadDigest.htm

[GBT-5791] China Bureau of Scientific Standards "Specifications for cartographic symbols on 1:5000 and 1:10000 topographic maps" *[1:5000 and 1:10000 dixingtu tushi GB/T 5791-93]*. Beijing: Biaozhun Chubanshe, 1993.

[GBT-7929] China Bureau of Scientific Standards "Specifications for cartographic symbols on 1:500, 1:1000 and 1:2000 topographic maps" *[1:500, 1:1000 and 1:2000 dixingtu tushi GB/T 7929-95].* Beijing: Biaozhun Chubanshe, 1995.

[GBT-12319] China Bureau of Scientific Standards "Symbols, abbreviations, and terms used on Chinese Charts" *[Zhongguo haitu tushi GB/T 12319-1998].* Beijing: Biaozhun Chubanshe, 1998.

[GBT-13923] China Bureau of Scientific Standards "Classification codes for national land information" *[Guotu jichu xinxi shuju fenlei yu daima GB/T 13923-92].* Beijing: Biaozhun Chubanshe, 1993.

[HEMCO] Hellenic Mapping and Cadastral Organization http://www.okxe.gr/

[Kavouras-Kokla, 2001] Kavouras, Marinos and Kokla, Margarita. "Ontology-Based Fusion of Geographc Databases." Athens: Dept of Rural and Surveying Engineering, Natl Technical University of Athens.http://www.survey.ntua.gr/main/labs/photo/laboratory/news/pdf/S43.%20Marinos%20Kavouras,%20Margarita%20Kokla.pdf

[TW-bianma] Taiwan Ministry of the Interior "Basic Topographic Map Materials Topographic Feature Classifications Code Table"*[Jiben dixingtu ziliaoku dixing ziliao bianma biao]*. http://ngis.moi.gov.tw/doc/stan/proper%20noun..htm

[Uitermark, 1999] Uitermark, Harry, et al. "Ontology-Based Geographic Dataset Integration," in Bohlen, et al, editors, "Proceedings of the International Workshop on Spatio-Temporal Database Management, 1999." Berlin: Springer, 1999: 60-78.

[Wache, 2001] Wache, Holger, et al. "Ontology-Based Integration of Information--A Survey of Existing Approaches." Seattle: Proceedings of the IJCAI Workshop: Ontologies and Information Sharing, 2001.