SYSTEM FOR ORGANIZING AND ACCESSING WEB RESOURCES USING FACETED INDEXING

by

Francis J. Devadason *
Neelawat Intaraksa **
Ponprapa Patamawongjariya**
Kavita Desai **

ABSTRACT

Organizing and providing access to the resources on the internet has been a problem area in spite of the availability of sophisticated search engines and other software tools. There have been several attempts to organize the resources on the WWW. Some of them have tried to use traditional Library Classification Schemes such as the Library of Congress Classification, the Dewey Decimal Classification and others. However there is a need to assign proper subject headings to them and present them in a logical or hierarchical sequence to cater to the need for browsing. This paper attempts to describe an experimental system designed to organize and provide access to web documents using a faceted pre-coordinate indexing system based on the Deep Structure Indexing System (DSIS) derived from POPSI (POstulate based Permuted Subject Indexing) of Bhattacharya, and the facet analysis and chain indexing system of Ranganathan. A prototype software system has been designed to create a database of records specifying Web documents according to the Dublin Core and input a faceted subject heading according to DSIS. Synonymous terms are added to the standard terms in the heading using appropriate symbols. Once the data is entered along with a description and URL of the Web document, the record is stored in the system. More than one faceted subject heading can be assigned to a record depending on the content of the original document. The system stores the surrogates and keeps the faceted subject headings separately after establishing a link. Search is carried out on index entries derived from the faceted subject heading using chain indexing technique. If a single term is input, the system searches for its presence in the faceted subject headings and displays the subject headings in a sorted sequence reflecting an organizing sequence. If the number of retrieved headings is too large (running into more than a page) then the user has the option of entering another search term to be searched in combination. The system searches subject headings already retrieved and look for those containing the second term. The retrieved faceted subject headings can be displayed and browsed. When the relevant subject heading is selected the system displays the records with their URLs. Using the URL the original document on the web can be accessed. The prototype system developed under Windows NT environment using ASP and web server is under rigorous testing. The database and indexes management routines need further development.

*** Director, Asian Institute of Technology, Pathumthani, Thailand.**
**** Computer Science & Information Management, Asian Institute of Technology, Pathumthani, Thailand**

0 Introduction

Some of the recent findings about the Internet are that (a) it is huge (it has exploded to more than a billion Web pages), (b) it is growing very fast, (c) search engines cover only a small fraction of the web, (d) even the combined indexes of the major search engines cover only a small fraction of the Web and (e) even the most dedicated surfer using the best search system would be able to find barely one-third of the pages (Dahn, 2000). According to the survey by Roper Starch Worldwide (NUA Internet surveys report, 2001-03-01), poor search engines and information overload are causing web-rage among Internet users. On an average, users get angry and frustrated after 12 minutes of fruitless searching. For 7 percent of respondents, it only takes 3 minutes before web-rage strikes. The overwhelming amount of information on the Web is turning people away from the medium and back to pre-Internet information resources.

The main reason is that the quality of the resources retrieved are doubtful, there is no simple way to index and organize the resources on the Internet and there are no central agencies like the national libraries to get a submission of the document/resource to grant copyright protection and to assign an International Standard Identification Number to each of the resources. Several search engines using different methods of indexing the resources, and ranking them according to "their own ranking algorithms including the payment of fee as a major factor" are being used, and their number is steadily growing. Even after browsing through the first 5 to 7 pages of the retrieved records one is tempted to believe that there may be some thing more relevant on the 19th page of the ranked results. Meta-search facilities that offer several engines to be used in parallel to search for information are available. However, "the goal of information systems is to create order in a collection of documents so that an information searcher need not scan the entire collection in an attempt to find information of interest" (Fugmann 1999).

There have been several attempts to organize the electronic resources including the use of traditional library classification schemes such as, Dewey Decimal Classification, (Mundie, D. A, 1999), Library of Congress Classification, Universal Decimal Classification as well as special subject classification schemes for Medicine, Computing etc (McKiernan, 1997). Moreover several subject-based information gateway projects have been on-going (Kerr, and MacLeod, 1998; Mitchel, and Mooney, 1996) to provide organized access to the electronic resources. Use of the classification schemes though help in organizing electronic documents broadly and help browsing, they cannot be used to organize the resources precisely and function as effective retrieval tools. Analysis of some of the websites using standard classification schemes for organizing the resources (Williamson, N.J. 1997) (McIlwaine, & Williamson, 1999) identified the fact that the application of classification at the sites was often superficial and poorly executed. This becomes an acute problem as the electronic documents are mostly on recent, developing and yet to be developed subject fields and facets. Relying on such classification schemes may not fully serve the purpose as e-documents get updated fast and may change their subject with each modification and updating. Moreover, assignment of class numbers is needed for documents in a library as it is necessary to restore the documents to the shelves in the classified sequence once they are borrowed and returned or their arrangement disturbed otherwise. In other words class numbers are used to mechanize the arrangement of documents on the shelves. In the case of documents on the web, it is only their surrogates that are organized and the sequence of these surrogates are not disturbed by

the users. In fact there is no need for assignment of class numbers to restore their sequence in a mechanical way. What is needed is a tool that provides a standard frame-work/ formula for the e-document developers to fill-in as a subject heading which will have the capacity to produce an organizing sequence when sorted using the assigned subject heading as key as well as provide the necessary index terms or keywords or search terms to provide access. Such a heading can be incorporated as a Meta tag in the resources and used for both organizing and indexing the resources. This paper describes a prototype of such a system that can provide organized access to networked resources.

2 Faceted Indexing

2.1 Categories of POstulate Based Permuted Subject Indexing) POPSI/ DSIS

Kaiser's "Systematic indexing" (Kaiser, J. 1911) was perhaps the first category based subject indexing system. The Deep Structure Indexing System (Devadason, F.J. 1986) is based on the POstulate-based Permuted Subject Indexing (POPSI) (Bhattacharyya, G. 1979) derived from the Chain Indexing system of Ranganathan especially his concept of facet analysis as applied to subject headings (Ranganathan, S.R. 1964). There have been studies of the relevance of facet analysis to search and organize the resources on the web (Ellis and Vasconcelos 1999). The component ideas in a subject heading, can be deemed to fall into any one of the Elementary Categories or Facets: Discipline (Base), Entity (Core), Property and Action. Each of these may admit of sub-divisions: species/type, and part. The Entity may have constituents too. In the case of `bamboo'; `arundinarisae', `bambusa lineata', `Brazilian bamboo' are species/types; `bud' `flower', `leaf', `rhizome', `culm', `root' are parts; `ash', `lignin', `peptin', `cellulose' etc., are constituents. Apart from the elementary categories, a special component called Modifier is recognized. For example: "red" in 'red rose', "concrete" in 'concrete bridge'. A modifier generally creates a species/type of the concept modified. Modifiers can be Common Modifiers like Form, Time, Place and Environment and Special Modifiers, which can be based on any of the elementary categories. Generally common modifiers can modify a combination of two or more category occurrences in a subject heading. There are also two other kinds of special modifiers. Independent modifiers can modify the focus independently without depending on any other modifier. Dependent modifiers cannot modify the focus directly but can modify only another modifier of the focus. For example: Temperature, High temperature, Very high temperature. When concepts are represented by terms there may be cases wherein a term (composite term) may have to be broken down (factored) into two or more constituent terms and each one of them identified as belonging to one or the other of the elementary categories. In Kaiser's systematic indexing, system agriculture is represented as a combination of land + cultivation corresponding to his categories of concrete and process. A composite term is considered as a synonymous term to the combination of the factored constituent terms.

2.2 Syntax of the Subject Headings

The basic rule of syntax for formulating subject headings is DISCIPLINE (BASE) followed by ENTITY (CORE OBJECT) which is followed by PROPERTY and/or ACTION. PROPERTY and/or ACTION may be further followed by PROPERTY and/or ACTION as

the case may be, followed by COMMON MODIFIERS. The SPECIES/TYPES and/or MODIFIERS and/or PARTS and/or CONSTITUENTS for each of the Elementary Categories follow immediately the manifestation to which they are respectively SPECIES/TYPES or MODIFIERS or PARTS or CONSTITUENTS without the manifestation of any other Elementary Category intervening. The rules of syntax give rise to a context-dependent sequence of the components in the subject heading.

2.3 Indicators of the Structure

Certain numeric codes have been prescribed to indicate the categories and their subdivisions to which individual concepts belong. One set of these codes are shown below:

0 Form Modifier	9 Discipline/Base
2 Time Modifier	8 Entity/Core Object
3 Environment Modifier	.2 Property
4 Place Modifier	.1 Action

SUBDIVISIONS

.3 Constituent
.4 Part
.5 Modifier of Kind 1 including Phase Relation Modifier
.6 Species/Type, including those created by Modifiers of Kind 2

In the subject headings the indicators precede the components to which they are indicators. The indicators for property and action as well as the sub-divisions (species, part etc.) are attached with the indicators for the elementary categories to which they are respectively property or action or sub-divisions.

2.4 Formulation of Subject Headings

Taking the title of the resource as the starting point, each of the specific subjects dealt within the resource are expressed in natural language. Each of the specific subjects (topics) may warrant a separate subject heading. Each of the component ideas corresponding to each of the elementary categories that are implied is explicitly stated to form expressive titles. Let one of the expressive titles be "In Leather technology, dry salt curing of pig skin in Thailand". This is then analyzed to identify the elementary categories and their sub-divisions to which each of the components in the expressive title belong. All composite terms are factored into their constituent terms and identified as belonging to one of the elementary categories. The component terms are written down as a formalized expression following the rules of syntax, as given below:

(Discipline/Base) Leather Technology (Entity/Core) Pigskin (Action) Dry salt curing (Place Modifier) Thailand.

Each of the component terms in the subject heading is then analyzed to find out its super-ordinate terms. This is done by finding out "of which the concerned component is a species/ type, or part or constituent" in the context of the subject as a whole. This process is continued

with each of such super-ordinates recognized in the process till it ends up with the concept of the elementary category. For this purpose terminological sources such as thesauri, glossaries, dictionaries etc., are to be used. Each of the super-ordinates thus recognized are fixed prior to the concerned term successively giving rise to a 'modulated' subject heading as follows:

(D/B) Leather technology (E/C) Hide and skin (Part of E) Skin (Type of E) Pig skin (A) Beam-house operation (Sub-action) Curing (Type of A) Salt curing (Type of A) Dry salt curing (Common modifier) Thailand.

NOTE: The reason for making each of the super-ordinates to precede the respective component terms is to endow the subject headings with the capacity to produce an organizing sequence effect resembling the sequence of class numbers when sorted alphabetically along with similarly formulated other subject headings. Moreover this will facilitate search using any of the super-ordinate terms also.

Each of the component terms in the subject heading is replaced with standard terms and the synonyms are attached to the standard terms with an appropriate symbol such as an equal to sign (=). For this purpose vocabulary control tools such as thesauri, classauri (Bhattacharyya, 1982; Devadason and Ramanujam, 1982; Devadason, 1985) are used. Appropriate indicators for the elementary categories, their sub-divisions, and common modifiers of different kinds are inserted in the appropriate places. The resulting subject heading is as follows:

Leather technology 8 Hide and skin 8.4 Skin 8.6 Pig skin 8.1 Beam-house operation 8.1.4 Curing 8.1.6 Salt curing 8.1.6 Dry salt curing 4 Thailand

2.5 Organizing Sequence in Alphabetical Arrangement

A set of subject headings formulated according to this system when sorted alphabetically resembles an organizing (classified) sequence as illustrated below:
Subjects according to Colon Class numbers:

L''aN5	= Medicine, Bibliography, 1950's
L''k1, N3	= Medicine, Dictionary.
L''v1'N	= Medicine, History, 19^{th} century
L''v44'N	= Medicine, History, India, 19^{th} century
L	= Medicine
L, 0;2	= Medicine, Anatomy
L,0;3	= Medicine, Physiology
L,0;4	= Medicine, Disease
L,0;4:3	= Medicine, Disease, Diagnosis
L,0;4:6	= Medicine, Disease, Treatment
L,9C	= Medicine, Child
L,9C;4	= Medicine, Child, Disease
L,9C;42	= Medicine, Child, Disease, Infectious disease
L,9C;42:6	= Medicine, Child, Disease, Infectious disease, Treatment
L,9C,4;4	= Medicine, Child, Respiratory system, Disease

L,9C,45;4 = Medicine, Child, Respiratory system, Lung, Disease

The same subjects with headings formulated according to the indexing system and sorted alphabetically:

Medicine 0 Bibliography 2 Nineteen fifties
Medicine 0 Dictionary
Medicine 0 History 2 Nineteenth century
Medicine 0 History 4 India 2 Nineteenth century
Medicine 1
Medicine 8.1 Physiology
Medicine 8.2 Anatomy
Medicine 8.2 Disease
Medicine 8.2 Disease 8.2.1 Diagnosis
Medicine 8.2 Disease 8.2.1 Treatment
Medicine 8 Child
Medicine 8 Child 8.2 Disease
Medicine 8 Child 8.2 Disease 8.2.6 Infectious disease
Medicine 8 Child 8.2 Disease 8.2.6 Infectious disease 8.2.1 Treatment
Medicine 8 Child 8.4 Respiratory system 8.2 Disease 8.2.6 Infectious disease 8.2.1 Treatment
Medicine 8 Child 8.4 Respiratory system 8.4 Lung 8.2 Disease 8.2.6 Infectious disease 8.2.1 Treatment

It may be seen that the two sequences are almost similar. In other words it is possible to bring an organizing sequence effect in the alphabetic arrangement of subject headings. Once the sequence is obtained, the indicators may be suppressed in the display or changed to suitable punctuation marks.

3 Access System for Web Resources

A prototype system for web resources using the POPSI/ Deep Structure Indexing System has been built which has the following functions:

System functions
The system staff namely, the administrator, manager, cataloger, indexer and the system users who access and search, can perform the following functions:

System Staff:

- ?? Select resources from sites recommended by users
- ?? Add new resources to the database
- ?? Summarize the resource descriptions
- ?? Revise the resource summaries or abstracts
- ?? Index the resources
- ?? Maintain subject headings
- ?? Revise and store resource descriptions as per standard metadata and create inverted index files

?? Monitor and review, link check and maintain the system

System Users:

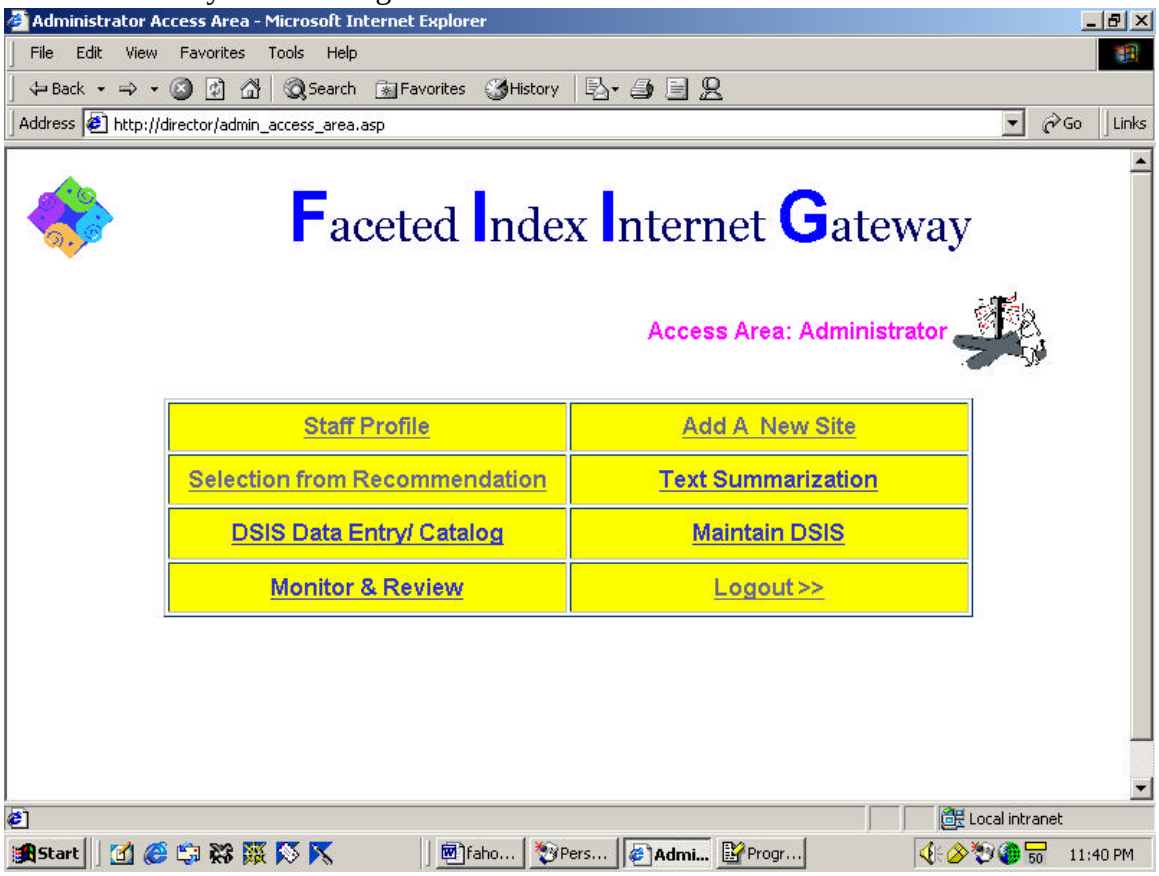
?? Browse the subject headings displayed in hierarchies;
?? Search the database with the use of search words;
?? Display the retrieved records in a subject hierarchic order;
?? Recommend new sites; and
?? Access original resources using the link provided in the retrieved records

System components

The main components of the system are:

?? Staff maintenance module;
?? Automatic text summarization and meta data module;
?? Cataloging and Indexing module;
?? Database and Index (DSIS) generation and maintenance module;
?? Search and Browse module; and
?? System maintenance module.

Figure 1, the Opening screen of the system for the administrators, shows the major functions of the system staff as given below:



The Staff Profile function allows the administrators to add new staff (cataloguer, indexer) and to maintain the staff profile and password authentication. The Selection from Recommendation allows the system staff to select the sites recommended by the users, evaluate them and add them to the system. The following screen (Figure – 2) shows the list of recommended sites and whether they are selected or not.

Figure – 2: Recommended Sites Listing

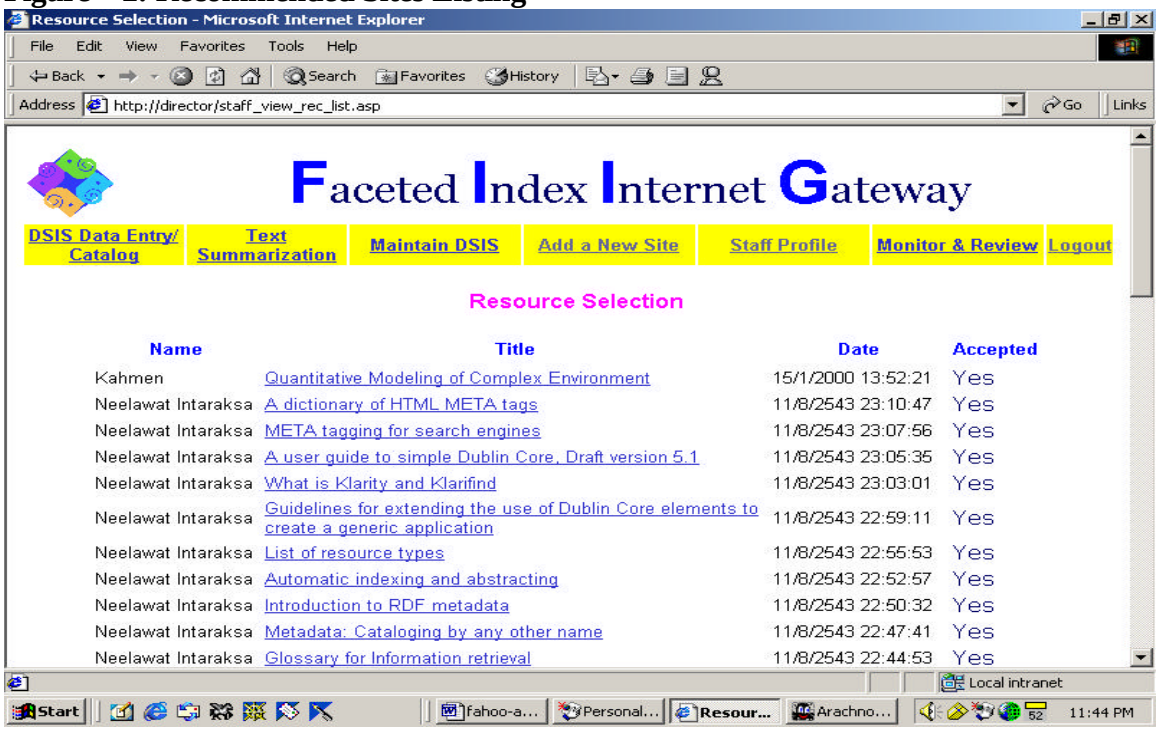
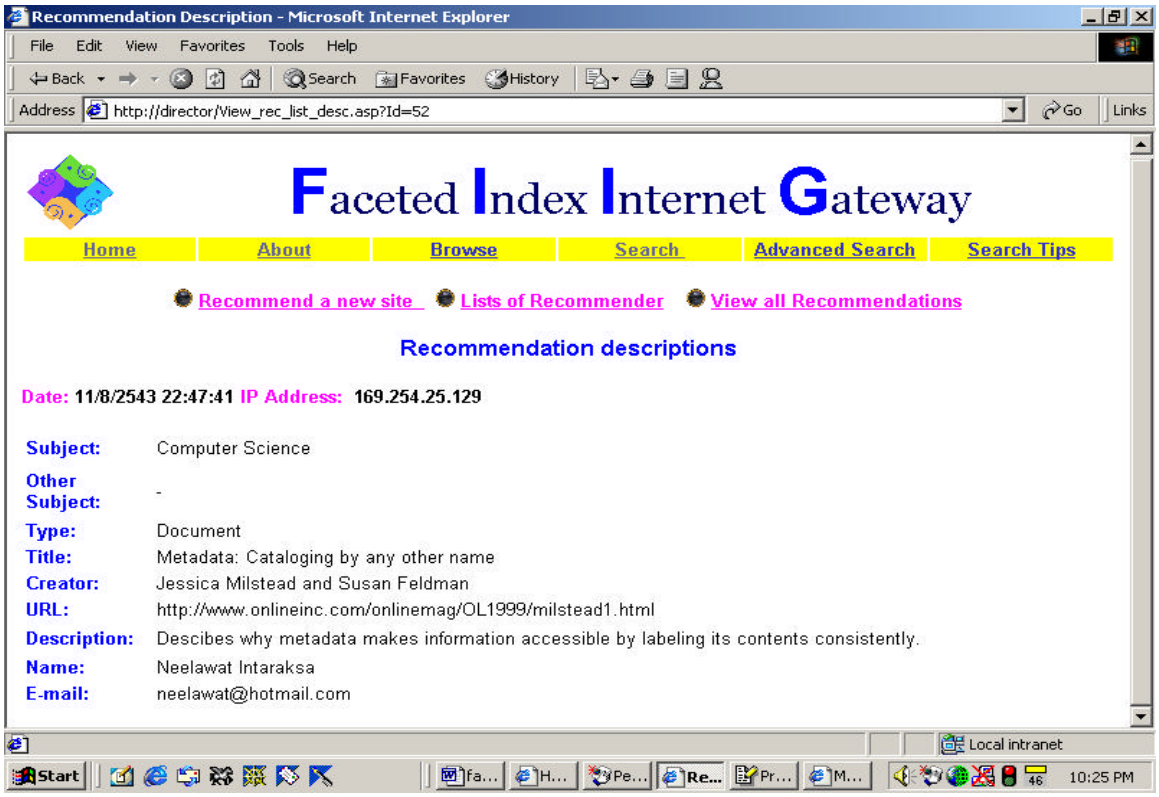


Figure – 3 below shows the description of the site as recommended by the users showing as far as possible the meta data identified and supplied by the user.



The selected resource can be added to the system, and the record is displayed as per the Dublin Core data elements with a summary of the text formed automatically as shown below:

Figure : 4 The Data Elements

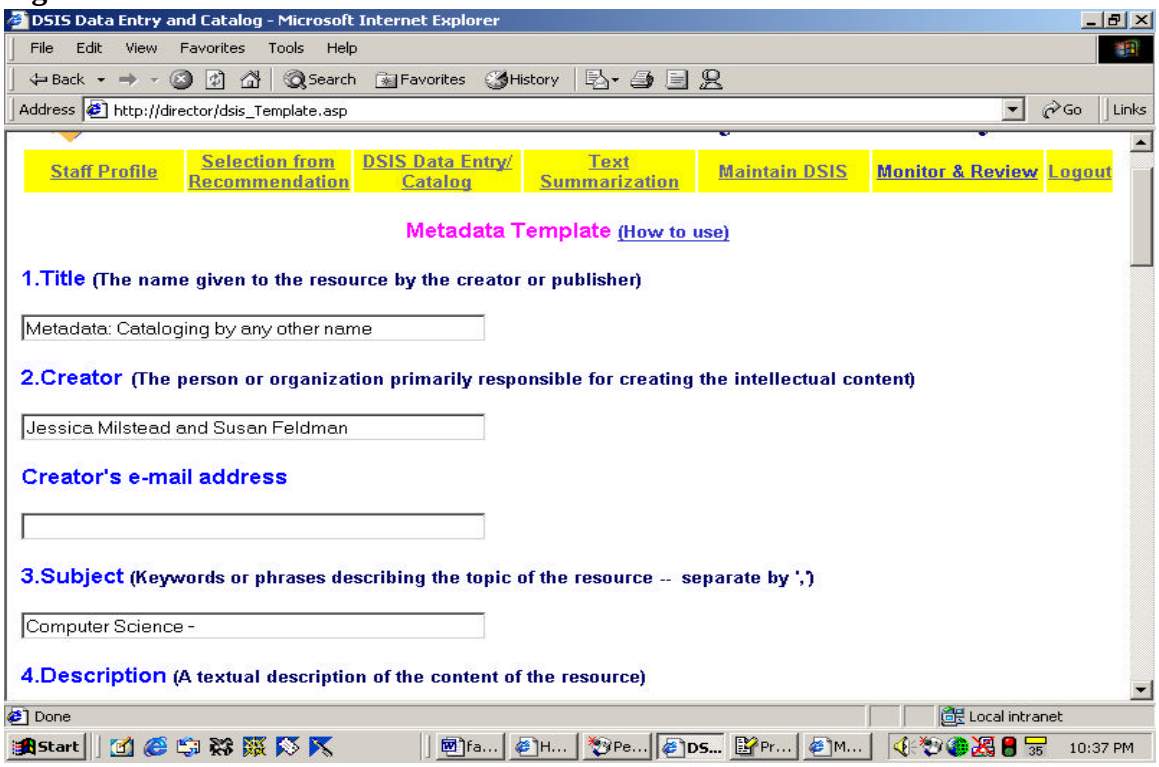


Figure 5: Data Elements (Contd.)

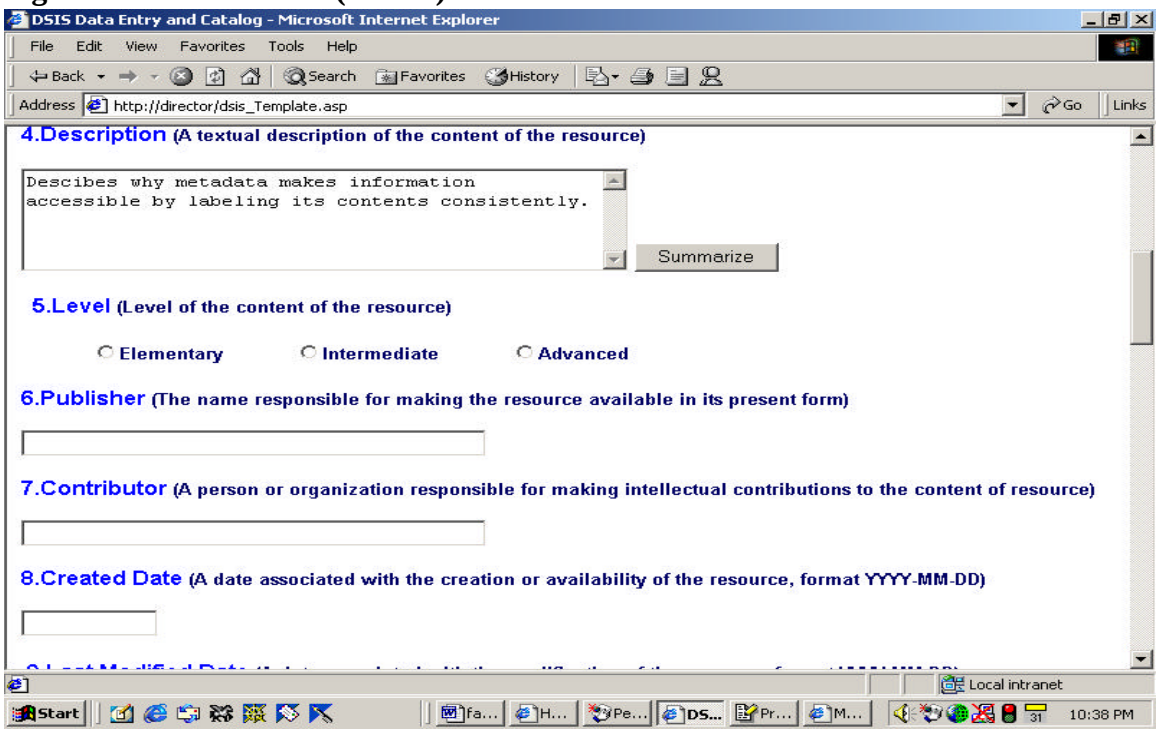


Figure 6: Data Elements (Contd.)

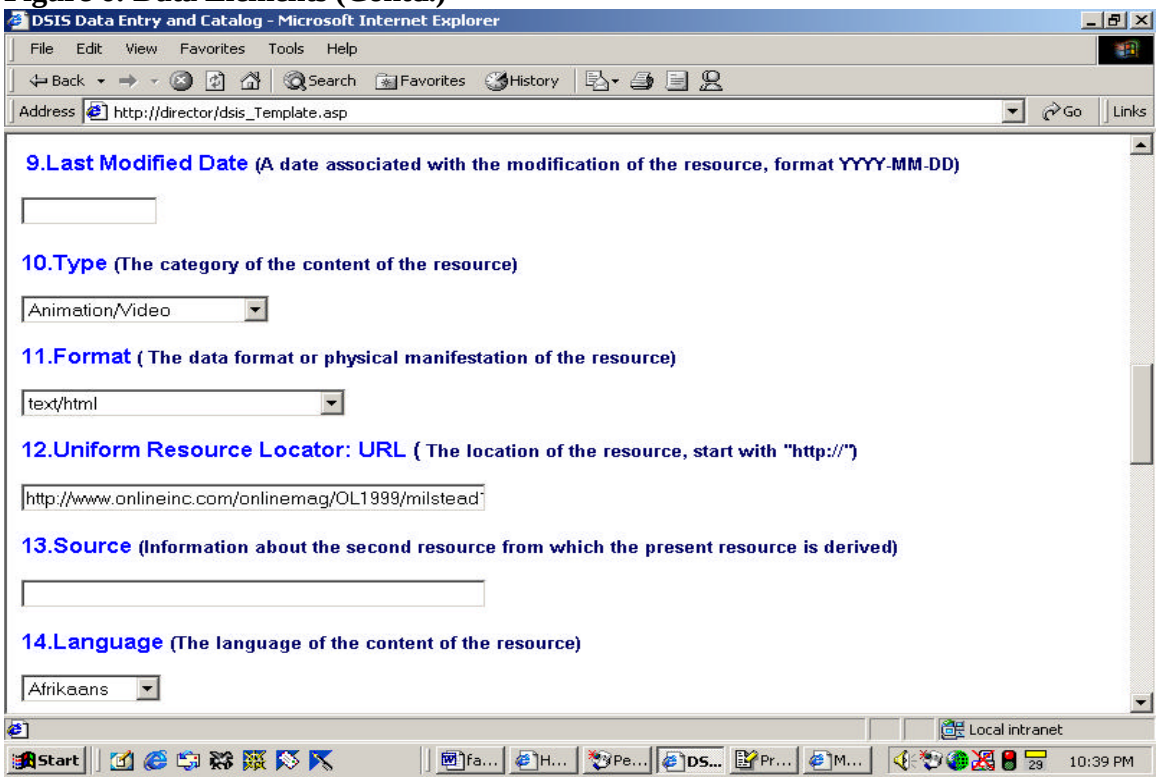


Figure 7: Data Elements (Contd.)

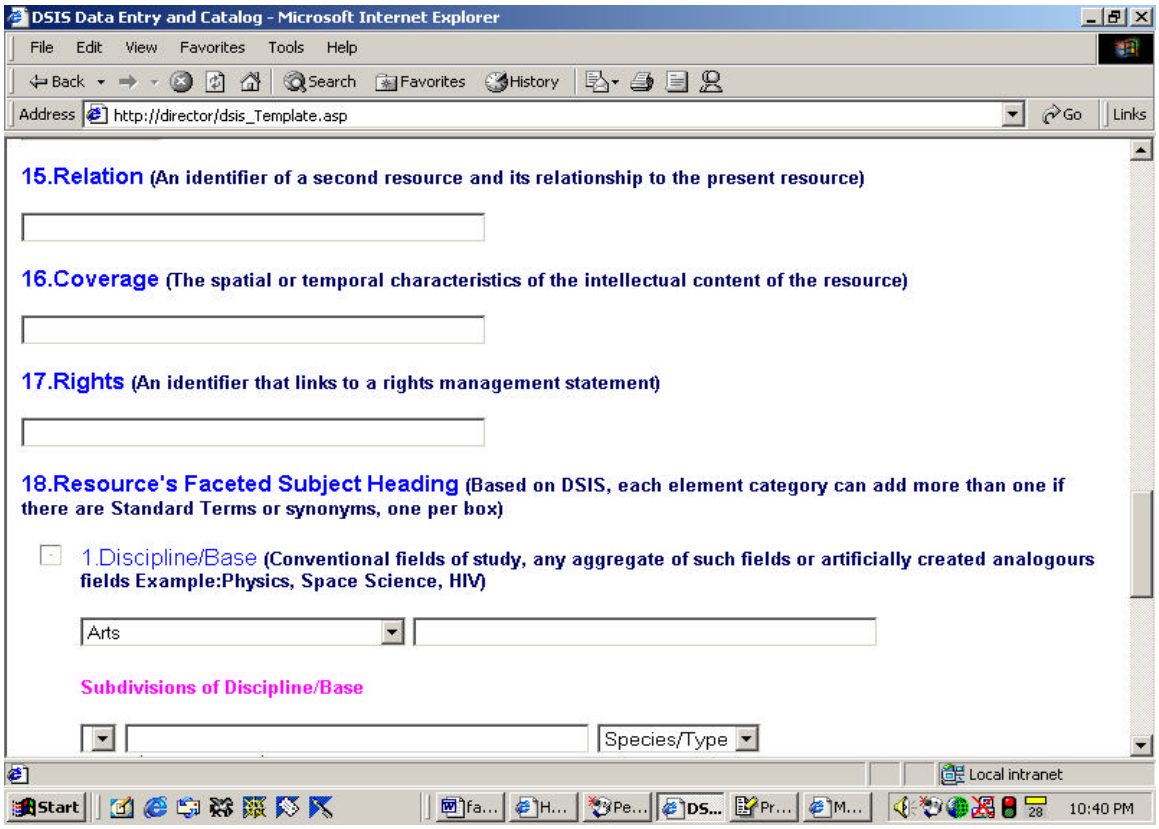
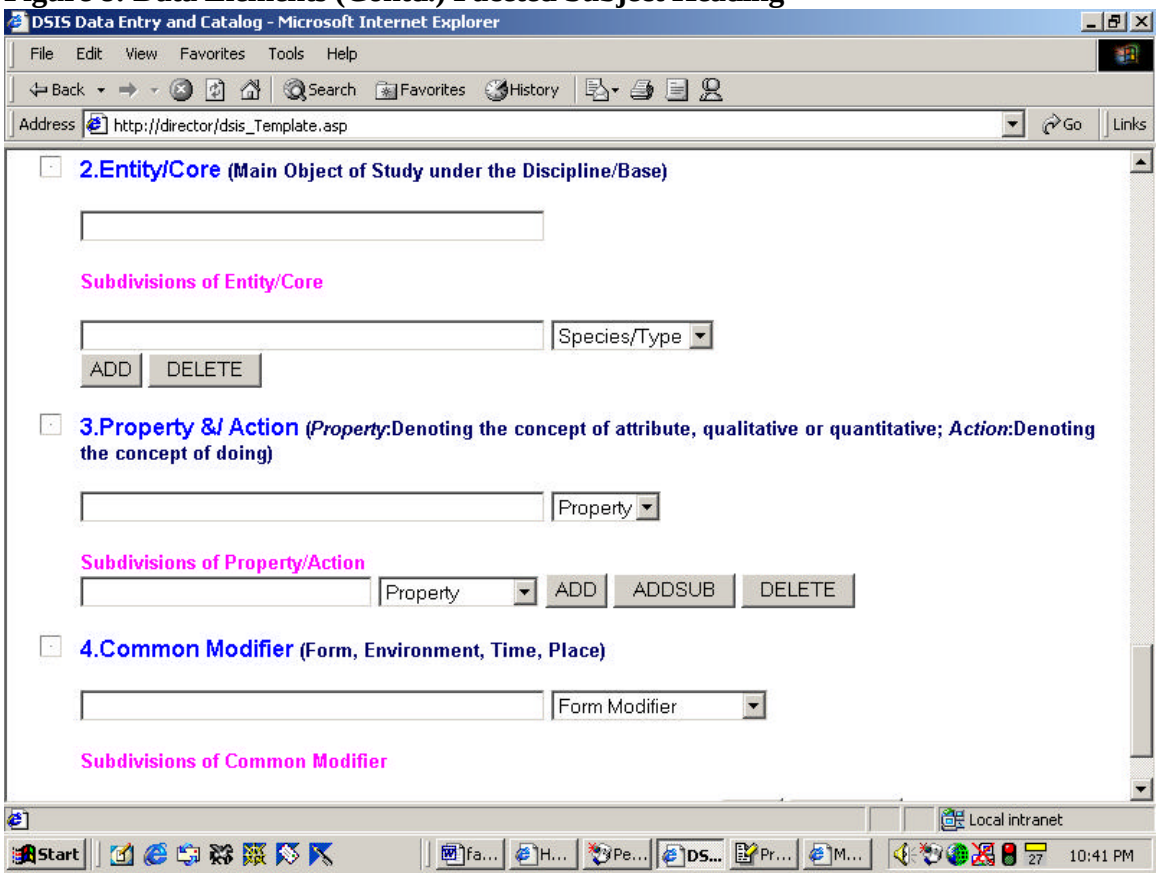

DSIS Data Entry and Catalog - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://director/dsis_Template.asp
Go Links
15.Relation (An identifier of a second resource and its relationship to the present resource)
16.Coverage (The spatial or temporal characteristics of the intellectual content of the resource)
17.Rights (An identifier that links to a rights management statement)
18.Resource's Faceted Subject Heading (Based on DSIS, each element category can add more than one if there are Standard Terms or synonyms, one per box)
1.Discipline/Base (Conventional fields of study, any aggregate of such fields or artificially created analogours fields Example:Physics, Space Science, HIV)
Arts
Subdivisions of Discipline/Base
Species/Type
Local intranet
Start
10:40 PM

Figure 8: Data Elements (Contd.) Faceted Subject Heading



The automatic text summarization module requires three values to be input viz.,

a) minimum frequency for a word to be taken as significant;
b) minimum number of significant words to form a cluster; and
c) maximum distance between two significant words in a cluster to ex tract the sentence to form the summary

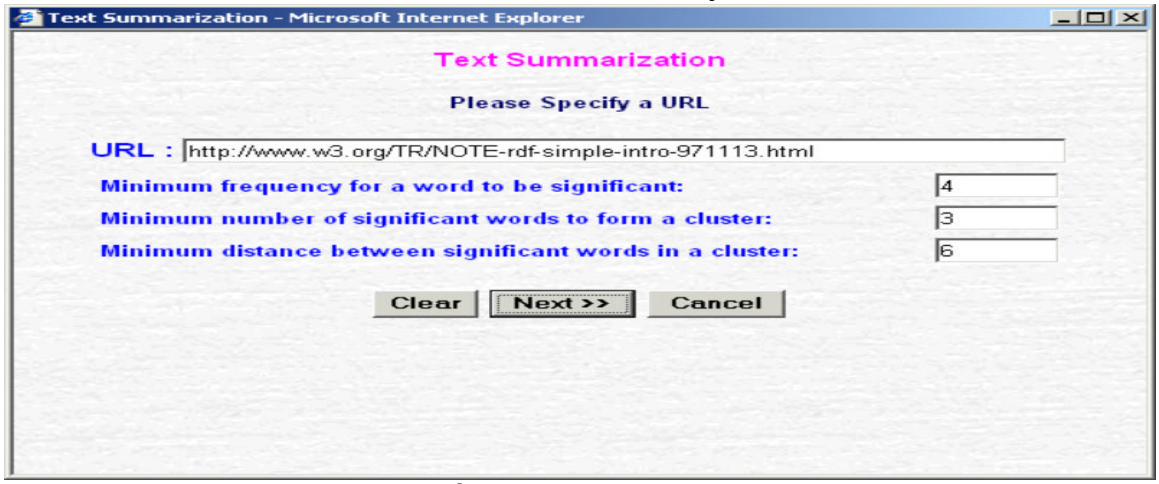
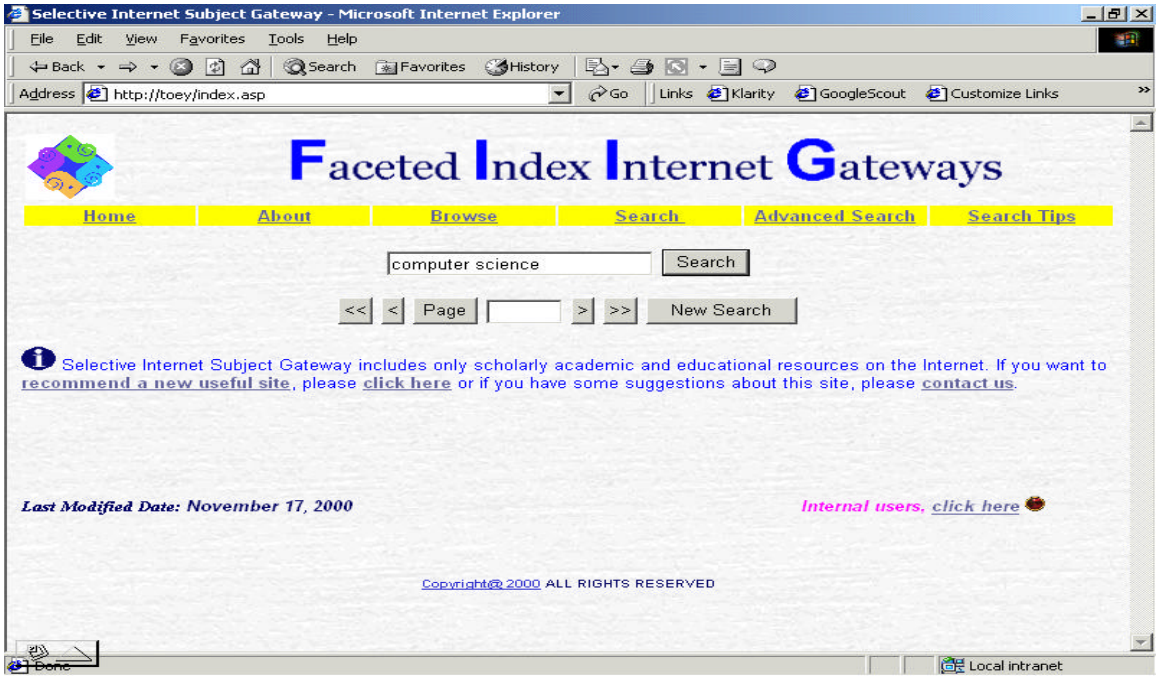


Figure – 9: Summarization of Text

The summarizer uses a list of stop words to remove common words in the beginning of the text processing. The size of the summary may be modified by changing the values supplied. It may be necessary to edit the summary to resolve any dangling sentences. It is possible to add a new site directly by specifying the URL. The system automatically access the resource, pulls out the metadata (cataloging elements), fills up the template, takes input for summarization and adds a summary.

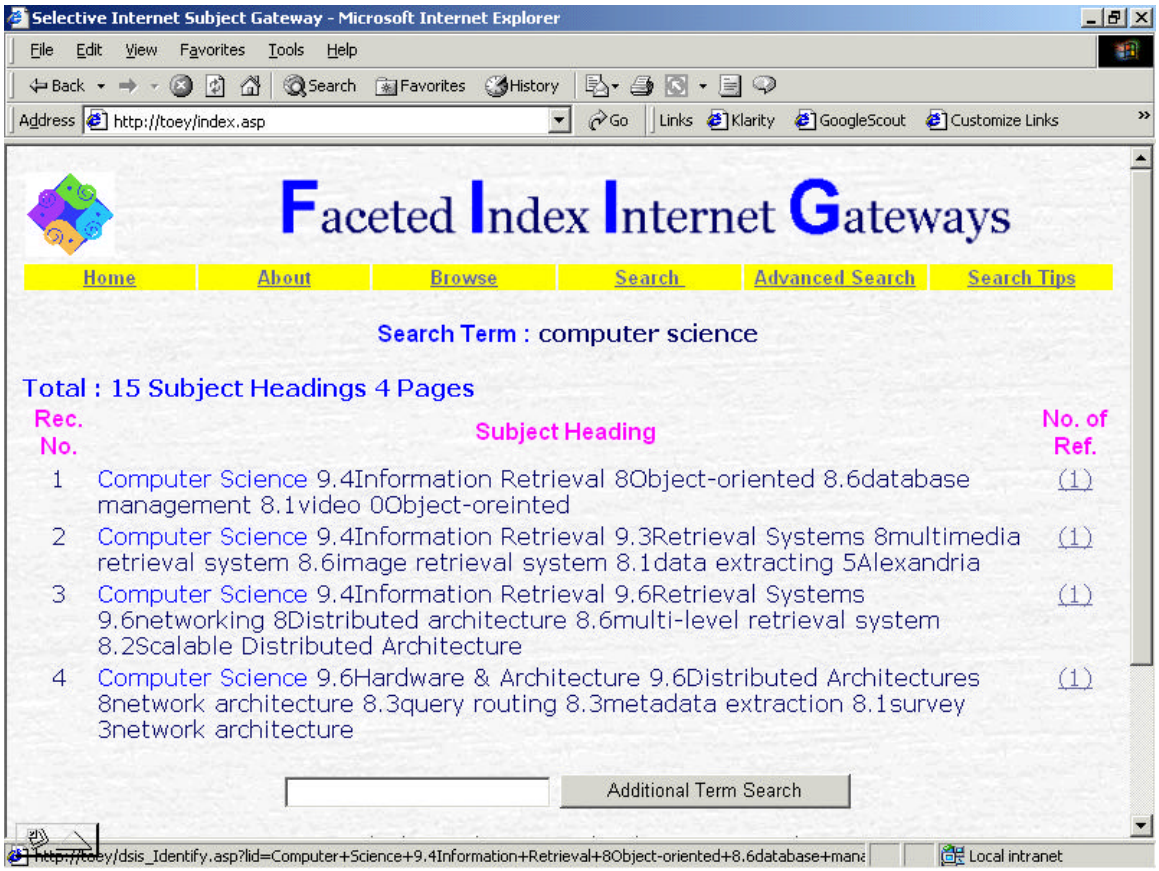
Figures 7 & 8 above show the portion of the input template that allows the faceted subject heading to be input manually. The Discipline / Base can be selected from the list displayed in the selection window. The Categories Property and Action are repeatable so also the addition of species and parts. The input template accommodates all these. Moreover more than one faceted subject heading could be assigned to a resource. Once the data is entered the record can be browsed and edited and corrected. Desired record can be retrieved by entering some of the input data and edited and modified. Once the records are entered and admitted to the system, the system stores the records in the data base and prepares the required indexes from the faceted subject heading for searching The following Figure - 10 shows the search screen. Once a search term(s) is input the system searches for faceted subject headings having the term(s) and displays in sorted sequence as shown in Figure - 11 next.

Figure 10: Search Screen



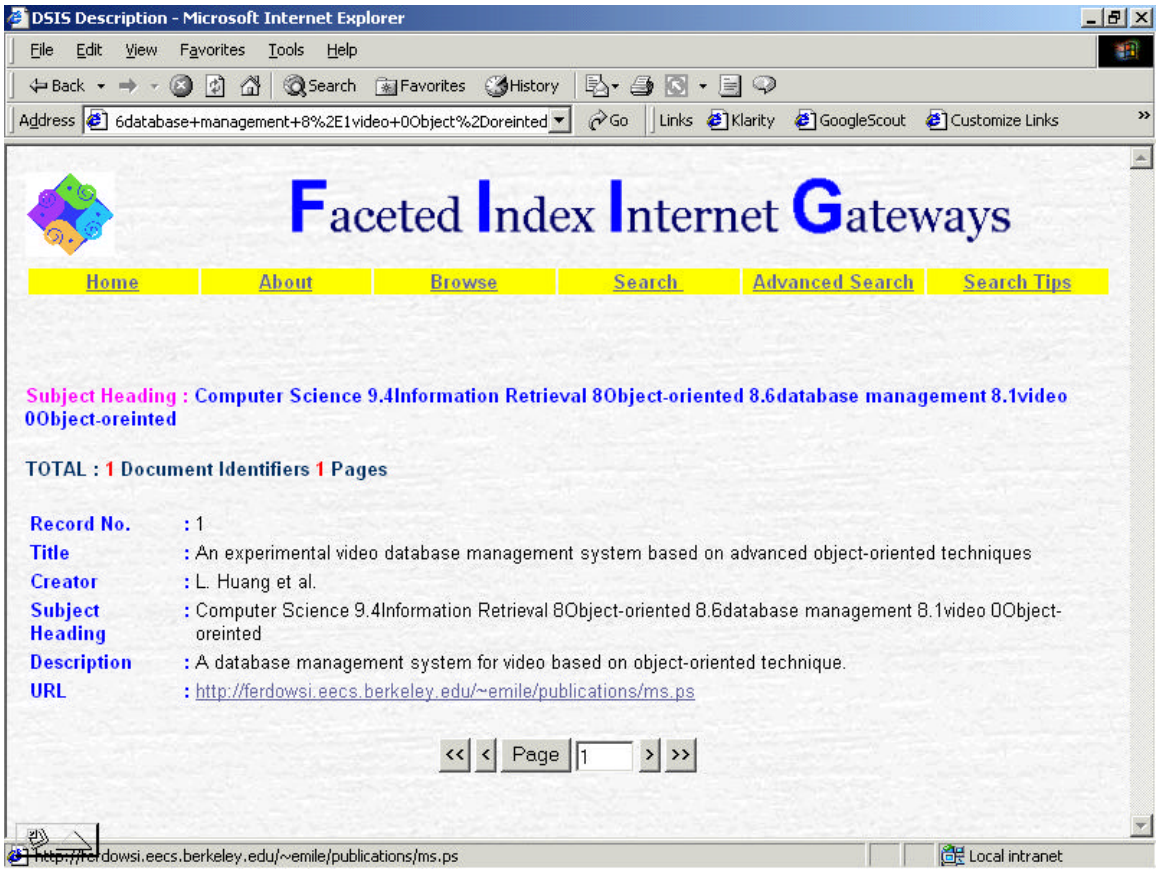
If the subject headings span more than one page, you can go browse through the pages to select the subject heading that best describes your query or input another term to be searched in combination. The faceted subject headings containing both the terms would be selected and displayed along with the number of records that would be retrieved for each. The records can be displayed by clicking this number at the right hand side. Clicking on the URL can

Figure 11: Retrieved subject headings in Organized Sequence



You may browse through the pages to select the subject heading that best describes your query or if the retrieved subject headings span more than one page you may input another additional term to be searched in combination. The faceted subject heading containing both the terms would be selected and displayed along with the number of records that would be retrieved for each. The records can be displayed by clicking this number on the right hand side. The record display will show the URL which can be clicked to access the resource on the net.

Figure – 12: Retrieved Record Display



It is possible to change the display to produce an organized index of different levels to make browsing easy. By bringing the required category (along with its sub-divisions) to the beginning of the subject chain, the organizing effect can be changed. The sequence can be varied according to the user. It can be presented as discipline-based if the user of the system is an academic. It can be changed to entity (object) based or action (process) based. When resources are collected for indexing and to build search files, it may be possible to exercise control over synonyms and standardize the terms used. It may also be possible to have a switching from non-standard terms used by the searcher to the standard ones. When any term used by the searcher is poly-hierarchic, then the subject chains in which the term occurs (only the super-ordinates in the category to which the term belongs) can be displayed, perhaps along with the term denoting the discipline or base in reverse order and the user can be asked to select the appropriate chain representing his/her query. It is possible to incorporate synonyms due to semantic factoring so that the system searches for the combination of factored terms to be present in faceted subject headings. It is possible to sort the headings and present higher level indexes for easy browsing. It is also possible to resolve the meaning of homonyms by displaying the full faceted subject headings having the term and asking the user to select the subject heading that represents the intended meaning of the searcher

4 Limitations and Labor Intensiveness

An indexing approach of this analytical type using facet analysis perhaps is best suited for scholarly academic resources having educational and research value. Moreover there may be cases wherein it would be difficult to recognize the categories in certain subjects in spite of illustrative examples. Provision of such subject chains for a resource as a whole and for the significant sub-units would be time consuming as vocabulary control tools and terminological sources of different types may have to be consulted and used. Normally it is expected to take about three months to train an indexer. To begin with an indexer would be able to provide the faceted subject heading for 5 resources a day which may increase to 10 per day in a weeks time. The optimum would be about 20 resources per day after working for about two months.

5 Conclusion

As mentioned by Weinberg (1996) "it is hoped that the systems for organizing information developed earlier will not be ignored, that their design flaws will not be replicated, and that our increasing knowledge of human factors will be incorporated into systems for indexing the Internet". Further work in adding modules for automatic text summarization and index terms suggestion as well as management functions for the administrator, cataloguer etc., are ongoing.

6. References

BHATTACHARYYA (G), POPSI: Its fundamentals and procedure based on a general theory of subject indexing languages. Library Science with a slant to Documentation. 16, 1;1979. p 1-34.

BHATTACHARYYA (G) Classaurus: Its fundamentals, design and use. In Proceedings of the 4th International Study Conference on Classification Research. Augsburg. June 28- July 2, 1982. Frankfurt, Indeks Verlag. p 139-148.

DAHN (MICHAEL) Counting angels on a pinhead: Critically interpreting web size estimates. Online. 24, 1; 2000 p 40.

DEVADASON (F J) and KOTHANDA (RAMANUJAM M). Computer aided construction of 'alphabetic' classaurus. In Proceedings of the 4th International Study Conference on Classification Research. Augsburg. June 28- July 2, 1982. Frankfurt, Indeks Verlag. p 173 – 182.

DEVADASON (F J). Online construction of alphabetic classaurus: A vocabulary control and indexing tool. Information Processing and Management. 21,1; 1985. p 11- 26.

DEVADASON, (F J). Computerized Deep Structure Indexing System. FID/CR Report 1986. Indeks Verlaag, Frankfurt. p 21.

ELLIS, D and VASCONCELOS (A). Ranganathan and the Net: using facet analysis to search and organize the world wide web. Aslib Proceedings. 51,1;1999. p 3-10.

FUGMANN (R). The empirical approach in the evaluation of information systems. Knowledge Organization: International Journal. 26,1;1999, p 3.

KAISER (J). Systematic indexing. The Card System Series. 2; 1911. Pitman & Sons, London.

KERR (LINDA) and MACLEOD (RODDY). Subject-based Information Gateways. Ariadne.10 Dec., 1998.
http://www.ariadne.ac.uk/issue18/subject-gateways/

MCILWAINE (I C) and WILLIAMSON (N J). International trends in subject analysis research. Knowledge Organization. International Journal. 16,1;1999. p 27.

MCKIERNAN (GERRY). Hand-made in Iowa: Organizing the web along the Lincoln highway. D Lib Magazine. 1997 Feb.
http://www.dlib.org/dlib/february97/02mckiernan.html

MITCHEL (STEVE) and MOONEY (MARGARET). INFOMINE: A model web-based academic virtual library. Information Technology and Libraries. 15,1;1996.
http://lib-www.ucr.edu/pubs/italmine.html

MUNDIE (DAVID A). Organizing computer resources: Or, How I learned to stop worrying and love the DDC. 12 May 1999.
http://ivory.lm.com/~mundie/CyberDewey/organizing_computers.html

Nua Internet Surveys - Search engines cause ire among Net users. 2001-03-01.
http://www.nua.com/surveys/

RANGANATHAN (S R). Subject heading and facet analysis. Journal of Documentation. 20,3;1964 p 109 - 119.

TINKER (AMANDA J) et al. The Dewey Decimal Classification and the transition from physical to electronic knowledge organization. Knowledge Organization. International Journal. 26,2;1999. p 80-96.

WEINBERG (BELLA HASS). Complexity in indexing systems - abandonment and failure: Implications for organizing the Internet. ASIS 1996 Annual Conference. 19 - 24 Oct, 1996.
http://www.asis.org/annual-96/ElectronicProceedings/weinberg.html

WILLIAMSON (N J). Knowledge structures and the Internet. In: Knowledge Organization for Information Retrieval: Proceedings of the 6th International Study Conference on Classification Research, University College, London, 16-18 June 1997. The Hague, Netherlands: International Federation for Information and Documentation, 23- 27.