A Digital Library Prototype for Access to Diverse Collections *

다양한 장서 접근을 위한 디지털 도서관의 프로토타입 구축

Won-Tae Choi **

Contents

ABSTRACT

This article is an overview of the digital library project, indicating what roles Koreas diverse digital collections may play. Our digital library prototype has simple architecture, consisting of digital repositories, filters, indexing and searching, and clients. Digital repositories include various types of materials and databases. The role of filters is to recognize a format of a document collection and mark the structural components of each of its documents. We are using a database management system (ORACLE and ConText) supporting user-defined functions and access methods that allows us to easily incorporate new object analysis, structuring, and indexing technology into a repository. Clients can be considered browsers or viewers designed for different document data types, such as image, audio, video, SGML, PDF, and KORMARC. The combination of navigational tools supports a variety of approaches to identifying collections and browsing or searching for individual items. The search interface was implemented using HTML forms and the World Wide Web's CGI mechanism.

초 록

본 논문은 다양한 유형으로 구성되어 있는 디지털 도서관의 장서가 어떠한 역할을 수행하는지를 나타내는 디지털 도서관의 구축에 관한 것이다. 본 연구에서 구축된 디지털도서관의 프로토타입은 디지털 리포지토리, 필터, 색인 및 검색, 클라이언트의 구조로 되어 있다. 디지털 리포지토리는 여러 가지 유형의 문서유형과 다양한 형태의 데이터베이스로 구성된다. 필터는 다양한 문헌의 포맷을 인식하고 문헌 각각의 조직적인 요소를 지능적으로 구분하는 역할을 수행한다.

본 시스템은 관계형 데이터베이스 관리 시스템인 ORACLE과 ConText를 이용하여 구성되었으며 새로운 객체의 분석 및 조직화, 색인기술의 적용을 용이하게 처리할 수 있다. 클라이언트는 여러 유형의 데이터 포맷(이미지, 오디오, 비디오, SGML, PDF, KORMARC 등)의 디스플레이를 위한 브라우저, 뷰어이다. 이용자는 이러한 도구들을 이용하여 문헌을 구분하고 각각의 아이템을 브라우징하고 탐색할 수 있다. 본 연구의 탐색 인터페이스는 HTML과 WWW의 CGI를 이용하여 구현되었다.

* The work described here is a joint effort of Byung Chul Lee and Yun Ho Kim and was funded in part by JoongWon Research Institute, KonKuk University.

** Professor, Dept. of Library and Information Science, Konkuk University.

접수일자 1998년 5월 20일

1. Introduction

Digital libraries basically store materials in electronic format and can manipulate large collections of those materials both effectively and efficiently. The key technological issues are how to store, index, search and display desired selections from and across large collections. While practical digital libraries must focus on issues of access costs and digitization technology, digital library research concentrates on how to develop the necessary infrastructure to effectively mass-manipulate the information on the network (Schatz and Chen 1996). The digital library projects use many contrasting approaches (Arms 1996, Fox et al. 1996, Smith 1996, Wactlar et al. 1996).

A traditional library is a single repository for materials from many sources to which a user comes seeking information. A repository is an organized collection in which documents and other objects are indexed for effective search. The library assembles its digital repositories by collections, each collection being a grouping of digital reproductions that form an integral whole.

This approach is consistent with archival practice, which provides context for individual documents by arranging them within a collection of related materials. The level of description for a collection depends on institutional priorities and resources available at a point in time. Digital repositories must be built with recognition that the level and structure of description will vary greatly.

2. Architecture

2.1 Overview

Our digital library prototype separates the access tools (index, dictionary, and thesaurus) from the digital repository that contains the resources themselves. This allows resources to be pointed directly from specialized indexes and thesaurus. Integrating heterogeneous collections, databases, and a commercial automated library system are the important features of our system.

Our system has a simple architecture, consisting of digital repositories, filters, indexing and searching, and clients. Figure 1 illustrates the basic architecture.

The digital repository includes various types of materials and databases. We are using a database management system (ORACLE DBMS and ConText) that supports user-defined functions and access methods. It allows us to easily incorporate new object

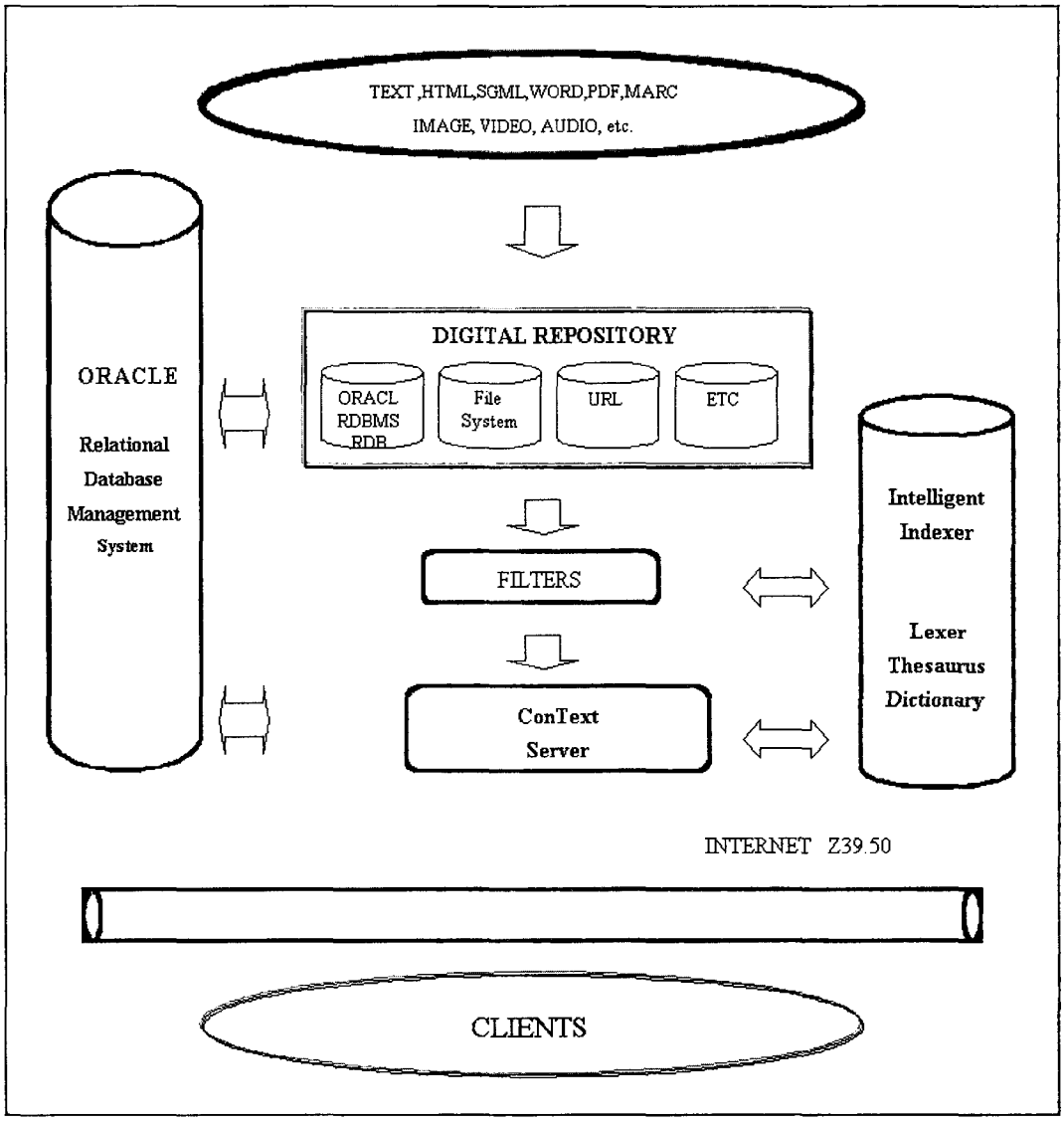

Figure 1. Basic architecture of digital library prototype

analysis, structuring, and indexing technology into a repository.

The search interface for our system is implemented using Hypertext Markup Language (HTML) forms and the World Wide Web's Common Gateway Interface (CGI) mechanism. The initial search interface for our digital repositories was made as simple as possible. Experienced searchers have requested the ability to formulate more precise searches. More complex query forms involving Boolean

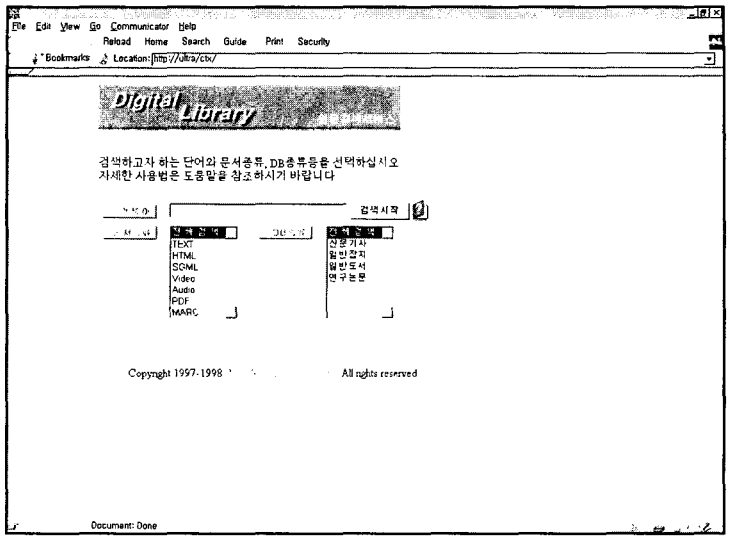

Figure 2. Search interface prototype

logic and searching for terms in specific fields are under test in other projects.

Figure 2 shows the search interface prototype. Users can use Boolean connectors to specify a phrase with different amounts of proximity or specify multiple phrases. Also Users can select the document type (audio, video, SGML etc.) and database (book, dissertations, journal etc.).

Digital repositories communicate with clients via several protocols, most notably the widely used HyperText Transfer Protocol (HTTP). The interfaces to external search engines, such as the online catalog, follow the Z39.50 protocol. Z39.50 is a standard whose purpose is to allow one computer operating in client mode to perform information retrieval queries against another computer acting as an information server.

2. 2 Digital repositories

To develop the appropriate system, we are creating a prototype set of information services called the KonKuk University Digital Library. This prototype repository includes various types of materials.

- Pictorial materials
- Audio materials
- Video materials
- Text materials reproduced as images
- Text materials reproduced as search-able text and images
- Korea Machine Readable Cataloging (KORMARC) and other MARC Files

The digital repository has three archi-

tecture, consisting of ORACLE RDBMS RDB, file system, URL. Text materials are stored ORACLE RDBMS RDB. For effective search, Multimedia materials are stored file system and URL information is stored URL system. The Pointer structure links text materials and other multimedia information each other.

The digital collections are being assembled collection by collection; these collections have different characteristics. For digital reproductions of original items, the greatest stability and public accessibility is obtained for images that reproduce manuscript documents, printed matter, and pictorial materials, and for searchable texts, including those that employ a Standard Generalized Markup Language (SGML) and Portable Document Format (PDF). Word files can be reproduced as a PDF format easily. Most of the current collections have a set of item-level bibliographic records, which can be indexed and searched.

For pictorial collections, the digital repository produces three image types: Graphics Interchange Format (GIF), Joint Photographic Experts' Group (JPG or JPEG), and Tagged Image File Format (TIFF). Sound recordings (audio) and videos in collections are offered in RealAudio and RealVideo format. The large files required producing audio and video formats, thereby launching the library is on a constant search for new and better compression and playback schemes. The files produced today will become obsolete more quickly than before.

For text materials reproduced as images we are using TIFF format. Images in the TIFF format are of higher resolution and large file size than GIF or JPEG format.

Some of the books, dissertations, and papers in the digital repositories can be coded in SGML with embedded links to images of the original pages, illustrations, and tables (ISO 8879 1986). The SGML representation captures features of documents that provide great potential for both convenient presentation and effective searching. SGMLs strength, in terms of retrieval, is that it reveals such deep document structure.

Some of the books, papers, and pamphlets in the digital repositories can be coded in Adobe's PDF format. PDF files are fully searchable and retain all the fonts, colors, and formatting of the original paper documents. PDF files are compact, cross platform and can be viewed by anyone with a free Acrobat Reader.

Some of the books and dissertations can be coded in KORMARC format. KORMARC stores a given holdings metadata in one record with four com-

ponents- a leader, a record directory, control fields, and variable fields. This structure, while not optimal for a relational database, is useful for specifying metadata I/O functions and for exchanging metadata records between digital libraries. The KORMARC standard contains fields, as well as a thesaurus-based field that permits references to specific thesauri, which are used to find terms that can be used to conduct data searches.

Integrating heterogeneous digital collections (audio, video, SGML, PDF, and KORMARC formats) is vital to our system. All information is differently stored in RDBMS, file, and URL file. Any numbers of collections, or information servers, are possible. Each is implemented as databases that supports user-defined functions and access methods.

2. 3 Filters

Collections consist of a single general type of material, but our collections (digital repository) in process incorporate several document types. It is easy to pour all this text into an indexing engine. It is more difficult to structure search options and present results in a way that is both efficient and comprehensible to users.

A primary role of the filter is to recognize a format of a document collection and mark the structural components of each of its documents. The filter processes a document and produces ASCII output. If the document is in a proprietary format, the program must recognize the format tags for the document and be able to convert the formatted text into ASCII text. The filter can be used for filtering documents in a variety of formats and act as their own indexing servers. We implemented the filters using C language. The output of filter processing can be indexed or processed through the ConTexts Linguistic Services.

In Window 95 and 98 environments, the filter extracts keyword, page and image information from heterogeneous documents (MS WORD, Excel, Power -Point, PDF, and CAD etc). We developed this filter using Window printer drive architecture. The function of filter extracts keyword, page, and font information from text. For image materials the filter extract keyword, page, and font information from image title. Also filter can extract the location and page information of image, audio, and video materials.

For example after adding an SGML document to the digital repository, we must index it for efficient retrieval. We developed procedures for generating

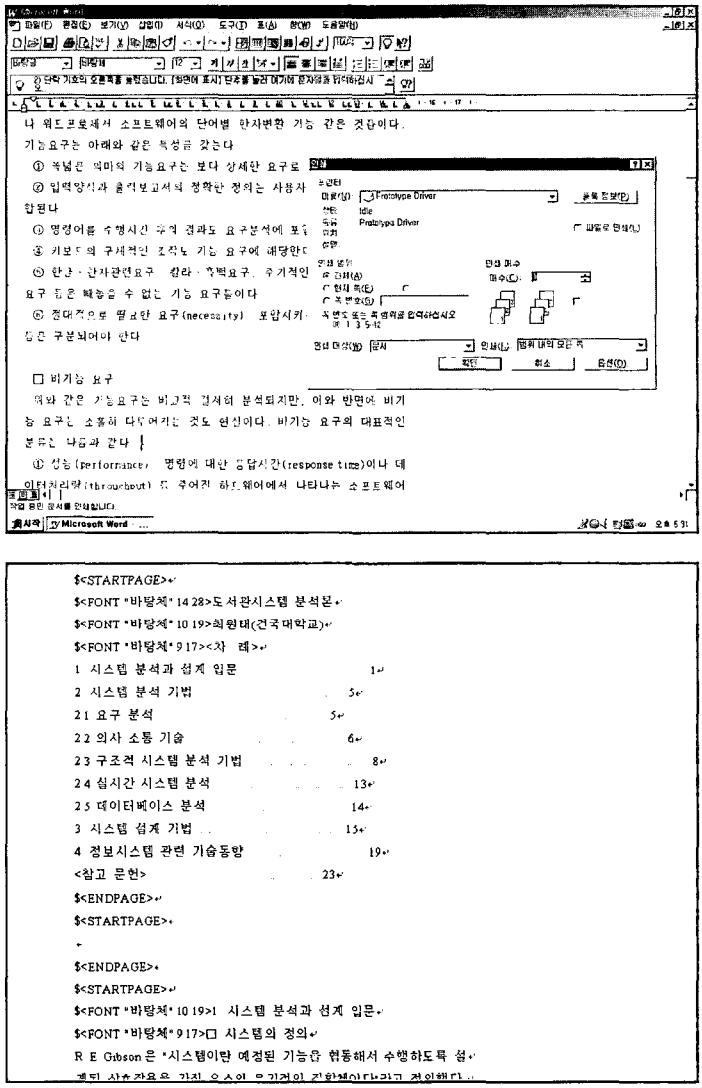

Figure 3. The example of filter driver and processing

collections of SGML materials. We process the heterogeneous SGML received from publishers. Tags differ from one publisher to another. We can federate some differences with simple syntactic transformations, such as HEAD or TTL or TITLE for the title tag. The filter also includes the HTML and KORMARC documents.

Our efforts extract semantics from documents using the scalable technology of concept spaces based on context. We then merge these efforts with indexing methodology to provide a single interface to

indexes of multiple collections.

2. 4 Indexing and searching

Indexing was originally developed for text documents and has long history. Each document is segmented into significant words, and generated tables that indicate which words occurred where in what documents. A user can search by specifying words; the system then supplies the results by looking up the word in the tables, and retrieving the documents containing it.

Indexing and searching for our system currently uses the filter and the free text search engine, ConText and ORACLE DBMS. The text indexing and list of hits (search results) is subdivided in the same way as for bibliographic searches. The use of a general indexing engine provides the simultaneous retrieval from indexes generated from textual material in different formats.

ConText supports both plain text and formatted text (i.e. Microsoft Word, WordPerfect, HTML etc.). In addition, ConText supports text that contains HTML tags. Regardless of the format, ConText requires the text to be filtered for the purposes of indexing texts or processing texts through the Linguistic Services, as well as highlighting the text for viewing.

ConText supports various types of functions.

- Exact word/phrase searching
- Logical combinations(and, or, not)
- Wild-card searching
- Expansions(linguistic stem, fuzzy match)
- Synonyms/thesauri(equivalent, broader and narrower terms)
- Proximity searching(words near each other)
- Weight terms
- Limit results(by score, by number of hits)

The search engine in use is ORACLEs SQL (Structured Query Language). ORACLE' s flexible query commands have been used to list the hits for queries. Figure 4 shows the search result interface of text data. The title displayed for each hit acts as a link to a full text materials display, which has a link to the item. Some items have the subject term. Each subject term is also a link that invokes a search for items on the same topic. Names of authors and other creators can link to searches for other works by the same person or organization.

Finding the right level of specificity for search terms is also a problem for users. One potential approach is to integrate thesaurus into the interface, by providing a

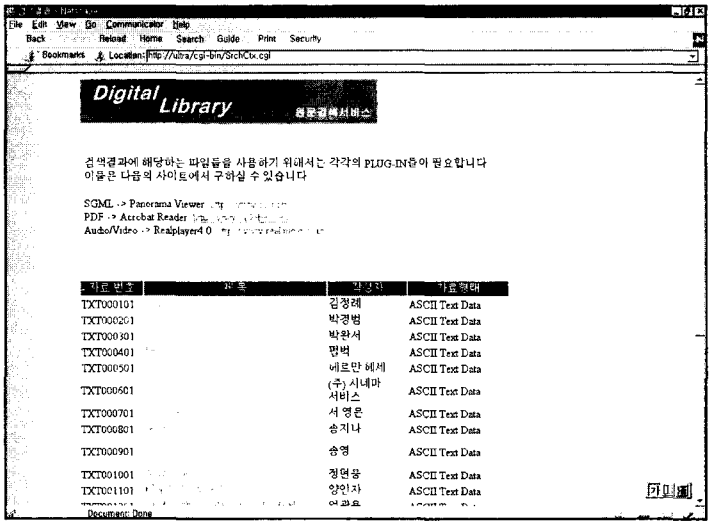

Figure 4. The display of search result list

browsable hierarchy of terms to assist selection, or by mapping commonly used terms into formal descriptor or to synonyms actually found in text documents.

2. 5 Clients

We are using several interoperable clients. These can be considered browser or viewer designed for different document data types, such as image, audio, video, SGML, PDF, and KORMARC. Clients support WWW browser plug-in architecture. To view special formats, users download free viewers and need to add viewers to web browser.

The combination of navigational tools supports a variety of approaches to identifying collections and browsing or searching for individual items. The majority of collections can be seen and read without special viewers. The web browser automatically displays the texts in these collections, which have been transcribed from original documents. The web browser allows you to search and browse all texts presented in HTML.

Users need no special viewers to view most photographs in collections. The web browser automatically shows images presented in GIF and JPG or JPEG, the formats in which most photographs are presented.

However, in some collections the repository provides sound recordings,

films, additional high-resolution images, and text with enhanced navigation. Sound recordings in collections are offered in RealAudio formats. Also Videos in collections are offered in RealVideo format. To view audio and video materials, both formats require special players (RealNetworks RealPlayer).

Some collections include pictures of pages and illustrations from original documents. To see pictures of original source document pages or illustrations, you will need a special viewer. To view TIFF images, users need to add a TIFF viewer to web browser.

The search interface for KORMARC was implemented using HTML forms and using the WWW's CGI. Using this interface it is possible to display the loan data linked with Automated Library System.

Figure 5 shows a portion of a KORMARC document as displayed in this viewer.

Some collections offer the added feature of a hot-linked table of contents and enhanced search and browse features. These collections, displayed in SGML, allow you to jump around within long selections and within complete books online.

In Korea there is currently no free or inexpensive SGML viewer that is satisfactory for general use. As an increasing array of documents is available in SGML format, we hope that a more powerful SGML viewer will become widely available or that WWW browsers will be extended to handle SGML. Figure 6 shows a portion of a SGML document as displayed in this viewer.

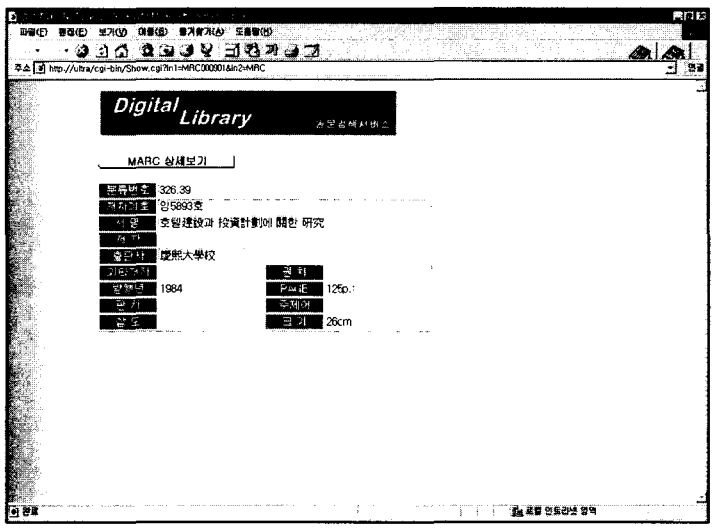

Figure 5. The example of KORMARC format display

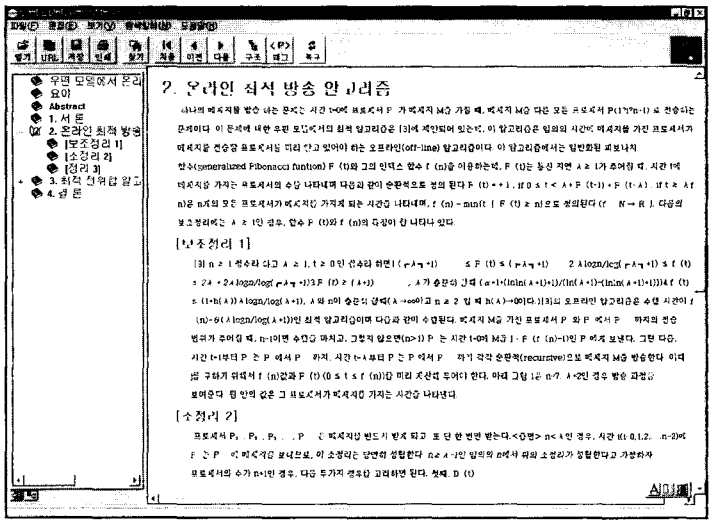

Figure 6. The example of SGML document display

A PDF viewer, for example, simplifies information requests about a multimedia document. If the document contains a PDF, viewing the PDF will activate the PDF viewer on it. Figure 7 shows a portion of a PDF document as displayed in this viewer.

Today, users of MS Windows can download a free viewer (Panorama Free, Adobe Acrobat Reader). This free viewer can present a structured table of contents derived from chapter and section headings. Headings act as direct links to corresponding sections of a document. However, to support the display of Korean documents with double-byte fonts there should be special processing. The internal search feature of Korean SGML viewer does not take advantage of the structural information encoded in the SGML format.

To provide access to a much greater range of the public, including users of other operating systems and proprietary browsers from an online service, the library has converted the documents into HTML. Also, it is clear that as the digital collection grows, new approaches will be needed. In the future it will be possible to view and retrieve the document of diverse collections in a few minutes.

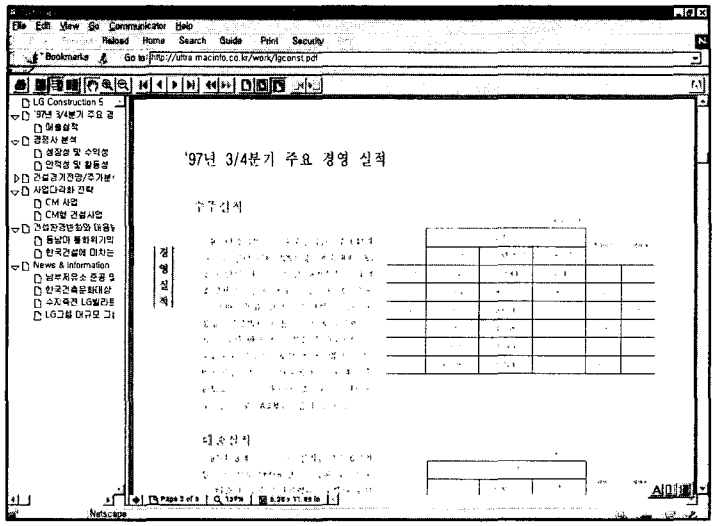

Figure 7. The example of PDF document display

3. Conclusion and future work

The library's efforts are based on a flexible, distributed, and modular architecture built using open standards, with an emphasis on the widest possible access. Integrating heterogeneous digital collections is vital to our system. To date, we have established a digital library prototype to validate diverse collections. Our indexing and searching techniques utilize the fine structure of the documents, so that user can search the documents for full-text retrieval using an ORACLE and ConText.

In this work, we have faced some problems of managing context and structure, multimedia indexing, and multilingual processing. We are proceeding on parallel paths: working to provide access to resources today, using today's technology in Korea and participating in cooperative longer-term efforts to develop a distributed architecture for digital libraries.

Another dimension of the challenge in building access tools is that a single interface is unlikely to satisfy the entire range of users. In the future, browsable finding aid documents will probably be used as the primary access aid for many collections where item-level cataloging is not feasible. We hope to not only develop a large and valuable digital library to support education and research, but also to show that it has proved to be of benefit, and that users indeed know how to use digital libraries.

References

Arms, Caroline R. 1996. Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress. (http://lcweb2.loc.gov/ammem/ammemhome.html)

Fox, Edward A. et al. 1996. A Scalable and Sustainable Approach to Unlock University Resources, D-Lib Magazine. (http://www.dlib.org/dlib/september96/theses/09fox.html)

ISO 8879-1986. Information Processing Text and Office Systems-Standard Generalized Markup Language (SGML), International Organization for Standardization.

Schatz, Bruce and Chen, Hsinchun. 1996. " Building Large-Scale Digital Libraries." IEEE Computer 29(5) : 22-26.

Smith, Terence R. 1996. "A Digital Library for Geographically Referenced Materials." IEEE Computer 29(5) : 54-60.

Wactlar, Howard D. et al. 1996. "Intel-ligent Access to Digital Video: Informedia Project." IEEE Computer 29(5) : 46-52.