DOCUMENT RESUME

ED 441 464 IR 057 735

| | |
|---|---|
| AUTHOR | Lim, Edward |
| TITLE | Southeast Asian Subject Gateways: Examination of Their Classification Practices. |
| PUB DATE | 1999-08-00 |
| NOTE | 9p.; 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/011-117e.htm. |
| PUB TYPE | Reports - Evaluative (142) -- Speeches/Meeting Papers (150) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Access to Information; Case Studies; *Classification; Foreign Countries; *Gateway Systems; *Information Retrieval; Search Strategies; *World Wide Web |
| IDENTIFIERS | Asia (Southeast); Metadata; Search Engines; Web Sites |

ABSTRACT

This paper discusses access to information about Southeast Asia on the Internet. The first section summarizes problems with the major Internet search engines and directories. Three broad categories of solutions to improve Internet searching and retrieval are suggested in the second section: (1) efforts to assist end users to use the search engines more effectively; (2) efforts to improve the capabilities and functionalities of search engines; and (3) use of metadata and subject/information gateways. The third section lists advantages of using classification schemes to aid information retrieval in a network environment. Internet use in Southeast Asia is described in the next section. The final section presents case studies of Southeast Asian subject gateways, including: gateways created outside the region; gateways created by commercial organizations; and gateways created by Southeast Asian libraries. (Contains 21 references.) (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

Conference Proceedings

**Code Number:**011-117_E
**Division Number:** IV
**Professional Group:** Classification and Indexing
**Joint Meeting with:** -
**Meeting Number:** 117
**Simultaneous Interpretation:** *No*

# Southeast Asian Subject gateways: examination of their classification practices

**Edward Lim**
*University Librarian*
*Monash University*
*Melbourne, Australia*



**Paper** 1

## The problem

The traditional library has always built collections or provided access to information resources designed to meet the needs of its primary clientele. That is to say, the information space within which its users have to navigate tend to be more focused, confined and finite. The Internet, on the other hand, with its almost infinite information space and lack of coherent organisation, poses serious navigational problems for researchers and other information seekers. To facilitate access to the large quantity of information available, a number of indexing services have sprung up to index a large proportion of the World Wide Web. Unfortunately, however, the major search engines or directory services like Alta Vista, Northern Light, HotBot, Excite and Lycos are not wholly satisfactory for effective information retrieval for a variety of reasons.

- They only index a fraction of the total number of documents available on the Web, and their individual coverage varies widely. Because no single search service is comprehensive, more than one search engine must be used for retrieval if high precision is required. Thus Lawrence and Giles have shown that by combining six engines about 3.5 times as much information can be retrieved as by one engine (Reuters, 1998).
- Many studies have shown that the quality of the search results using different search engines vary greatly.
- Most of the search engines undertake full text indexing, but they do not necessarily index the entire contents of the documents they encounter. Frequently only the first few

lines are indexed. Furthermore, since there is a lack of vocabulary control, keyword searches can result in hundreds if not thousands of items being retrieved - much of which might not be relevant.
- Most of the search engines are almost useless for searching and retrieving non-textual documents and objects like audio, video and executable programs.
- Most search engines cannot search and retrieve password controlled sites.
- Because many Web documents do not undergo the authoritative accreditation process, which is usually found in the print environment in peer reviewed journals or in the monographs published by prestigious publishers, it is frequently difficult to separate the wheat from the chaff. Thus the problem is not only one of retrieving relevant information but also one of retrieving quality information.

Added to the problem of locating resources using any one of the major search engines is the difficulty that users have in formulating good search strategies. Reference librarians have long known that users usually express their requests in a fairly imprecise way, and it takes great skill to elicit the exact requirements of users. The problem with unmediated access, which is the direction to which libraries are being driven by financial and economic considerations, is that left to their own devices and without adequate training and guidance, users will have great difficulty in finding the information that they want from the Internet.

## The solutions

A number of solutions have been suggested to improve Internet searching and retrieval. They fall into three broad categories.The first relates to efforts to assist end users to use the search engines more effectively. This can be done through the provision of information literacy programs, printed guides, or direct assistance by librarians. The second category relates to efforts to improve the capabilities and functionalities of search engines, e.g. improving the harvesting or data gathering and indexing process, allowing natural language searching, adding the capability to mix boolean operators, increasing the limits on the number of search keys, improving the ranking algorithm, supporting truncation and wild card searching, indexing more fields, and so on. The third category involves more human intervention. This category can be divided into two sub-categories:

- The first sub-category includes the use of metadata to improve the harvesting and retrieval process.
- The second sub-category relates to attempts to build subject or information gateways which will assist users to discover quality information quickly and effectively. A subject gateway, in the context of network-based resource access, has been defined as "some facility that allows easier access to network-based resources in a defined subject area. The simplest types of subject gateways are sets of Web pages containing lists of links to resources." (Kirriemuir, 1998)

The Development of a European Service for Information on Research and Education Project, DESIRE (1998) has described the major characteristics of subject gateways as follows:

In the traditional information environment human intermediaries, such as publishers and librarians, filter and process information so that users can search catalogues and indexes of organised knowledge as opposed to raw data and disparate information. Subject gateways work on the same principle - they employ subject experts and information professionals to select, classify and catalogue Internet resources to aid search and retrieval for the users. Users are offered access to a database of Internet resource descriptions which they can search by keyword or browse by subject areas. They can do this in the knowledge that they are looking at a quality controlled collection of resources. A description of each resource is provided to help users assess very quickly its origin, content and nature, enabling them to decide if it is worth investigating further.

It is interesting that the problems identified in the Internet environment have always been

problems faced in the information science field - problems relating to the organisation of knowledge, problems caused by the vast quantity of published information (information explosion), problems of identifying, categorising and retrieving information when required, and problems that are intrinsic to complex subjects which are difficult to describe even using the traditional classification schemes. It is certainly ironic that to deal with the complex problems of information storage and retrieval, computing professionals and others are rediscovering the traditional or classical tools long used by librarians to organise and retrieve information.

## Reasons for classifying

Steps to improve the retrieval of information now combine the use of robot based search engines as well as the creation of subject gateways listing useful Web sites in specific subject areas. To improve the browsing capability of these subject gateways, some form of hierarchical browsing becomes necessary, and this in turn has led to the adoption of library classification schemes to provide the subject hierarchy.

Vizine-Goetz (1998, p.93) states that "The knowledge structures that form traditional classification schemes hold great potential for improving resource description and discovery on the Internet and for organizing electronic document collections."

The DESIRE (1999) project has produced a report outlining some of the advantages of using classification schemes to aid information retrieval in a network environment. These include:

- A classification scheme can facilitate browsing for inexperienced users.
- The hierarchical nature of classifications allows searching to be broadened or narrowed when required.
- A classification scheme gives context to search terms.
- A classification scheme permits multilingual access to a collection because the notation used is independent of any language
- The logical division of a classification scheme provides a mechanism for partitioning and manipulating the results sets.
- The use of an approved classification scheme can facilitate browsing across multiple databases.
- An established classification scheme is likely to be kept up to date.
- Library classification schemes have the potential to be well-known to their users.
- Many classification schemes are now available in machine readable form.

An examination of the practices of a number of Web sites shows that the vast majority arrange their subjects in alphabetical order, using subject descriptors, sometimes based on the controlled vocabulary of a specific subject thesaurus. Others make use of a hierarchical classification scheme. The DESIRE (1999) project has identified the classification schemes used as:

- Established schemes like the Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) or the Library of Congress Classification (LCC).
- National general schemes, e.g. Nederlandse Basisclassificatie (BC); Sveriges Allmäma Biblioteksförening (SAB).
- Subject specific schemes, e.g. Iconclass for art resources; National Library of Medicine (NLM) scheme.
- Home-grown schemes.

## The Internet in Southeast Asia

It should be stressed that while the use of the Internet has spread quite considerably in Southeast Asia, usage on a per capita basis is still not as great as in the developed countries .

By way of contrast the number of Internet users in Australia as at December 1998 was reported to be 4.36 million, representing 24.2% of the total population (NUA Ltd, 1998). Even the number of Internet Service Providers (ISP's) in Southeast Asia is fairly small. Most of these ISP's are regulated. The number of libraries that have Internet access is also very limited. Most of these are university, college or national libraries.

## Case studies

The statistics point to the fact that one should not expect too many subject gateways to be developed within Southeast Asia. Basically, Southeast Asian subject gateways can be divided into three broad groups according to their creators.

- Gateways created outside the region
- Gateways created by commercial organisations
- Gateways created by Southeast Asian libraries
- In this paper, some typical examples in each of these categories will be examined.

### *International*

It is probably not surprising that the largest subject gateway on Southeast Asia (Southeast Asian Studies WWW Virtual Library, 199?) is actually located outside the region at Leiden University in the Netherlands via the IIAS (International Institute for Asian Studies) gateway (Gateway to Asian Studies, 199?). This gateway was developed:

- to publicise Dutch research on Asian studies on the Internet by providing information relating to Dutch networked information on Asian studies;
- to provide a "window to networked information on the Internet world wide, with an emphasis on the World Wide Web".

The Southeast Asian Studies WWW Virtual Library is actually a division of the Asian Studies WWW Virtual Library, maintained by Dr T. Matthew Ciolek of the Australian National University (Asian Studies WWW Virtual Library, 1994-1999).

The classification scheme used appears to have been developed in-house. The country links show great inconsistency in the categorisation of Web sites. For example, the Philippines site is maintained by IIAS, and has therefore the same look and feel of the Southeast Asian WWW Virtual Library Homepage.

On the other hand, the Singapore and Malaysian sites, because they are maintained by different organisations, have their own idiosyncratic arrangements. The Malaysian site is maintained by the Malaysian Timber Council (Asian Studies - Malaysia, 1997) and is the least professional of all from a librarian's perspective. The Singapore site is maintained by the South/Southeast Asia Library Service (SSEALS) of the University of California, Berkeley (Singapore WWW Virtual Library, 1998) and has a more professional look. With the requisite software plug-in to a web browser, it is also possible to view the Chinese characters on that site.

### *Commercial*

Most of the major Web sites relating to Southeast Asian topics are maintained by commercial organisations. For example, the Internet Service Provider Jaring maintains a gateway to Malaysian information resources (Malaysia {HYPERLINK mailto:homep@ge} , 1999). It provides hotlinks to a large number of Malaysian web sites (Hotlinks, 1999). The Hotlinks site is very professionally designed, and provides an A-Z index of the links as well as a classification of the various resources.

### *Libraries*

Most of the Subject Gateways provided by libraries are those by university, college or national libraries, largely because these libraries tend to be better resourced and supported than public or school libraries. In spite of this, the sites do not appear to be better classified than those established by commercial sites. The three sites examined in this paper are those at Universiti Sains Malaysia (USM), the National University of Singapore (NUS) and the Asian Institute of Technology (AIT) in Thailand.

The USM site (Virtual Subject Libraries, 1996-1999) classifies the sites according to the academic schools existing in that university, viz. Biology, Chemical Sciences, Education, Humanities, Industrial Technology, Management, Social Sciences, Pharmaceutical Sciences, Housing, Building and Planning.

The NUS site (Subject Guides to Web Resources, 1999) also organises the Web resources roughly by faculty, viz. Area Studies, Arts & Humanities, Business, Engineering, Law, Library & Information Science, Social Sciences, Medicine and Science. In addition, it provides an alphabetical list of resources mainly by form or type, e.g. Electronic journals, Citation guides, Indexes, Internet site reviews, Listservs, Newspapers, Patents, Reference shelves, Travel, Universities and Colleges.

The AIT site has a page on Asian Information as well as Thai information (Guide to Thai Information on the Internet, 1998). It is clear from the brief descriptions of the above subject gateways that they are very poorly organised from a classifier's viewpoint.

- The subject groupings are not very logical or consistent. For example, the Southeast Asian WWW Virtual Library web site subdivides topics by broad subject groupings such as Arts & Humanities, Science, Social Studies, but for some unknown reason specifically excludes Government and Politics and Business (from the Arts & Humanities or Social Studies categories) and lists them separately.
- Even when subject descriptors are used, they are not always arranged alphabetically by topic.
- The arrangement appears to be random in many instances, and the reasons why the subjects are so arranged are often not made clear.
- In many instances, the groupings after the first couple of levels abandon even the fiction of logic.
- The contents appear to be a hotchpotch of web sites and individual titles of books and journals.
- The sites can be characterised more by what they do not do rather than by what they do.

## Conclusion

The above analysis shows that Southeast Asian gateways have still not adopted many of the mechanisms or technologies for facilitating efficient and effective access to Web resources. Although some of these techniques are still at the experimental level, many are based on the methodologies which have been and continue to be used by librarians to organise knowledge.

Apart from the use of traditional classification schemes, the methods or techniques used to improve access include the provision of alphabetical indexes, the creation of mechanisms for cross-searching or cross-browsing, and the ability to undertake automatic classification. CyberStacks ™, for example, which classifies its resources using LCC, provides two alphabetical indexes to the classified approach - the Cross-Classification Index and the Title Index. The Cross-Classification Index is actually no different from the traditional classified index used in the Classified Catalogue of old. CyberStacks™, however, proposes to improve the search and browse capability of this tool by introducing a structured thesaurus of subject headings in place of the currently loosely created subject descriptors (McKiernan, 1997).

While classification schemes provide a convenient means of browsing the resources in specific

subject areas, the proliferation of similar subject gateways has led to an identified need for improving resource discovery without requiring the user to go through the time consuming effort of searching several gateways in each of these areas. It has been pointed out that what is required is some mechanism to allow users to execute "a single cross-search of several subject gateways in these areas" and to have the results of this search provided in a single cumulative listing (Kirriemuir et al., 1998). In spite of the problems involved in cross-searching, a demonstrator system has been developed, using the Common Indexing Protocol, which allows various remote databases to "forward knowledge" (i.e. knowledge of the data that they each hold in advance of the end user's query being processed), and to provide 'query routing' from "a single initial database server on to others that are likely to hold relevant information."

While cross-searching across subject gateways might provide one solution to bringing together all relevant information, the truth is that many subject gateways have been set up to encourage browsing using hierarchical classification systems rather than searching. To solve the problem of "cross-browsing" or searching across different subject gateways, which use different classification schemes, writers have suggested the following solutions:

- mapping the different classification schemes of subject gateways;
- using an agreed thesaurus scheme as an index to a number of classification schemes;
- using the WWW Consortium's Resource Description Framework (RDF) to provide a common framework for the exchange of machine-readable information on the Web.

Classification is obviously a labour intensive operation. As a result there are a number of experiments to develop automatic classification mechanisms (DESIRE, 1999; Dolin et al., 1998). The most important project in this area is probably OCLC's Scorpion project, which seeks "to address the challenge of applying classification schemes and subject headings cost effectively to electronic information" (Thompson et al., 1997).

Thus, although there are many techniques and methodologies which can be used to improve access to information resources on the Internet, many Southeast Asian libraries and organisations, unfortunately, still do not have the expertise, knowledge or resources to make use of these. This situation is not unexpected, as many Southeast Asian libraries continue to work within very tight financial constraints and suffer from a lack of trained staff; and so their ability to make use of the more sophisticated techniques and mechanisms available is limited.

If there is an area that IFLA wishes to make a contribution, perhaps this is the one. There is an obvious need to conduct workshops to train Southeast Asian librarians to use many of the classification techniques to improve and enhance access to their Web subject gateways.

## References

Asian Studies - Malaysia (1997). Malaysian Timber Council. Available at:{HYPERLINK http://www.mtc.com.my/Virtual-Library/Malaysia.html} . (Accessed 03.05.1999).

Asian Studies WWW Virtual Library (1994-1999). Australian National University. Available at: {HYPERLINK http://coombs.anu.edu.au/WWWVL-AsianStudies.html/} (Accessed 03.05.1999).

DESIRE (1998) Homepage. Available at: http://www.desire/org/ (Accessed 27.04.1999).

DESIRE (1999). "The role of classification schemes in Internet resource description and discovery." Available at: http://www.ukoln.ac.uk/metadata/desire/classification/. (Accessed 27.04.99).

Dolin, R. et al. (1998). "Using automated classification for summarizing and selecting heterogeneous information sources." D-Lib Magazine, January. Available at: http://www.dlib.org/dlib/january98/dolin/01dolin.html/. (Accessed 15.04.1999).

Gateway to Asian Studies (199?). International Institute for Asian Studies. Available at: http://iias.leidenuniv.nl/wwwvl/index.html/. (Accessed 15.04.1999).

Guide to Thai Information in the Internet (1998). Asian Institute of Technology. Available at: http://emailhost.ait.ac.th/Asia/infoth.html/. (Accessed 27.04.1999).

Hotlinks (1999). Jaring. Available at: http://www.mymalaysia.net.my/hotlink/. (Accessed 03.05.1999).

Internet Service Providers in SEA (199?). SunSite. Available at: {HYPERLINK http://sunsite.nus.sg/SEAlinks/isp.html} . (Accessed 27.04.1999)

Kirriemuir, John et al. (1998) "Cross-searching subject gates; the query routing and forward knowledge approach." D-Lib Magazine, January. Available at: http://www.dlib.org/dlib/january98/01kirriemuir.html/. (Accessed 15.04.1999).

McKiernan, Gerry (1997). "Hand-made in Iowa: organizing the Web along the Lincoln Highway." D-Lib Magazine, February. Available at: http://www.dlib.org.dlib/february97/02mckiernan.html/. (Accessed 15.04.1999).

Malaysia {HYPERLINK mailto:Homep@ge} (1999). Jaring. Available at: {HYPERLINK http://www. jaring.my/}welcome.html. (Accessed 03.05.1999).

NUA Ltd. (1998). "How many online?" Available at: {HYPERLINK http://www.nua.ie/surveys/how_many_online/asia.html}. (Accessed 03.05.1999).

Reuters (1998). "Search engines fall short." Available at: {HYPERLINK http://www.news.com/News/Item/0,4,20728,00.html/}. (Accessed 03.05.1999).

Singapore WWW Virtual Library (1998). South/Southeast Asia Library Service, University of California, Berkeley. Available at: http://library.berkeley.edu/SSEAL/SouthAsia/WWWVL/singapore.html/. (Accessed 03.05.1999).

South-East Asia Information (199?). SunSite. Available at: {HYPERLINK:http://sunsite.nus.sg/asiavc.html/ }. (Accessed 27.04.1999).

Southeast Asian Studies WWW Virtual Library (199?). International Institute for Asian Studies. Available at: http://iias.leidenuniv.nl/wwwvl/southeas.html/. (Accessed 15.04.1999).

Subject Guides to Web Resources (1999). National University of Singapore. Available at: http://www.lib.nus.edu.sg/webcont.html/. (Accessed 27.04.1999).

Thompson, Roger et al. (1997). "Evaluating Dewey concepts as a knowledge base for automatic subject assignment." Available at: http://orc.rsch.oclc.org:6109/Eval_dc.html/. (Accessed 19.04.1999).

Virtual Subject Libraries (1996-1999). Universiti Sains Malaysia. Available at: http://www.lib.usm.my/bysubject/vsl2.html/. (Accessed 15.04.1999).

Vizine-Goetz, Diane (1998). "OCLC investigates using classification tools to organize Internet data." In: Clinic on Library Applications of Data Processing: 1997. Visualizing subject access for 21st century information sources; ed. by Pauline Atherton Cochrane and Eric H. Johnson, with the editorial assistance of Sandra Roe. Urbana-Champaign: Graduate School of Library & Information Science, University of Illinois at Urbana-Champaign, pp. 93-105.

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

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

# NOTICE

## REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").