Semantic Interoperability and Cultural Specificity:
Examples from Chinese, Japanese, Mongolian and Uighur

Merrick Lex Berman
China Historical GIS
Harvard Yenching Institute
November 2003

Abstract:

Creating a database of historical Chinese administrative units for the CHGIS project required the definition of specialized feature classifications. Although limited to the domain of Chinese language materials, these classifications revealed variability over time, region, and ethnic or cultural dimensions. To explore the variable usage of feature classification terminology across linguistic and regional boundaries, the official cartographic feature classification systems for China, Taiwan, and Japan were integrated with the CHGIS historical types by mapping the types from all of the separate lists to those found in a designated control vocabulary.

Here the process is extended to include traditional Mongolian and Uighur geographic feature types. The culturally specific terms found in the Mongol and Uighur entries demonstrate the general truth, which is that the majority of existing geographic classification systems remain highly unstructured, ad-hoc agglomerations of terms, thrown together for the expedient purposes of each individual project. Not until sufficient accumulation of geographic terms across many languages and disciplines is done, will we be able to propose the deep structure of a truly generic ontology for geographic features.

1.1 Overview of Ontologies

The specific approaches to semantic interoperability among geographic feature classification systems addressed in this paper will make more sense if we begin by stepping back a few paces to take an overview of the general terminology and concepts being used in the field today. The bulk of this overview is drawn from the work of a leading center of expertise, Ontogeo, where many years of research have already been completed. **[Ontogeo]** The gist of Ontogeo's research has been to discover the steps in the process of integrating various classification systems, and to come up with a formal language and methodology for doing so.

Ontologies have been recently defined as "theories that use a specific vocabulary to describe entities, classes, properties, and functions related to a certain view of the world. They can be a simple taxonomy, a lexicon or thesaurus, or even a fully axiomatized theory." **[Fonseca, 2002]** What differentiates an ontology from a thesaurus is that the ontology is not only devised as a means of relating existing terms to one another, but seeks to define specific concepts within a domain of knowledge, and to define the relationships between those concepts. **[Beck, 2002]**

1.2 Informal to Formal Ontologies

The ontologies which we commonly come across have been formulated for various purposes, and don't conform to any particular standard. Therefore we differentiate them according to several types ranging from the informal to the formal.

The simpler taxonomical types are called informal ontologies, and generally exist as a list of terms, arranged in hierarchical relationships. If natural language definitions are provided for each term in the hierarchy, they are called terminological ontologies. When the relationships between each term have been explicitly stated in formal logic or as axioms that can be processed programmatically, we are dealing with formal ontologies.

Unfortunately or fortunately for us, most of the ontologies which are currently in use for describing geographic features are very informal, ad-hoc collections of terms which were devised for a specific purpose. My position, explained in the following sections, is that we can take better advantage of informal ontologies at first, as we build up a network of related concepts, and that this work is prerequisite to establishing finer distinctions between classes and individual concepts.

1.3 Granularity

In addition to their degree of formality, ontologies are typically classed according to granularity, from the general to the specific. The most general ontologies, or top level ontologies, deal with concepts not specific to any one domain. These might deal with time, for example, providing a framework for concepts related to the measurement of time. Domain ontologies are those related to a specific domain of knowledge.

Also included in the granularity scale are application ontologies, which essentially combine the definitions provided within a specific domain ontology with a set of functional rules for performing a process or task upon the classes and individual terms in the domain ontology. Related to the application ontology is the task ontology, which defines specific tasks and their sequences or procedures.

Finally, the type of ontology which is derived from the common elements occurring across multiple ontologies is called a meta ontology, or is alternately referred to as a core ontology or generic ontology.

2.1 To Merge and Integrate Ontologies

Before discussing the actual processes involved in semantic mediation, we need to define the components being input and output. Each ontology being input, or analyzed and decomposed in terms of the others, is called a source ontology. After the analysis and comparison is completed, the final output which combines elements of all the source ontologies is referred to as the shared ontology.

Shared ontologies are roughly differentiated into two types: merged ontologies and integrated ontologies. In a merged ontology, individual terms from several source ontologies have been either mapped as equivalent, or have been slightly altered to establish equivalence. This is a case in which the actual terms used, or the meanings associated with them may have been changed, therefore the resulting terms in the merged ontology might not be the same as those found in the source ontologies. Merged ontologies encompass cases of partial compatibility, where some sections or terms of the source ontologies have been unified, or cases of unification, in which each class and term from the source ontologies have been forced to become fully compatible with the others. In both cases the resulting merged ontologies contain distortions of the structural elements of the source ontologies and are no longer functional according to their original hierarchical arrangements.

Integrated ontologies are those in which each class and individual term of the source ontologies have been preserved intact, but are rearranged into a new all-encompassing hierarchy along with some additional concepts and relationships to create a functional whole. With true integration, the original terms are not changed or distorted and are useable as separate components or as integrated members of the new ontology.

An example of true integration was completed by Ontogeo using DIGEST, **[Digest]** CORINE, **[CORINE]** and CADASTRE **[HEMCO]** as the source ontologies. The Ontogeo approach tackles the problem of semantic mediation between multiple ontologies using a formal concept analysis and development of an underlying concept lattice that contains all of the classes and individual terms from the source ontologies. **[Kavouras, 2001]**

2.2 Integration Process

The methodology of integration begins with the extraction of terms, definitions, and relationships from the source ontologies. The extraction of semantic information from the sources raises a host of problems in itself, not the least of which are caused by the original format of the various source ontologies, and the degree to which each class and term has been defined and related to one another. In the test case scenario described later, the extraction process has been greatly simplified, (perhaps oversimplified!), to demonstrate an alternative process that creates loose mappings and indirect semantic equivalence as a first step.

When all of the extracted concepts have been obtained, they undergo a rigorous comparison, which seeks to determine the degree of semantic similarity and heterogenity between the concepts. The semantic conflicts or distance between concepts in the source ontologies is complicated by issues of scope, in which the source ontologies were either more general or more specific than one another. An example would be a general ontology which makes such distinctions as continental land mass, ocean, lake, and island, as compared to a domain ontology for harbors which includes distinctions between fixed piers, floating piers, bouys, shipping lanes, pleasure boat anchorages, etc.

Further difficulties arise in dissimilar relationships among terms in the source ontologies. One source may allow multiple inheritance, while another allows only single inheritance. One source may approach geographic features as types of land cover and land use, while another may approach the same features as types of parcels defined in cadastral terms.

Last but not least we must try to sort out the semantic differences between the concepts, and this proves in many ways to be the hardest task of all. The fact is that within a particular language there is a great degree of ambiguity in the definition of terms related to geographic features, and when the task is expanded to include changes in concepts and how they are applied over time, or in particular dialects or regions it becomes murkier still. Once we acknowledge the difficulty of the task, we can only be struck numb by the seemingly hopeless complications that arise when we try to work across languages, cultures, centuries, and continents! A concrete example of this is given in the following section 3, where source ontologies from four Asian languages are integrated using an English language source as the controlled vocabulary.

2.3 Methodology of Semantic Comparison

Before turning to the testbed example, let's consider the parameters of the semantic comparison process as defined by Ontogeo. Drawing from existing models developed in computational linguistics, concepts are compared in terms of equivalence, overlap, relatedness, and disjoint. **[Kavouras, 2003]** Of these parameters it is obviously easier to test for equivalence and disjoint as true and false statements, than it is to define overlap and relatedness.

For example, we can make a straightforward conclusion about whether the term wooded area is equivalent to forest, and is disjoint with airport. Less clear is what we mean to associate a geological classification of rock as an overlap with geographic terms for boulder or mountain. And when we discuss relatedness we quickly run into problems of how to characterize relationships as well as multiple inheritance. Should we classify the term pagoda, for example, as equivalent to (or a sub-type of) tower, religious site, historic site, building, ornamental architecture, library, or museum? In fact, some pagodas may function as examples of all of these things simultaneously.

Clearly the existence of pagodas, as enigmatic as they may be when we try to classify them, must be considered as valid geographic features, at least in the Chinese and Japanese contexts. In Chinese we find, in addition to the generic word ta (pagoda), that Buddhist sites occasional define them with narrower terms: bao ta (relic pagoda) and jing ta (classic pagoda). However, due to my own lack of expertise, I don't know if similar narrower terms exist in Tibetan, Japanese, or any other language. Surely such terms do exist, but until they have been made available in a way for me to discover them, I have no way of incorporating any of them into a hierarchy of terms made to accommodate relic pagoda and classic pagoda.

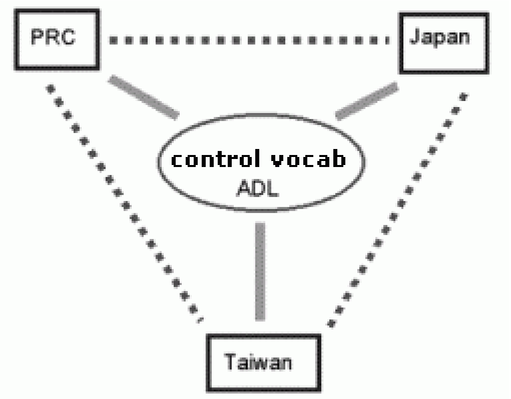
If the hierarchy of relationships that we devise is too specific, then a future mapping of possibly related terms may go undiscovered. In other words, if the Tibetan chorten, or Sanskrit stupa, were later designated as religious site, if we had carefully decided to classify relic pagoda and classic pagoda as sub-types of pagoda, and pagoda as a sub-type of tower, then we might never encounter them in the same context. Our pagodas would only be found when looking at terms related to towers, and our chortens found only when searching for religious sites.

The results of an experimental crosswalk dealing with these terms, described in the next section, lead me to believe that we should AVOID trying to make such finer distinctions about degrees of overlap and relatedness among terms too early in the comparison process. It makes more sense to absorb more examples, from the widest number of cross-linguistic and cross-cultural sources as possible, before working on the deeper structure of the concept lattice, because—like high energy physicists wondering what particles will next appear in the cloud chamber—we simply do not yet have an adequate grasp of the fundamental particles. Therefore in the first draft, I would advocate a looser and more flexible hierarchy in the resultant shared ontology, making use of multiple inheritance, as being much more desirable than a highly structured network of associations that formalize the overlap and relatedness among all the concepts.

3.1 Comparison Testbed

For the comparison of geographic features found in several unrelated domain ontologies, an English language control vocabulary (Alexandria Digital Library Feature Type Thesaurus) was used as a core ontology. **[ADL-Feature]** Several source ontologies were then analyzed, and each or the terms found in the source ontologies were mapped to a roughly equivalent term in the core ontology. The source ontologies included official cartographic feature standards lists from the People's Republic of China (PRC), Japan, and Taiwan; and a specialized list of historical features developed for the China Historical GIS project. **[Berman, 2002]**

The PRC list was compiled from 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]** Historical features from CHGIS were added to the PRC types. This harmonized list was then mapped 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 added separately.

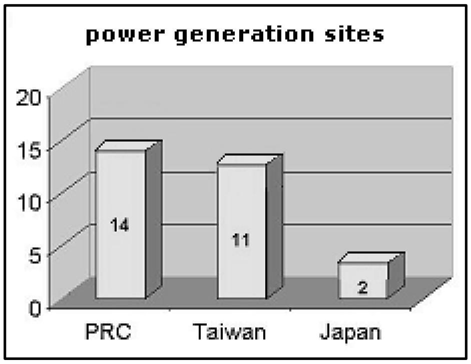


The figure on the left summarizes the process—first, each type found in the separate lists are matched to one or more types in the control vocabulary list (solid gray lines). Once all of the terms are mapped to at least one equivalent term in the control vocabulary, the terms in source ontologies acquire indirect semantic equivalence with one another (dashed black lines).

This process can be repeated with any number of source ontologies. Each time a new set of terms is added, new groups of indirectly related terms are discovered. By taking this approach we can grow the shared ontology organically over time, and each additional source that is added can acquire indirect semantic equivalence with all the others.

3.2 Examination of All Concepts Mapped to a Single Control Vocabulary Type

When the first mapping was done, using the source lists from China, Japan, and Taiwan, it was immediately apparent that the number of types from each list that could be mapped to a particular control type serves as a quantifiable measure of differences in scope between the source ontologies. For example, the ADL type "power generation sites," was mapped to:



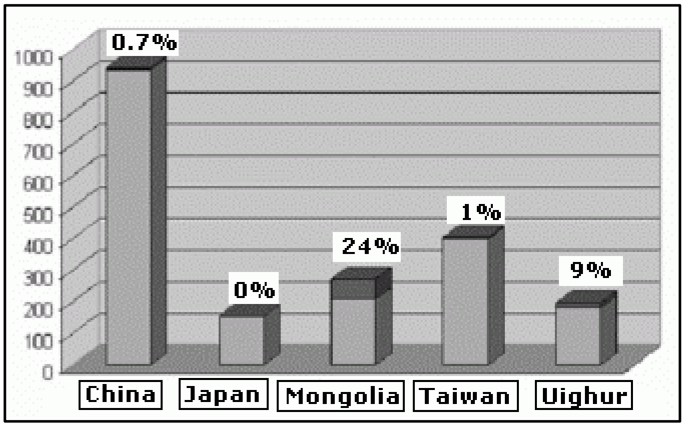
14 terms in the China list (1.5%)
11 terms in the Taiwan list (2.7%)
2 terms in the Japan list (1.3%)

Clearly the scope of the PRC list is wider and more detailed for concepts related to power generation sites than the Japanese list. Is this due to the fact that the Chinese list (945 terms) is simply larger than the Japanese list (155 terms)? This cannot be the only reason, because the percentage of the lists assigned to power generation is actually comparable (1.3 – 1.5%). If we examine only the total number of terms in each and the percentage mapped to power generation sites, then the Taiwan list would definitely appear to be more focused on power generation—with twice percentage of the total number of terms (2.7%) dedicated to that control type—than either of the other two lists.

This suggests an axiom: domain weighting of terms within a single source ontology can be measured by examining the percentage of the list dedicated to each control type in the shared ontology as compared with the percentages of the other source lists dedicated to the same control type.

In the example, the domain weighting is probably due to domestic politics and cartographic generalization more than anything else, since the lists of terms were largely derived from the official cartographic standards. But if the axiom holds true then it should apply equally well to other sources created for entirely different purposes.

This was tested by the addition of two completely unrelated and unofficial source ontologies for geographic features in the Mongolian and Uighur languages. Neither of these added sources contained terms mapped to the control type power generation sites. Therefore the domain weighting for this type was 0%. However, if we look at the control type mountains, the picture is dramatically different:



The weighting of terms related to the type mountains in the Uighur source was ten times greater than that of the China, Japan, Taiwan lists, while the Mongolian source was twenty-five times greater! The domain weighting among the sources shows a clear preponderance of mountain terms in the Mongolian source ontology.

The results indicate not only that domain weighting is measurable in the way suggested, but also that the more atypical or idiosyncratic particular classes and terms are within a particular source ontology, the more easily those idiosyncracies can be detected.

3.3 Evaluation of Equivalence – Disjoint

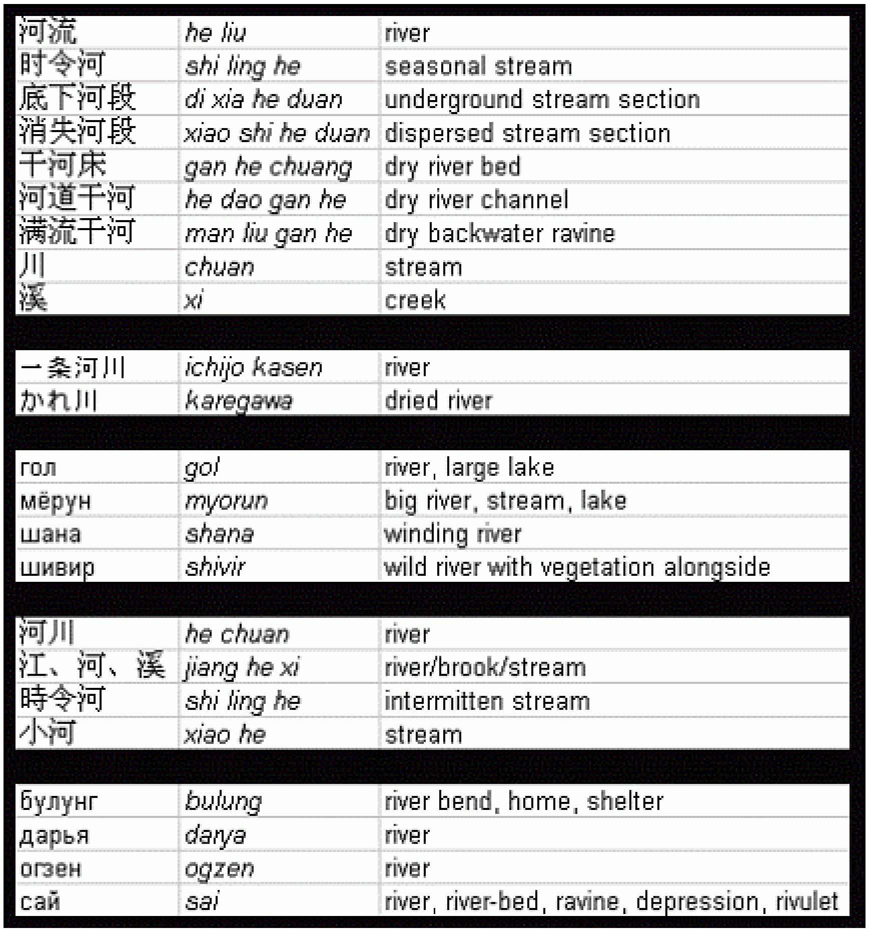
As mentioned in the previous sections, the methodology used for mapping each term in the source ontologies to a type in the control vocabulary has been deliberately over-simplified. In the mapping of any particular term to the control type mountains, for example, no time was spent trying to evaluate degrees of overlap or relatedness. The only criterion used was whether or not the source ontology term had sufficient semantic equivalence to justify a mapping. Therefore, all of the following terms were mapped to the control type mountains from the Mongolian list:

Serben khada - mountain with sharp drop on one side of the divide
Shiva - mountain with jagged toothed peaks
Altay – mountains with snow-covered peaks

To take another example, all of the following terms were mapped to the control type rivers:

Gol – river
Myorun – big river
Shana – bend in the river

In order to evaluate the parameters of equivalence, overlap, relatedness, disjoint among the terms that had all been mapped to a single control type, we examined all of the possible terms for rivers. These included 9 terms from the PRC list (1%), 2 terms from the Japan list (1.3%), 4 terms from the Mongolian list (1.5%), 4 terms from the Taiwan list (1%), and 4 terms from the Uighur list (2%). The control type rivers was chosen because of the generally equivalent percentages represented from each source.



河流	*he liu*	river
时令河	*shi ling he*	seasonal stream
底下河段	*di xia he duan*	underground stream section
消失河段	*xiao shi he duan*	dispersed stream section
千河床	*gan he chuang*	dry river bed
河道千河	*he dao gan he*	dry river channel
满流千河	*man liu gan he*	dry backwater ravine
川	*chuan*	stream
溪	*xi*	creek
一条河川	*ichijo kasen*	river
かれ川	*karegawa*	dried river
гол	*gol*	river, large lake
мёрун	*myorun*	big river, stream, lake
шана	*shana*	winding river
шивир	*shivir*	wild river with vegetation alongside
河川	*he chuan*	river
江、河、溪	*jiang he xi*	river/brook/stream
時令河	*shi ling he*	intermitten stream
小河	*xiao he*	stream
булунг	*bulung*	river bend, home, shelter
дарья	*darya*	river
огзен	*ogzen*	river
сай	*sai*	river, river-bed, ravine, depression, rivulet

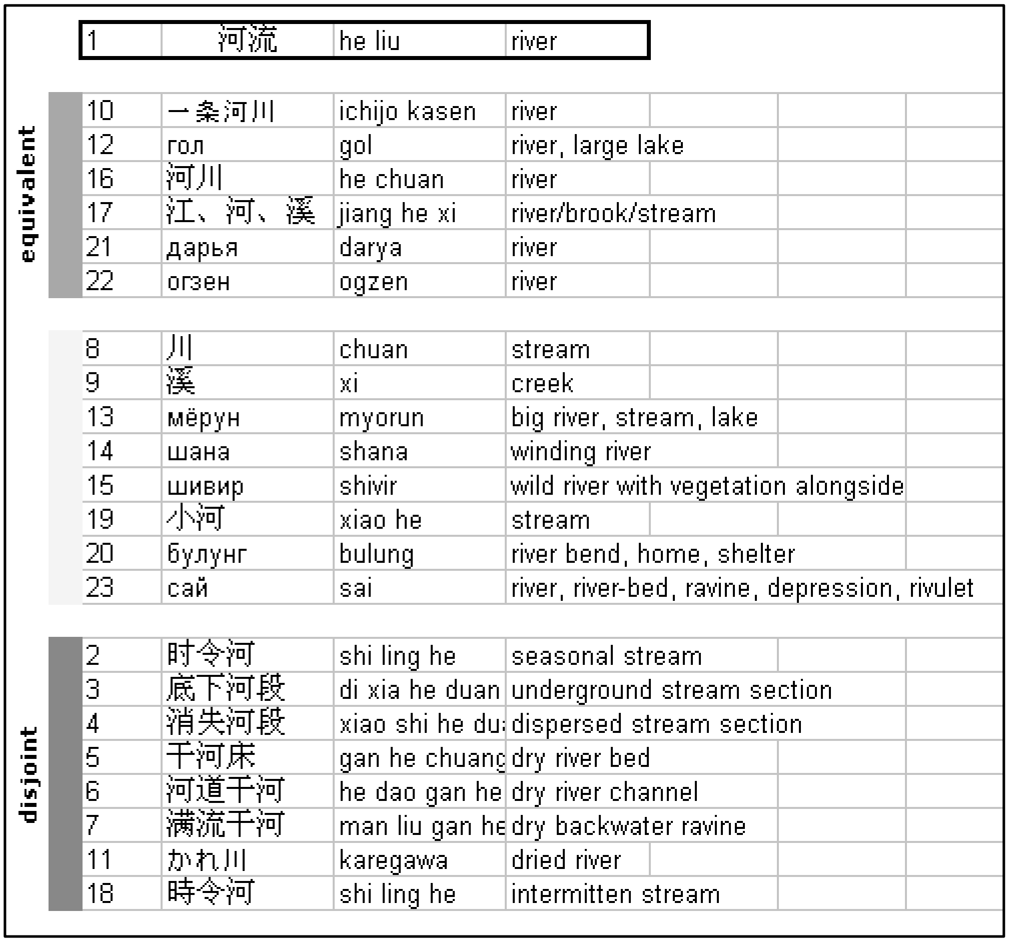
The next step was to attempt to select one term and evaluate its parameters on the equivalent – disjoint scale to all the other terms, including those drawn from the same source ontology.

Starting from the first term, he liu (river) in the PRC list, we found direct equivalent matches in all of the other lists (ichijo kasen, gol, he chuan, jiang he xi, darya, ogzen).

Determination of disjoints was also fairly straightforward. However, when attempting to evaluate overlap and relatedness, the issue became immediately unclear. To what degree does a dry river bed overlap river? To what degree is a dry river bed related to river? Is this a distinction that we actually want to measure?

The more time spent on the problem of trying to determine whether any of these types must be designated as overlapping or related to the other types, the more convinced I became that is was not necessary.

After all, the equivalence and disjoint were extremely straightforward: he liu (river) is equivalent to gol (river) and darya (river); while he liu (river) is disjoint from karegawa (dried river) and from bulung (river bend). And as for all of the items that had neither equivalence to nor disjoint from he liu, why shouldn't they fall into a temporary subset of general overlap or relatedness? The following table shows all terms mapped to the control type rivers rearranged according to their equivalence – disjoint to he liu:



	1	河流	he liu	river
equivalent	10	一条河川	ichijo kasen	river
	12	гол	gol	river, large lake
	16	河川	he chuan	river
	17	江、河、溪	jiang he xi	river/brook/stream
	21	дарья	darya	river
	22	огзен	ogzen	river
	8	川	chuan	stream
	9	溪	xi	creek
	13	мёрун	myorun	big river, stream, lake
	14	шана	shana	winding river
	15	шивир	shivir	wild river with vegetation alongside
	19	小河	xiao he	stream
	20	булунг	bulung	river bend, home, shelter
	23	сай	sai	river, river-bed, ravine, depression, rivulet
disjoint	2	时令河	shi ling he	seasonal stream
	3	底下河段	di xia he duan	underground stream section
	4	消失河段	xiao shi he du	dispersed stream section
	5	千河床	gan he chuang	dry river bed
	6	河道千河	he dao gan he	dry river channel
	7	满流千河	man liu gan he	dry backwater ravine
	11	かれ川	karegawa	dried river
	18	時令河	shi ling he	intermitten stream

The utility of this arrangement is obvious, for if we were to store the equivalent and disjoint mappings in a related table, it would be a simple matter to create this arrangement for any term that had been evaluated for equivalence and disjoint. Indeed, the addition of new source ontologies into the mix need not require a mapping to each and every term but only to those terms for which the evaluator was confident! Because, if we simply extend the mappings of equivalence in algebraic terms (if a = b, and b = c, then a = c), any term that was subsequently mapped as equivalent to darya, can by extension be known as equivalent to he liu!

If the determining factor of whether or not to add equivalent – disjoint values is based on confidence, then it becomes even more essential to allow some room for uncertainty. As shown above, the middle section contains terms of uncertain relatedness to he liu. They have indirect semantic equivalence, because they have all been mapped to the control type rivers, but they need not have any specific relationship to each other beyond that. In my view this loose mapping allows us the flexibility to expand the shared ontology in a simple and straightforward manner, while allowing for more specific equivalent – disjoint relationships to be established in a relational table as needed. In this way we can prepare the groundwork for formal concept analysis based on an ever-widening network of semantic interoperability, and at the same time avoid the creation of concept lattice structures that need to be deconstructed and remade every time a new source ontology is integrated.

URL: http://chgis.fas.harvard.edu/tools/xwalk

Sources:

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

[Beck] Beck, Howard and Helena Sofia Pinto. ”Overview of Approach, Methodologies, Standards, and Tools for Ontologies.” Agricultural Ontology Service, UN FAO, 2002. http://www.fao.org/agris/aos/Documents/BackgroundAOS.html

[Berman, 2002] Berman, Lex. “Multilingual Geographic Feature Classification Index for China and Japan.” Presented at Electronic Cultural Atlas & PNC Joint Meeting, Osaka, Japan, Sep 2002. http://www.fas.harvard.edu/~chgis/work/docs/papers/lex_pnc_osaka_081602.pdf

[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

[Fonseca] Fonseca, Frederico, Egenhofer, Max, Agouris, Peggy, and Gilberto Camara. ”Using Ontologies for Integrated Geographic Information Systems.” In *Transactions in GIS* 6(3), 2002. http://www.spatial.maine.edu/~fred/fonseca_TGIS_2002.pdf

[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, 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

[Kavouras, 2003] Kavouras, Marinos. "Unified Ontological Framework for Semantic Integration." Presented at Intl Workshop on Next Generation Geospatial Information, Boston, October 2003. http://ontogeo.ntua.gr/publications/boston-kavouras-presentation.pdf

[ONTOGEO] Geospatial Ontology Research Group http://ontogeo.ntua.gr/