

Digital Libraries on the Internet

Taly Sharon and Ariel J. Frank

Bar-Ilan University

{taly,ariel}@cs.biu.ac.il

Abstract

The Internet and the Web have been growing in leaps over the past few years, accelerating the problem of information explosion, a well-known phenomenon to all of us. Indeed, search engines (SEs) that have popped up everywhere enable us to access the cyberspace, but they flood us with vast amounts of irrelevant information. Nonetheless, considering the vast amount of information, the Web is considered by many to be the world's ultimate virtual library - but is this solution the right one?

In any case, the Web and the SEs do not substitute the classical, loved libraries. Looking backwards, libraries can be classified into 3 types:

- 1) Analog/Paper Library (PL) - the classical paper library with its card catalog.
- 2) Automated/Hybrid Library (AL) - an analog library with a computerized catalog.
- 3) Digital Library (DL) - a computerized library in which most of the information is digital.

The problems of our regular libraries are well known and need not be detailed here. On the other hand, it is less clear to us what a digital library is and what are its various characteristics.

First, we classify the digital libraries into three categories:

- 1) Single Digital Library (SDL) - the regular classical library implemented in a fully computerized fashion.
- 2) Federated Digital Library (FDL) - this is a federation of several independent libraries, centered on a common theme, on the network.
- 3) Harvested Digital Library (HDL) - this is a virtual library providing summarized access to related material scattered over the network.

Consequently, we compare the various types of libraries. Moreover, we further discuss aspects of DLs such as personal vs. public libraries and multimedia libraries. Following that, we focus on a comprehensive comparison between HDLs and SEs on the Web.

To demonstrate, we show exemplary digital libraries. In particular, we focus on the Katsir HDL, based on the Harvest system, which is currently being developed in Bar-Ilan University as cooperation between the Mathematics and Computer Science department and the department of Information Studies.

1 Introduction

The Internet and the Web have been growing in leaps and bounds over the past few years, accelerating the problem of information explosion, a well-known phenomena to all of us. According to Nature [1], the publicly indexable Web contains an estimated 800 million pages as of February 1999, encompassing about 15 terabytes of information or about 6 terabytes of text after removing HTML tags, comments, and extra white-space.

Indeed, the growing amount of Search Engines (SEs) that have popped up everywhere, reaching more than 2400 different SEs, enable us to access the cyberspace, but they also flood us with vast amounts of irrelevant information. Search engine coverage, relative to the estimated size of the publicly indexable Web, has recently decreased substantially, with no engine indexing more than about 16% of the estimated size of the publicly indexable Web. It is interesting to note that 83% of the Web sites contain commercial content and only 6% contain scientific or educational content [1].

The article is structured as follows. This section presents the resource repository hierarchy, defines the notion of the library and the development from paper to digital libraries. The next section classifies digital libraries, compares between the different types and introduces the logical harvesting model. The section following provides additional aspects regarding digital libraries. A concluding discussion ends the article.

1.1 Resource Repositories Hierarchy

Both Search Engines (SEs) and Digital Libraries (DLs) are Internet Resource Discovery (IRD) Tools. We introduce a resource repositories hierarchy (depicted in figure 1) with two major paradigms: search engines and digital libraries, where each branches to categories. SEs can be classified into three categories: Basic-SE, Directory, and Meta-SE. All the categories support search user interfaces, but with significant differences in their construction method:

- 1) Basic-SE/Index - a tool that uses an automatic robot/crawler to gather metadata on items.
- 2) Directory/Catalog/Guide - a tool that uses human judgement to collect and catalog items.
- 3) Meta-SE - a tool that holds no database of its own, but rather queries Basic-SEs upon a user request.

A detailed discussion about digital libraries, including DL categories, will be presented in section 2.

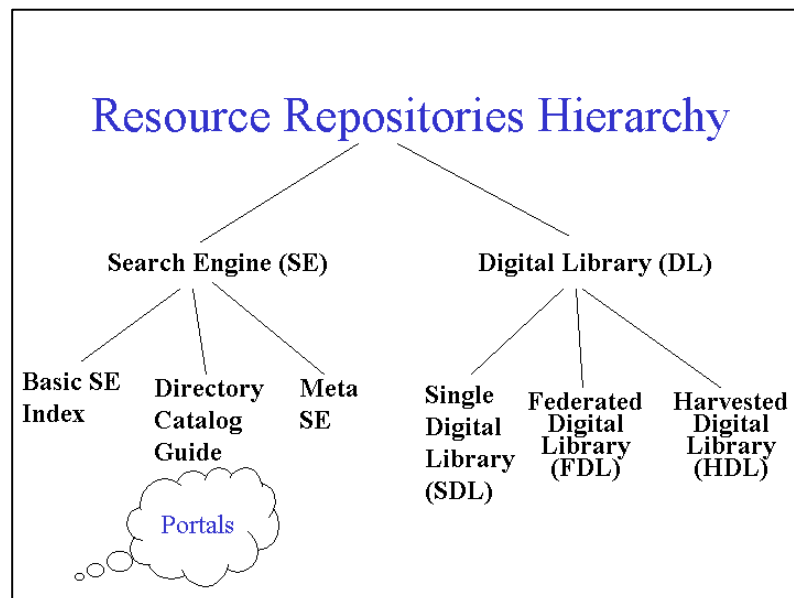


Figure 1: Resource Repositories Hierarchy

1.2 What is a library

Before we delve into digital libraries, we define the notion of a library in general and of a digital library specifically. We define a library as having six major characteristics:

1) Collection of data objects -

A library holds a collection of data objects, also called holdings, items, resources, or just material. The items can be:

- Books and journals.
- Documents (e.g., HTML pages)
- Multimedia objects (such as pictures or images, tapes or video files, etc.).
- The library objects can be available locally in the library, or indirectly, by using a network to access them.

2) Collection of metadata structures -

A library contains a collection of metadata structures, such as catalogs, guides, dictionaries, thesauri, indices, summaries, annotations, glossaries, etc.

3) Collection of services -

A library provides a collection of services, such as:

- Various access methods (search, browse, etc.) for different users.
- Management of the library (purchase, shelf arranging, computerization, communication).
- Search (query formulation), browse, and consultation.
- Logging/statistics and Performance Measurement Evaluation (PME).
- Selective Dissemination of Information (SDI) or as it is called Push mode on the Internet.

4) Domain focus -

A library has a domain focus and its collection has a purpose. For example: art, science, or literature. Also, it is usually created to serve a community of users, and therefore is finely grained. For example: academic, public, special, school, national, or state library.

5) Quality control -

A library uses quality control in the sense that all its material is verified and consistent with the profile, or stereotype, of the library. The material is filtered before it is included in the library, and also its metadata is usually enriched (e.g., annotated), etc.

6) Preservation -

Libraries and archives have served as the central institutional focus for preservation, and both types of institutions include preservation as one of their core functions. The purpose of preservation [2] is to ensure protection of information of enduring value for access by present and future generations. Preservation includes regular allocation of resources for persistence, preventive measures to arrest deterioration of materials, and remedial measures to restore the usability of selected materials.

1.3 From Paper to Digital Libraries

In any case, the Web and the SEs do not substitute the classical, loved libraries. Looking backwards, libraries can be classified into 3 types [3]:

- 1) Paper/Analog Library (PL) - the classical paper library with its card catalog.
- 2) Automated/Hybrid Library (AL) - a paper library with a computerized catalog.
- 3) Digital Library (DL) - a computerized library in which most of the information is digital.

No one questions or disputes the long and lasting contribution of existing classical libraries [4]. The concept of the paper library and the various services it provides are well established. The idea is that DLs should provide all these and more [5,6]. We use the term 'integrated services' in DLs to allude to that. These integrated services will add services that are made possible by use of the digital medium such as: varied search techniques resulting in focused results, faster provision of relevant resources, and access also to multimedia resources.

The problems of our regular libraries are well known and need not be detailed here. On the other hand, it is less clear to us what a digital library is and how it works - this is the subject of this paper.

2 Digital Libraries

2.1 Classifying DLs

We classify the digital libraries into three categories (see figure 1): Stand-alone Digital Library (SDL), Federated Digital Library (FDL), and Harvested Digital Library (HDL). We now detail:

1) Stand-alone Digital Library (SDL)

This is the regular classical library implemented in a fully computerized fashion. SDL is simply a library in which the holdings are digital (i.e., electronic – scanned or digitized). The SDL is self-contained – the material is localized and centralized. In fact, it is a computerized instance of the classical library with the benefits of computerization. Examples of SDLs are the Library of Congress (LC) and its National Digital Library (NDL) (<http://www.loc.gov>), and the Israeli K12 Portal Snunit (<http://www.snunit.k12.il>).

2) Federated Digital Library (FDL)

This is a federation of several independent SDLs in the network, organized around a common theme, and coupled together on the network. A FDL composes several autonomous SDLs that form a networked library with a transparent user interface. The different SDLs are heterogeneous and are connected via communication networks. The major challenge in the construction and maintenance of a FDL is interoperability (since the different repositories use different metadata formats and standards). Examples of FDLs are the Networked Computer Science Technical Reference Library (NCSTR) (<http://www.ncstrl.org>) and Networked Digital Library of Theses and Dissertations (NDLTD) (<http://www.ndltd.org>).

3) Harvested Digital Library (HDL)

This is a virtual library providing summarized access to related material scattered over the network. A HDL holds only metadata with pointers to the holdings that are “one click away” in Cyberspace. The material held in the libraries is harvested (converted into summaries) according to the definition of an Information Specialist (IS). However, a HDL has regular DL characteristics, it is finely grained and subject focused. It has rich library services, and has high quality control preserved by the IS, who is also responsible for annotating the objects in the library. The HDL harvesting model is further detailed in section 3. Examples of HDLs are the Internet Public Library (IPL) (<http://www.ipl.org>) and the WWW Virtual Library (<http://www.vlib.org/>).

2.2 Comparison

To emphasize the different aspects of this DL categorization, let us get into the various DL types. In SDL and FDL, the items are electronically purchased or fully digitized/scanned. These items are stored in the local repository (in SDL), or in separate SDL repositories accessed using a network protocol (in FDL). Each SDL (and all together in a FDL) holds a huge repository containing both the items and some metadata structures to enable efficient retrieval. This material is updated every now and then, in a process similar to the one in classical library. It is important to note that composing a FDL out of SDLs requires interoperability capabilities, and the use of a common protocol.

In contrast to the SDL and FDL, the HDL's items are gathered from the network. These items are scattered on many servers, and accessed via direct retrieval using standard protocols such as HTTP, FTP, etc. The HDL holds only metadata on the items, and therefore its repository is small and compact. Because the items that belong to the HDL can be updated any time by their authors, their summaries have to be dynamically refreshed in the HDL using computerized procedures that are triggered automatically or initiated explicitly by the IS. An interesting point is that the profile of a HDL can be changed by the IS to enhance the library contents.

2.3 Harvesting Model for HDLs

We will now describe our developed logical model [7] for constructing HDLs (see figure 3). The model includes processes (represented by circles), data repositories (represented by rectangles) and auxiliary repositories (represented by parallelograms).

The initiating IS invokes the Harvester with the DL harvesting request. The Harvester generates the initial DL profile and passes this as the harvesting query to the Locator component. The Locator uses various network search techniques to enrich the initial collection of URLs to be harvested. The next component to be invoked is the Gatherer. It uses each top-level URL, in a recursive manner, to gather all referenced resources from the network providers, and passes them to the Filtering component.

The Filtering component is responsible for blocking the non-relevant documents from reaching the focused DL. It uses various levels of filtering that all remaining documents have to pass to be considered relevant. A first level, for example, can use 'regular expressions' to match query keywords with the URL string tokens. A second level can use statistical techniques on the document itself, based on keyword counts and frequencies. A third level might use a Categorizer to classify the document and check if it belongs to the gathered DL categories. More levels or any geared combination of levels can ensure a cleaner DL devoid of 'noises'.

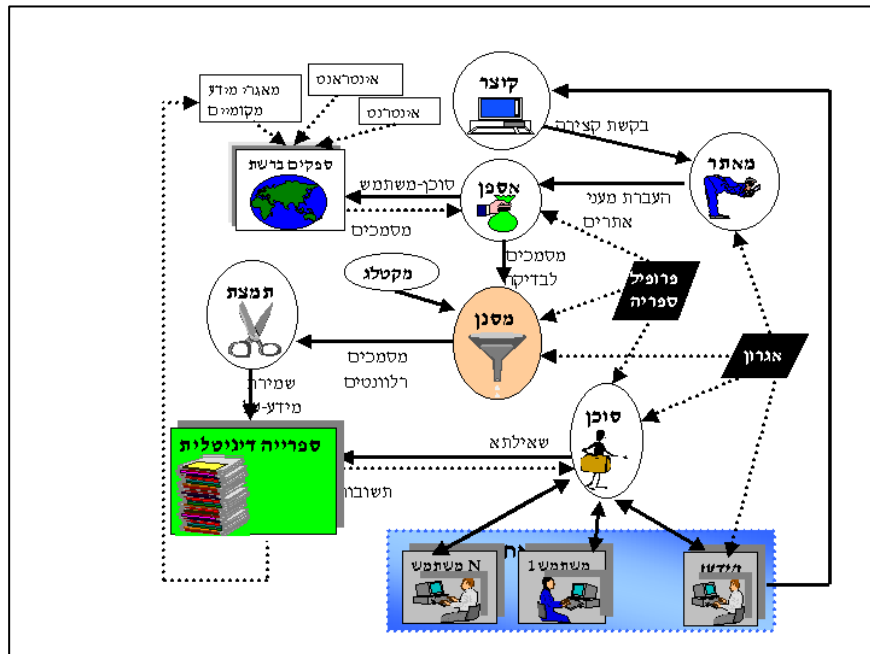


Figure 3: Logical model for HDLs

All relevant documents are passed now to the Summarizer. It extracts a summary of the document, and passes a stream of summaries to the Broker. The Broker indexes the summaries and organizes the DL.

The IS builds for the DL a relevant topics-tree, possibly using advanced IR tools for categorization and clustering. The Retriever provides the DL user with a user-friendly interface.

2.4 Implementation of Harvest/Katsir System

To demonstrate HDLs, we focus on the Katsir HDL (<http://bicsir.cs.biu.ac.il:8088/katsir/>), based on the Harvest system (<http://www.tardis.ed.ac.uk/harvest>), an initial/partial implementation of the harvesting model for HDL. Katsir [7] is currently being developed in Bar-Ilan University (BIU) as cooperation between the Mathematics and Computer Science department and the department of Information Studies.

The Harvest system, developed mainly at University of Colorado, USA, is an integrated set of tools to gather, extract, organize, search, cache, and replicate relevant information across the Internet. With modest effort users can tailor Harvest to digest information in many different formats, and offer custom search services on the Internet. Moreover, Harvest makes very efficient use of network traffic, remote servers, and disk space.

The Katsir system is based on the Harvest system. Katsir can be directed to build a focused DL, based on both local and networked harvested materials. Katsir's enhancements over the original Harvest system are:

1) Hebrew support

Full Hebrew support for both resources and interfaces was added to Harvest.

2) Improved user interface

While Harvest's user interface supports keywords query, Katsir enable to user to retrieve information by keywords or conduct a guided tour by following a topics-tree that enables hypertext access to relevant materials.

3) Improved Graphical User Interface for the IS

While the IS interface to Harvest is through Unix, using file editors to define harvesting options, Katsir implemented a GUI interface to define DL profile. The interface includes URLs for initialization, harvesting options, ports selection, etc. It also employs the standard browsers Bookmarks/Favorites utilities to define the initial DL Harvesting URLs.

The initial Katsir system was developed and implemented in an educational environment as a response to the unique requirements of the Israeli educational system. This project was part of a drive to enhance and assimilate information and telecommunication technologies in Israel, based on the public Internet.

Both researchers and students of the departments of Information Studies and of Mathematics and Computer Science at BIU (<http://www.cs.biu.ac.il>) were involved in the development of Katsir. Their cooperation in the framework of the demonstrative Gilo High-School project was complementary in knowledge and experience, and fruitful in its results. Moreover, the project itself was planned and executed with full cooperation of the Gilo user community. This begun in requirements analysis, in deciding the topics to be harvested, in training the users in accessing the system, in conducting the control and evaluation process, and in reviewing and implementing the conclusions reached as part of the working methods of Katsir.

3 Additional Aspects

3.1 Personal DLs vs. Public DLs

All libraries can be personal or public, but this issue is mostly interesting in the implementation of HDL. HDL can support the harvesting of both personal and public DLs. A personal DL is one constructed for personal use by a seemingly only self-interested person. Here, the person is both the harvester and its user. The personal and library profiles could be related or even be the same here. A public DL is geared to a wider range of audience. The harvester here is the IS, but the users are many. The library profile supports the library (users) stereotype and is maintained by the library IS. A public DL supports a large audience and has to have a higher level of integrated services, including support for pull and push modes.

3.2 Textual DLs vs. Multimedia DLs

Most DLs nowadays are mainly textually oriented libraries. However, a major advantage of being digital is the option of supporting libraries that contain Multi-Media (MM) resources. One major problem with MM DLs is how to summarize MM resources [8]. The methods for summarizing MM resources are entirely different from textual ones; they include color, texture, shape, objects, etc. for visual information [9,10]; pitch, rhythm, etc. for vocal information [11]. Also, the repository generated is different, requiring special tools, such as multimedia and object oriented databases.

Another hard problem is intelligent search and retrieval of MM resources [12]. Usual IR tools are textually oriented. Some techniques for MM contents search and retrieval have been developed for use in MM-SEs [13]. These techniques includes query by example, and the tools are sketching or giving an image as example to the query for images and video [9], or based on singing a melody or humming a few notes for audio [11].

There are several implementation of MM DL (see <http://sunsite.berkeley.edu/ImageFinder/>), few examples on the Web are:

- 1) Webseek (<http://disney.ctr.columbia.edu/webseek/>).
- 2) Metaseek (http://mahler.ctr.columbia.edu:8080/cgi-bin/MetaSEEk_cate).
- 3) Virage (<http://www.virage.com>).
- 4) Meldex (<http://www.nzdl.org/musiclib>).

However, most of these techniques are low-level using query by example or similarity measures, while DLs should also use higher-level methods based on object specification and recognition methods.

3.3 Composing DLs

Another promising idea here is the composition of existing DLs to construct a coarser-grained DL. For example, we might have finely-grained DLs on the categories of Spaniels and Briards (see figure 2). These two can be composed to a coarser-grained DL on dogs. If we had additional DLs on cats, we could then combine them to a mammal DL. And similarly, further compose a DL on the category of animals and so on. The DLs participating in a composed DL can be physically distributed of course. The DL composition process uses techniques of open networking, shared file systems, replication servers, and resource caching.

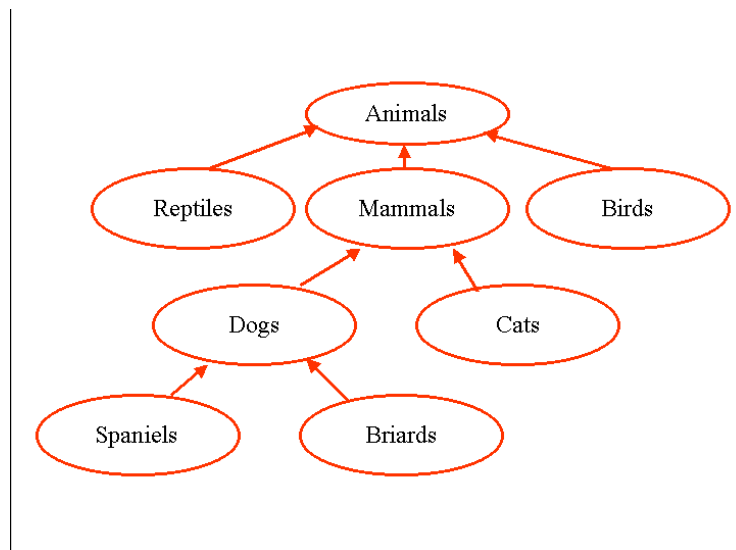


Figure 2: Composition of DLs

4 Discussion

Considering the vast amount of information, the Web is considered by many to be the world's ultimate virtual digital library - but is this solution the right one?

We already confronted SEs with DLs to realize the differences between them. We can compare the different categories of SEs with the different categories of DLs (see figure 1). Basic-SE is similar to all DLs in the basic user interface, IR tools, and network access. Furthermore, Basic-SE is similar to HDL since they both hold metadata repositories rather than full items. A directory is even closer to a DL than a Basic-SE, since it is humanly compiled and therefore has quality control. But let's not forget it does not have domain focus and DL integrated services. Meta-SE is somewhat similar to FDL in the sense that both generate on-the-fly queries to other SEs/DLs to answer user queries. We elaborate more on the differences between SEs and DLs in the following section.

4.1 Search Engines vs. Digital Libraries

The Search Engine (SE) paradigm and the DL one are really located at the extremes of a spectrum of data repositories and types of search (see figure 1). There are two sides to each of these coins: the data repository construction side (see figure 4) and the user information search side (see figure 5). We will now discuss and contrast these aspects.

As regards to the construction of SEs, this is a complex undertaking. It is clearly a long-term effort that is (eventually) supported by commercial companies. The SE aims to build a quantitative global repository that represents as much information available on the Internet as possible or at least a large amount of it. The SE maintains various data structures, to represent its repository, like indices, directories, and catalogs. It also provides an elaborate user interface for search purposes. The SE continuously employs various types of robots to search out and index pages on the Internet and to dynamically update its provided repository.

Let's look now at the user side of SEs. Assume that a user needs some information on a certain topic that currently interests him. So he summons on a whim of a second his favorite SE to search for any relevant information. The SE is invoked with an ad-hoc query composed of a supposedly appropriate combination of keywords. The SE will certainly return a lot of information (with low precision and recall), which is bound to overload the user. He will then have to tediously sift through it all and manually filter the supplied references. The relevant information found will then be immediately used or temporarily kept in a cache for a short-term period of use.

Measure	Search Engine	(Harvested) DL
Effort	Complex Undertaking	Medium
Emphasis	Quantitative	Qualitative
Content	Global/Shallow	Focused/Annotated

Repository	Huge	Small
Maintenance	Continuously Updated (Robots)	Dynamically Updated

Figure 4: Search Engines vs. Digital Libraries - Server Side

Measure	Search Engine	(Harvested) DL
Interest	Sudden	Lasting
Query	Ad-hoc	Souder
Use	Short Term	Medium Term
Information Returned (coverage/recall)	A lot	Modest
Quality (precision)	Noisy	Clean
Sift/Filter	Manual	Not much needed
Distribution	Pull Mode	Both Pull/Push Mode

Figure 5: Search Engines vs. Digital Libraries - Client Side

Consider now the process of harvesting (i.e., constructing) a DL. A user, say an Information Specialist (IS), realizes a well-thought out need to build a qualitative data repository on an important focused topic. He decides to invest by harvesting and maintaining a long-term DL, described by a set of specific categories. So he interacts with an IS interface to carefully define his DL harvesting request. The DL is then harvested and made available to its users. It supports transparent user access methods using various data structures to enable efficient keywords search, touring a DL via a topics tree, and DB/SQL oriented views of the DL contents (see figure 6). The contents of the DL are continuously kept current and the DL can be annotated and enhanced with additional relevant material.

Let's check now on the use of DLs. A serious user will tend to often need information on a topic included in his areas of interest. There is a good chance then that he already has access to a relevant DL, previously harvested. So he invokes the high-level DL interface and chooses an appropriate way to search this DL. The DL will return a reasonable amount of information (with high precision and recall) that the user can readily digest. The returned results will be made available at three levels of detail: first, a high-level summary (metadata) [14,15]; then, if requested, an additional abstract; and finally, if relevant, the referenced resource itself will be fetched and presented. Not much sifting will be necessary in any case. The relevant information can be further annotated by the user and later rediscovered whenever needed.

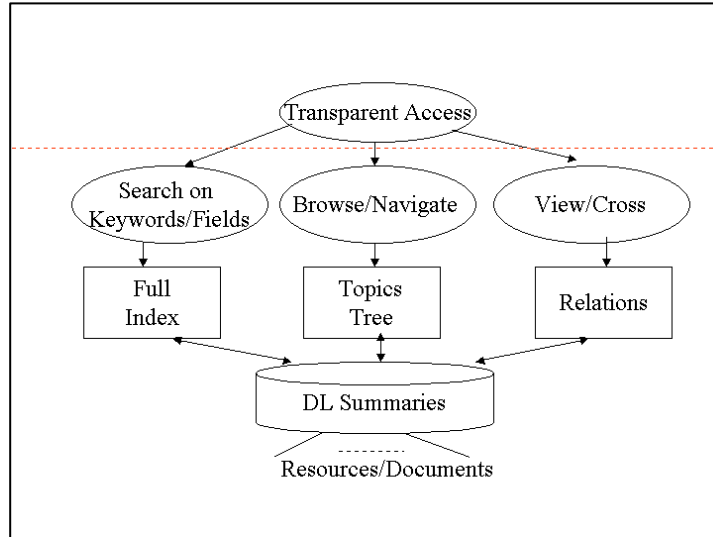


Figure 6: User access methods and data structures in HDL

So, to summarize, SEs necessitate a huge organizational effort, provide the user with too much noisy information, but are useful for a one-time shot for quickly needed information. DLs, on the other hand, require a modest support effort, provide the user with focused information, but have to be made available beforehand while excelling in quality and ease of use. It is important to note that these two paradigms are neither conflicting nor exclusive, but are complementary in nature.

5 Conclusion

Digital libraries and search engines on the Internet are similar in many ways yet differ in others. The direction they are all going seems somewhat alike, yet, more research should be carried out to determine the real trends. Further research can also probe into the versatile types of SEs and DLs and their generations. More exploration into additional aspects discussed here, like multimedia, composition of libraries, and DL profiles should take place to promote these issues for the benefit of the millions of users surfing the net.

Bibliography

- [1] Hedstrom M., Digital Preservation: a Time Bomb for Digital Libraries, <http://www.uky.edu/~kiernan/DL/hedstrom.html>
- [2] Lawrence S., Giles L., Accessibility and Distribution of Information on the Web, *Nature*, 400, 107-109, 1999, <http://www.wwwmetrics.com/>
- [3] Chen H., A. L. Houston, "Digital Libraries: Social Issues and Technological Advances", *Advances in Computers*, Academic Press, vol. 48, pp. 257-314.
- [4] Tock S., Can the Library Survive in a Digital Age?, <http://comm1.uwsp.edu/302x/jan98projects/tock.htm>
- [5] Kessler J., *Internet Digital Libraries*, Artech House, Boston, 1996.
- [6] Lesk M., *Practical Digital Libraries*, Morgan Kaufmann, San Francisco, 1997.
- [7] Hanani U., A. Frank, Intelligent Information Harvesting Architecture: an Application to a High School Environment, *Online Information 96*, London, December 1996, pp. 211-220.
- [8] Schauble P., Multimedia Information Retrieval, *Kluwer Academic*, Boston, 1997.
- [9] Gupta J., Visual Information Retrieval, *Communication of the ACM*, 40(5), 1997.
- [10] Smith J.R., Chang S., Searching for images and videos on the World-Wide Web, August 1996, <http://www.ctr.columbia.edu/webseek/paper/>.
- [11] Rodger J. McNab, Lloyd A. Smith, David Bainbridge and Ian H. Witten, The New Zealand Digital Library MEDLody inDEX, *D-Lib Magazine*, May 1997, <http://www.dlib.org/dlib/may97/meldex/05witten.html>
- [12] Maybury M., Intelligent Multimedia Information Retrieval, *AAAI/MIT Press*, 1977.
- [13] Kumar P.S., Babu G. P., Intelligent multimedia data: data+indices+interface, *Multimedia Systems* 6:395-407, 1998.
- [14] Lassila O., Web Metadata: A Matter of Semantic, *IEEE Internet Computing*, July/August 1998, 2, 4, 30-37.
- [15] Rust G., Metadata: the Right Approach, *D-Lib Magazine*, July/August 1998, <http://www.dlib.org/dlib/july98/rust/07rust.html>