# KOCIS-KD : A Model for Interoperating Korean Diaspora Information Collections

# KOCIS-KD: A Model for Interoperating Korean Diaspora Information Collections

**Douglas Avison BLACK**
*Consultant to East Rock Institute*

## 1. Introduction

Two basic problems occur in information retrieval in pursuing both general and specific inquiries. One is finding and/or creating information repositories and the other is organizing information to sort out what is relevant to the inquiry. Both problems are particularly acute when trying to do comparative analysis such as is often important for cross cultural study, particularly diaspora study. The design of the *Korean Online Cultural Information System for Korean Diaspora* (KOCIS-KD) addresses both these problems.

The development of KOCIS-KD began under the direction of Dr. Hesung Koh more than thirty years ago[1], and continues under her direction as a project of East Rock Institute (ERI) in New Haven, Connecticut, USA, which Dr. Koh founded and directs. KOCIS-KD (at its creation simply KOCIS), was unique in that it used digital storage and retrieval, but more importantly, it used a subject classification and information indexing scheme particularly relevant to Korean and Korean

[1] KOCIS began as the pilot project to computerize the classification system of Human Relations Area Files (HRAF) also in New Haven, Connecticut. Hesung Koh, as HRAF's Director of Information System and also later as Director of Research expanded the classification system to be multi-faceted and added generic and specific keyterm indexing and data quality control indexing to meet the needs of comparative studies of East Asian cultures. HRAF Automated Bibliographic System (HABS) and HRAF Cultural Information System (HACIS) are now implemented with Korean and Korean diaspora data and called KOCIS.

Diaspora studies. Primary features of the scheme are its multi-faceted classifications and its embodiment of the core theory of diaspora culture, with its triangular aspects of diaspora, homeland, and host country factors. Other key facet divisions are culture, subject, time, place, persons, organizations and particularly important, key terms, and data quality control. Key terms have the added dimension of being either culturally generic or culturally specific.

Describing in detail the scholarly relevancy of KOCIS-KD features to Korean Diaspora studies is better left to Dr. Koh. As a software development consultant to ERI, my intention in this paper is to discuss the significance of the KOCIS-KD design, especially in light of current, powerful developments in information technology and science. We will find that many early aspects of KOCIS-KD design were prescient of current developments and thus are particularly well positioned to take advantage of them. We will also find that still unique aspects of KOCIS-KD design can take these developments even further, particularly in ways that are especially valuable to Korean Diaspora studies.

The new developments in information technology and science that we are concerned with focus on developments in 1) the use of and construction of classifications and indexing schemes, particularly thesauri and ontologies, and 2) interoperability between information sources across the Internet and the World Wide Web, particularly as they relate to the development of the emerging Semantic Web. These developments provide both new opportunities for sharing information among scholars and their respective organizations and institutions and new responsibilities to participate in the sharing according to standards.

There are two other important areas of information technology and science developments that are not discussed in depth in this paper but need to be noted because of their importance to developing a global network for Korean Diaspora studies. These developments are reflective of the prescient concepts embedded in the 30 year plus history of KOCIS and thus on the suitability of the KOCIS model to take advantage of them. The first development is in the area of internationalization. Solutions for building multi-lingual software and web applications include the Unicode character set, efficient Unicode encoding schemes such as UTF-8, the multi-lingual thesaurus standard ISO 569, and automatic

locale sensitivity and resource property files in leading software application architecture models such as Sun's J2EE architecture. The second area is software development for organizing of complex objects such as multi-media websites or field data collections, including semantic web functionality. Very recently, interoperating object repository architectures and implementations, such as Fedora[2], are gaining unique momentum at world leading universities. Fedora is an open source, free, and robust product developed at Cornell University and the University of Virginia and supported by growing expertise at several other leading universities including Yale University. Such applications make a production level development of an interoperative Korean Diaspora global network modeled on KOCIS-KD finally a practical vision.

An example inquiry demonstrating the nature of the problem that *KOCIS-KD*'s design addresses might be trying to determine if there is a correlation between the degree of ethnic identity in a diaspora community and the economic status of the persons in the community. A further aspect of this inquiry might be to try to determine if the correlation is causative and, if so, in what direction.

Without prior knowledge, it is easy to imagine reasons for both positive and negative correlations between these factors. If there is a relationship, it is easy to imagine reasons why the correlation might be causative in either direction. Perhaps a strong sense of ethnic identity increases prejudice of the host culture and decreases opportunity for assimilation causing a negative correlation between strong ethnic identity and economic opportunity. On the other hand, perhaps a strong sense of ethnic identity encourages family and community support and stimulates the host culture to establish norms and laws protecting the rights of diaspora persons. Or perhaps, having achieved economic success, for reasons independent of ethnic identity, persons are inclined to strengthen their ethnic identity because they are more secure from prejudice. The inquiry is reasonable but the answers are elusive.

To begin such an inquiry one has to find markers of ethnic identity and markers of economic success within the diaspora community of interest. Determining and finding such markers is difficult enough but

[2] See Fedora project website: http://www.fedora.info/.

that's not all. If one looks only at one diaspora community, how do you know that the result isn't primarily related to the particular host culture? Or perhaps there are significant effects of different homeland cultures, even across the same ethnic population, depending on the location and time period of emigration or even of current events.

In brief, one must solve the two basic problems with which we began. One must find repositories of diverse information sources related primarily by the meaning of the inquiry, and one must be able to sort the information by markers of diverse factors. How the *KOCIS-KD* design solves these problems is the topic of this paper.

## 2. Organizing Information

### The Basic Goals of Information Query and Retrieval

Five basic goals of any information query and retrieval system can be identified.

1. Maximize retrieval of all relevant resources.
2. Minimize retrieval of irrelevant resources.
3. Reveal the overall conceptual dimensions of the subject matter.
4. Provide an efficient browsing system.
5. Refer to further information.

The first two goals I cite assume that the user has a strong sense of what they hope to find. These two goals are to maximize the completeness of the results and to minimize the retrieval of irrelevant results. Measurements of these goals are commonly referred to as *recall* and *precision*, and in most situations they are inversely related, such that strategies that increase recall decrease precision and vice versa.

The third and fourth goals assume that the user might not have a strong sense of what they are looking for. These goals require a revealing organization of the overall subject matter and an efficient browsing process.

The fifth goal is often overlooked in literature discussing query and retrieval systems but it is a constant part our everyday information processing and in fact integral to many common everyday systems. For instance, papers written for conferences usually aim at fulfilling all five of the above goals, with the fifth goal supplied by references. On the web, the

major search engines focus on this fifth goal to the exclusion the other four goals.

KOCIS-KD aims at the first four goals through the use of a multi-faceted thesaurus as we will discuss below. KOCIS-KD aims at the fifth goal, which equates to interoperability between information collections, by planning to extend the use of this thesaurus to operate on the emerging Semantic Web.

## Material Types

All five goals are affected by the type of information, often called *material type*, in the collection. The KOCIS-KD data collection is currently about 200 full text articles with images from the ERI's Korean and Korean American Studies Bulletin. Prior versions of KOCIS-KD not yet integrated into the current version include significant bibliographic records. Ultimately in addition to integrating the independent bibliographic data we plan to include full size monographs and multi-media data including images, sound recordings, and videos and interactive demonstrations and learning tools. The question we ask here is how should this information be organized and queried?

## Query Systems

### Word-based Searching

The most common electronic querying systems available are keyword and text in context systems. I distinguish between keyword and text in context where by keyword I mean some open or closed set of words available in some type of metadata field such as author, title, or even keywords, versus any text, in un-fielded, full text, discursive articles, monographs, etc. Often both of these forms of searching are confusingly called keyword searching in the literature. I will use the term *word-based* searching.

Word-based searching has two basic problems. First, finding words is not equivalent to finding relevant concepts. The word may be a *homonym* of the intended word and have an entirely different meaning, such as *running* to mean the recreational or competitive sport and *running* as in

*running* water. Because of this problem word-based searching reduces both the recall and precision measurements of our first to basic goals.

Second, word-based searching reveals nothing about the conceptual dimensions of the subject of an information collection and provides no system for browsing the collection, thus failing at our third and fourth main goals.

### Subject Indexed Searching

The general alternative to word-based searching is subject indexed searching or a combination of indexed and word-based searching. KOCIS-KD uses such a combination. There are several forms of indexed searching. The simplest kind is that such as in back-of-the-book indexing, which provides an unrestricted list of terms and links to locations in a data collection. A more sophisticated form uses authority lists of terms with references from an unrestricted list of terms, such as an authority list of author names in a preferred format, with references from other formats or other names used for the same person. Still more sophisticated is a more completely structured set of relationships between preferred terms, as well as a referring list of non-preferred terms.

The thesaurus structure is the most common form used for such structured relationships. There are both single and multi-faceted forms of thesauri. KOCIS-KD uses a multi-faceted thesaurus.

## Single and Multi-Faceted Thesauri

### Multi-Faceted Thesauri and Intersecting Perspectives

Each facet in a thesaurus has a well defined set of possible standard relationships, with the fundamental structural aspects being a hierarchical set of concepts and a flat list of terms that refer a user to the branches in the hierarchy. I will discuss details and standards for the structure within a single facet later. I will also later discuss forms of structured classification schemes other than the thesaurus, particularly ontologies, which are important to KOCIS-KD and especially to interoperability. First I will discuss the value of multiple facets in a single thesaurus or just as purposely applied to other classification forms including ontologies. Facets are vital to KOCIS-KD and to Korean Diaspora studies.

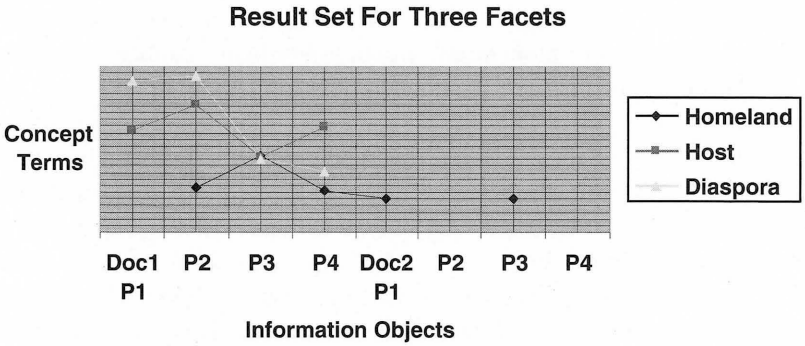
The purpose of multiple facets is to provide for indexing and locating information according to multiple perspectives, and, particularly, to do so simultaneously. For instance, to find information in a Korean Diaspora information collection one would want at a minimum to find information simultaneously by culture and by subject. The facets are considered to have *orthogonal* relationships with each other. The use of the term orthogonal here doesn't quite have the mathematical meaning of perpendicular but it is related. Each facet is considered to be a separate axis that can intersect with the other facets. Facets, unlike perpendicular lines can intersect more than once or not at all. Each axis represents a dimension of the overall subject. The facets intersect where information is indexed by categories from two or more facets. Just as facets define a dimension of the overall subject, so do particular combinations of intersections.

Classification hierarchies should be divided into separate facets when many levels of the hierarchy would need to be broken into sibling branches each qualified by an alternative value of the same subject concept. If one tried to combine culture and subject into a single hierarchy, one would need to create sibling branches combining every culture category with every subject category. For instance starting down one branch one could have Economy/Korean Americans, Economy/Korean Japanese, Economy-Manufacturing/Korean Americans, Economy-Manufacturing/Korean Japanese, etc. and down another branch Values/Korean Americans, etc. Such a structure almost immediately becomes completely unwieldy.

A multi-faceted system on the other hand is based on conceptually bounded separate hierarchies, the intersections of which occur in the indexing of the information and can be found by queries against this indexing. For the purpose of multi-faceted searching, the relationship between facets in the query is always a *conjunctive* relationship, or in other words an AND relationship in commonly used query languages, whereas the relationship between categories across branches in each hierarchy is by default a *disjunctive* or *OR relationship*. These defaults can be altered for particular purposes.

What are the important facets of Korean Diaspora information? Dr. Koh's work provides a foundation for this analysis. I suggest that such analysis ought to be a fundamental background to approaching any study of Korean Diaspora. We have not yet done justice to Dr. Koh's concepts in

our current implementation of KOCIS-KD. KOCIS-KD uses the following facets: Subjects, Cultures, Time Periods, Places, Persons, Organizations, Key Terms, and Citation Data. Cultures are further qualified by Homeland, Host and Diaspora culture. Another key concept of KOCIS-KD is data quality control. We do not have a separate facet in the current implementation for data quality control. However data quality can be controlled for by the intersection of other facets, particularly intersections with the Citation Data facet and potentially with intersections with the Persons and Organizations facets.




## Single Facet Relationships

Each facet has a single hierarchical structure pertaining to a single perspective on a field of study. For instance in KOCIS-KD there is a facet for Subjects that is a hierarchical organization of cultural subjects such as Religion and Philosophy, Economy, Polity, Values, etc. with each branch dividing into sub-categories to as deep a level as necessary.

There are both term based and concept based thesauri. KOCIS-KD is a concept based thesauri. The KOCIS-KD thesaurus was guided by widely used published standards for thesaurus construction. These standards are compatible. They are the American National Information Standards Organization (NISO) Z39.19 thesauri standard and the International Standards Organization, ISO 2788 monolingual thesauri standard and ISO 5964 multilingual thesauri standard.

Each branch in a thesaurus facet has a restricted set of possible properties and possible relationships with other branches. The most basic

property is the *Preferred Term.* This is the identifying term for the concept. Each concept branch will also have a *Scope Note.* At minimum this is a descriptive explanation of the concept. The concept can have additional descriptive aspects that can be entered either as part of a single scope note property or as additional properties or as sub properties of the scope note. This can include a usage note, describing how the concept should be used in indexing, a history note describing how the definition and use of the concept for indexing has varied over time and a source note describing validating resources and arguments for the concept definition and its use, etc.

Each concept has a set of relationships to other concepts in the thesaurus. These are the broader concepts and narrower concepts in the hierarchical structure of the thesaurus and related concepts for relationships that cut across the hierarchy. There are forms of thesauri that allow multiple broader concepts and forms that allow only one or one primary broader concept and any number of additional broader concepts. There are several display and navigation issues that are resolved by allowing only one primary broader term and this is the approach KOCIS-KD uses.

Each concept also typically has a set of entry level vocabulary terms. These terms provide users references to the concept so that, for instance, a user who may think of the term psychology can enter that term and be referred to the concept Psychocultural. The terms from all the concepts are grouped together into a single index for the thesaurus. The entry level vocabulary are often called *non preferred* terms to distinguish them from the preferred identifying term for the concept or *used for* terms referring to the relationship of the concept to the term in that the concept is "used for" indexing subjects of related to the entry level term. This same relationship is typically called a "use" relationship when referring to how the entry level term relates to the concept, in that the indexing of subjects relating to the entry level term *use* particular concepts.

### View Based and Full Text Searching

The most common web search interface is some form of an entry box for search terms which when executed yields a list of abbreviated results with links to a more complete display of the information object, such as a document or a citation record. The problem with this type of interface in a

thesaurus based classification scheme is that the returned list bares only a hap-hazard relation to organization of the information according to the thesaurus. An alternative search interface was promoted by A.S. Pollit with the HIBROWSE system called view-based. KOCIS-KD uses the essence of a *view-based* interface. (Pollit) (Hyvönen)

KOCIS-KD does not implement all the potential advantages of a view-based searching but its basic design is compatible with implementing these features. The basic concept of view-based searching is that information objects are found by navigating the thesaurus and especially by navigating the intersections of the thesaurus facets. The term view is used as a synonym of *facet.*

View based searching has the following advantages: 1) Information objects can be showed in proximity, hierarchical, and sequential order that relates to the conceptual organization of the classification system. 2) The thesaurus navigation doesn't require the user to guess at terms used to identify indexed concepts. 3) The number of hits (information object links) can be pre-calculated and branches with no hits can be grayed out or not shown preventing returning empty result sets.

Of course for all the value of a thesaurus indexed information collection, there is still value for words in context searching. View-based searching does not exclude combining thesaurus navigation with words in context searching.

**View Based and Full Text Indexing**

There are several options for, but little literature and analysis on, how to create the links between index terms and information objects. In building a global network of information this will be one of the most important features to work out. KOCIS-KD uses a system that is rather unique but is a natural extension of its use of a thesaurus and of view-based searching. It includes the information objects in the thesaurus. As we will see later in this paper, this system is particularly compatible with broader and ultimately more powerful classification structures of topic maps and ontologies for which KOCIS-KD is designed to take advantage of as the future world of interoperability becomes a reality.

Indexing can be executed as pointers in one of two directions. Most commonly pointers to concepts in the thesaurus are embedded in the

information objects, perhaps by storing codes in embedded or attached XML metadata elements. However there are strong arguments for taking the opposite approach by storing pointers to information objects from the thesaurus concepts. A third alternative could use an external object to store two way links much like a joining table in a relational database.

The problem with either storing indexing in the information objects or using joining tables is that the first corrupts the object and both create a maintenance problem, indeed, a huge maintenance problem in a global network, especially if one wishes to allow the same information to be indexed with different ontologies or even with conflicting indexing by different indexers using the same ontology.

The solution of storing the indexing in the thesaurus or ontology doesn't completely eliminate the maintenance problems with a global network, but the problem becomes the same problem as the basic interoperability problem of mapping between ontologies and of maintaining customized branches of single ontologies. Further in both networked and non-networked systems storing indexing in the thesaurus means the same tools used to build and maintain the thesaurus or ontology and the same underlying data structures can also be used for indexing. Using thesaurus or ontology tools to index information is a *view-based* indexing procedure that is a complement to *view-based* searching.

Storing the indexing in the thesaurus does alter the concept of facet usage somewhat. In its canonical form, the orthogonal facets of a multi-faceted thesaurus intersect at points in a field of information objects. However to store indexing in the thesaurus we have to incorporate the information objects as another facet. In the canonical form there should be no relationships drawn between concepts in separate facets, certainly not narrower or broader relationships. We need to loosen this restriction and allow relationships to be drawn to the information object facet. Those relationships are our indexing.

For instance, a particular document or a particular work of art are examples of information objects. They might be indexed by ontology facets of media type, time period, place, culture, subject matter. In a traditional multi-faceted thesaurus: culture would never be a kind of subject, time never a kind of place, subject never a kind of media type, etc. But in an ontology with information objects as a facet, a document might

indeed be a classified under a time period, a work of art under a culture, etc. This is a liberating construct. The field on which facet intersections occur is now created by the intersection. It might make sense that a place can be a "kind" of time such as *Uzbek Soviet Socialist Republic* being classified under "soviet era". Graying the line between conceptual organization of information and information itself more accurately models reality.

### Inductive Bottom-Up Versus Deductive Top-Down Thesaurus Construction

Another virtue of merging the information into the thesaurus or ontology is it creates a mechanism for inductive, bottom-up construction of the classification scheme. Since the origins of Dr. Koh's work related to KOCIS-KD, a principle has been pursued that the classification scheme should not be static, but rather the scheme should be continually modified while using it in the indexing process. For instance, one facet of KOCIS-KD is Time Periods. One could create a regular division of time by centuries or years or some other regular measured unit. But regularly divided time units would not be nearly as informative as historically significant periods and one-time event markers. However, a universally complete classification of historical periods and events, especially in a specialized field such as Korean Diaspora studies, is probably not possible, certainly not one that predicts the future. The concepts need to be dynamically developed.

Bottom up thesaurus construction in the KOCIS-KD means examining the information objects and deriving candidate classification concepts and terms from the information context. Since the information objects are stored in the thesaurus the mechanics of doing bottom up construction are simplified.

### Using Keyterms in Bottom-Up Thesaurus Construction

One of Dr. Koh's original insights in the history of the development of KOCIS-KD is the potential significance of keyterms. Particularly important is the concept that a keyterm can be either culturally generic or culturally specific. It also turns out isolating keyterms provides an excellent mechanism for inductive thesaurus construction. Keyterms provide an

excellent list of candidate terms identifying concepts that might be appropriately integrated into the thesaurus. For example a term such as *Soviet era* in literature on Kazakhstan Koreans might be a candidate for a time period concept.

**Human Indexing Versus Machine-Aided and Automatic Indexing.**

Creating indexed databases with enough indexed information to make the whole effort worthwhile is a problem. To build a worthwhile information collection with human indexing requires a cadre of experts working over many years. Various software tools can be used to do rule-based machine-aided indexing but these systems only help human indexers not eliminate them. Experts must still define the rules, such as if certain word combinations occur in certain proximity, then assign this concept, and this can only be relied on to provide candidate associations. Humans must still read or otherwise evaluate the content. Other systems make attempts at artificial intelligence and machine comprehension of natural language but success here is still a long way off and probably could never do as good a job as human experts.

**Distributed Indexing Over the Semantic Web**

Interoperability and the Semantic Web, however, could make human indexing feasible. This could be especially true in a discipline bounded network such as a global Korean Diaspora studies network. Instead of a trained and centralized cadre of indexing experts, anybody and everybody could contribute to the indexing. With authority and authentication and semantic web techniques indexers could be ranked by various qualities on their level of expertise and the indexing they do can be filtered out or included as needed or preferred by searchers. Further, and particularly if indexing of the same information by different indexers is encouraged, statistical methods can be used to validate the quality of indexing. For instance, if eight out of ten indexers associate a particular document with a particular classification category and the other two indexers each use a different category then probably there is greater validity to the choice of the eight. However if one of the other two is a preeminent expert in the field then a different weighting scale would have an effect. Searchers could select custom filtering of indexing such as using only indexing done by

experts, only indexing done by their own university department members or even only their own indexing.

There would be several ways to create incentives to encourage people to do indexing. One would just be to appeal to many persons impulse to voice an opinion. Another would be to encourage indexing to be done by students as an assignment. But there could also be built in positive return in doing indexing, such as using it as a means of bookmarking while browsing.

Such a system would not eliminate the continuing value of more centralized experts. It would be best to have a core indexed information set. Further, if a wide body of people were participating in indexing, there would be demand for the expert trainers and consultants.

### Distributed Thesaurus Construction

The same semantic web techniques which are used for mapping between thesauri or ontologies can be used to map between versions of the same thesaurus and to plug-in and remove branch extensions. In this way individual scholars or university departments for example could use customized versions of KOCIS-KD that fit their own specific research needs while at the same time gaining the benefit of the overall core classification scheme.

The ability to plug-in branch thesaurus extensions would also encourage contributions to a dynamically growing ontology. With semantic web techniques these contributions can be properly filtered out from the core or incorporated into the core with the oversight of authorized experts.

### Beyond Thesauri: Toward Ontologies

Thesauri are one standard form of classification structure. They are probably the most widely used form. Most library classification systems such as Library of Congress Schedules and several major online databases systems such as Medline, and PsychInfo are based on thesauri structures. It is important that a classification structure is systematic and functional and much research has been done examining what structures are most functional. It would not be wise to create an ad-hoc structure for any information collection indexing system.

Thesauri are not the only standard classification structures. In fact the literature on such schemes demonstrates that there is considerable confusion and overlap of definitions. Different disciplines, particularly those of library science and those of computer and information science (Lancaster), tend to sometimes pursue redundant paths. The most current literature recognizes the redundancy and tries to delineate some order. I expect that the various disciplines will successfully merge the main thrusts of the efforts. With understanding that the lines are not clear cut it is possible to provide an approximate scaling of classification systems from less to more powerful and flexible. To some extent this same lineage approximates the historical development of such systems, with the most powerful systems being *topic* maps and *ontologies*. Perhaps more revealing is that the order of increasing power could also be represented as a set of encompassing forms, such that uncontrolled keyterms are a subset of authority lists, which are a subset of taxonomies, which are a subset of thesauri, which are a subset of topic maps, which are a subset of ontologies.

Since classification is sometimes only part of what a scheme does, a more descriptive and commonly used term for all such schemes is *Knowledge Organization System*, often referred to by the acronym *KOS*. The following table provides a non-exhaustive list of KOS in approximate order of power and flexibility.

| Classification Type | Fundamental Structure | Specification Standards | DTDs or Schemas | Markup Syntax |
|---|---|---|---|---|
| Uncontrolled Keyterms | Flat list | | | List in a single metadata field |
| Authority Lists | Flat list | | | List in a metadata field with a validation mechanism. May have referring list of uncontrolled keyterms. |
| Taxonomies | Hierarchical | Discipline specific | | Predates computers. Uses print formatting. |
| Thesauri | Hierarchical with index; single or multi-faceted | Z39.19; ISO 2788; ISO 5964 | No official general schema or DTD; TML is one. SKOS | Probably most frequently uses SGML or XML today. Predates |

| | | | | |
|---|---|---|---|---|
| | | | expresses a thesauri as an in RDF-XML as an ontology. | computers and often relies only on print formatting. |
| Topic Maps | Hierarchical with index; single or multi-faceted | ISO/IEC 13250 | TMCL | SGML; XML |
| Ontologies | Hyperlinked web; each link is of a specific type; the most basic of which create a hierarchical structure. | W3C recommendations for RDF and OWL | RDFS; OWL | RDF-XML |

## 3. Finding Information Resources and Interoperability

KOCIS-KD currently has a simple, interoperable feature, which automatically uses the concept terms and related terms to search the web with the major web search engines. The feature is hard-coded in that the format of the query language for the various web search engines is explicitly coded in the application. Nonetheless the feature shows the promise of interoperating classification schemes. Because the search uses terms and their relations from the thesaurus the user has a predefined menu of conceptual relationships with which to search.

Classification and indexing schemes, and as we will see, most powerfully, ontologies, are fundamental to an emerging new world of interoperability. A shallow level of interoperability occurs today with the use of major search engines such as Google and Yahoo. These are largely restricted to finding web sites and pages using a combination of their own classification schemes and indexing procedures and scanning web pages for meta fields and terms in context. However they do not interact with other ontologies or with databases. They fail to find what is often called the *deep web* and they fail to properly filter for precision.

A theoretical solution would be a universal classification system. However, it is an obvious truth that the world would never agree on a universal scheme. Nor in fact is it possible that a world-wide scheme could be specific enough for use in a particular subject area like Korean Diaspora studies. The alternative to a system by which one classification and indexing system can be interpreted in terms of another system in

automated or semi-automated processes.

This is the world of interoperability. Like classifications schemes, developments toward interoperability have begun independently in many different disciplines and have taken somewhat different approaches and use somewhat different terminology, but perhaps the most currently prominent development is that of the *Semantic Web* supported by the W3C web standards consortium and led by Tim Berners Lee the usually cited founder of the World Wide Web. The term *Semantic Web* implies an Internet web where the semantics, the meanings, of language can be understood by automated or semi-automated processes even when the terminology differs both within and across languages.

The core underlying goal of these developments is more complete and more relevant results (higher recall and precision) from querying and browsing information sources, especially on the World Wide Web. The unique targeted solution of these developments is automated and semi-automated brokering of querying, browsing, and result processing across systems around the world. The solution depends on semantic translation of distinct *vocabularies*. Currently, most interoperating systems are based primarily on hard coded mappings between systems. The Semantic Web is different in that it aims for automatic logically deduced mappings.

Rather than requiring either a fully expressive universal *vocabulary* or hard coded mappings, Semantic Web techniques use a very simple language called RDF-XML, *Resource Description Framework-XML*, to express very simple relationships which form building blocks from which distinct domain-specific vocabularies can be built. This simple language can be extended by using an extension language called RDFS, *RDF Schema*. Because the descriptions are built from simple grammatical structures, they can be decoded by computer programs.

A simple RDF building block, or fact, is expressed as a triple of *subject, predicate*, and *object* or alternatively described as *resource, property, value*. One triple can be the object of another triple and sets of triples sharing the same subject can be combined. But ultimately there is nothing more complicated than sets of things of three. Because of this, RDF doesn't require the complicated mappings that are necessary between distinct schema for relational databases, object-oriented databases, hierarchical XML documents, or even flat metadata fields or authority

lists.

Each subject is uniquely identified by a URI. Often this *Universal Resource Identifier* is a URL, *Universal Resource Locator*, or in other words web-page address such as *http://www.kockiskd.org/ontology/subjects/psychocultural.* In any case, with the help an XML mechanism called *name-spacing* it is possible to assign practically infinite universally unique identifiers with little effort.

With properly assigned predicates, the structure of atomic triples means that a deeply hierarchical organization of resources, such as in a thesaurus can be broken down into a flat list of triples and vice versa, a flattened list can be built back up into a deep hierarchy. Besides functioning as an identifier, the URIs can function as addresses. I like to equate the communication protocol of the Semantic Web to the IP packet sending system of the Internet itself. To begin a mapping process between one ontology and another it is only necessary to send or discover a single triple from an ontology somewhere on the Internet. In the way that IP packets can come from anywhere and be re-ordered at their destination, RDF-XML triples can do the same.

## 4. Conclusion

The design of the KOCIS-KD website has several features that make it a useful model and foundation for use in a global information sharing network for Korean Diaspora studies. These include:

- A multi-faceted information system classification system pertinent to Korean Diaspora studies.
- A core standard thesaurus structure.
- A modification of thesaurus structure incorporating the information object links in the thesaurus. This allows using thesaurus relationships and tools to index the information.
- A process for building the thesaurus dynamically and inductively by examining the data, especially keyterms.
- A simple implementation of using classification relationships for searching the Web for additional resources outside the core information collection.

- A plan for distributed indexing.
- A plan for distributed extensions and modifications to the core classifications.
- A plan for mapping to other world wide classification systems in the emerging Semantic Web

## References and Bibliography

Aitchison, Jean, David Bawden, and Alan Gilchrist. (1997) Thesaurus Construction and Use: A Practical Manual. 4th. ed. London: Aslib, 2000.

ANSI/NISO Z30.19 Guidelines for the construction, format, and management of monolingual thesauri / Developed by the National Information Standards Organization: approved August 23, 2003 by the American National Standards Institute. NISO Press. Bethesda, Maryland. 2003.

Berners, Lee Tim et al. T*he Semantic Web. A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities.* Scientific American. May 2001. http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Fedora Project Website. http://www.fedora.info/.

Hyvönen E. et al.: Application of Ontology Techniques to View-Based Semantic Search and Browsing. In C. Bussler, J. Davies, D. Fensel, R. Studer (eds.): The Semantic Web: Research and Applications. Proceedings of the First European Semantic Web Symposium (ESWS 2004), Springer-Verlag, LNCS 3053, 2004. http://www.cs.helsinki.fi/u/eahyvone/publications/promoottori.pdf

Hyvönen, E. et al. A Portal for Publishing Museum Collections on the Semantic Web. Proceedings of ECAI/PAIS 2004, Valencia, Spain, 2004. http://www.cs.helsinki.fi/u/eahyvone/publications/MuseumFinlandPAIS2004.pdf

ISO Standard 2788-1986 *Documentation - Guidelines for the establishment and development of monolingual thesauri.* 1986. International Organisation for Standardisation, Technical Committee ISO/TC 46, Documentation.

ISO Standard 5964-1985 *Documentation - Guidelines for the establishment and development of multilingual thesauri.* First edition 1985-02-15. International Organisation for Standardisation, Technical Committee ISO/TC 46, Documentation.

Koh, Hesung Chun. T*he HRAF Automated Bibliographic System Today* BEHAVIOR SCIENCE RESEARCH VOLUME 13, NUMBER 2 1978 Human Relations Area Files, 2054 Y. S., New Haven, Conn. 1978. http://www.kociskd.org/project

Koh, Hesung Chun. *What is KOCIS-KD?* East Rock Institute. March 1, 2003. http://www.kociskd.org/project

Koh, Hesung Chun. *A Social Science Bibliographic System: Orientation and Framework.* Human Relations Area Files. http://www.kociskd.org/project

Koh, Hesung Chun. KOCIS Analysis Structure. East Rock Institute. 2003. http://www.kociskd.org/project

Lancaster, F.W. Indexing and Abstracting in Theory and Practice, 3rd ed. London. Facet Publishing. pages x-xiv. 2003.

Miller, Ken and Matthews, Brian. *Having the Right Connections: the LIMBER Project.* Journal of Networked Information. Volume 1. Issue 8. Networked Knowledge Organization Systems. 2/5 2001. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Miller/#9

*Moench E. et al.* Semantic Miner - Ontology-Based Knowledge Retrieval. Journal of Universal Computer Science, Vol. 9, Issue 7, pages 682-696. 2003. http://www.jucs.org/juc_9_7/semantic_miner_ontology_based.

Pollit, A. S. The key role of classification and indexing in view-based

searching. IFLA '97 Copenhagen Aug 31 - Sept 3 1997 63rd IFLA General Conference Booklet 4, Section on Classification and Indexing Session 95 Paper 009-CLASS-1-E. http://www.ifla.org/IV/ifla63/63polst.pdf

Unicode Home Page http://www.unicode.org/.

W3C Semantic Web Activity. http://www.w3.org/2001/sw/.

## Abstract

The information by which we try to understand Korean Diaspora are, like the diaspora itself, vast, complex, and spread around the world. The Internet gives us more access to this information than ever before, but how do we find it and how can we organize it to understand it? The solution is a cooperatively built interoperative system with two primary characteristics: a relevant classification, indexing, and retrieval scheme; and applying this scheme to a large amount of complex data on the Internet. *KOCIS-KD* is a project to create a portal and model to facilitate comparative analysis and area study for Korean Diaspora studies. This paper discusses the importance of KOCIS-KD's design for a global network for Korean Diaspora studies.