Chapter 18

DESIGN AND IMPLEMENTATION OF DIGITAL LIBRARIES

Xiuqi Li and Borko Furht

Abstract

As the Internet and the World Wide Web expanded so fast, digital libraries has become a very hot topic. Since 1992 a lot of studies have been done and some achievements have been made. This chapter is a survey of these studies. We first discuss designing digital libraries, including definition of digital libraries, infrastructure requirements for digital libraries, research issues related to digital libraries, and the architecture of digital libraries. Then a project, Digital Library Initiative, is introduced as an example of implementing digital libraries.

1. INTRODUCTION

Because of World Wide Web, access to the Internet has become part of our daily life. A huge number of people search the Internet every day. More and more people need to search indexed collections. But the commercial technology for searching large collections, developed in the US government sponsored research projects in 1960s, has not changed much. A new revolution in information retrieval technology has been spurred by this public awareness of the net as a critical infrastructure in 1990s. [1]

Many people believe that a Net Millemmium, where the Net forms the basic infrastructure of everyday life, is coming. "For this transformation to actually occur, however, the functionality of the Net must be boosted beyond providing mere access to one that supports truly effective searches"[1]. All kinds of collections must be indexed and searched effectively, including those for small communities and large disciplines, for formal and informal communications, for text, image and video repositories, and those across languages and cultures. A fundamentally new technology is needed to support this new search and indexing functionality – this is “digital libraries.”

Basically the purpose of digital libraries is to bring the efficient and effective search to the Net. However, in a real digital library, searching is not enough. The main activities of users can be classified into five categories: locating and selecting among relevant sources, retrieving information form them, interpreting what was retrieved, managing the filtered-out information locally, and sharing results with others. “These activities are not necessarily sequential, but are repeated and interleaved” [2].

There is no single definition for digital libraries. And as times goes by, we know more and more about digital libraries, the definition evolves. From information management point of view, digital libraries are systems that combine the machinery of digital computing, storage and communication, the content, and software needed to reproduce, emulate, and extend the services of collecting, cataloging, finding and disseminating information offered by traditional libraries based on paper and other materials. From the user point of view, digital libraries are systems that provide a community of users with coherent access to a large, organized repository of information and knowledge.

When designing and implementing digital libraries, there are several aspects needed to considerate [3]:

- Interoperability: how to confederate heterogeneous and autonomous digital libraries to provide users with a coherent view of the various resources in these digital libraries
- Description of objects and repositories: describe digital objects and collections to facilitate the use of mechanisms such as protocols that support distributed search and retrieval and provide the foundation for effective interoperability
- Collection management and organization: incorporating information resources on the network into managed collections, rights management, payment and control, non-textual and multimedia information capture, organization, storage, indexing and retrieval
- User interfaces and human-computer interaction: user behavior modeling, display of information, visualization and navigation of large information collections, linkage to information manipulation/analysis tools, adaptability to variations in user workstations and network bandwidth
- Economic, social and legal issues: rights management, economic models for the use of electronic information, and billing systems to support these economic models, user privacy

Since 1992 digital libraries emerged as a research area, there has been a lot of work done. Some achievement has been made, especially in description of objects and repositories, collection organization, and user interfaces. And a lot of digital libraries have been developed, in U.S.A, European, Australia and Asia.

United States is the leader of digital library research area. National Science Foundation (NSF), Advanced Research Projects Agency (ARPA), and National Aeronautics and Space Administration (NASA) jointly funded a digital library research project called Digital Library Initiative (DLI). It is divided into two phases, which are called NLI I and NLI II, respectively. NLI I began in 1994 and ended in 1998. The total budge was US$ 25M. It focused on dramatically advancing the means to collect, store, and organize information in digital forms, and making it available in user-friendly ways for searching, retrieval and processing through communication networks. Six universities participated in this initiative. They are Carnegie Mellon University, Stanford University, University of California at Berkley, University of California at Santa Barbara, University of Illinois at Urbana-Champaign, and University of Michigan. Each university focused on a specific area. Carnegie Mellon University focused on interactive on-line digital video library system, University of California, Berkeley on environmental and geographic information, University of Michigan on earth and space sciences, University of California, Santa Barbara on spatially referenced map information, Stanford University on interoperation mechanisms among heterogeneous services, and University of Illinois at Urbana-Champaign on federating repositories of scientific literature. Compared to DLI I, DLI II is a broader and larger effort. Besides NASA, ARPA and NSF, National Library of Medicine (NLM), Library of Congress (LOC), National Endowment for the Humanities (NEH) and Federal Bureau of Investigation (FBI) also sponsored this project. It has begun this summer. There are 24 projects approved. Based on NLI I, NLI II will emphasize on human-centered research, content and collections-based research, system-centered research, development of digital libraries testbed for technology testing, demonstration and validation, and as prototype resources for technical and non-technical domain communities and will plan testbeds and applications for undergraduate education [4].

2. DESIGN OF DIGITAL LIBRARIES

When designing digital libraries, first we need to answer these questions:

1. What is Digital Library? How a Digital Library differentiates from an information repository or from World Wide Web? How many Digital Libraries will there be and how they will inter-link? How might this look to users [5]?

2. What will be the infrastructure for Digital Library? What is context of a Digital Library? What is the relationship between Digital Library and intellectual property management including publisher concern?

3. How can a Digital Library be evaluated?

The third question is the most difficult to answer. Although metrics for traditional library such as precision and recall can be directly applicable to some aspects of digital library and have been widely accepted, the digital library is much more complex and there are much more to be considered. “Metrics are required to deal with issues such as the distributed nature of the digital library, the importance of user interfaces to the system, and the need for systems approaches to deal with heterogeneity among the various components and content of the digital library” [5]. There is a group working on this issue, called D-Lib Working Group on Digital Library Metrics.

There are mainly four kinds of research issues in digital libraries: interoperability, description of objects and repositories, collection management and organization, and user interfaces and human-computer interaction. We present these issues in detail in Section 2.3.

As for the architecture of digital libraries, different researchers gave different solutions. We will introduce a commonly accepted architecture, which is described in Section 2.4.

Since video has its special characteristics that are quite different from text, additional issues need be addressed in a digital video library system than a text-only digital library. These issues include video storage, video compression, video indexing, and video retrieval. We discuss these issues in Section 2.5.

2.1 DEFINITION OF DIGITAL LIBRARIES

Before defining Digital Libraries, we introduce several fundamental assumptions:

- The digital libraries are not a bounded, uniform collection of information.
- There will be increasing diversity of information and service providers.
- There is more than just searching in digital libraries.

Especially we should notice the last point. As shown in Figure 1, the main activities of users can be classified into five categories: locating and selecting among relevant sources, retrieving information from them, interpreting what was retrieved, managing the filtered-out information locally, and sharing results with others. These activities are not necessarily sequential, but are repeated and interleaved [2]. Users can move freely in the circle to get their work done. In general, users will be involved in multiple tasks at the same time. They will need to move back and forth among these tasks and among the five areas of activity. They need to find, analyze, and understand information of varying genres. They need to re-organize the information to use it in multiple contexts, and to manipulate it in collaboration with colleagues of different backgrounds and focus of interest.

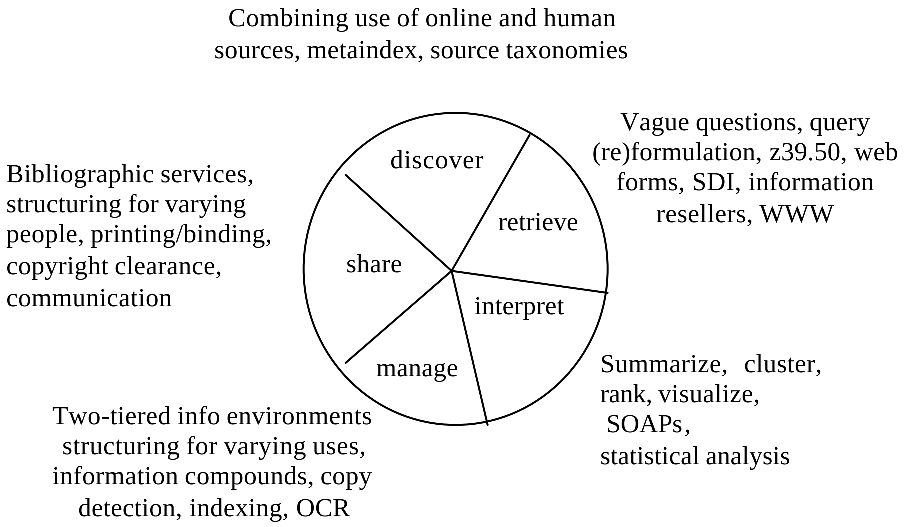

Combining use of online and human sources, metaindex, source taxonomies
Vague questions, query (re)formulation, z39.50, web forms, SDI, information resellers, WWW
Bibliographic services, structuring for varying people, printing/binding, copyright clearance, communication
discover
retrieve
share
interpret
manage
Summarize, cluster, rank, visualize, SOAPs, statistical analysis
Two-tiered info environments structuring for varying uses, information compounds, copy detection, indexing, OCR

Figure 1. The main activities of digital library users.

There is no single definition for Digital Libraries. The definition evolves as research progresses and we learn more about digital libraries. Some of the current definitions are:

- Digital libraries are systems that combine the machinery of digital computing, storage and communication, the content, and software needed to reproduce, emulate, and extend the services of collecting, cataloging, finding and disseminating information offered by traditional libraries based on paper and other materials. A full service digital library must not only fulfill all essential services provided by traditional libraries but also make good use of the advantages of digital technology.

- Digital libraries are viewed as systems providing a community of users with coherent access to a large, organized repository of information and knowledge. This organization of information is characterized by the absence of prior detailed knowledge of the uses of the information. The ability of the user to access, reorganize, and utilize this repository is enriched by the capabilities of digital technologies [3].

- The concept of a “digital library” is not merely equivalent to a digitized collection with information management tools. It is rather an environment to bring together collections, services, and people in support of the full life cycle of creation, dissemination, use, and preservation of data, information, and knowledge [6].

From the definitions above, it can be concluded that researchers have stretched the definition of digital libraries. More people are recognizing that digital library is not a topic only in computer and information science, but advances in digital library also depend on efforts from legal community.

Digital libraries are libraries extended and enhanced through digital technology. Important aspects of a library that may be extended and enhanced include:

- the collection of the library
- the organization and management of the collections
- access to library items and the processing of the information contained in the items
- the communication of information about the items.

The purposes of digital libraries are:

- to speed up the systematic development of the means to collect, store, and organize information and knowledge in digital form, and of digital library collections,
- to promote the economical and efficient delivery of information to all parts of society,
- to encourage co-operative efforts which leverage the considerable investment in research resources, computing and communications network,
- to strengthen communication and collaboration between and among the research, business, government and educational communities,
- to contribute to the lifelong learning opportunities of all people.

Figure 2 shows a digital library service model. Digital libraries distribute a rich and coherent set of information services (including selection, organization, access, distribution, and persistence) to users reliably and economically. These services are enabled by a suite of tools that operates on objects consisting of content packages, related metadata, service methods, and means of management.

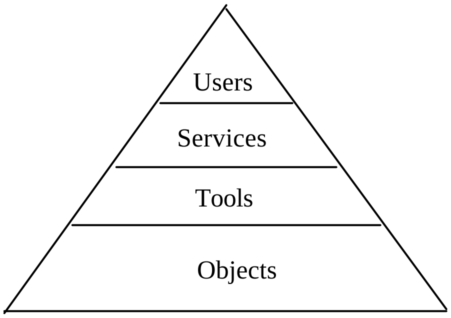


Figure 2. Digital libraries service model.

As for the relationship between digital libraries and NII, digital libraries provide the critical information management technology for the NII, and at the same time represent its primary information and knowledge repositories. In other words, digital library is the core of the NII. The information services, search facilities, and multimedia technologies constitute the digital libraries technologies. Like other NII technologies, they must provide for dependability, manageability, ease of use, interoperability, and security and privacy [3].

Notice in most cases, we use the plural term “digital libraries” meaning that we do not expect to see a single digital library. Each information repository is managed separately, possibly with different technologies, and hence each constitutes a digital library [3]. However we should integrate "virtually" separate libraries into a single one.

2.2 INFRASTRUCTURE REQUIREMENTS FOR DIGITAL LIBRARIES

Each single organization can create its own digital library. To share information across these libraries, it is necessary and important to have a common infrastructure facilitating such sharing. The same infrastructure can also be supportive to sharing of technologies used to build the digital libraries.

The infrastructure for digital libraries should include the following components:

- Shared information representation models, service representation models, and access protocols. These will facilitate the sharing of information and services across digital libraries [3].
- Information "content" sharing agreements. This will take the form of communities of organizations that agree to share their collections. Initially, the sharing may be free, but eventually the community will institute common charging schemes. The communities will also provide rules for having additional members join [3].
- Resource directories. The infrastructure should describe available information resources and relative models and protocol and characterize the contents.
- Coordination forum. The goal of this forum is to coordinate national research and development activities [3].

Among these components, to establish common schemes for the naming of digital objects, and the linking of these schemes to protocols for object transmission, metadata, and object type classifications is the most urgent need. Naming schemes for digital objects that allow global unique reference is the basis for facilitating resource sharing, linkages, and interoperation among digital library systems and for facilitating scale-up of digital library prototypes.

Another essential requirement is a public key cryptosystem infrastructure, including the development of a system of key servers and the definition of standards and protocols. This is necessary to support digital library needs in areas such as security and authentication, privacy, rights management, and payments for the use of intellectual property [3]. Only after these problems are addressed, is it possible for commercial publishers and other information suppliers to make large amounts of high-value copyrighted information broadly available to digital library users. This in turn will restrict the development of research prototypes and may be a distorting factor in studies of user behavior.

2.3 RESEARCH ISSUES IN DIGITAL LIBRARIES

There are five key research issues in digital libraries. They are (i) interoperability, (ii) description of objects and repositories, (iii) collection management and organization, (iv) user interfaces and human-computer interaction, and (v) economic, social, and legal issues.

1. Interoperability

The more technical interoperability research involve protocol design that supports a broad range of interaction types, inter-repository protocols, distributed search protocols and technologies (including the ability to search across heterogeneous databases with some level of semantic consistency), and object interchange protocols [3]. The various services provided by digital libraries must be interoperable. Existing Internet protocols are obviously inadequate for this. New protocols and systems are needed. This incurs the question of how to deploy prototype systems and how to make the tradeoffs between advanced capabilities and ubiquity of access. Managing this contradiction will have a critical influence in the development of digital libraries.

2. Description of Objects and Repositories

Description of objects and repositories are necessary to provide users a coherent view of information in various digital libraries. Objects and repositories must be described in a consistent fashion to facilitate distributed search and retrieval of diverse sources. Interoperability at the level of deep semantics will require breakthroughs in description as well as retrieval, objects interchange, and object retrieval protocols [3].

Issues here include the definition and use of metadata and its capture or computation from objects, the use of computed descriptions of objects, federation and integration of heterogeneous repositories with disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for automatic rating, ranking and evaluation of information quality, genre, and other properties [3]. Knowledge representation and interchange, the definition and interchange of ontologies for information context, and the appropriate roles of human librarians and subject expert in the digital library context are also important.

3. Collection Management and Organization

The central problems here are policies and methods for incorporating information resources on the networked into managed collections, rights management, payment, and control. The relationship between replication and caching of information and collection management in a distributed environment, the authority and quality of content in digital libraries, ensuring and identifying the attributes of contents, enhanced support of textual information and support of nontextual and multimedia information capture, organization, storage and retrieval all call for research. The preservation of digital content for long periods of time, across multiple generation of hardware and software technologies and standards is essential in the creation of effective digital libraries and need careful examination.

4.User Interfaces and Human-Computer Interaction

Among lot of issues in user interfaces and human-computer interaction, some are central problems. These issues include: display of information, visualization and navigation of large information collections, and linkages to information manipulation/analysis tools, the use of more sophisticated models of user behavior and needs in long-term interactions with digital libraries, more comprehensive understanding of user needs, objectives, and behavior in employing digital library systems, and adapting to variations in the capabilities of user workstations and network connections in presenting appropriate user interfaces.

5. Economic, Social and Legal Issues

Digital libraries are not simply technological constructs; they exist within a rich legal, social, and economic context, and will succeed only to the extent that they meet these broader needs. Rights management, economic models for the use of electronic information, and billing systems to support these economic models will be needed [3]. User privacy and complex policy issues concerning collection development and management, and preservation and archiving are also needed. Existing library practice may be helpful to solving these problems. We need to better understand the social context of digital documents, including authorship, ownership, the act of publication, versions, authenticity, and integrity.

2.4 ARCHITECTURE FOR INFORMATION IN DIGITAL LIBRARIES

In this section we discuss architecture for information presentation in Digital Libraries. We first present two solutions for the architecture, and then examine the core that supports infrastructure of one of these solutions.

2.4.1 Architecture of Digital Library Systems

According to [7], the key components in a digital library system are user interfaces, repositories, and handle system and search system, as shown in Figure 3.

User Interfaces

Each user interface has two parts. One is for the actual interactions with users. The other is **client services** that allow users to decide where to search and what to retrieve, interpret information structured as digital objects, negotiate terms and conditions, manage relationships between digital objects, remember the state of the interaction, and convert the protocols used by the various part of the system.

Repository

"Repositories store and manage digital objects and other information." A digital object is a data structure whose principal components are digital materials, or **data**, and **key-metadata**. The key-metadata includes a globally unique identifier for this digital object, called a **handle**; it may also include other metadata. The data can be elements or other digital objects [8]. There may be many repositories of various types, like modern repositories, legacy databases, and Web servers, in a large digital library. **Repository Access Protocol (RAP)** is the interface to this repository. Features of RAP include: (i) explicit recognition of rights and permissions that need to be satisfied before a client can access a digital object, (ii) support for a very general range of dissemination of digital objects, and (iii) an open architecture with well defined interfaces [7].

Handle System

Handles are general-purpose identifiers that can be used to identify Internet resources, such as digital objects, over long periods of time and to manage materials stored in any repository or database.

Search System

When a digital library system is designed, it is assumed that there will be many indexes and catalog that can be searched to discover information before retrieving it from a repository. These indexes may be independently managed and support a wide range of protocols [7].

Based on the work in [7,10], services offered by digital libraries are decomposed into four parts: collection service, naming service, repository service, and indexing service. Repository service provides from simple deposit and access to digital objects to sophisticated management, aggregation and marshaling of the information stored in the repository [10]. With the index service, digital objects that may be distributed across multiple repository servers are discovered via query. The index service also provides metadata, which is used by other services, and the capabilities of its query mechanisms. The collection service provides the means for aggregation of sets of digital objects into meaningful collection [10]. Collections are created by a collection server by reading its metadata and applying its collection definition criteria to define which objects belong to these collections. A user interface gateway offers searching for and access to objects within local collections and make query routing decisions with collection service and index service together based on factors such as content, cost, performance and the like. This decomposition facilitates the extensibility of digital libraries. New services can be easily added as a component.

Handle system, digital object, and the common repository access interface (RAP) forms the core infrastructure of a digital library [10].

It should be noticed that digital library is a distributed system; the four components in Figure 3 may be physically located in many places.

An agent-based architecture for digital libraries is presented in [9]. It is presented in Section 3.1.3.

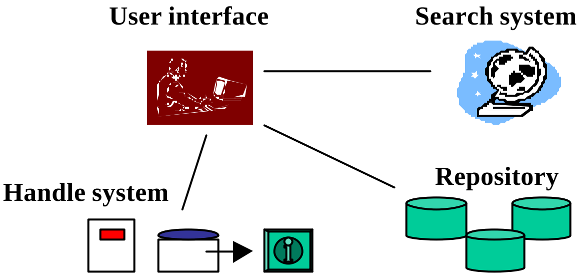


Figure 3. Major components of a digital library system.

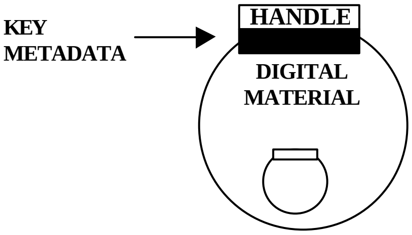


Figure 4. A digital object.

2.4.2 Digital Objects

The digital object

A digital object is a fundamental unit of the digital library architecture [7]. It consists of two components: key-metadata and digital material, as illustrated in Figure 4.

Key-metadata

The key-metadata is the information stored in the digital object that is needed to manage the digital object in a networked environment – for example to store, replicate, or transmit the object without providing access to the content. It includes a handle, an identifier globally unique to the digital object, terms and conditions, and other optional metadata.

Digital material
The digital material (or **data**) can be a set of sequences of bits or other digital objects. It is used to store digital library materials. For instance, a digital object may store a text with SGML mark-up.

Note that because of the characteristics of information, a digital object could be embedded into another digital object, which is called **MetaObject** like Metadata for digital objects.

2.4.3 Handle and the Handle System

In digital libraries, there are various items, such as people, computers, networks, repositories, databases, search systems, Web servers, digital objects, and many more. To keep tracking of these items, a systematic approach to identification is needed.

Handles are a set of general purpose identifies. In the digital library system, handles are used to identify digital objects and repositories. However handles can also be used to identify almost any Internet resource. A **handle system** is a distributed system that stores handles and associated data that is used to locate or access the item named by the handle.

Handles are different from the widely used Uniform Resource Location (URL) in that they identify resources by name, while URLs identify Internet resources by location.

Handles are names that persist for long periods of time, but the resource that they identify may change its form, may be stored in many locations, move its location, or otherwise be altered with time [7].

cnri.dlib/february96-arms	URL	http://www.dlib.org/dlib/februrary96/02arms.html
	RAP	repository.dlib.org

Figure 5. A handle record.

An illustrative example of handles is given in Figure 5. The handle is "cnri.dlib/july96-arms", identifying an article in D-Lib Magazine. Two fields of handle data stored in the handle system for this item indicates that this article can be found in two locations. Each data field contains two parts, a data type and the data. The first data field is of type "URL"; the associated data is a conventional URL. The second is of type "RAP", indicating that the item can be accessed using the protocol known as RAP; the data is the address of the repository in which the item is stored [7].

Note that the handle for this article remains the same forever. But the handle data may change with time. If this article is moved or duplicated in another repository, the data part of the handle recorded will be changed. The handle itself, however, will remain unchanged.

Resolving a handle is presenting a handle to the handle system and receiving as a reply information about the item identified. Usually users send a name (handle) to the handle system to find the location or locations of the digital object with that name.

Naming Authorities

Handles are created by naming authorities, administrative units that are authorized to create and edit handles [7]. A naming authority's name is composed of one or more strings separated by periods. For example,

cnri.dlib
loc.ndlp.amrlp
10.12345

In the handle system, there are two mechanisms to control that have permission to create naming authorities and create and edit handles: individual administrators and administrative groups. The latter are considered as more flexible and convenient.

"Each naming authority has at least one administrator or administrative group with full privileges for that naming authority, including permission to create a sub-naming authority. The administrator creates permissions for administration of handles within that naming authority, and can also create new naming authorities. Administrators can delegate privileges to other administrators, including the privilege of creating sub-naming authorities. "[7] Naming authorities are created hierarchically.

2.4.4 The Repository

A **repository** is a network-accessible storage system in which digital objects may be stored for possible subsequent access or retrieval. The repository has mechanisms for adding new digital objects to its collection (**depositing**) and for making them available (**accessing**), using, at a minimum, the **repository access protocol**. The repository may contain other related information, services and management systems.

Repositories have official, unique names, assigned or approved to assure uniqueness by a global naming authority. A repository name is not necessarily the name of a particular host. It may correspond to a set of hosts at different physical locations.

Each repository agrees on a protocol, called **Repository Access Protocol,** allowing deposits and access of digital objects or information about digital objects from that repository. RAP is used to provide only the most basic capabilities. It may change over time. Repositories may support other more powerful query languages allowing users to access objects that meet meaningful criteria.

(i) Access to a digital object (ACCESS_DO)

Access to a digital object will generally invoke a service program that performs stated operations on the digital object or its metadata depending on the parameters supplied with the service request [8]. There are three service requests, **metadata**, **key-metadata** or the whole **digital object**.

When a user accesses a digital object through **ACCESS_DO**, he receives a **dissemination**, the result of the service request, and information such as the key-metadata of the digital object, the identity of the repository, the service request that produced the result, the method of communication (if appropriate) and a transaction string corresponding to an entry in the transaction record. The transaction string is distinctive to the repository. In addition, the dissemination may contain an appropriately authenticated version of some portion of the properties record for that object, including the specific terms and conditions that apply to this use of the digital object and the materials contained therein.

(ii) Deposit of a digital object (DEPOSIT_DO)

There are several forms of DEPOSIT_DO. It could be taking data, a handle, and perhaps other metadata as arguments, and producing a stored digital object and properties record from these arguments." Or it may take a digital object as argument, probably with additional metadata, and simply deposit it. Also it possibly will take only data and certain non-key-metadata, automatically request a handle from a handle server, and then simultaneously store the object and register the handle.

The DEPOSIT_DO command could be used to replicate an existing digital object at additional repositories, or to directly modify an existing mutable digital object.

(iii) Access to reference services (ACCESS_REF)

This command provides a uniform and understood way to identify alternate means of accessing a specified repository and/or information about objects in that repository. Two possible responses are (i) No information, and (ii) a list of servers, protocol-name pairs, with the interpretation that each server, speaking the named protocol, will provide information about the contents of the repository [8].

2.5 DESIGN OF DIGITAL VIDEO LIBRARY SYSTEM

Video poses unique problems because of the difficulties in representing its contents. It is well known that image takes up much more space than the representation of the original text. Video is not only imagery but consists of 30 images per second [11].

Besides this, there are also other problems caused by introducing video into a digital library. In this section we will address these problems and overview the general architecture of a digital video library.

2.5.1 Research Issues in a Digital Video Library

Video compression
Video is quite different from text. From a presentation point of view, video data is huge and involves time dependent characteristics that must be adhered to for coherent viewing. Because of the storage and network limitation, before video is presented to users, it has to be compressed.

Video indexing
There have been sophisticated parsing and indexing technologies for text processing in various structured forms, from ASCII to PostScript to SGML and HTML. Video contains abundant information, conveyed in both the video signal (camera motion, scene changes, colors) and the audio signal (noises, silence, dialogue). But this information for indexing is inaccessible to the primarily text-based information retrieval mechanism. A common practice today is to log or tag the video with keywords and other forms of structured text to identify its content [11].

Video Segmentation
Since the time to scan a video cannot be dramatically shorter than the real time of the video, a digital video library must be efficient at giving users the material they need. To make the retrieval of bits faster, and to enable faster viewing or information assimilation, the digital video library will need to support partitioning video into small-sized clips and alternate representations of the video [11].

One of the issues relates to the implementation of the partitioning. In text documents, there are chapters, sections, subheadings, and similar conventions. Analogously, video data have scenes, shots, camera motions, and transitions. Manually describing this structure in a machine-readable form is obviously tedious and infeasible.

In addition to trying to size the video clips appropriately, the digital video library can provide the users alternate representations for the video, or layers of information. Users could then cheaply (in terms of data transfer time, possible economic cost, and user viewing time) review a given layer of information before deciding upon whether to incur the cost of richer layers of information or the complete video clip. For example, a given half hour video may have a text title, a text abstract, a full text transcript, a representative single image, and a representative one minute “skim” video, all in addition to the full video itself. The user could quickly review the title and perhaps the representative image, decide on whether to view the abstract and perhaps full transcript, and finally make the decision on whether to retrieve and view the full video [11].

Video Retrieving and Browsing

The basic service, offered by the digital video library, is easy and efficient information searching and retrieval. The two current standard measures of performance in information retrieval are *recall* and *precision*. Recall is the proportion of relevant documents that are actually retrieved, and precision is the proportion of retrieved documents that are actually relevant. These two measures may be traded off one for the other, i.e., returning one document that is a known match to a query guarantees 100% precision, but fails at recall if a number of other documents were relevant as well, or returning all of the library’s contents for a query guarantees 100% recall, but fails miserably at precision and filtering the information [11]. The goal of video retrieval is to get the most out of both recall and precision.

It is possible that when a general-purpose digital video library is created, precision has to be sacrificed to ensure that the material the user is interested in will be recalled in the result set. Then the result set probably becomes fairly large, so the user may need to filter the set and decide what is important. Three principle issues with respect to searching for information are:

- How to let the user quickly skim the video objects to locate sections of interest
- How to let the user adjust the size of the video objects returned
- How to aid users in the identification of desired video when multiple objects are returned [11].

2.5.2 Architecture of a Digital Video Library System

The Digital Video Library System is a complex system composed of the software components shown in Figure 6. These components are described below.

Video Storage System (VSS)

The Video Storage System stores video segments for processing and retrieving purposes. In order to provide intelligent access to portions of a video, the Video Storage System must be able to deliver numerous short video segments simultaneously.

Video Processing System (VPS)

The Video Processing System consists of video processing programs to manipulate, compress, compact, and analyze the video and audio components of a video segment. It also contains a component to recognize keywords from the sound track of video segments.

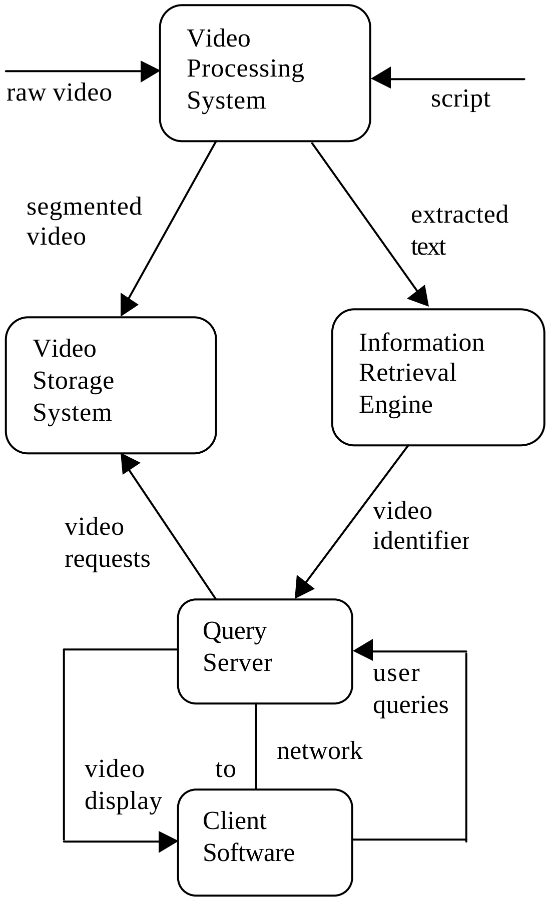


Figure 6. Software components of a digital video library.

Information Retrieval Engine (IRE)
The Information Retrieval Engine is used to store indices extracted from video segments and other information about the video segments, such as sources, copyright, and authorization. The Information Retrieval Engine can support both free-text and Boolean queries.

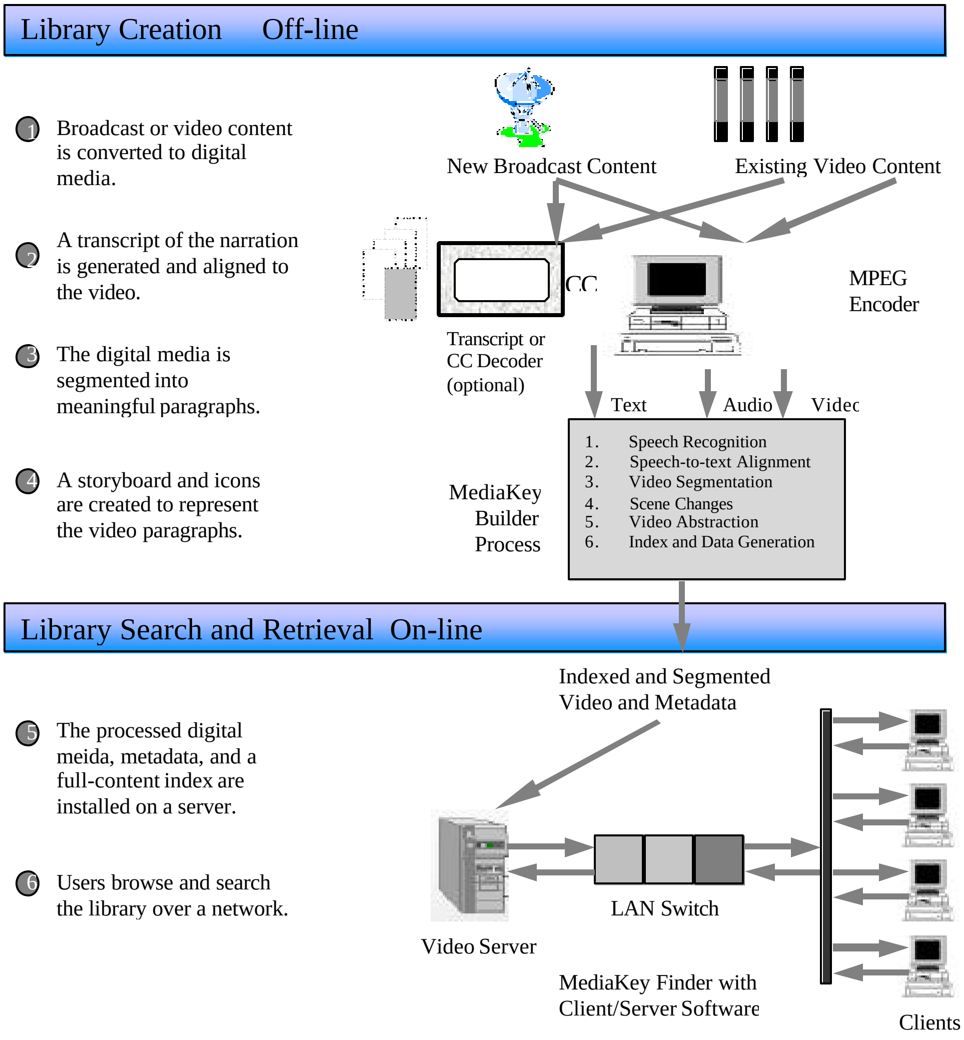
Client
The client is a graphical user interface residing on the user's computer. It includes interfaces for conducting structured and free text searching, hypertext browsing and a simple video editor.

Query Sever (QS)
The Query Server processes video queries from the remote client and communicates with the Information Retrieval Engine and Video Storage System to enable users to extract video data and create multimedia representations of the information of interest.

As seen in the Figure 6, these components are tightly interrelated and support three very different Digital Video Library System functions:

- The creation of the Digital Video Library System archive.
- The processing of video in the Digital Video System to build automatic indices.
- The access of the Digital Video Library System by users.

The system architecture of a digital video library is shown in Figure 7.



1. Broadcast or video content is converted to digital media.
2. A transcript of the narration is generated and aligned to the video.
3. The digital media is segmented into meaningful paragraphs.
4. A storyboard and icons are created to represent the video paragraphs.
5. The processed digital meida, metadata, and a full-content index are installed on a server.
6. Users browse and search the library over a network.

1. Speech Recognition
2. Speech-to-text Alignment
3. Video Segmentation
4. Scene Changes
5. Video Abstraction
6. Index and Data Generation

Figure 7. System architecture of a digital video library.

3. DIGITAL LIBRARIES INITIATIVE PROJECT

Since 1994, six research projects developing new methodologies or technologies to support digital library have been funded through a joint initiative of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA). These are collectively referred to as the Digital Library Initiative I and the total budget was US$ 25 M. Since then, digital libraries research has been regarded as a national challenge in the United States which is similar to the High Performance Computing and Communications Program (HPCC). Digital Library Initiative I ended at the end of 1998 and Digital Library Initiative II has begun this summer.

3.1 DIGITAL LIBRARIES INITIATIVE PROJECT --- PHASE I (DLI I)

The focus of Initiative I is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval and processing via communication networks – all in user-friendly ways. Six universities were involved in the Initiative I. Each is specialized in one specific topic.

Carnegie Mellon University
http://informedia.cs.cmu.edu

The Informedia interactive on-line digital video library system created by Carnegie Mellon University and WQED/Pittsburgh enable users to access, explore and retrieve science and mathematics materials from video archives.

University of California, Berkeley
http://elib.cs.berkeley.edu

This project produced a prototype digital library with a focus on environmental information. The library collected diverse information about the environment to be used for the preparation and evaluation of environmental data, impact reports and related materials.

University of Michigan
http://www.si.umich.edu/UMDL

This project conducted coordinated research and development to create, operate, use and evaluate a test bed of a large-scale, continually evolving multimedia digital library. The content focus of the library was earth and space sciences.

University of California, Santa Barbara
http://alexandria.sdc.ucsb.edu

Project Alexandria developed a digital library providing easy access to large and diverse collections of maps, images and pictorial materials as well as a full range of new electronic library services.

Stanford University
http://www-digilib.stanford.edu

The Stanford Integrated Digital Library Project developed the enabling technologies for a single, integrated "virtual" library that will provide uniform access to the large number of emerging networked information sources and collections--both on-line versions of pre-existing works and new works that will become available in the future.

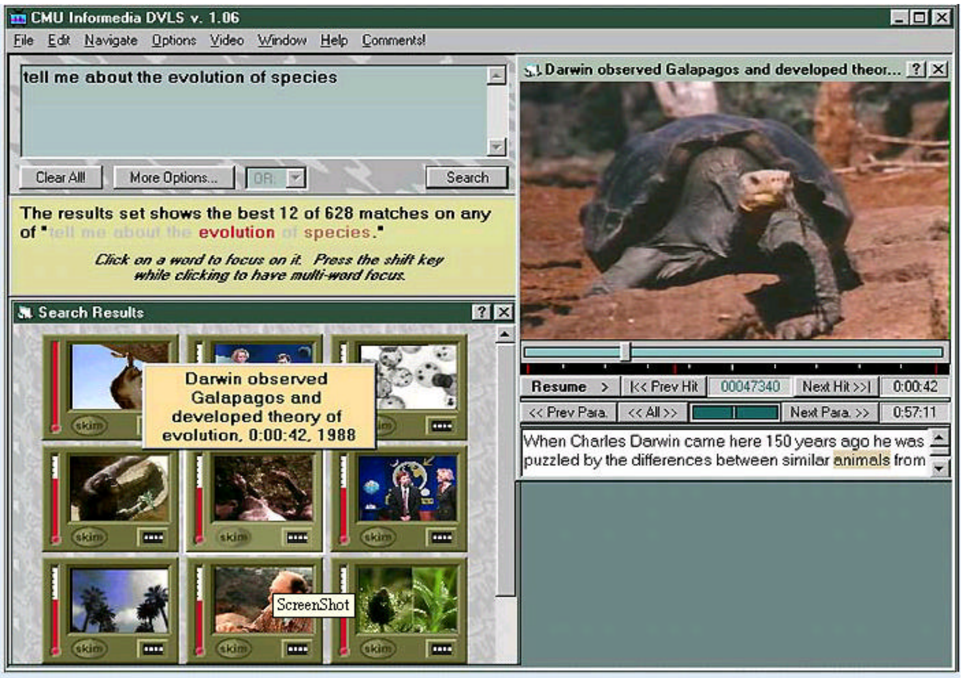
University of Illinois in Urbana-Champaign
http://dli.grainger.uiuc.edu/national.htm

This project draws on the new Grainger Engineering Library Information Center at the University of Illinois in Urbana-Champaign and the Artificial Intelligence Research Lab at the University of Arizona, http://ai.bpa.arizona.edu. This project is entered around journals and magazines in the engineering and science literature. The initial prototype system includes a user interface based on a customized version of Mosaic, software developed at the university under NSF sponsorship to help users navigate on the World Wide Web.

3.1.1 Carnegie Mellon University Informedia Digital Library Project

The Informedia Digital Video Library Project at Carnegie Mellon University is a large digital library of text, images, videos and audio data available for full content retrieval. It integrates natural language understanding, image processing, speech recognition, and video compression. The Informedia System allows a user to explore multimedia data in depth as well as in breadth.

Figure 8 is an example of how these components are combined in the Informedia user interface. An overview of the structure of the Informedia system is shown in Figure 9.



tell me about the evolution of species

The results set shows the best 12 of 628 matches on any of "tell me about the evolution of species."

Click on a word to focus on it. Press the shift key while clicking to have multi-word focus.

Darwin observed Galapagos and developed theor...

Darwin observed Galapagos and developed theory of evolution, 0:00:42, 1988

When Charles Darwin came here 150 years ago he was puzzled by the differences between similar animals from

Figure 8. The user interface of Informedia digital library.

The Informedia Library project is primarily used in education and training. Besides this, another application is News-on-Demand. News-on-demand monitors the evening news from the major networks and allows the user to retrieve stories in which they are interested. The News-on-demand application focuses on the limits of what can be done automatically and in limited time [17]. While other informedia prototypes are designed to be educational test beds, the News-on-Demand system is fully automated.

Currently, the Informedia collection contains approximately 1.5 terabytes of data, which is 2,400 hours of video encoded in the MPEG 1 format. Around 2,000 hours of CNN news broadcasts beginning in 1996 forms that the main body of the content. The remaining result from PBS broadcast documentaries produced by WQED, Pittsburgh, and documentaries for distance education produced by the BBC for the British Open University. The subject of the majority of these documentaries is mathematics and science. Besides these, there is also a small quantity of public domain videos, typically from government agency sources.

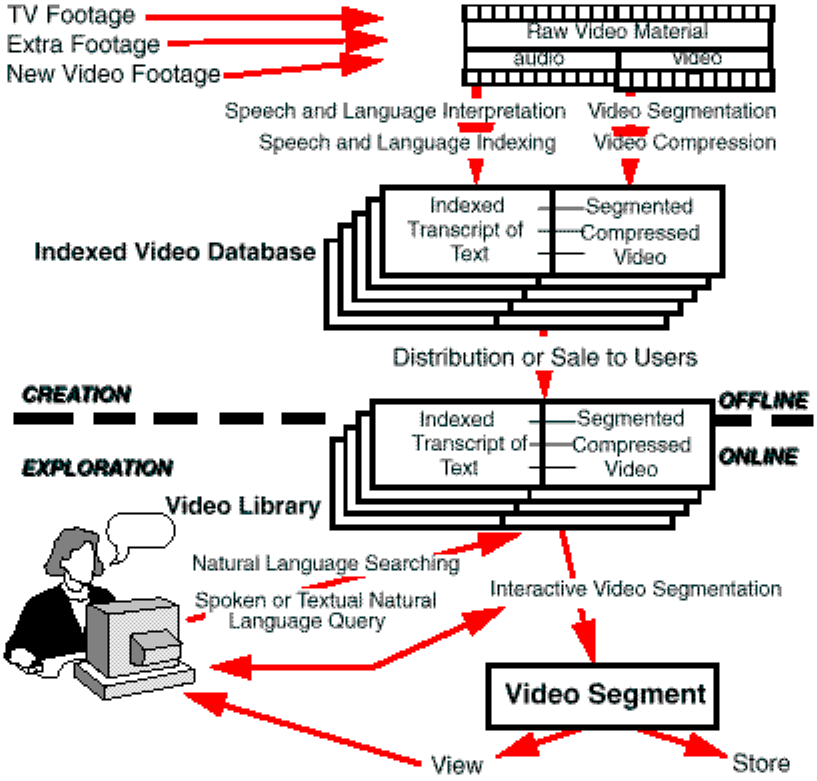

TV Footage
Extra Footage
New Video Footage
Raw Video Material
audio
video
Speech and Language Interpretation
Speech and Language Indexing
Video Segmentation
Video Compression
Indexed Video Database
Indexed Transcript of Text
Segmented Compressed Video
Distribution or Sale to Users
CREATION
EXPLORATION
OFFLINE
ONLINE
Indexed Transcript of Text
Segmented Compressed Video
Video Library
Natural Language Searching
Spoken or Textual Natural Language Query
Interactive Video Segmentation
Video Segment
View
Store

Figure 9. Overview of the Informedia Digital Video Library.

The metadata created by Informedia is extensive and automatically derived. It is an important resource for digital library researchers. Metadata for the Informedia collection includes:

1. **Transcripts** - textual forms of the audio tracks derived from:

- Closed captioning for the CNN data.
- Manual transcripts for the documentary material.
- Automatically derived transcripts from the Sphinx II speech recognizer for all of the data.

2. **Transcript alignment** - Sphinx II derived transcript to video time alignment for all three forms of transcription.

3. **Video OCR** - text regions identified and extracted from video imagery, converted to text via OCR.

4. **Face Descriptions** - human faces detected in video, described by Eigen Face representations.

5. **Geocodes** - latitude and longitude associated with video segments, derived from place names identified in the transcript and Video OCR data, computed from a gazetteer of world locations.

6. **Stills** - representative bit map or JPEG images selected from every automatically identified shot break (change of camera view).

7. **Segments** - video sequences representing single topic stories.

8. **Filmstrips** - collections of stills representing a segment.

9. **Topics** - automatically identified subjects of segments.

10. **Skims** - automatically created video abstracts comprised of concatenated sub-sections of segments creating a shortened version of the video for previewing [17].

3.1.2 University of California Berkley SunSITE Digital Library Project

This SunSITE testbed provides public access to important datasets pertaining to the environment, including environmental documents and reports, image collections, maps, sensor data and other collections [18]. At the mean time, this testbed serves as the foundation for research efforts in computer vision, database management, document analysis, natural language processing, and storage management. It is also used in School of Information Management and Systems of UC Berkeley for user assessment and evaluation and for information retrieval research. Researchers in College of Environmental Design of UC Berkeley use the testbed for Geographic Information Systems (GIS) experiments.

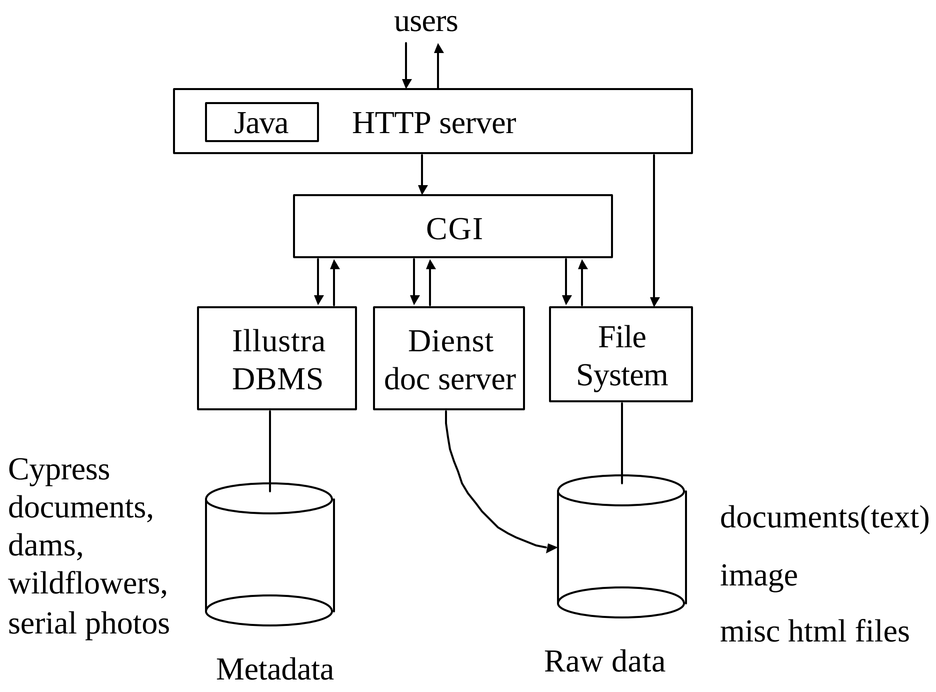


Figure 10. SunSITE digital library architecture: data access.

Software System Architecture

All access to the testbed is provided via the HTTP protocol for public and project members [18]. As shown in Figure 10, the interaction between WWW clients and other software systems is provided through the Common Gateway Interface (CGI) mechanism. Foremost among these systems is the relational database server, enabling forms-based access to nearly all data in the Berkeley Digital Library Project. Other methods besides forms are available for accessing the data, such as clickable maps and sorted lists. These and many others are available via the Access Matrix, which provides a top-level access point to all the data in the testbed.

Collections in the Testbed Database

The collection began as a testbed for research in computer science and information technology; it has since become a valuable repository of environmental and biological information. As of early 1999, the collection represents about a half terabyte of data, including over 70,000 digital images, nearly 300,000 pages of environmental documents, and over a million records in geographical and botanical databases. All of these data are accessible in online searchable databases; they are also freely available for the purpose of research and experimentation. An example is shown in Figure 11.

Figure 12 is an example of content-based searching. In this example, user can search for any picture based on specified colors.

3.1.3 University of Michigan Digital Library Project

As a large-scale effort, the University of Michigan Digital Library (UMDL) provides information services for research and education, in university and high school environments. The wide range of users and uses incurred scale and heterogeneity problem. These issues are addressed in the UMDL by designing an open, distributed system architecture where interacting software agents cooperate and compete to provide library services. The distributed architecture promotes modularity, flexibility, and incremental development, and accommodates diversity in current and future library environments. However, distribution also presents difficult problems in interoperability, coordination, search, and resource allocation. The activities are coordinated in the UMDL by dynamically forming agent teams to perform complex library tasks [19].

Agents

The architecture of UMDL, shown in Figure 13, is based on the concept of a software agent. An agent represents an element of the digital library (collection or service), and is a highly encapsulated piece of software that has the following special properties:

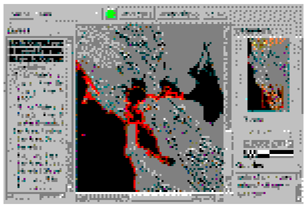
- **Autonomy**: the agent represents both the capabilities (ability to compute something) and the preferences over how that capability is used. Thus, agents have the ability to reason about how they use their resources. In other words, an agent not have to fulfill every request for service, only those consistent with its preferences. A traditional computer program does not have this reasoning ability.

- **Negotiation**: since the agents are autonomous, they must negotiate with other agents to gain access to other resources or capabilities. The process of negotiation can be, but is not required to be, stateful and will often consist of a "conversation sequence", where multiple messages are exchanged according to some prescribed protocol, which itself can be negotiated [9].

- **Botanical Data**

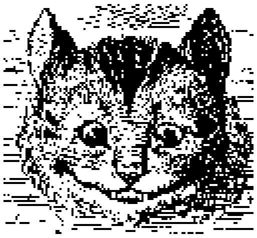
The **CalFlora Database** contains taxonomical and distribution information for the 8000+ native California plants. The **Occurrence Database** includes over 300,000 records of California plant sighting from many federal, state, and private sources. The botanical databases are linked to our CalPhotos collection of California plants, and are also linked to external collections of data, maps, and photos.

Geographical Data



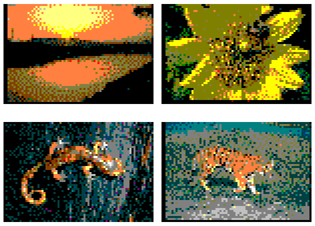
Much of the geographical data in our collection is being used to develop our web-based **GIS Viewer**. The **Street Finder** uses 500,000 Tiger records of S.F. Bay Area streets along with the 70,000-record USGS GNIS database. **California Dams** is a database of information about the 1395 dams under state jurisdiction. An additional 11 GB of geographical data represents maps and imagery that have been processed for inclusion as layers in our GIS Viewer. This includes Digital Ortho Quads and DRG maps for the S.F. Bay Area.

Documents



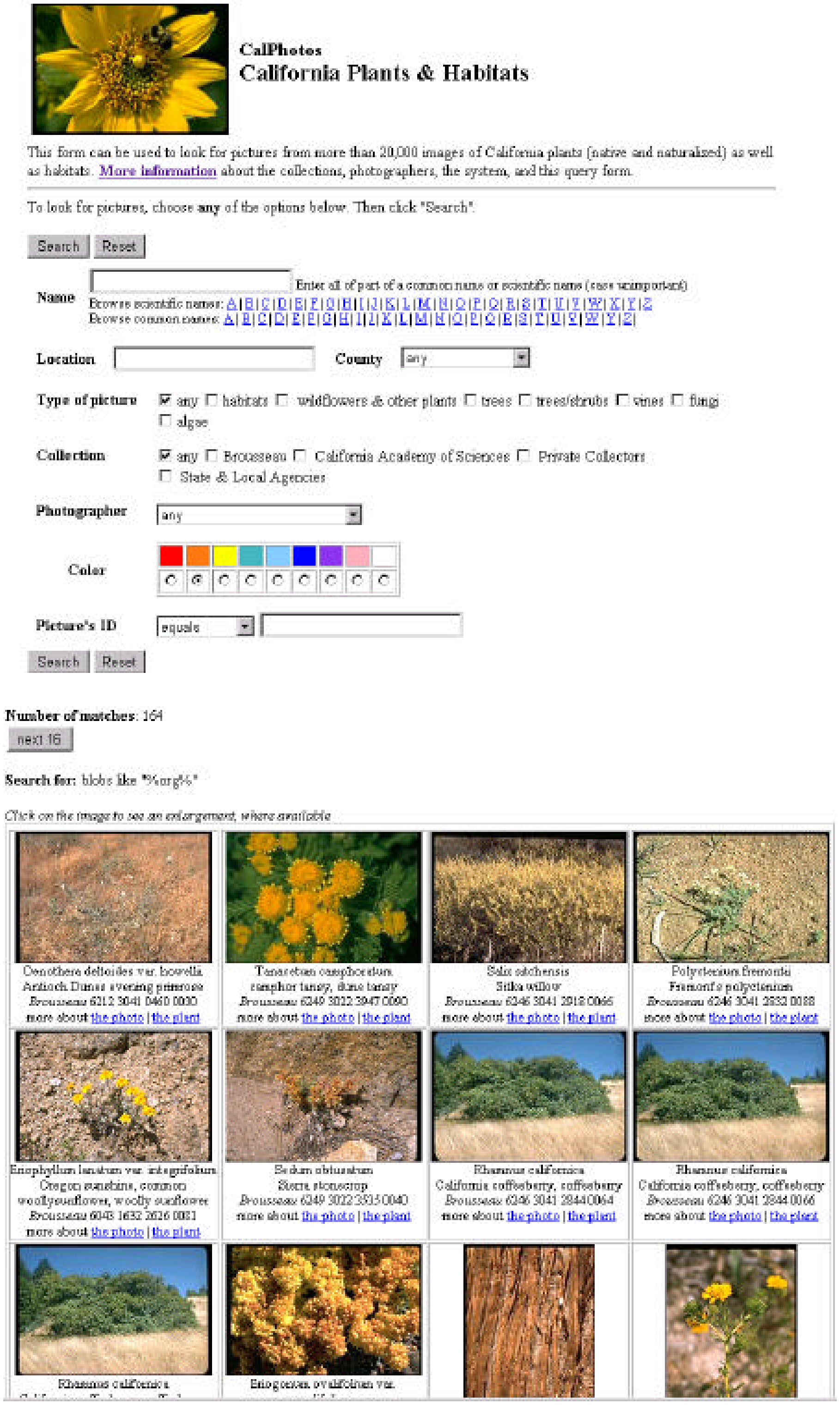
Most of the 300,000 pages of digital documents are environmental reports and plans that were provided by California State agencies. The most frequently accessed documents include **County General Plans** for every California county and a survey **of 125 Sacramento Delta fish species**. In addition to providing online access to important environmental documents, the document collection is the testbed for the **Multivalent Document research**.

Photographs



The photo collection includes 17,000 images of **California natural resources** from the state Department of Water Resources, several hundred **aerial photos**, 17,000 photos of **California native plants** from St. Mary's College, the California Academy of Science, and others, a small collection of **California animals**, and 40,000 **Corel stock photos**. These images are used within the project for **computer vision research**

Figure 11. An example of the repository of environmental and biological data.



CalPhotos
California Plants & Habitats

This form can be used to look for pictures from more than 20,000 images of California plants (native and naturalized) as well as habitats. More information about the collections, photographers, the system, and this query form.

To look for pictures, choose **any** of the options below. Then click "Search".

Name Enter all of part of a common name or scientific name (case unimportant)
Browse scientific names: A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z
Browse common names: A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | Y | Z |

Location **County** any

Type of picture ☑ any ☐ habitats ☐ wildflowers & other plants ☐ trees ☐ trees/shrubs ☐ vines ☐ fungi ☐ algae

Collection ☑ any ☐ Brousseau ☐ California Academy of Sciences ☐ Private Collectors ☐ State & Local Agencies

Photographer any

Color ○ ● ○ ○ ○ ○ ○ ○ ○

Picture's ID equals

Number of matches: 164

Search for: blobs like "%org%"

Click on the image to see an enlargement, where available

Oenothera deltoides var. howellii Antioch Dunes evening primrose *Brousseau* 6212 3041 0460 0030 more about the photo | the plant	Tanacetum camphoratum camphor tansy, dune tansy *Brousseau* 6249 3022 3947 0090 more about the photo | the plant	Salix sitchensis Sitka willow *Brousseau* 6246 3041 2912 0066 more about the photo | the plant	Polyctenium fremontii Fremont's polyctenium *Brousseau* 6246 3041 2832 0088 more about the photo | the plant
Eriophyllum lanatum var. integrifolium Oregon sunshine, common woollysunflower, woolly sunflower *Brousseau* 6043 1632 2626 0081 more about the photo | the plant	Sedum obtusatum Sierra stonecrop *Brousseau* 6249 3022 3515 0040 more about the photo | the plant	Rhamnus californica California coffeeberry, coffeeberry *Brousseau* 6246 3041 2844 0064 more about the photo | the plant	Rhamnus californica California coffeeberry, coffeeberry *Brousseau* 6246 3041 2844 0066 more about the photo | the plant
Rhamnus californica	Eriogonum ovalifolium var.		

Figure 12. An example of content-based search and results of search.

Autonomy, implying local or decentralized control, is critical to scalability of UMDL. Negotiation is complementary to autonomy, in that autonomous agents must be capable of making binding commitments for the system to work.

There are three types of agents:

- **UIAs (User Interface Agents)** provide a communication wrapper around a user interface. This wrapper performs two functions. First, it encapsulates user queries in the proper form for the UMDL protocols. Second, it publishes a profile of the user to appropriate agents, which is used by mediator agents to guide the search process [9].

- **Mediator agents**, there are many types of mediator agents, performing all tasks that are required to pass on a query from a UIA to a collection, monitor the progress of a query, transmit the results of a query, and perform all ways of translation and bookkeeping. Currently, there are two classes of mediators in UMDL. "Registry agents capture the address and contents of each collection. Query-planning agents receive queries and route them to collections, possibly consulting other sources of information to establish the route."[9] Another special type of mediators, facilitators, mediates negotiation among agents.

- **CIAs (Collection Interface Agents)** provide a communication wrapper for a collection of information. CIAs perform translation tasks similar to those performed by the UIA for a user interface, and publish the contents and capabilities of a collection in the conspectus language. The conspectus is a normalized description of content. It provides interoperability for various search and retrieval methods through a common representation over collections. It is written in a language defined by UM, which is called UCL (UMDL Conspectus Language) [9].

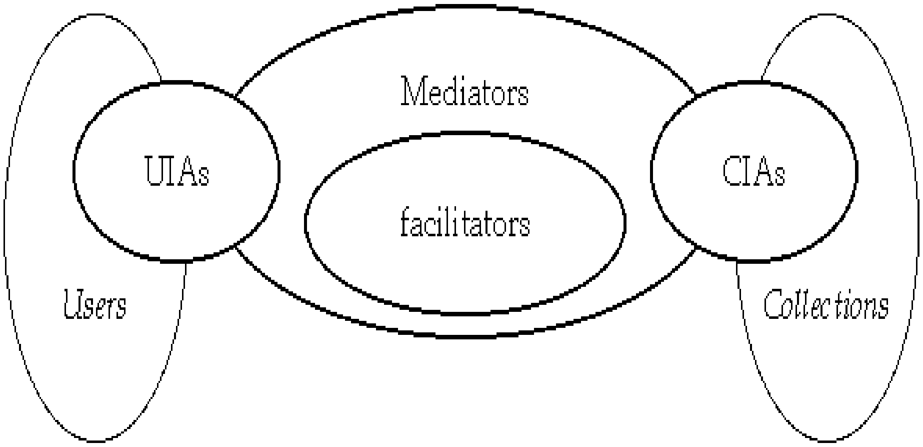


Figure 13. UMDL architecture.

Agent Teams

Complex UMDL tasks requires the coordination of multiple specialized agents working together on behalf of users and collection providers [20]. While the scope and nature of the desired tasks will continually evolve, a fundamental requirement of agents is that they can form teams. Agents must therefore be capable of describing their capabilities in a way that other agents understand, and communicating these descriptions to other agents.

UMDL agents communicate at three distinctive levels of abstraction. At the lowest level, agents utilize network protocols like TCP/IP to transport messages among themselves. The interpretation and processing of these messages is dictated by task-specific protocols. At the second level, agents communicate in more widely accepted language such as Z39.50. The capabilities of a specialized agent will remain untapped unless the agent can make its abilities and location known, and participate in the team-formation process. We thus define special protocols, shared by all UMDL agents, for the team formation and negotiation tasks. These UMDL protocols represent the third level of abstraction in agent communication.

The UMDL protocols are designed to allow agents to advertise them and find each other based on capabilities. A special agent **called Registry Agent** maintains a database that contains information about all the agents in UMDL, including their respective content and capability descriptions.

Figure 14 shows the interaction between agents when there is a search by author.

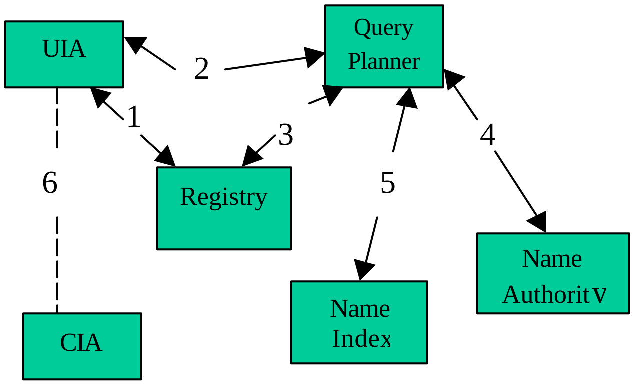


Figure 14. Interaction among agents when searching by author.

The main part of the testbed collections of UMDL is earth and space science. Commercial content focuses on timely journal literature and reference resources. These resources include:

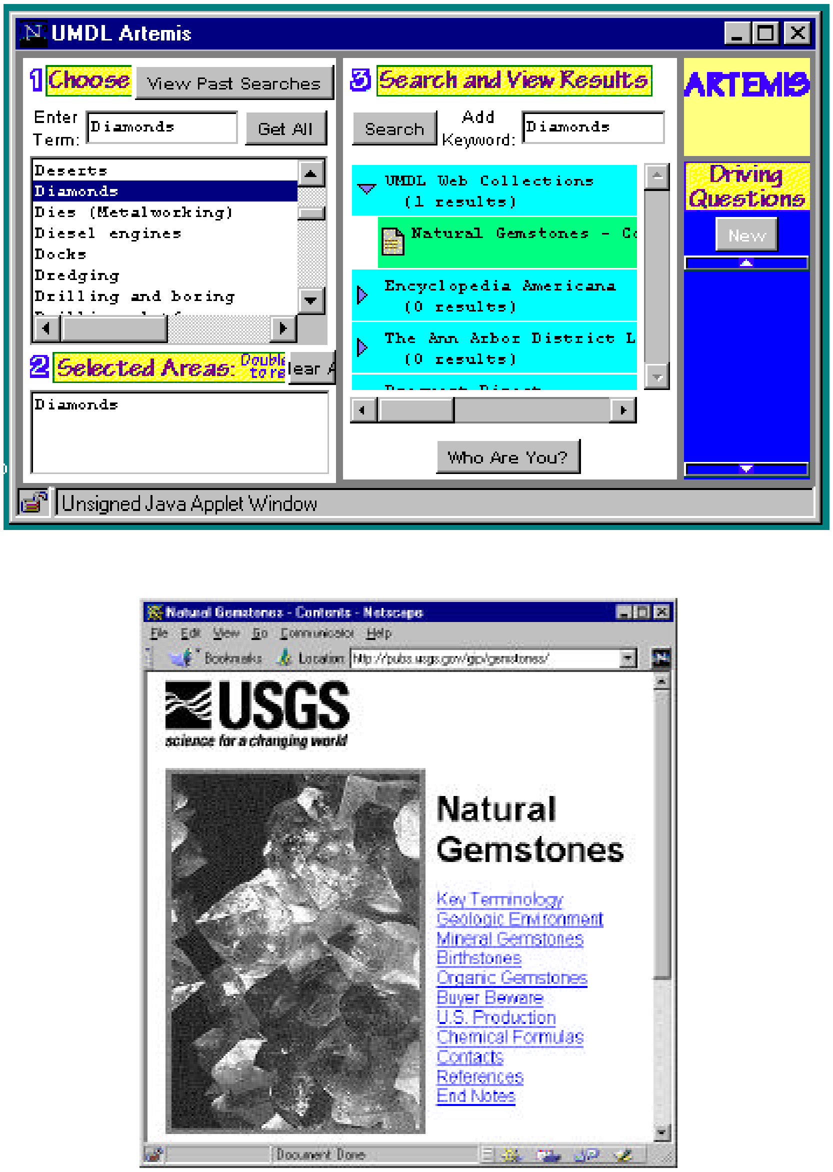
- Encyclopedia of Science and Technology (McGraw Hill)
- 200 core and popular journals (UMI)
- Encyclopedia American (Grolier’s)
- 50 scientific journals (Elsevier)
- Encyclopedia Britannica

The main tasks supported by UMDL user interface Artemis/Recommendation System are:

- Generalized "Subject Area searches
- "Keyword searches
- Recommendation of other web sites

The special function of Artemis is that users can make comments on the page and give it a rating.

A searching example is illustrated in Figure. In this example, the search for gemstones was created in Figure 15a and obtained results are shown in Figure 15b.



UMDL Artemis

1 Choose | View Past Searches

Enter Term: Diamonds | Get All

Deserts
Diamonds
Dies (Metalworking)
Diesel engines
Docks
Dredging
Drilling and boring

2 Selected Areas:

Diamonds

3 Search and View Results

Search | Add Keyword: Diamonds

UMDL Web Collections (1 results)
Natural Gemstones - Co
Encyclopedia Americana (0 results)
The Ann Arbor District L (0 results)

Who Are You?

Unsigned Java Applet Window

Natural Gemstones - Contents - Netscape

Location: http://pubs.usgs.gov/gip/gemstones/

USGS science for a changing world

Natural Gemstones

Key Terminology
Geologic Environment
Mineral Gemstones
Birthstones
Organic Gemstones
Buyer Beware
U.S. Production
Chemical Formulas
Contacts
References
End Notes

Document: Done

Figure 15. (a) Searching for gemstones, (b) Results of search.

3.1.4 University of California Santa Barbara Digital Library Project

The Alexandria Project's goal is to build a distributed digital library for materials that are referenced in geographic terms, such as by the names of communities or the types of geological features found in the material.

Figure 16 illustrates the basic ADL (Alexandria Digital Library) architecture, which derives a traditional library's four major components.

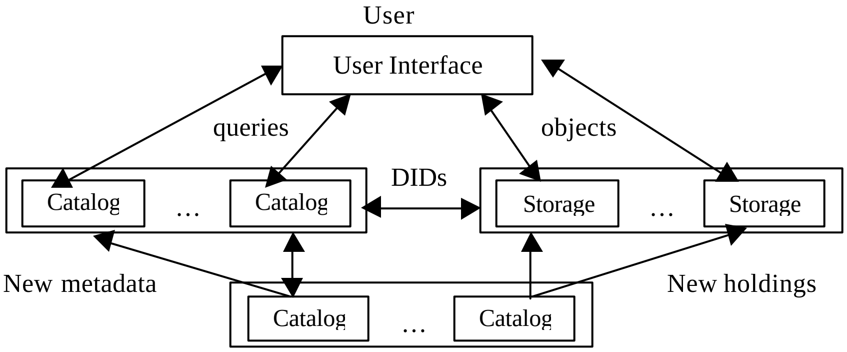


Figure 16. Architecture of ADL.

The **storage** component maintains and serves up the digital holdings of the library. These correspond to the ``stacks" of physical holdings (books, journals, etc.) in a traditional library. The **catalog** component manages, and facilitates searches of, the metadata describing the holdings, analogous to a traditional library card catalog. Catalog metadata are associated with storage objects by unique object identifiers, analogous to traditional library call numbers. The **ingest** component comprises the mechanisms by which librarians and other authorized users populate the catalog and storage components. Finally, The **user interface** component is the collection of mechanisms by which one interacts with the catalog (to conduct a search) or the storage (to retrieve objects corresponding to search results).

The Web prototype architecture is shown in Figure 17.

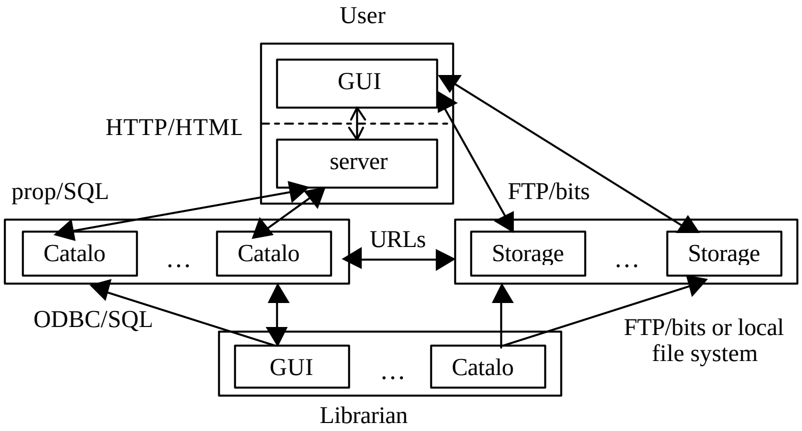
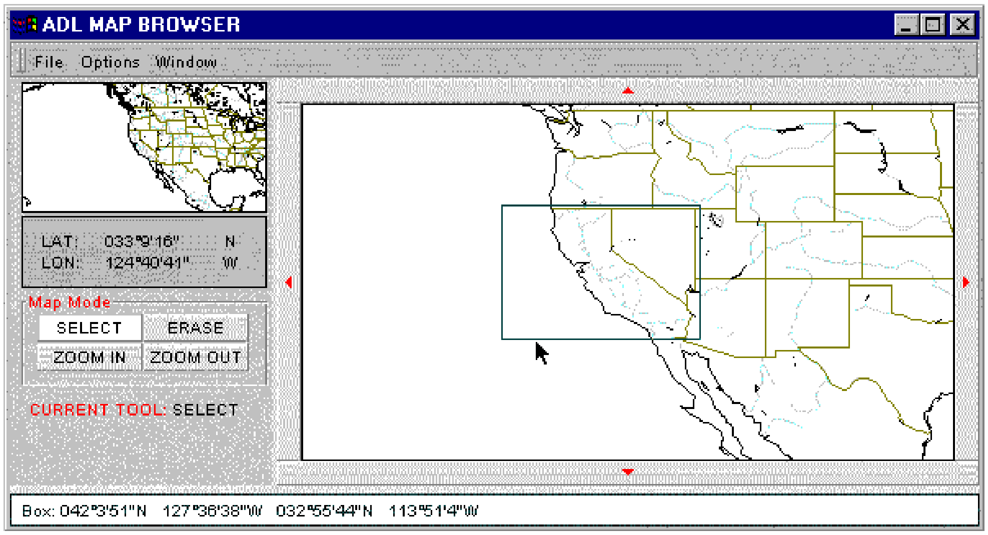


Figure 17. ADL Web prototype architecture.

ADL uses wavelets for image processing and texture for content-based retrieval. Besides, ADL also investigated parallel computation to address various performance issues, including multiprocessor servers, parallel I/O, and parallel wavelet transforms, both forward (for image ingest) and inverse (for efficient multi-scale image browsing).

Based on a traditional map library housed in the Map and Imagery Laboratory (MIL) in the Davidson Library at UCSB, ADL's holdings focus on collections of geographically referenced materials, including maps, satellite images, digitized aerial photographs, specialized textual material (such as gazetteers), and their associated metadata.

The user interface of ADL consists of several components. The major components are map browser, search options, workspace and metadata browser. Their screens are shown in Figures 18-21.

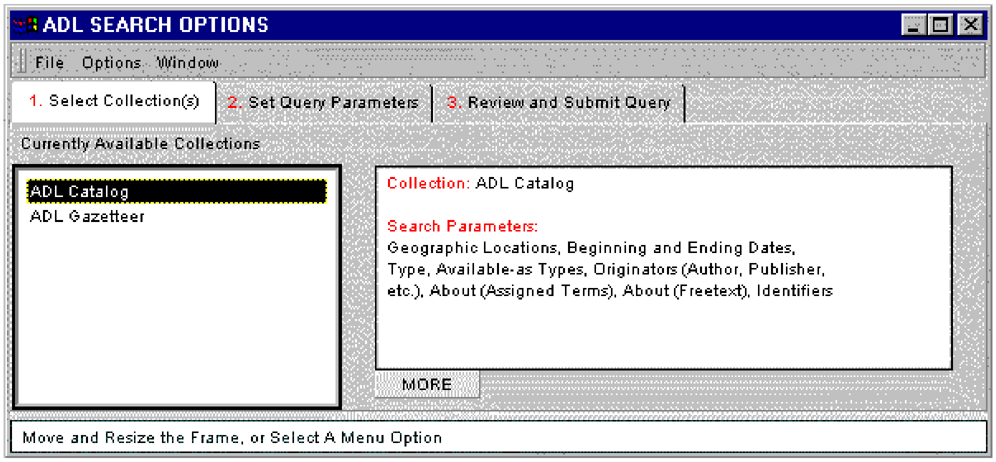


LAT: 033°9'16" N
LON: 124°40'41" W

CURRENT TOOL: SELECT

Box: 042°3'51"N 127°36'38"W 032°55'44"N 113°51'4"W

Figure 18. ADL map browser.



Currently Available Collections

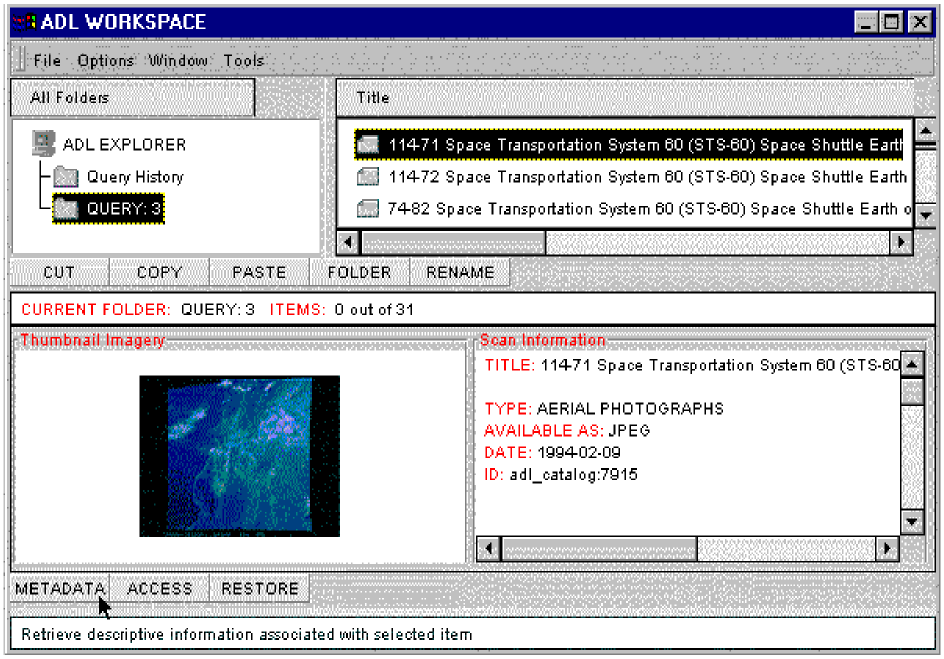
ADL Catalog
ADL Gazetteer

Collection: ADL Catalog

Search Parameters:
Geographic Locations, Beginning and Ending Dates, Type, Available-as Types, Originators (Author, Publisher, etc.), About (Assigned Terms), About (Freetext), Identifiers

Move and Resize the Frame, or Select A Menu Option

Figure 19. ADL search options.



ADL EXPLORER
Query History
QUERY: 3

114-71 Space Transportation System 60 (STS-60) Space Shuttle Earth
114-72 Space Transportation System 60 (STS-60) Space Shuttle Earth
74-82 Space Transportation System 60 (STS-60) Space Shuttle Earth o

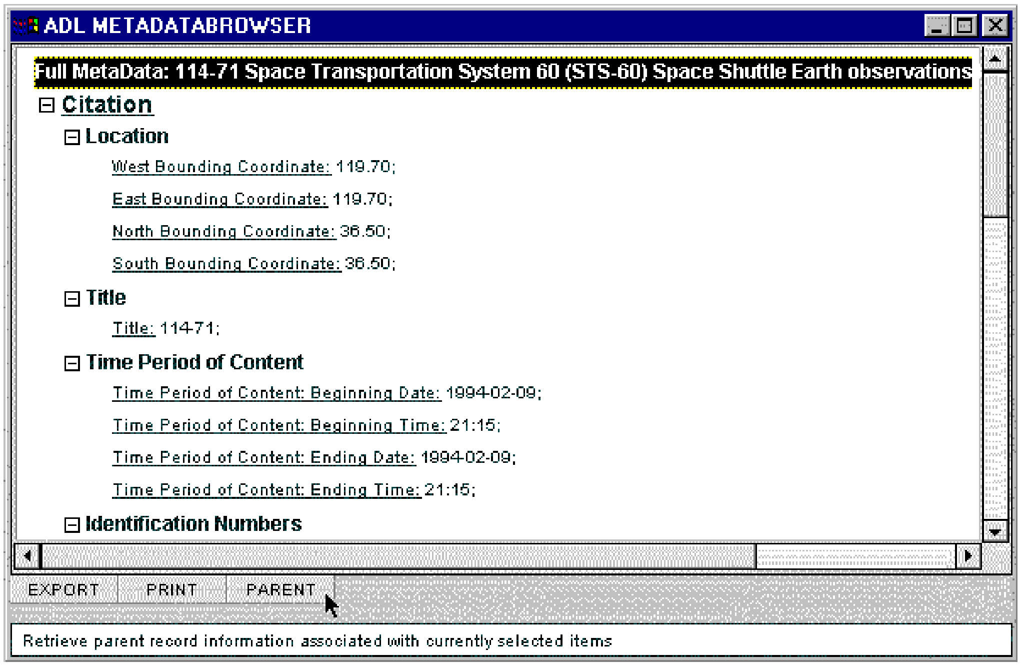
CURRENT FOLDER: QUERY: 3 ITEMS: 0 out of 31

TITLE: 114-71 Space Transportation System 60 (STS-60

TYPE: AERIAL PHOTOGRAPHS
AVAILABLE AS: JPEG
DATE: 1994-02-09
ID: adl_catalog:7915

Retrieve descriptive information associated with selected item

Figure 20. ADL workspace.



Full MetaData: 114-71 Space Transportation System 60 (STS-60) Space Shuttle Earth observations

⊟ Citation

⊟ Location

West Bounding Coordinate: 119.70;

East Bounding Coordinate: 119.70;

North Bounding Coordinate: 36.50;

South Bounding Coordinate: 36.50;

⊟ Title

Title: 114-71;

⊟ Time Period of Content

Time Period of Content: Beginning Date: 1994-02-09;

Time Period of Content: Beginning Time: 21:15;

Time Period of Content: Ending Date: 1994-02-09;

Time Period of Content: Ending Time: 21:15;

⊟ Identification Numbers

Retrieve parent record information associated with currently selected items

Figure 21. Metadata browser.

3.1.5 Stanford University Digital Library Project

Stanford digital library project focuses on **interoperability**. They developed the "**InfoBus**" protocol – **Digital Library InterOperating Protocol (DLIOP)**, which provides a uniform way to access a variety of services and information sources through "proxies" acting as interpreters between the InfoBus protocol and the native protocol. The InfoBus is implemented on top of a CORBA-based architecture using Inprise's Visibroker and Xerox' ILU. The second area is the legal and economic issue of a networked environment.

Figure 22 shows an example of three protocol domains. The first one is the local domain, which is a local network used by an information-services provider such as a company, a university, or even an individual. The second one is Telnet service domain, where clients log in to remote machines. The third one is HTTP, the protocol used for the WWW [21].

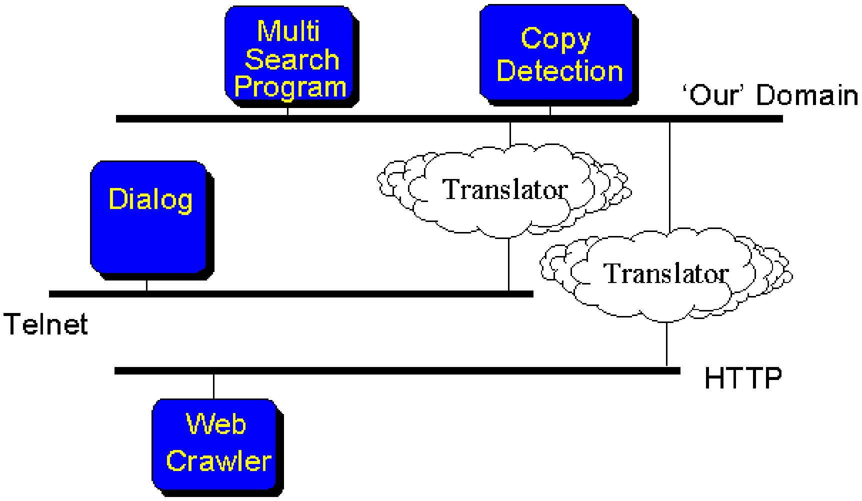


Figure 22. Interoperation across protocol domains.

The services in all the domains are accessible through their respective protocols. The service-interaction protocols in the local domain are locally controlled. The Dialog information service is an example of a Telnet-based information provider. The WebCrawler, a search engine that indexes documents on the WWW and returns their URL in response to queries, is an example of an HTTP-based service.

Dialog presents a teletype interface, through which the user first follows a standard login sequence (`Please logon:`), then selects one of the many databases offered through Dialog (`begin 245`). Users search the database through a proprietary query language (`select Library/ti`), then examine the results, and last terminate the sessions (`logout`). One possible abstraction of this process is that an open session operation is followed by open database, search, and quit operations. This abstraction can also be applied to WebCrawler, as shown in Figure 23.

The basic idea of Stanford InfoBus is Library Service Proxy. **Library-Service Proxy (LSP)** objects are created. Method calls on an LSP object invoke each interface element (`open session, open database`, and so on), and the method performs the appropriate operation on the corresponding service [21]. Figure 24 shows how LSPs can be used as the building blocks for the translators in Figure 23. The translator clouds are full of LSPs, each representing one service. A common interface thus makes two quite different services accessible from the local domain.

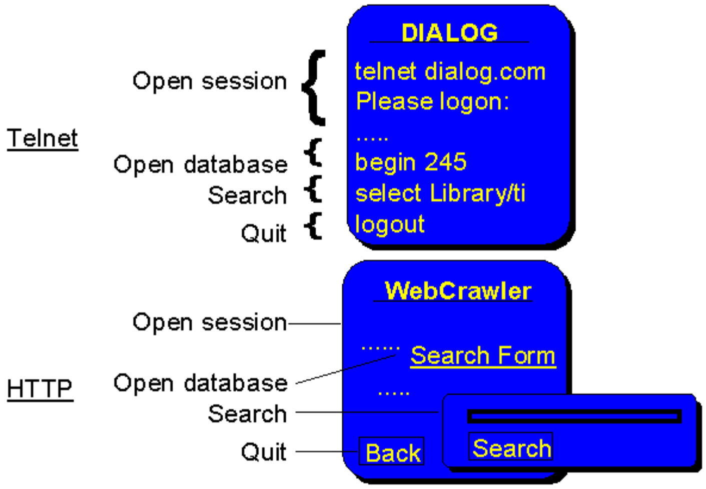


Figure 23. Glue for service access.

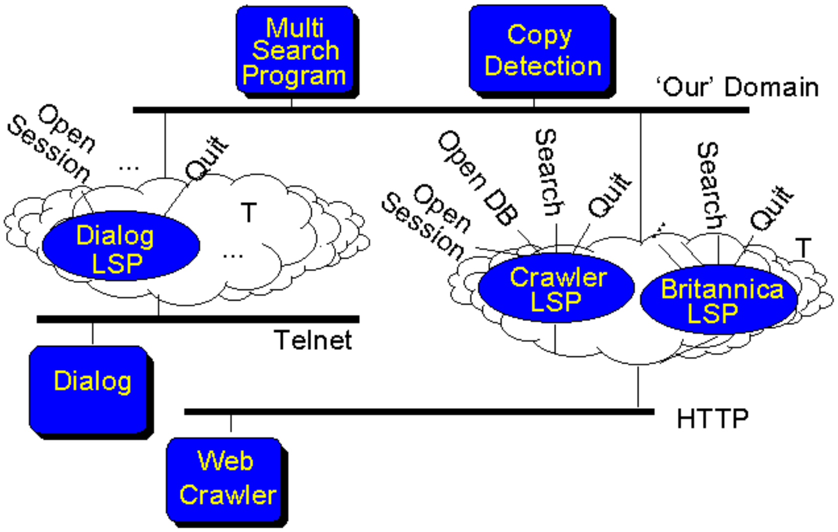


Figure 24. InfoBus idea—library service proxy.

There are many projects in Stanford digital library related to InfoBus. The collection of Stanford Digital Library is primarily computing literature. However, it has a strong focus on networked information sources, meaning that the vast arrays of topics found on the WWW are accessible through this project. The user interface DLITE is illustrated in Figure 25. It runs next to a Netscape browser.

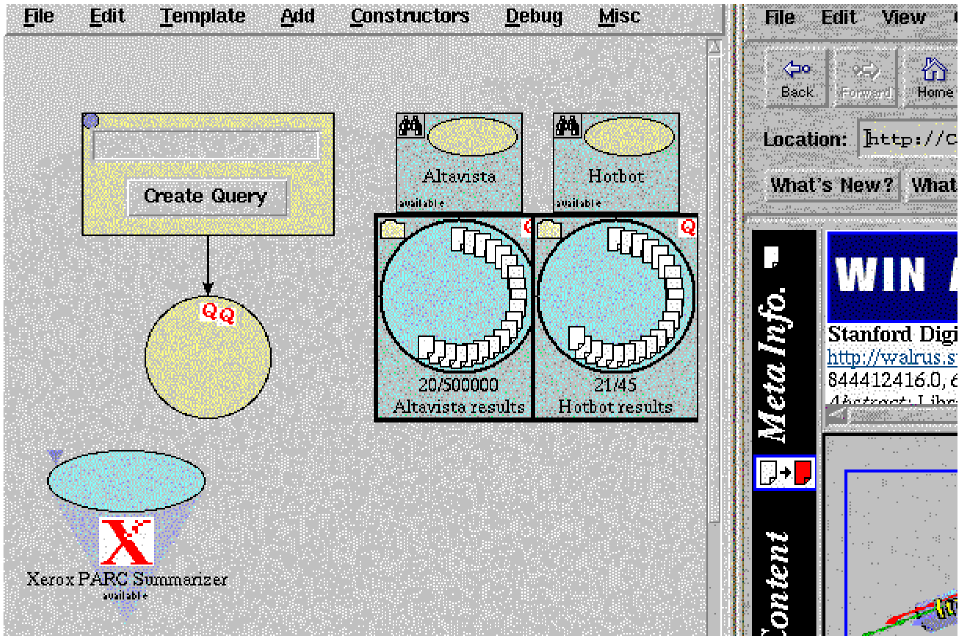


Figure 25. DLITE user interface.

3.1.6 University of Illinois in Urban-Champaign DeLiver Digital Library Project

The UIUC Digital library research effort was centered on building an experimental testbed containing tens of thousands of full-text journal articles from physics, engineering, and computer science and make them accessible over the WWW, often before they are available in print. The UIUC DLI Testbed, DeLiver, was emphasized on using the document structure to provide federated search across publisher collections. The sociology research included the evaluation of its effectiveness under use by over one thousand UIUC faculty and students, a user community an order of magnitude bigger than the last generation of research projects centered on search of scientific literature [22]. The technology research investigated indexing the contents of text documents to enable federated search across multiple sources, and testing this on millions of documents for semantic federation.

The structures of documents in the testbed are specified by Standard Generalized Markup Language (SGML). Their research efforts extract semantics from documents using the scalable technology of concept spaces based on context frequency. Then these efforts were merged with traditional library indexing to provide a single Internet interface to indexes of multiple repositories.

They developed a **Distributed Repository Model**, which is shown in Figure 26.

The UIUC Testbed (DeLIver) provides enhanced access over the Internet to the full text of selected engineering journals, using SGML document structure to facilitate search. Access to these materials is currently limited to UIUC faculty, students, and staff. The Testbed collection gathers articles directly from publishers in SGML format. These articles include texts and all figures, tables, images, and mathematical equations. The testbed collection presently comprises around 40,000 articles from journals in electrical engineering, physics and civil engineering.

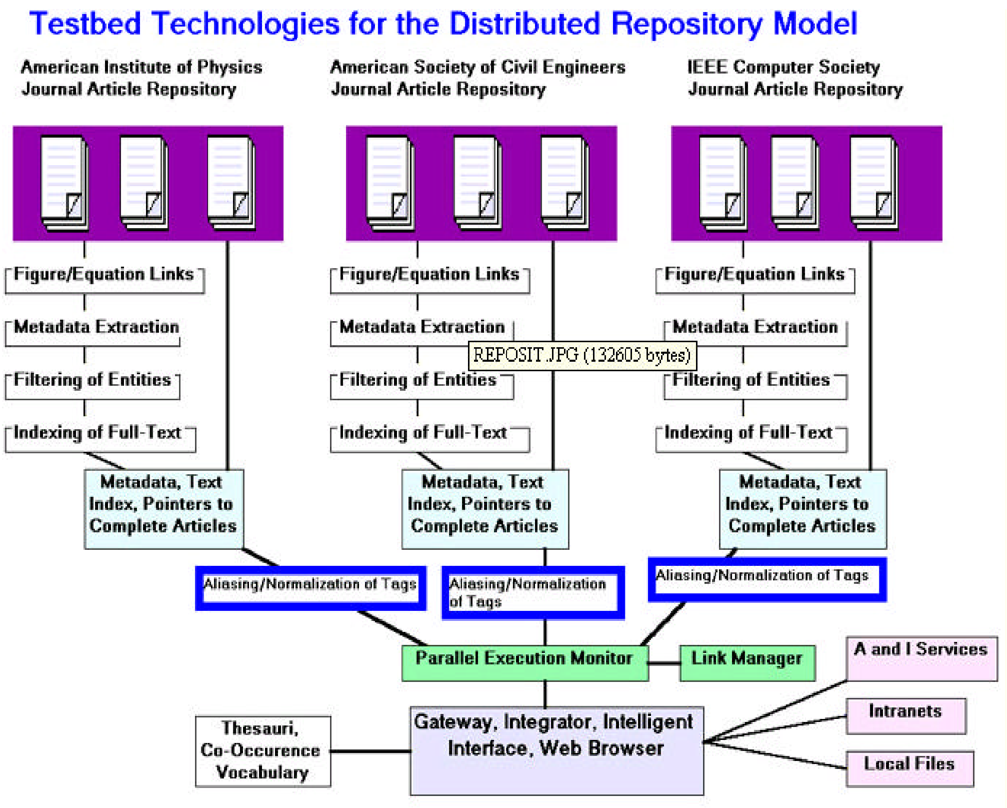

Testbed Technologies for the Distributed Repository Model
American Institute of Physics Journal Article Repository
American Society of Civil Engineers Journal Article Repository
IEEE Computer Society Journal Article Repository
Figure/Equation Links
Metadata Extraction
Filtering of Entities
Indexing of Full-Text
Metadata, Text Index, Pointers to Complete Articles
REPOSIT.JPG (132605 bytes)
Aliasing/Normalization of Tags
Parallel Execution Monitor
Link Manager
A and I Services
Intranets
Local Files
Thesauri, Co-Occurence Vocabulary
Gateway, Integrator, Intelligent Interface, Web Browser

Figure 26. Testbed distributed repository model.

3.2 DIGITAL LIBRARIES INITIATIVE PROJECT -- PHASE II (DLI II)

The Digital Libraries Initiative - Phase 2 is an interagency program sponsored by the:

- National Science Foundation (NSF)
- Defense Advanced Research Projects Agency (DARPA)
- National Library of Medicine (NLM)
- Library of Congress (LOC)
- National Endowment for the Humanities (NEH)
- National Aeronautics & Space Administration (NASA)
- Federal Bureau of Investigation (FBI)

in partnership with:

- Institute of Museum and Library Services (IMLS)
- Smithsonian Institution (SI)
- National Archives and Records Administration (NARA)

The primary purposes of this initiative are to provide leadership in research fundamental to the development of the next generation of digital libraries, to advance the use and usability of globally distributed, networked information resources, and to encourage existing and new communities to focus on innovative applications areas. Since digital libraries can serve as intellectual infrastructure, this Initiative looks to stimulate partnering arrangements necessary to create next-generation operational systems in such areas as education, engineering and design, earth and space sciences, biosciences, geography, economics, and the arts and humanities. It will address the digital libraries life cycle from information creation, access and use, to archiving and preservation. Research to gain a better understanding of the long term social, behavioral and economic implications of and effects of new digital libraries capabilities in such areas of human activity as research, education, commerce, defense, health services and recreation is an important part of this initiative.

The special interests in this initiative are:

Research in the following areas:

- ***Human-Centered Research***

Human-centered digital libraries research seeks to further understanding of the impacts and potential of digital libraries to enhance human activities in creating, seeking, and using information and to promote technical research designed to achieve these goals.

- ***Content And Collections-Based Research***

Content and collection-centered digital library research focuses on better understanding of and advancing access to novel digital content and collections.

- ***Systems-Centered Research***

Systems-centered digital libraries research focuses on component technologies and integration to realize information environments that are dynamic and flexible; responsive at the level of individual, group, and institution; and capable of adapting large, amorphous, continually growing bodies of data to user-defined structure and scale.

Testbeds and Applications

This focuses on development of digital library testbeds for technology testing, demonstration and validation, and as prototype resources for domain communities - technical and non-technical. Applications projects are expected to result in enduring information environments for research, learning, and advancing public use in creative ways.

Planning Testbeds and Applications for Undergraduate Education

Projects funded to date under DLI2 include the ones listed below. As more projects are funded, they will be added to the list below:

- **A Patient Care Digital Library**: Personalized Search and Summarization over Multimedia Information, Columbia University

- **Informedia-II**: Integrated Video Information Extraction and Synthesis for Adaptive Presentation and Summarization from Distributed Libraries, Carnegie Mellon University

- **The Alexandria Digital Earth Prototype (ADEPT)**, University of California at Santa Barbara

- **Stanford Digital Libraries Technologies**, University of California at Berkeley, the University of California at Santa Barbara, and Stanford University

- **Re-inventing Scholarly Information Dissemination and Use**, University of California at Berkeley, the University of California at Santa Barbara, and Stanford University

- **An Operational Social Science Digital Data Library**, Harvard University

- **Security and Reliability in Component-based Digital Libraries**, Cornell University

- **Founding a National Gallery of the Spoken Word**, Michigan State University

- **A Digital Library for the Humanities**, Tufts University

- **A Software and Data Library for Experiments, Simulations and Archiving**, University of South Carolina

- **Digital Workflow Management**: Lester S. Levy Collection of Sheet Music, Johns Hopkins University

- **A Multi-tiered Extensible Digital Archive of Folk Literature**, University of California at Davis

- **The Digital Athenaeum**: New techniques for restoring, searching, and editing humanities collections University of Kentucky

- **Data Provenance, University of Pennsylvania DL of Vertebrate Morphology using a new High Resolution X-ray CT Scanning facility**, University of Texas at Austin

- **Using the Informedia Digital Video Library to Author Multimedia Material**, Carnegie Mellon University
- **High-Performance Digital Library Classification Systems**: **From Information Retrieval to Knowledge Management**, University of Arizona

- **A Distributed Information Filtering System for Digital Libraries**, Indiana University Bloomington

- **Automatic Reference Librarians for the World Wide Web**, University of Washington

- **Tracking Footprints through a Medical Information Space**: **Computer Scientist-Physician Collaborative Study of Document Selection by Expert Problem Solvers,** Oregon Health Sciences University and Oregon Graduate Institute of Science and Technology

- **Image Filtering for Secure Distribution of Medical Information**, Stanford University

- **Using the National Engineering Education Delivery System as the Foundation for Building a Test-Bed Digital Library for Science, Mathematics, Engineering and Technology Education**, University of California, Berkeley

- **Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science**, Old Dominion University

- **Virtual Skeletons in 3 Dimensions: The Digital Library as a Platform for Studying Web-Anatomical Form and Function**, University of Texas at Austin

4. CONCLUSIONS

In this chapter we examined the design and implementation of digital libraries. There is no single definition for digital libraries and the definition evolves as the research goes on. The common consensus is that they provide their users with a coherent view of heterogeneous autonomously managed resources. There are a lot of research issues waiting for resolution. These issues are classified as five major kinds, namely interoperability, description of objects and repositories, collection management and organization and user interface and human-computer interaction and economic, social and legal issues.

A commonly accepted architecture of digital library is based on digital objects and handle system and common repository access interface (RAP). Handle is a general-purpose unique identifier for Internet resources, including digital objects. Handle system is a distributed system that manages handles. Access and deposit of digital objects is conducted according to Repository Access Protocol (RAP).

When designing a digital video library system, we have to consider special issues related to characteristics of video such as video compression, video indexing, video segmentation and video retrieval.

Digital Library Initiative is one of the earliest efforts in digital library research in digital library area. It consists of two phases. DLI I just ended last year. It focused on the basic issues of digital library, particularly efficient searching technical documents on the Internet. Each participant was concentrated on one specific research areas, created its own testbed and tested the ideas on the testbed. Based on Phase I, Phase II will be a broader effort and will emphasize research and practices on human-centered system. So far there has been 24 funded projects going on.

We have made some achievement, especially in areas such as description of objects and repositories, user interface and interoperability. But digital libraries are much complicated systems. It is basically international. It is not a topic only existing in computer and information science. It is involved in many communities, including social, legal and political communities. Joint efforts are necessary for solutions to safeguarding digital contents and users and providing users convenient services at the same time. There is still a long way for it to achieve maturity and become commercial products.

References

1. B. Schatz and H. Chen, "Digital Libraries: Technological Advances and Social Impacts," *Computer*, Vol. 32, February 1999.
2. A. Paepcke, “Digital Libraries: Searching Is Not Enough – What We Learned On-Site,” *D-Lib Magazine*, Vol. 2, No. 2, May 1996.
3. C. Lynch and H. Garcia-Molina, “Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the May 18-19, 1995,” *IITA Digital Libraries Workshop*, August 1995.
4. NSF Announcement, “Digital Libraries Initiative – Phase 2,” Announcement Number NSF 98-63,1998.
5. B. M. Leiner, “From the Editor: Metrics and the Digital Library,” *D-Lib Magazine,* Vol. 4, No. 7/8, July/August 1998.
6. S. M. Griffin, “NSF/DARPA/NASA Digital Libraries Initiative: A Program Manager’s Perspective,” *D-Lib Magazine*, Vol. 4, No. 7/8, July/August 1998.

7. W.Y. Arms, E. A. Overly, M. Restoj, and C. Blanchi, “An Architecture for Information in Digital Libraries,” *D-Lib Magazine*, Vol. 3, No. 2, February 1997.
8. R. Kahn and R.Wilensky, “A Framework for Distributed Digital Object Services,” *D-Lib Magazine*, Vol. 1, No. 5, May 1995.
9. W.P. Birmingham, “An Agent-Based Architecture for Digital Libraries”, *D-Lib Magazine*, Vol. 1, No.7, July 1995.
10. B. M. Leiner, “The NCSTRL Approach to Open Architecture for the Confederated Digital Library,” *D-Lib Magazine*, Vol. 4, No. 12, December 1998.
11. M. Christel, S. Stevens, T. Kanade, M. Mauldin, R. Reddy, and H. Wactlar, “Techniques for the Creation and Exploration of Digital Video Libraries”, Chapter in the book Multimedia Tools and Applications, Ed. B. Furht, Kluwer Academic Publishers, Norwell, MA, 1996.
12. B. Scheatz and H. Chen, “Building Large-Scale Digital Libraries,” *Computer*, Vol.29, May 1996.
13. V. Ogle and R. Wilensky, “Testbed Development for the Berkley Digital Library Project,” *D-Lib Magazine*, Vol. 2, No. 8, August 1996.
14. Alexandria Digital Library User Interface Tutorial.
15. J. Frew, M. Freeston, R. B. Kemp, et al., “The Alexandria Digital Library Testbed,” *D-Lib Magazine*, Vol. 2, No. 8, August 1996.
16. C. Lichti, C. Falousos, H. Wactlar, M. Christel, and A. Hauptmann, “Informedia: Lessons from a Terabyte+, Operational, Digital Video Database System,” *Proc. of Very Large Database Conference*, New York, August 1998.
17. Scott Stevens, “Carnegie Mellon University: The Informedia Digital Video and Spoken Language Document Testbed,” *D-Lib Magazine*, Vol. 5, No. 2, February 1996.
18. V. Ogle and R. Wilensky, “Testbed Development for the Berkeley Digital Library Project,” *D-Lib Magazine*, Vol. 2, No. 7/8, 1996.
19. D. E. Atkins, W. P. Birmingham, E. H. Durfee, E. Glover, T. Mullen, E. A. Rundenstteiner, E. Soloway, J. M. Vidal, R. Wallace, and M. P. Wellman, “Building the University of Michigan Digital Library: Interacting Software Agents in Support of Inquiry-Based Education,” http://ai.eecs.umich.edu/people/wellman/pubs/Building-UMDL.html, 1999.
20. D. E. Atkins, W. P. Birmingham, E. H. Durfee, E. J. Glover, T. Mullen, E.A. Rundensteiner, E. Soloway, J. M. Vidal, R. Wallace, and M. P. Wellman, “Toward Inquiry-Based Education Through Interacting Software Agents,” *Computer*, Vol. 29, May 1996.
21. A. Paepcke, S. B. Cousins, H. Garcia-Molina, S. W. Hassan, S. P. Ketchpel, M. Roscheisen, and T. Winograd, “Using Distributed Objects for Digital Library Interoperability,” *Computer*, Vol. 29, May 1996.
22. http://dli.grainger.uiuc.edu.