Interoperable Metadata Framework to facilitate Retrieval of Educational Resources from the Internet

M.Y Chuttur
School of Library and Information Science,
Indiana University, USA.
mchuttur@indiana.edu

Abstract

The Internet is in doubt the largest network ever with a large collection of useful resources. Major related concern however lies in how to store, organize and retrieve information from this ever growing network. This paper focuses exclusively on educational resources available on the Internet, highlighting the main limitations faced by users in actually getting access and retrieving information they desire. Information search process over the Internet is explained and use of controlled Meta data for efficient retrieval over the Internet is presented. Semantic interoperability as a key issue is also discussed in this paper.

Keywords: Metadata, Semantic Interoperability, Search Engines, Internet, Education.

Introduction

With the advent of cheaper computers and networking facilities, it has become a very common practice for university academics today to use dedicated web servers on which they can post their lecture notes to students. This considerably increases accessibility of educational resources via the Internet whereby students can use a web browser to locate any materials of interest and eventually download them from any location and any time at their convenience. Besides lecture notes, these servers can also contain publications, thesis, student presentations, reports, past exam papers, assignments or any other materials that may be of relevance to the education service delivered by the particular university.

Similarly, professional bodies who are involved in enhancing education and research, have adopted the Internet as a very appropriate means to reach a wide range of audience to support their purpose. Popular organizations like the **IEEE (www.ieee.org), ACM (www.acm.org), and DLib.org (www.dlib.org)** have got dedicated portals through which students/academics/researches can have access to a large collection of information that aid in their education process. Other educational websites, like **www.citeseer.com, www.w3schools.com** etc. freely provide large collections of research publications and/or online tutorials that help in providing a centralized service for a large number of research papers and training materials in the field of computing. This strongly helps interested students and researchers to stay current with ongoing developments in the IT field.

In this way, Internet users are faced with a multitude of services offered to them via different websites that is university websites, educational websites and professional body's websites, from which they are going to locate and download any resources of interest.

Unfortunately, with the Internet growing at a very fast rate, together with new websites being created almost every day and more and more educational materials being added frequently, internet users are not always able to keep track of all the web site addresses (URLs) that are available to them. Furthermore the process of accessing each of these numerous websites to browse for educational materials often turn out to be tedious and very time consuming. Furthermore, the amount of irrelevant materials returned as a result of using search engines, also contributes to frustration on the part of users. Very often, huge amount of time and resources (bandwidth, internet connection) would have been wasted in accessing individual websites or using search engines in looking for educational materials on the Internet.

The above problems actually highlight the main limitations of the Internet, i.e. *uncontrolled distribution of resources* and *inefficient retrieval of required information, to play a key role in the dissemination of educational resources from existing rich content distributed educational web servers.*

This paper focuses on these limitations and provides for a framework that can assist students, academics and researchers to readily and efficiently obtain relevant information from different university/educational web servers using a single interface. The proposed solution exploits the features of search engines and the use of **controlled Metadata** with special attention given to semantic interoperability among different web servers. . The paper is organized as follows; current information search process and related limitations over the Internet are described, various measures proposed so far to cater for these limitations are discussed and finally a framework that enhances information access and retrieval over the Internet is proposed. The paper ends with a conclusion after a short discussion of the implications of such a framework.

Information search over the Internet

Currently Internet users will access information using the following three main methods:

- Selection from classification links (examples are education, entertainment, travel, health etc.) provided by websites.
- Browsing through different hyperlinks available on websites being visited
- Searching for specific information using *'user formulated query strings'* via search engines (Google, Yahoo, AltaVista etc.)

Currently most Internet users will prefer to use search engines to look for desired information and this can be very clearly explained by the fact that it not only provides an intuitive way of looking for information which is that of getting an answer from a question (user query) but it also provides a summary of available links with a small contents description leaving the user sufficient freedom of deciding whether the materials available are relevant or not.

Inherent features like centralizing many web pages via a single web interface and providing small description of available resources make search engines a very useful concept to look for information on web repositories. However compared to its ability in providing for a single interface for information discovery, search engines are not always very efficient in retrieving desired information such that users are most of the time drowned with irrelevant links and materials (information overload). Closer studies for this inefficiency revealed that the amount of irrelevant materials returned can be explained by either the *(1) user not being able to formulate their query correctly* or *(2) search engines not being able to retrieve available information properly.* Both cases are discussed below.

1 Problem in User Query Formulation

It is well known that users ask question in the common vocabulary they are most familiar with, for example, a person who has been constantly exposed to the word ‘module’ will use the same word to look for related information instead of the words ‘course’ or ‘subject’. But since search engines have got their own way of storing information in their databases, for a query formulated by the word ‘module’, no hit at all will be returned if either of the words ‘subject’ or ‘course’ is used in the underlying storage scheme. This seriously poses a problem on the part of the user in that he/she may not be aware of the word to use to make a search as a result of synonyms.

2 Search Engines Operations and Limitations

Different search engines can have their own method of collecting information over the Internet. They all provide a single interface through which they accept user formulated query strings to look for possible matches within their databases and then to return a series of links believed to be of relevance to the need of the user. The databases are created and maintained by either human or small programs referred to as spiders, robots or crawlers. But in general, typical operations involved are as follows:

2.1 Spiders/robots or crawlers as Gatherer

Search engine periodically send small programs over the internet with a list of links to go to. These small programs are often called spiders, crawlers or robots as they will jump from links to links available over the internet, while at the same time collecting sufficient information to allow indexing of different web pages they visit. This information is sent back to an Indexer.

2.2 Indexing collected information

The indexer is concerned in looking for the most accurate description of the web sites visited so far by the spiders/robots and crawlers and to keep that information (in the form of small description and keywords) together with either copies of the pages or links pointing to the corresponding web pages in the search engine database. Various techniques are used for indexing including *Human/Automatic indexing, full text indexing, Frequency count indexing* and *Meta data indexing.*

Human Indexing will involve reading of each web page collected and looking for specific words (key words) that most likely provide information as to the contents of the corresponding web page. Since it represents a huge task, this practice is restricted to categorizing web sites instead. Automatic indexing by specialized software is used instead. The latter analyses a web page and automatically looks for keywords/index to describe the page.

Full text indexing will involve using the whole text present in the web page to describe the document. This method poses a serious problem of storage space in search engine database and hence is rarely used.

Instead, texts within documents are processed by ignoring stop words (the, a, is, etc.) that do not help in differentiating web pages and to use only useful nouns that contain meaning. These nouns can further undergo stemming to reduce grammatically related words to the same form. An example is for the word ‘drug’ which can exist as drug, drugs and drugged. A frequency count (frequency count indexing) over these nouns usually provides indication as to the topic of concern within each web page. For example a page with a high occurrence of the words ‘nuclear’ and ‘physics’ will most probably relate to the subject of nuclear physics. Figure 1 shows the different operations involved in text processing.

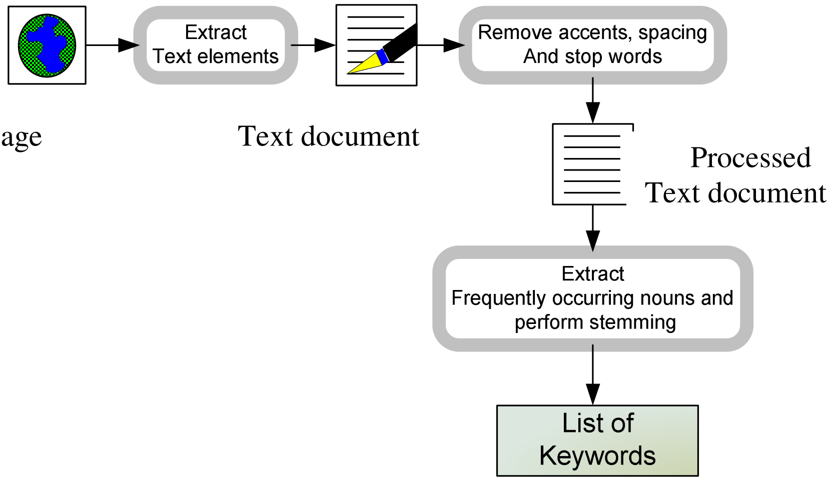
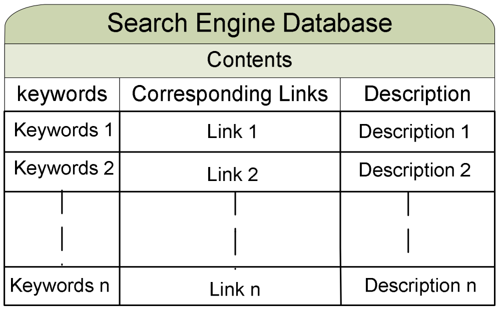


Figure 1: Text processing performed by Indexer

But because, text processing involves additional resources, search engine indexers very often obtain their keywords from <title> or <head> tags present in web pages. The <Meta> tag (Meta data indexing) further help in providing a description of the contents of individual web pages.

The consequent indexing results are then stored in the search engine database. Additional information stored can be the web page itself or links pointing to the corresponding web page together with a small description of the web page. Figure 2 gives an example of the possible contents of a search engine database.



Search Engine Database		
Contents		
keywords	Corresponding Links	Description
Keywords 1	Link 1	Description 1
Keywords 2	Link 2	Description 2
|	|	|
Keywords n	Link n	Description n

Figure 2: Possible Contents of Search Engine Database

2.3 User Interaction

When a user requests for a particular type of information, the search engine uses the query string and tries to find the best match in its database. The result is returned in the form of a Web page containing a list of different possible relevant links with a small description. Figure 3 summarizes the steps involved in getting information for a user on the Internet.

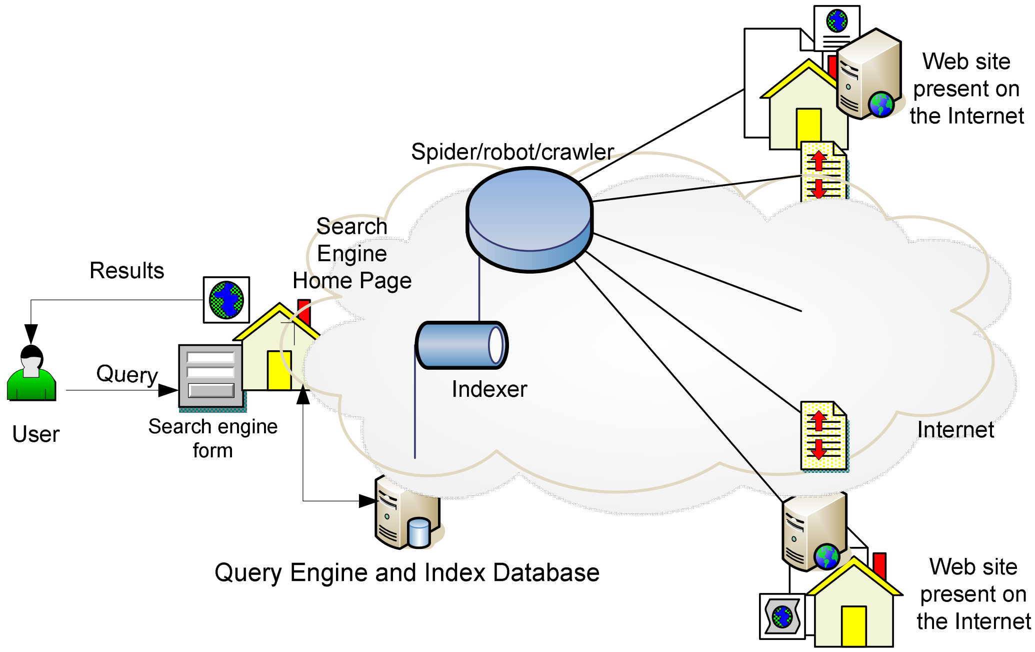

Web site present on the Internet
Spider/robot/crawler
Search Engine Home Page
Results
Query
User
Search engine form
Indexer
Internet
Query Engine and Index Database
Web site present on the Internet

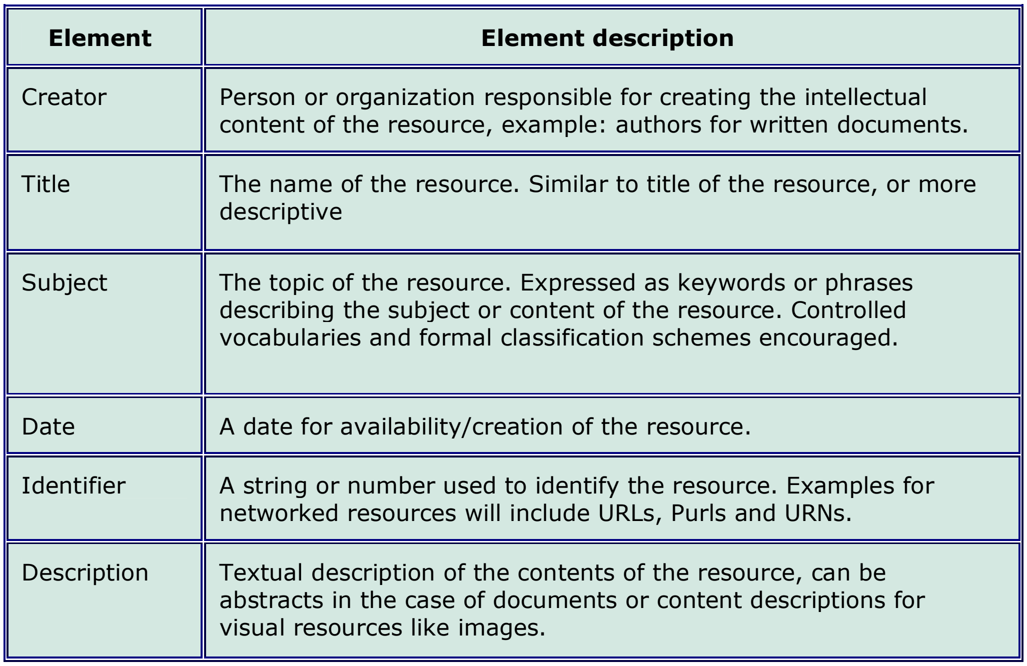
Figure3: Typical Information Searching Process by Search Engine over the Internet.

2.4 Irrelevant Results, Why?

So far the results obtained from Search Engines are not totally relevant. Actually Search Engines use mainly keywords to evaluate whether a given web page is relevant or not. Typically a method that involves frequency count will count the number of occurrence of these keywords and eventually decide on the content of the document under process. Though appearing to be trivial, this technique works very well in specialized fields like medical science whereby a strict vocabulary is used. But in most other areas, vocabulary used can differ largely with each web page using either different terms (module, subject and course) or same terms but with different meanings (polysemy). Furthermore, since search engines will look for high frequency words, some website authors deliberately use certain words like 'free', 'download', etc. to a high extent so as to attract search engines to their websites. Others completely modify the description in the <Meta> tag making the contents and indexed keywords completely irrelevant to each other.

Increasing results relevance: Use of controlled Meta Data

Meta data gives information regarding the data present within a given web page. Its purpose is to be as concise and accurate as possible to allow people to know at a glance the contents of each web page. The simplest and easiest way to use Meta data is to embed them within the same web page. If properly used, Meta data can improve the relevance of returned results by search engines to a high extent. So far <Meta> tags in HTML documents were freely used, hence causing confusion, but with the advent of carefully designed Meta data schemes, authors have to abide to certain rules and format to help in the understanding of their web pages. Popular Meta data schemes are the Dublin Core, IEEE LOM and Australia's EDNA. These schemes clearly define the different details, through specific elements that every author should include in their web pages, so as to ensure accurate description of the contents of the materials they are placing over the internet. An example of different elements used in Dublin Core is shown in Figure 4.



Element	Element description
Creator	Person or organization responsible for creating the intellectual content of the resource, example: authors for written documents.
Title	The name of the resource. Similar to title of the resource, or more descriptive
Subject	The topic of the resource. Expressed as keywords or phrases describing the subject or content of the resource. Controlled vocabularies and formal classification schemes encouraged.
Date	A date for availability/creation of the resource.
Identifier	A string or number used to identify the resource. Examples for networked resources will include URLs, Purls and URNs.
Description	Textual description of the contents of the resource, can be abstracts in the case of documents or content descriptions for visual resources like images.

Figure 4: Sample of Dublin Core Meta data Scheme identifiers

The use of these different elements provides for a solution to different terms used by different educational bodies to represent related educational materials, by having different possible term occurrences clearly specified. If Dublin Core scheme is used, this detail will be present in the **Subject** element. For example in the case of specifying materials relevant to a module(course/subject) dealing with “Ethical issues in Computer Science”, different universities may use different naming conventions like “Professional Issues in Computing”, “Ethical Issues in Computing”, “Legal issues in Computing” etc. Another typical example is for materials related to “Object Oriented Programming”, universities will use different terms like “Object oriented Software Development”, “Object Oriented Techniques”, “Object Oriented Software Engineering” and “Object Oriented Application Development” etc. Hence to avoid an individual from having to know all of these occurrences, Meta data elements can contain most commonly used keywords to relate to a given material, hence ensuring efficient retrieval of materials from the Internet.

Existing implementations that have proved the efficiency of using this method include Harvest for *PhysDoc*, which is a dedicated portal for resources related to the area of Physics. Currently there are hundreds of Meta data schemes to choose from and that number is ever increasing every day. Each of them proposes their own elements including the main purpose to represent contents of their web pages as accurately as possible. For example Dublin Core will use ‘DC.Description’ to provide for a textual description of the web page contents whereas IEEE LOM will use ‘General.Description’ while other Meta data schemes may use other terms like ‘New.ContentDescription’. Hence if a user is to access individual websites which is using controlled Meta data, the relevance of results obtained will be expected to be fairly high. The problem of knowing each and every URLs and accessing individual websites still persist however. A single interface, as in our case, should provide for an attractive solution. But due to the existence and use of different Meta data schemes, semantic interoperability need to be catered for, so as to ensure efficient retrieval of information from each of these web sites separately.

Semantic interoperability

An ideal definition for interoperability has not yet been formulated. The reason being that interoperability can be regarded at various levels and hence different definitions are expressed. For our purpose, semantic interoperability is taken as the ability to access web servers of different universities/educational bodies and to retrieve information that share the same concept from differently expressed Meta data schemes. For instance the proposal should be able to understand that 'DC.Description' for Dublin Core is the same as 'General.Description' for IEEE LOM and so on.

Ways for achieving semantic interoperability

Based on the definition for semantic interoperability (Meta data interoperability), there have been different proposed solutions during recent years. However none of them have proved to be very efficient. Different approaches and a brief description of each associated with their drawbacks are as follows:

1) Modifying existing web servers so that they interoperate.
This can be achieved by giving specific instructions to university or educational bodies as to the preferred metadata scheme to be used to represent the information on their web servers. An example would be to make all of them use the Dublin Core metadata specification. Another method would be to make use of highly structured mark up languages like XML to represent their data. However, this method imposes constraints on the information providers such that they are not able to freely represent their information. Hence, at some point of time, certain organizations may just ignore the standards and stick to what they view as more appropriate to them. Eventually the above model may fail to work.

2) Adoption of a middleware that links the functionalities of all the related websites.
The middleware basically processes all the information exchange parameters between the clients and the servers in order to maintain consistency among the different websites. Metadata and Protocol conversions also take place and popular implementation of middleware framework includes the OMG Common Object Request Broker Architecture (CORBA). Another implementation of such middleware is found at Stanford University with its Infobus Architecture. The assumption with this model however is that all the different clients and servers should be able to interoperate with the middleware infrastructure. In practical situations, this may not be possible except if the information providers choose to implement their system in a technology compatible with the proposed middleware. Eventually again, service providers are not totally free to use their own schemas and hence the model may fail at a given time.

3) Using mobile agents.
This technology attempts to provide clients with a powerful method of accessing servers and collect accurate results on behalf of the user. One proposal for achieving interoperability with agent technology is to have a wrapper enclosing the agent program such that mobile agents can autonomously access different resources equally among different heterogeneous networks. The wrapper is responsible to customize the agent in order to suit the server it is accessing. In this way, an agent may use metadata specific to a given server and hence make desired operations such as querying the index database while maintaining high consistency in information retrieved. However, agent technology is of a complex nature requiring careful considerations and few successful implementations have been under way so far. AgentSys is one such attempt. Furthermore for agent technology to be efficient there is need for an agent

environment with servers accepting and being able to negotiate communication with agents. Eventually, this adds further constraints on the information providers, which may not like the idea of having foreign programs called agents residing on their servers for security purpose.

4) Meta data cross mapping
Crosswalk strategy is another approach which involves the cross mapping of existing metadata schemas. It is an initiative of the UK office for Library and Information Networking (UKOLN). In this way, different metadata schemas are made to interoperate. However using this method assumes that all the metadata schemas can be mapped onto each other, which in practice is not true.

5) Common vocabulary mapping
A somehow similar method is to look for a set of common vocabulary that can map over all the different metadata schemas. Eventually those metadata elements that appear only in some metadata schemas would be ignored. This however provides limited coherence between interoperability for the different web servers and also, the fact of ignoring some metadata schemas may in some cases hide valuable information.

6) Use of server configuration files
Last but not least, each web server can be treated as a separate entity where distributed search is performed. This technique usually requires prior knowledge of the databases to search and involves the use of configuration files specific to each destination server. Hence when a user makes a query, the client system modifies the query expression in a format expected by the destination server so that the search operation can be performed seamlessly. Such an approach has been implemented in [13] and its success has shown its practicability.

Proposed Interoperable Meta Data Framework

After considering the various ways of achieving semantic interoperability among the different existing Meta Data schemes, a hybrid model employing Meta data mapping technique and specific configuration files appears to be the most appropriate solution. A description of the proposed methodology is as follows.

Methodology

Each website is required to have corresponding educational materials available on their web servers together with enough information regarding the contents of these materials. This is to be provided by the use of an appropriate Meta data scheme either embedded within HTML pages containing the educational materials available on their web sites (desired) or as a separate file, linking to the different resources. Details pertaining to the Meta data scheme explaining the relevance of each element used should be also available on the educational web sites.

Metadata gathering by specialized crawler

The proposed framework will make use of a crawler that will only consider information present in the Meta data elements present on university/educational web sites and send this information (including Meta data scheme used) back to an indexer.

Use of configuration files by Indexer

The latter will load a small Meta data specific configuration file so as to know where to look for what information. In this case, if Dublin Core is the scheme used, the indexer will know that ‘DC.Subject’ will provide keywords or phrases describing the contents of the material whereas if IEEE LOM is used, the same information will be obtained under ‘General.Keywords’. Use of controlled Meta data ensures accuracy of information collected, while using Meta data specific configuration files caters for seamless extraction of relevant and consistent information from disparate Meta data schemes. The indexer can then build its own database similar to that of a search engine. In this case however, additional column providing for synonyms or thesaurus can be generated using descriptions obtained from underlying Meta data used.

Search engine like interface for user query

The proposal will include a unique interface via an education portal very similar to that of a search engine allowing the user to formulate his/her query via keywords or phrases. Consequent submission of query to the interface will launch a search in the indexed database and due to the presence of various keywords or phrases that are most commonly used; the probability of high relevance of returned results should be high.

Figure 5 illustrates the operations of the proposed framework.

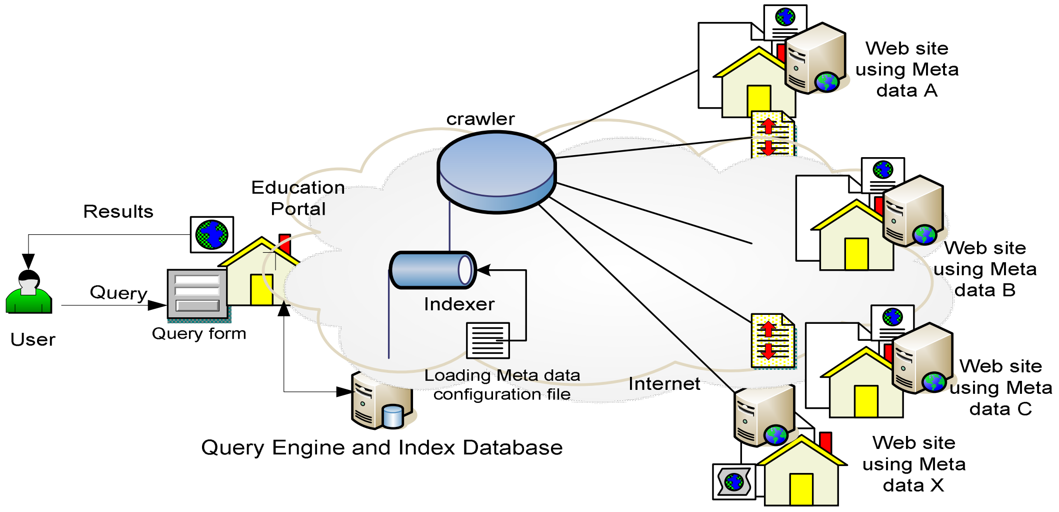
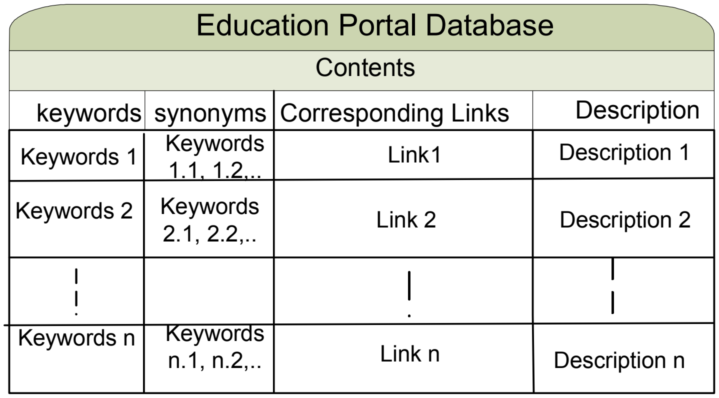

Web site using Meta data A
crawler
Education Portal
Results
Query
User
Query form
Indexer
Web site using Meta data B
Loading Meta data configuration file
Internet
Web site using Meta data C
Query Engine and Index Database
Web site using Meta data X

Figure 5: Proposed Framework for semantic Interoperability

The contents of the index database may look something like shown in figure 6.



Education Portal Database

Contents

keywords	synonyms	Corresponding Links	Description
Keywords 1	Keywords 1.1, 1.2,..	Link1	Description 1
Keywords 2	Keywords 2.1, 2.2,..	Link 2	Description 2
⋮		⋮	⋮
Keywords n	Keywords n.1, n.2,..	Link n	Description n

Figure 6: Contents of database containing details for other possible occurrences of given keywords

Benefits of Proposed framework

There are a number of benefits associated with the proposed framework. They are as follows:

- The unique education portal provide a single interface where an individual can perform a search in multiple databases (distributed search) without the need to worry about all the possible/existing web sites that may provide for the desired information. In this case, there is no need to recall each and every URL of available educational websites.

- Use of controlled Meta data provides for accurate description of the contents of materials stored on educational servers, while providing for synonyms or related phrases that can be used to describe the same materials. Example "Information Management" can be related to "Information System Management", "Strategic Information Management" and "Management Information System". A user may use the common vocabulary he/she is more familiar with, but the proposed framework will cater for related materials, and returned results will include documents containing other possible terminologies used as well.

- Semantic interoperability means that information gathered will be made available according to specific Meta data used by different web sites, such that relevance of the returned results is ensured compared to the use of other existing search engines such as Google and Yahoo.

- Moreover add on services can be made available via the education portal like document rating or web site rating so as to help users obtain feedback from results obtained by other users.

Ensuring Proposed Framework Success

The Internet being an 'open' network, it would be difficult to control each and every material that is being placed on web servers everyday. Hence to ensure the proper running of the proposed framework in this paper, it is essential to ensure that materials being posted on each educational web server contains controlled Meta data. Furthermore the descriptions of these Meta data need to be as accurate as possible.

In an academic environment like universities, this process may be catered by academics themselves. Other educational bodies may rely on authors of educational materials while administrators may cross check the relevancy of Meta data description and the corresponding materials (publications, thesis, paper etc).

Limitations and Future Works

The proposed framework assumes the prior knowledge of each Meta data scheme used by educational web servers from which the crawler collects useful information. In practice, this may not be possible due to the large number of Meta data schemes being created. Furthermore, the problem of scalability, whereby database size is of serious concern has not been dealt with. Another important issue that needs to be taken into consideration is the rights of authors (copyright). And finally, the proposal assumes a dedicated single interface via a unique Education Portal, but who administers and maintains that portal still need to be formulated.

Conclusion

This paper has identified the Internet as a key resource that can be better exploited in the field of Education. Currently an individual would face various difficulties in locating useful and relevant educational materials from the Internet causing considerable waste of time, resources and bringing frustration on the part of an individual in using such service. Careful study of information search process over the Internet revealed various limitations that have helped in proposing a framework for better utilization of this platform as an education sharing network.

Typically the framework highlighted the need of using controlled Meta data to accurately represent material contents on educational servers and to have a dedicated education portal that would allow for a single interface that an individual would use to access different web sites provided by universities or educational bodies. Use of synonyms in Meta data has been described as providing for a useful solution to cater for difficulty in query formulation by users as well.

Furthermore, our proposal stresses on semantic interoperability among different Meta data schemes to ensure seamless but accurate retrieval of relevant materials from web servers that may adopt different schemes.

In this way, information which would have been otherwise inaccessible can be made available with just a few mouse clicks.

References

[1] Gralla Preston, "*How the Internet Works",* Techmedia, Millenium Edition, 2000.

[2] Baeza-Yates R. and Ribeiro-Neto B., *"Modern Information Retrieval",* Addison Wesley', 1999.

[3] "Search engine-Introduction", Minnesota State Archive, URL: http://www.mnhs.org/preserve/ records/ SEintro.html, Accessed Jan 2005

[4] Chris Taylor, "An Introduction to Meta Data", University of Queensland Library, 2003

[5] Dublin Core Metadata Initiative, URL: http://dublincore.org

[6] IEEE learning Technology Standards Committee's Learning Object Meta-data Working Group, Approved LOM Working Draft 5 (WD5), URL: http://ltsc.ieee.org/wg12

[7] Galatis Helen, "EDNA Meta data" September 2002, Paper URL: http://www.edna.edu.au/ metadata, Accessed Jan 2005

[8] Severiens T, Hohlfield M., Zimmermann K., Hilf E., "*PhysDoc-A Distributed Network of Physics Institutions Documents",* D-lib Magazine, December 2000, Volume 6 Number 12, ISSN 1082-9873.

[9] Lynch C., Garcia-Molina H., "Interoperability, Scaling and the Digital Libraries Research Agenda: A report on the May 18-19,1995", IITA Digital Libraries Workshop, August 22, 1995, URL: http://diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html

[10] Miller P., "*Interoperability, What is it and Why should I want it?",* Ariadne Issue 24, URL http://www.ariadne.ac.uk/issue24/interoperability/intro.html

[11] Pe'rez I., Delgado J., Peig E., "*Metadata Interoperability and Meta-search on the Web",* Universitat Pompeu Fabra (UPF), Departament de Tecnologia, National Intitute of Informatics, 2001.

[12] UKOLN, *"Metadata, mapping between metadata format"*, URL: http://www.ukoln.ac.uk/metadata/ interoperability

[13] Warnick W., Scott R., Johnson L., Lederman A., Spence K., Allen V., "*Searching the Deep Web",* D-lib Magazine, January 2001, Volume 7 Number 1, ISSN 1082-9873

[14] XML (Extensible Markup Language), URL http://www.w3.org/XML

[15] Edwards J., Harkey D., Orfali R., "*Instant Corba",* John Wiley and Sons Inc., 1997.

[16] Cousins S.B., "*Reification and Affordances in a user interface for interacting with heterogeneous distributed applications",* PhD Thesis, Stanford University, August 1997.

[17] Karmouch A., Falchuk B., " *AgentSys: A Mobile Agent System for Digital Media Access and Interaction on the Internet",* University of Ottawa, Canada.