DOCUMENT RESUME

ED 441 416 IR 057 687

| | |
|---|---|
| AUTHOR | Chan, Lois Mai; Lin, Xia; Zeng, Marcia |
| TITLE | Structural and Multilingual Approaches to Subject Access on the Web. |
| PUB DATE | 1999-08-00 |
| NOTE | 12p.; In: IFLA Council and General Conference. Conference Programme and Proceedings (65th, Bangkok, Thailand, August 20-28, 1999); see IR 057 674. |
| AVAILABLE FROM | For full text: http://www.ifla.org/IV/ifla65/papers/012-117e.htm. |
| PUB TYPE | Reports - Descriptive (141) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Access to Information; *Classification; Computer Software Development; *Directories; *Indexing; *Information Retrieval; Information Services; Languages; *World Wide Web |
| IDENTIFIERS | Knowledge Management; *Search Engines |

ABSTRACT

This paper presents some of the efforts currently being made to develop mechanisms that can organize World Wide Web resources for efficient and effective retrieval, as well as programs that can accommodate multiple languages. Part 1 discusses structural approaches to organizing Web resources, including the use of hierarchical or classification-based formats and functional requirements of Web organizers. A report of a research project on developing a personalized knowledge organization and access mechanism, called Knowledge Class, is presented in Part 2, including purpose, objectives, and specifications of Knowledge Class, system design, design principles, the iterative design process, multilingual support, and plans for the near future. Part 3 describes a multilingual approach to subject access, including multilingual-oriented services of major search engines and characteristics of Web subject directories in a multilingual environment (i.e., alphabetical arrangement of categories, implementation of the literary warrant principle, flexibility in reflecting local interests, and limitation of language-dependent search features). (MES)

IFLANET
Search Contacts
International Federation of Library Associations and Institutions
Annual Conference

# 65th IFLA Council and General Conference

# Bangkok, Thailand, August 20 - August 28, 1999

IFLA 1999 BANGKOK
Conference Proceedings

**Code Number:** 012-117_E
**Division Number:** IV
**Professional Group:** Classification and Indexing
**Joint Meeting with:** -
**Meeting Number:** 117
**Simultaneous Interpretation:** *No*

# Structural and multilingual approaches to subject access on the web

**Lois Mai Chan**
*School of Library and Information Science, University of Kentucky
Lexington, Kentucky, USA*

**Xia Lin**
*College of Information Science and Technology, Drexel University, Philadelphia, Pennsylvania, USA*

**Marcia Zeng**
*School of Library and Information Science, Kent State University
Kent, Ohio, USA*

## Paper

## Introduction

Among the many challenges in discovering and mining useful resources on the World Wide Web are the sheer volume of what is available and language barriers. Mechanisms that can organize web resources for more efficient and effective retrieval are urgently needed, and there is an equally obvious and pressing need for programs that can accommodate multiple languages. In this presentation, which is in three parts, we will discuss some of the efforts currently being made on both fronts.

## Part I: Structural Approaches to Organizing Web Resources (Lois Mai Chan)

Many subject categorizing schemes have been developed to organize and manage web resources. They are known by various names such as subject guides, web guides, subject categories, subject directories, subject hierarchies, and so on. What many of these systems have in common is that they manifest the traditional classification principles of hierarchical structure, domain partition, subordination of the specific to the general, and array of related subjects. A survey of the hierarchical structures now functioning as web organizers shows considerable variation in complexity and sophistication, in subject scope and depth of coverage, and in the number of items they cover. They also vary in the classificational patterns on which they are based. In some cases attempts have been made to adapt existing schemes such as DDC, LCC, and UDC to the web environment. Diane Vizine-Goetz, in her research, has shown how, with appropriate improvements, such schemes may be used to enhance web subject retrieval (Vizine-Goetz). Systems that use hierarchical structure to organize web resources include:

1. subject guides devised by popular web search services such as ***Yahoo!***, ***Lycos***, ***InfoSeek***, ***Excite***, etc;
2. schemes devised by individual libraries to facilitate access to the web resources they have selected and included in their local systems; and,
3. Web organizers and directories based on existing schemes, for instance, OCLC's Netfirst based on DDC, CyberStacks and Scout Report Signpost based on LCC.

Using hierarchical or classification-based formats to organize web resources could have important advantages, among which are improved subject browsing facilities, potential multi-lingual access and improved interoperability with other services (Koch and Day). A hierarchical structure can be thought of as a conceptual map - perhaps of the entire universe of knowledge or perhaps of a particular domain therein. Such a map sorts informational content into related groups (and their subgroups) and thus allows searchers to confine their approaches to defined areas where similar material is concentrated.

Knowledge viewed through an organized structure is easier to perceive and to comprehend once perceived. In subject access, a hierarchical structural presents a logical path to the desired objects. Above all, it improves precision by first defining and narrowing the domain for searching. This advantage is evident even in hierarchical structures that offer only a broad level of categorization. The reason for the benefit may be that hierarchy, even at a broad level, exemplifies two functions of traditional classification: collocation (inclusion) and partition (exclusion). While collocation is predicated on inclusion, which is a fundamental property of classification; partition captures another fundamental property, exclusion. It is how well a hierarchical structure fulfills these two functions that determines its potential helpfulness in a search environment. Inclusion collocates objects and ideas that are alike. But, in a vast information domain, it is just as important to exclude information not wanted as to include what is being sought. Exclusion can be accomplished by dividing a large amount of information into smaller parts as a means of isolating the part that is most likely to be relevant (Chan 1995). The larger the information domain, the more important it is to find an effective and efficient way to define narrower domains for searching. One of the major causes of false hits in retrieval is homographs, that is, words that look the same but have different meanings. The advantage of searching within a specific domain is that terms are often ambiguous across several disciplines but seldom have multiple meanings within a particular discipline or subject domain.

There are advantages of using classification in the web environment where different conditions from those of the print environment prevail. In traditional systems, subject data (including classification numbers and indexing terms) are typically **attached to** their sources, either on documents themselves (call numbers on spines) or in surrogates (cataloging records or other metadata records such as the Dublin Core). In contrast, in the web environment, subject data often **are separate from**, or **reside outside** the resources themselves. Instead, such information can be stored in directories or other types of web interfaces that link subject data to the resources but do not affect them otherwise; individual links are made from the subject provisions in the web organizer to the resources through the urls. The advantage of "linking-to" rather than "storing-with" is flexibility. With a linked system, if a classification or other subject organization scheme is revised, it is only the links that may have to be changed:

the web pages and sites are not affected in any way. Re-classification is not a problem. Furthermore, the scope and the depth of any given system can be easily adjusted on the basis of literary warrant, whether the warrant be popular, consumer-oriented, or academic/scientific. For example, common categories found in popular subject guides include automobiles, entertainment, family, sport, and travel, while the most commonly found categories in academic web guides are humanities, social sciences, science, technology, and law. Furthermore, the web guides can also be easily adapted to local or regional needs, or modified for the needs of a specific clientele.

The use of hierarchical or classificatory structure on the Web is still relatively new. As web resources continue to grow, one may expect corresponding growth and refinement in ways to organize them. Now, it is perhaps not too early to consider some of the functional requirements of web organizers. The desirable characteristics may be summarized as follows; a scheme designed for organizing web resources should be: (a) intuitive, logical, and easy to use, with hierarchies and cross-references clearly displayed and with current and expressive captions; (b) flexible, adjustable, and expandable, to reflect rapidly changing and diverse environments; (c) useful in a wide range of settings, and applicable over a wide range of the number of sites to which it applies; and, (d) relatively easy to maintain and to revise.

The first question is whether to adapt a current classification scheme or whether to begin afresh. It is apparent from the current situation that those who design and build web organizers lean toward devices that are based on their own understanding of the needs and search habits of their users. What is at issue here is the difference between two methods for categorizing subject content. Familiar classification schemes, typically represent a top-down approach, starting with the whole universe or an entire discipline of knowledge, determining major classes on theoretical grounds, and subdividing them hierarchically into increasingly specific levels. This approach has generally been used whether the resulting scheme is custom-tailored for specialists or designed with a diverse population in mind. The alternative approach is a bottom-up operation that begins with specific terms or items (web pages in this context) which are then grouped and organized, first into a microcosm, finally, as coverage becomes fuller, into a macrocosm. In the web environment, where most subject guides have also been designed with the general public in mind, many recent efforts to categorize web resources seem to take the latter, bottom-up, approach.

The question of which approach is likely to prove more effective in the web environment does not have a definite answer. Either approach leads to a system that embodies domain partition, general/specific delineation, and array of related topics - features that are considered important for effective retrieval from a very large resource collection. What seems most likely is that time will show that top-down systems are especially suitable for highly structured established fields; bottom-up systems, on the other hand, may be particularly well suited to the mass of varied and fluctuating material that makes up so much of the Web. It seems likely, also, that the bottom-up approach works especially well for personalized or customized web organizers, several of which have emerged in recent months. An example is Northernlight's "Custom Search Folders," a device which categorizes the results of particular searches into broad categories.

The second part of this presentation consists of a report on a research project on developing a personalized knowledge organization and access mechanism.

## Part II: Knowledge Class (Xia Lin and Lois Mai Chan)

### *Purpose, Objectives, and Specifications of Knowledge Class*

The purpose of this research project is to create and test a device called "Knowledge Class," designed for customizing knowledge organization and access, to supplement and complement existing devices for Web users. In a widely cited paper published in *Scientific American* (March 1997), Clifford Lynch suggests: "Combining the skills of the librarian and the computer scientist may help organize the anarchy of the Internet." In our project, we have been exploring the possibility of combining existing methods of knowledge organization with advanced Web technology to create an easy-to-use framework for individual Web users. Preliminary results have been reported in recent literature (Lin and Chan 1997). In this presentation, we will briefly summarize the major characteristics and report on the latest progress.

Knowledge Class contains two basic components: an organizing framework and an interface for access to and retrieval of web resources. The organizing framework is a classified mini-thesaurus, consisting of a hierarchically structured collection of terms on a specific topic or discipline of interest or concern to an individual user. The interface serves as an interactive mechanism between the user and the terms in the organized framework as well as between the user and web resources. Through this device, the user can initiate searches by selecting the display terms or by using pre-stored search strategies, which often contain synonyms and can also connect to sites previously discovered by clicking on links with pre-stored urls.

In Knowledge Class, we try to recapture some of the advantages of traditional methods for efficient and effective information storage and retrieval and apply them to the web environment. Specifically, three aspects are considered:

1. classification principles for organizing information and displaying subject relationships;
2. controlled vocabulary features, particularly the control of synonyms and homographs for the purpose of improving recall and precision; and,
3. search strategies formulated and pre-stored for the purpose of optimizing search results and current awareness.

We set out to design Knowledge Class in such a way that it:

- organizes concepts and terms on a specific subject or topic into a logical structure showing subject relationships;
- facilitates browsing of subject terms and their relationships;
- stores useful search terms and strategies so they are available for future use;
- allows the addition of synonyms for better recall and qualifiers to resolve ambiguities or distinguish among homographs;
- initiates searches using pre-stored terms and strategies in a chosen search engine; and,
- stores urls of specific sites for future use

In other words, we hope to take information service one step further, beyond what has been available so far. In online retrieval, a great deal of emphasis has been put on retrieval results, and rightly so. But, after retrieval, there is also the need for organizing related information and "storing" it in a sense for future use and re-use. This can be done by providing the means for re-visiting the sites and, equally important, for retracing the steps used to find the resources in the first place.

Improving both subject browsing and precision of retrieval are our two main goals. In the first stage of our work [Lin and Chan 1997], we introduced the mini-thesaurus-like device. We emphasized that: (1) a knowledge structure can be built on principles of classification and bibliographical organization; (2) the knowledge structure could be seamlessly integrated with search engines for access to web resources; and, (3) an easy-to-use graphical interface could be constructed to support user interactions not only with the organizing structure but with the relevant resources discovered and retrieved through search engines.

### *System Design for Knowledge Class*

An advantage of conducting research on the Web is that prototype systems can be designed and tested incrementally in the real environment. We started with simple HTML coding to experiment with the idea of Knowledge Class as we initially envisioned it. During implementation and testing, we continuously revised the functions and added new features to it. As we learned more and understood more about its performance, we began to implement it in more sophisticated and robust system languages such as JavaScript and Java. It is this learning-by-doing process that has helped the evolution of Knowledge Class.

### *Design Principles*

From the very beginning, we set up several goals for the design of Knowledge Class. The project started with a search for a device or a system that would provide an optimal balance between automatic and manual indexing in building the organizing frame. Our first design principle was to maximize the benefits of both manual and automatic indexing. Secondly, we wanted to design an easy-to-use interface for Knowledge Class. The system should be usable and useful to a broad range of users. Librarians and information specialists may want to create knowledge classes for their clients. End-users may want to use Knowledge Class to replace simple bookmarking functions of browsers. School teachers may use knowledge classes to cover topics they teach and students may use them to explore class topics and to expand their knowledge by adding more search terms to the knowledge classes and linking them to web resources. We want all these users to be able to use the system with a minimal learning curve. Thirdly, we want users to be free from having to learn detailed syntax of query construction, to be free from memorizing each search engine's homepage, and to be free from having to construct complex search strategies. While Knowledge Class provides a mini-thesaurus for users, what makes it really useful is its connection to the search engines. The system should do as much work behind the scene as possible. It should connect to search engines directly, add synonyms automatically to search queries, and provide different search strategies for different terms. Most of all, the system should make all these transparent to users so that they can focus on semantics and the content of the topics when they use Knowledge Class.

### *Iterative Design Process*

The design of Knowledge Class went through three stages. Firstly, a basic frame was designed in HTML to include four windows. The first window displays all the branches in a knowledge class. The second window is for individual branches in an expandable/contractible tree structure; only one branch is shown at a time. The third window is the main window for displaying search results. The fourth window is for displaying and switching search engines. The four windows are on one html page and can be easily loaded onto web browsers.

In the second stage, we worked with a group of library science students at the University of Kentucky. Each student developed a knowledge class using the basic framework we provided. During this stage, we found that different search strategies needed to be developed for different types of searches. For example, some of the terms need to be searched as individual words; others would be much better searched as a phrase, and still others need to be searched with additional contextual terms taken from higher levels in the hierarchy of the knowledge class. Through many trials and tests, a coding system was developed to facilitate assignment of a specific search strategy to each term. An entry in a knowledge class typically looks like:

--, mutual funds, mutual-funds Investment-trusts Unit-trusts, http://www.brill.com, 1

There are five parts in this entry, each separated by a comma. The first, the number of dashes indicates the hierarchical level of this term. The second is the display term (what will be shown on the tree structure). The third is the search terms; it can include many terms that are synonymous or related to the display term. The fourth is a direct link; if it is present, a "link" icon is displayed to allow the user to click on it to go directly to the page. The final number in the entry is the coded search strategy. The complete list of coded search strategies is discussed in Lin and Chan (1997).

In the third stage, we further improved the design by implementing a Java version of Knowledge Class. In this version, window structures were redesigned to make it easier to switch from one branch to another without reloading the entire page. Taking advantage of Java's graphical power, we placed in one succinct frame what used to be scattered in three separated windows: all branches in a knowledge class, tree-structures for each branch, and search engines of Knowledge Class. With the saved screen space we were able to add another level to the display - a list of all the knowledge classes we have created thus far. Another major improvement in this version is the separation of programming files and data files. In the early versions, JavaScript and mini-thesaurus entries had to be included on the same html page, making it difficult for the user to modify or change the mini-thesaurus without a good understanding of JavaScript. With Java, the programming part is completely compiled and separated from mini-thesaurus data. The user can create, add or modify any content and structure in the data file without any knowledge of the programs.

### *Multilingual Support*

While we were designing the data structure, we found another benefit in separating display terms from search terms. Our original consideration was to make the connection to search engines more flexible and make query construction easier. We found that this feature became especially useful in developing multilingual knowledge classes.

While constructing a knowledge class on Wales, one of our students developed a bilingual classified mini-thesaurus with terms in both English and Welsh. For pages displaying Welsh terms, she wanted the searches to be conducted in both languages. With the separation of display terms and search terms, this is easy to implement - she simply included both English and Welsh terms in the knowledge class, and the search engines would then search webpages in both languages. Our testing indicated that this is a very effective approach to providing multilingual support. An example of a multi-lingual knowledge class is Complementary & Alternative Medicine (CAM), shows the part for Chinese Medicine in Chinese. We developed this branch in both English and Chinese (GB), and provided links to switch from one to the other. In the Chinese version, each search term includes both English and Chinese equivalents. Thus, for search engines that accept Chinese GB coding, search results will include both English and Chinese pages. We found this knowledge class to be particularly helpful for researchers who have limited knowledge of a particular language but wish to access materials in that language. For example, American researchers in traditional Chinese Medicine typically know some Chinese, but they may not feel comfortable enough to browse or enter search queries in Chinese. Using this knowledge class, they can browse in the English version and then switch to the Chinese version for retrieval, or they can click on the English terms and still be able to retrieve relevance resources in Chinese. This feature makes multilingual access to web resources both possible and efficient.

Knowledge Class is an ongoing project, which we plan to continue to improve, making it a useful tool for subject access to web resources. We believe that, for effective retrieval, web resources need to be organized in terms of "information units," not by individual physical pages. It is analogous to cataloging in libraries: for the sake of manageability and efficiency we catalogue at monograph or journal levels, not at the individual chapter or article level. We are building Knowledge Class to become such information units. In the future, a "mega" search engine will only need to index at the level of these "information units." With this device, users will first find relevant information units and then gain access to individual web pages.

**Plans for the near future include:**

1. We hope to enlist more people to create knowledge classes on a wide variety of topics. We will provide free software to encourage cooperation. We particularly hope to involve more information professionals, and to have librarians, information specialists, library school students and faculty members participate in their creation. When more people are involved, an advisory committee could be formed to guide and review the process and to ensure the quality of knowledge classes in the collection.

2. We plan to develop written guidelines for both information professionals and end-users who are interested in using knowledge classes. For information professionals, the emphasis will be on how to apply the principles and techniques of classification and information retrieval to the creation of knowledge classes and how to adapt different search strategies for different entries. For end-users, the emphasis will be on how to modify an existing knowledge class to suit their personal purposes.

3. We plan to further improve the software. Currently, the data must be edited in a text-editing program and users cannot change the search strategies online. In the next version, the user will be provided with tools to add terms to the entries in the hierarchical structure, to add synonyms to the list of search terms, and to change search strategies, etc. An authoring tool will also be developed so that the complete knowledge class can be developed and tested in a graphical environment.

## Part III: Multilingual Approach to Subject Access (Marcia Lei Zeng)

The phenomenon of multiple languages used in representing data on the Web calls for ways to solve the problem of users having to deal with languages both known and unknown to them. In the past, most search engines were geared toward indexing pages in Western European languages. Almost all the search interfaces were in English and often highlighted news or other events of interest to a U.S. audience. As Internet connections become all-pervasive and intranets invade the corporate network, the scope of available data increases radically. Since 1998, the World Wide Web search engines havebeen involved in a competition of globalization and localization. Multilingual processing has emerged as a key issue in the evolution of search engine technologies.

### *Multilingual-oriented Services of Major Search Engines*

To serve multi-lingual and multi-cultural populations all over the world, major search engines, such as ***AltaVista***, ***Excite***, ***HotBot***, ***InfoSeek***, and ***Yahoo!*** , have developed new services functioning as regional search guides, summarized below:

**Domain Filtering.** Usually each country has its own top-level domain on the Internet, e.g., .uk for the United Kingdom. The easiest way to create a regional guide with regional content is to filter by domain. Results are usually taken from the main listings but filtered by domain. Typical services are ***Global Excite*** (including Australia, Chinese, France, Germany, Italy, Japan, Netherlands, Sweden, and U.K.), ***InfoSeek International*** (covering Brazil, Denmark, Germany, Spain, France, Italy, Japan, Mexico, Netherlands, Sweden, and United Kingdom), and ***Lycos in***: (covering Germany, UK, France, Netherlands, Italy, Switzerland, Belgium, Sweden, Spain, Japan, and Korea.)

**Domain Detection.** In this case, the search engine detects the country a visitor comes from and presents a custom front page that is usually tailored with some specific information.

**Mirror Sites**. Mirror sites are the search engine sites physically located outside the United States. They may be more responsive, since they are isolated from the heavy U.S. traffic and the problems of crossing oceans and long distances.

**Language Specific Search.** Some services transcend national boundaries and instead are aimed at those speaking a common language. ***AltaVista*** and ***Northern Light*** both offer such services as searching the documents in particular languages. This is different from domain filtering (searches are limited within a country domain code, such as .uk) because it is completely content based. ***AltaVista*** stores information from pages in different languages in one index, regardless of character sets in which it is written.

**Multilingual search.** ***AltaVista*** also offers multilingual search through its "One World" technology that especially appeals to those speaking Asian languages. Basically, ***AltaVista*** translates whatever page it finds into Unicode, which can store characters for all languages. The searcher can request translation of a search query, or a whole webpage, from or into the language he would like to search or read.

**Regional Interfaces**. Creating a regional interface can be as simple as presenting the same search engine's look-and-feel in the appropriate language for a particular country. There are various ways to provide this service. In the case of subject directories, users sometimes can see a complete translated page from English, with no change of contents or order of categories. In other cases, users may see a bilingual display of the directory, for example, a subject directory in both English and Japanese. To display a text-based Japanese directory would require a local character code set be loaded in a client site machine. In order to avoid such a requirement, some directories provide an image/graphic-based display. Regional interfaces may also have different content focuses and displays.

**Localized Subject Directories.** Instead of having a set of regional interfaces which might be the products of transliterated or translated versions of a global or US version subject directory, localized subject directories provide tailored versions which reflect local interests. This is achieved by using local languages for the whole directory, defining and naming categories based on local convention, presenting categories according to local interests and including categories which assemble local focuses. ***World Yahoo!*** offers 19 versions of its directory, covering the Americas, Pacific Rim, and Europe.

### *Web Subject Directories in a Multilingual Environment*

At the introduction of this paper, Professor Chan listed major issues to be considered in devising a useful Web organizer. They are: scope of subject matter and depth of hierarchy, defining and naming categories, logical structure, clearly defined facets, citation order, cross classification, alphabetical index, terminology of captions, and notation. Among the well-known search engines and web subject directories, ***Yahoo!*** has been the leader of web organizers, and has successfully applied classification structure in its entire service. About one year ago, other major search engines also adopted subject directory methodology using their "folk classification" structure. An analysis of these services based on the issues that Professor Chan outlined revealed various approaches used by these services. This section of the paper will discuss some characteristics related to the multilingual services of a few search engines. Most examples are from ***World Yahoo!*** subject directories and were retrieved on February 12, 1999. In fact, many of the phenomena found in the following discussion also exist in other search engines' services, such as ***Northern Light***'s subject directory and ***InfoSeek***'s primary categories.

### 1. Alphabetical arrangement of categories

***World Yahoo!*** offers nearly 20 versions of its unique directory for various countries and regions all over the world. The directory divides all web resources into 14 main categories and has virtually included all subject matter. Some names/captions and coverage of the main directories in non-English versions (e.g., ***Yahoo!*** France) may differ from the global version, (also known as the US version.) Because no notation is used in the ***Yahoo!*** classification, alphabetical order becomes the natural and only arrangement of all categories and their sub-categories. No systematic system/schedule or logical order of the categories is applied. A complete browsing process is always needed when locating a particular topic on ***Yahoo!***. This unavoidably causes an inconsistency of orders in non-English versions of ***Yahoo!*** directories. In other words, although all regional directories may have the same 14 major categories, the Spanish, French, Italian, and German versions would have different orders for the categories, according to their own alphabets. For non-Roman language, such as Chinese, there seemed to have different arrangement systems: not alphabetical, nor systematic.

### 2. Implementation of Literary Warrant Principle

Web subject directories basically follow the principle of literary warrant. The depth of hierarchies in a web directory depends on the amount of web source information in a particular area. ***Yahoo!***, may divide sub-categories into three hierarchical levels (e.g., *Arts:Design Arts:Color Theory*) or nine levels (e.g., *Business and Economy: Companies: Computers: Software: Internet: World Wide Web: HTML Editors: MS Windows: HTML Assistant*). The literary warrant principle also leads to the decisions regarding the inclusion and exclusion of subordinates within a subject area. When using the regional directories of ***Yahoo!***, a user has the choice of limiting the results only from a selected region. For example, when using the ***Yahoo! UK&Ireland*** regional directory, one can request that a search be limited within UK only. It is very common, at this point, to see different depth of hierarchies and different number of subordinate categories in the final results, because they are decided by the practical situation of websites at that region. (You may check displays under the "religions: Faiths and Practices" in the various Yahoo regional directories.) Subject areas relating to culture, society, political and legal systems, business, health, etc. represent the most dynamic treatment guided by literary warrant principle.

### 3. Flexibility in reflecting local interests

While trying to keep a unique and standardized classification structure, web subject directories have also shown many possible ways to reflect local interests. First, particular main categories may be presented in a significant position when needed. Usually, all main categories are displayed according to the alphabetical order instead of a logical order. However, during the World Cup period, ***Yahoo! France*** unsurprisingly moved the SPORT category to the forefront, showing World Cup in a very significant location. Second, subordinates displayed under each main category vary from country to country and from time to time. (Please note the differences in displaying subordinates of *Arts & Humanities* in the ***Yahoo!*** regional directories.) For example, under a main category *Arts and Humanities*, regional directories prioritized subordinates to be displayed in significant places. These subordinates were chosen from dozens in the classification. The choices vary at different regional services. It is important to note that the subordinates listed under the main categories may not be the immediate subordinates of them. (Please check examples of "fashion", "literature", and photography" in ***Yahoo!*** regional directories.) The priority given to these "grandchildren"-level subordinators reflects an emphasis of local interests and indicates the flexibility that hierarchical levels and "belonging" relationships can be broken down when a subject/topic is more important than its logical position in a classification structure. Another interesting phenomenon to be noticed is the treatment of names/captions of a category. (Please check the subordinates displayed under *Business and Economy* in various ***Yahoo!*** regional directories.) It would also be very interesting to observe how names/captions are treated. In the formal presentation of the category *Business & Economy*, *"Employment"* instead of *"Jobs"* and *"Finance and Investment"* instead of *"Finance"* or *"Investing"* are used. An examination of the directory yields more similar examples. This means that the list of selected subordinates shown in the main page does not follow a restricted rule to display them according to their "official" name or captions.

Third, in the Web subject directories, there are many cross-classification treatments. For example, *"Taxes"* is listed under 93 categories when searched in all ***Yahoo!***, 122 categories when searched in *Canada only* sites, five in *UK only*, two in *Australia only*, one in *Singapore only*, and Zero in *HongKong only*. Whether *"Taxes"* is an important local interest can be determined from its listing under the main categories. In the above example, *"Taxes"* is given a significant place under the Business & Economy of the ***Yahoo! UK&Ireland*** directory. It is also listed under the Government category in USA-oriented ***Yahoo!*** directory, together with Military, Politics, and Law, which indicates the importance of this issue among current US government activities.

### 4. Limitations of language-dependent search features

With the exception of ***Yahoo!***, which accepts submissions by webpage creators and which has its staff to evaluate the descriptions of websites manually, most search engines employ a language-dependent automatic treatment of websites to rank or cluster resources based on meta tags (such as subject terms, keywords, abstracts tags in the element), page titles and word frequencies. The limitation of such automatic weighting and clustering approaches in a non-English environment is obvious. Non-English webpages may supply metadata and title in English, but the search and display based on these elements will result in non-English documents being mixed with English documents. In most cases, without installing the character code sets, a web browser will not enable reading, say, East Asian characters. Therefore, such a display of mixed languages only wastes a user's time since no content of those links could be read or understood.

Furthermore, many search engines use word frequency as a major parameter in the automatic identifying and classifying of webpage content. ***AltaVista***'s "Refine" feature uses automatic clustering theory based on word co-occurrence. By analyzing words that co-occur with the searched words in a document, documents are automatically clustered. The results are displayed through a list showing terms grouped according to co-occurrence count, or as a map showing terms and their relationships. A user can further refine a search strategy by including or excluding particular groups of words so that a higher precision of search can be achieved. However, this feature is limited to documents in English and a few Western languages; e.g. for Chinese, although ***AltaVista*** allows language-specific search, it only applies to form basic simple queries, not to the next 'refine' step.

***Northern Light*** provides a feature known as "Custom Search Folders" for refining search strategies. The service claims that its folders are not pre-set, one-size-fits-all, like other Web directories. Rather, every time a search performed on ***Northern Light***, it creates a series of Custom Search Folders based on the individual search. A user can select the subjects, types, sources, and languages he wants to explore. Based on the number of documents in each folder and their relevance to a query, the search engine determines and suggests which Custom Search Folders will be most helpful to a user. Nevertheless, only five Western languages are served at present.

## Conclusion

The road towards a fully functional cross-lingual subject access is both optimistic and sophisticated. Many other technical issues as well as social and cultural issues also need to be addressed; these include character encoding support, user interface linguistic translation, support of culture-specific data formats (date, currency, etc.), user interface graphical modification (color, images), foreign products support (e.g. databases), and operating system compatibility. In summary, there has been an increasing need for effective mechanisms to organize web resources for exploration, discovery, and retrieval. Multilingual approach to subject access, as demonstrated by major Web search engines and subject directories, has explored various ways of implementing hierarchical or classificatory structure. These new services also have progressed further beyond the traditional classification's conventions. Having the advantage of storing a classification outside the resources or their surrogates, these web-based services can be very flexible in arranging and displaying categories and their relationships, and reflecting local interests in a subject directory. The principle of literary warrant is fully functional in the practices of Web subject directories. There are still many limits when the subject classification structure and automatic clustering methodologies are used in multilingual processing. How to ensure globalization and localization in a cross-language and cross-culture environment at the same time? The question remains unanswered by the available technologies and theories.

## References

Batty, David. (1998) WWW -- wealth, weariness or waste: controlled vocabulary and thesauri in support

of online information access. D-Lib Magazine (http://www.dlib.org/dlib/november98/11batty.html).

Chan, Lois Mai. (1995). Classification, present and future. Cataloging & Classification Quarterly, 21(2), 5-17.

Koch, Traugott, Michael Day, and others. The role of classification schemes in Internet resource description and discovery. ({hyperlink http://www.ukoln.ac.uk/metadata/dsire/classification/)}

Lester, Dan. (December 1995). Profile of a Web database," Database 46-50

Lin, X. & Chan, L. M. (1997). Knowledge Class - A dynamic structure for subject access on the web. Proceedings of the 8th ASIS SIG/CR Classification Research Workshop. (November 1, Washington, D.C.). pp. 31-40.

Lynch, Clifford Lynch. (1997). Searching the Internet. Scientific American 276(3), 52-56.

Search engine watch. Compiled by Danny Sullivan. Retrieved February 12, 1999 from the World Wide Web: {hyperlink http://searchenginewatch.com/ }

Vizine-Goetz, Diane. Using library classification schemes for Internet resources

(http://www.oclc.org/oclc/man/colloq/v-g.htm)

**Latest Revision:** *June 1, 1999*