Chapter 15: Digital Libraries
Section 15.4: Document Models, Representations, and Access
Section 15.5: Prototypes, Projects, and Interfaces
Section 15.6: Standards
Section 15.7: Trends and Research Issues

December 20, 1999
Summary

- Introduction
- Document Models, Representations, and Access
- Prototypes, Projects, and Interfaces
- Standards
- Trends and Research Issues
- Conclusion
Introduction

Digital library (DL) can be defined by:
- the combination of a collection of digital objects (repository);
- descriptions of those objects (metadata);
- a set of users (patrons or target audience or users);
- systems that offer a variety of services such as capture, indexing, cataloging, search, browsing, retrieval, delivery, archiving, and preservation.
Information retrieval (IR) is essential for the success of DLs, so they can achieve high levels of effectiveness while at the same time affording ease of use to a diverse community.

A significant portion of the research and development efforts related to DLs has been in the IR area.

This presentation reviews some of these efforts, organizes them into a simple framework, and highlights needs for the future.
Without documents there would be no IR or DLs. Hence, it is appropriate to consider definitions of ‘document’, and to develop suitable formalizations, as well as to articulate research concerns.
Because DLs can be constructed for a particular institution or nation, it is likely that the expansion of DLs will increase access to documents in a variety of languages.

There are issues of character encoding:
- Unicode provides a single 16-bit encoding scheme suitable for all natural languages;
- Downloading fonts from a special server or gateway is a less costly implementation.
The next crucial problem is searching multilingual collections:

- The simplest approach is to locate words or phrases in dictionaries and to use the translated terms to search in collections in other languages.

- It is likely that research in this area will continue to be of great importance to both the IR and DL communities.
Multimedia documents’ streams usually must be synchronized in some way, and so it is promising that a new standard for handling this over the Web has been adopted.

IR has been applied to various types of multimedia content:
- Columbia University: a large image collection from the Web can be searched on content using visual queries;
- IBM: *Query By Image Content (QBIC)* system for images and video was developed.
Better handling of multimedia is at the heart of future research on many types of documents in DLs.

Very powerful representation, description, query and retrieval systems may be required to properly handle the complexity of multimedia collections.
Structured documents are streams with one or more structures imposed.

Metadata can be represented as an SGML document and SGML content can be included in the base document and/or be kept separately.

Structure is often important:
- in documents, when one wants to add value or make texts ‘smart’ (SGML);
- in retrieval (OpenText);
- at the level above documents, which makes searching necessary and possible.
Most DLs are spread across computers, that is spanning physical and/or logical spaces.

Dealing with collections of information that are distributed in nature is one of the common requirements of DL technology.
There are two approaches to tackle this problem:

- The first one is to develop a description language for each DL and to build federated search systems that can interpret it;
- The second one is to make each DL support a powerful protocol aimed at effective retrieval.

The first course is supported by BioKleili system and the second one by Computer Interchange of Museum Information (CIMI).
Figure 15.2 Architecture of the BioKleisli system (adapted from [829, 128]).
Federated search is the support for finding items that are scattered among a distributed collection of information sources or services, typically involving sending queries to a number of servers and then merging the results to present in an integrated, consistent, coordinated format.
A variety of approaches has been adopted:

- Collecting the required information, often through Web crawling of various sorts;
- Focusing on intelling search;
- Fusing of results;
- Segmenting the collection and/or its indexes so that most searches only look at a small number of the most useful sources of information.
Document Models, Representations, and Access

- DLs must manage intellectual properties. These services must support agreed-upon principles, copyright practices, contracts and other agreements and laws.

- A key to the implementation of policies for access management is having trusted systems.

- Stronger mechanisms are crucial in DLs to:
  - protect intellectual property rights;
  - control the types of access afforded to different user groups.
Prototypes, Projects and Interfaces
- Internacional Range of Efforts -

- Since each nation wishes to share highlights of its history, culture, and accomplishments with the rest of the world, developing a DL can be very helpful.

- Examples:
  - European ERCIM program and UK initiatives;
  - New Zealand, Australia and Singapore;
  - IBM Digital Library;
  - Networked Digital Library of Theses and Dissertations (NDLTD).
Prototypes, Projects, and Interfaces - Usability -

- Key to the success of the DLs is having usable systems. This is a serious challenge!
- Simple library catalog systems were observe in 1986 to be difficult to use.
- A 1997 study at Virginia Tech of four digital library: ACM; NCSTRL; NDLTD; IEEE-CS;
- The participants were 48 Virginia Tech students. 38 graduate students, 8 undergraduate, 2 other graduate studies;
The study concluded that many systems have serious usability problems:

- Pre-test found that very few users had worked with a DL;
- Post-test showed that user expectations and priorities changed over the short test period;

Features derived from user feedback and evaluators observations: clear overview; search criteria for simple search; search criteria for advanced search; fast searching and retrieval; examples searches; ability to download a fraction of the article; save queries for future refinement.
Standards

- Since there are many DL projects worldwide, involving diverse research, development, and commercial approaches, it is imperative that standards are employed so as to make interoperability and data exchange possible.
- At the heart of supporting federated DLs is agreement on protocols for computer-computer communication.
Standards - Protocols and Federation -

- The standard Z39.50 was designed to search remote bibliographic collections;
- One example of widespread utilization was the WAIS system (based on Z39.50), very popular before WWW emerged;
- The application of Z39.50 to DLs was demonstrated in the CIMI project;
- Dienst is another standard which has been considered in regard to NDLTD.
In the broadest sense, metadata can describe not only documents but also collections and whole DLs along with their services.

MARC has been widely used, although there are some concerns with it:

- It usually involves working with binary records which must be converted for interchange;
- There are a number of national versions with slight differences, as well as differences in cataloging practices that yield the MARC records (USMARC).
Dublin Core (DC) includes 15 core elements that can be used to describe any digital object:

- content (Title, Subject, Description, Source, Language, Relation and Coverage);
- intellectual property issues (Creator, Publisher, Contributor and Rights);
- digital objects (Data Type, Format and Identifier).
Warwick Framework deals with packages and connections between packages.

Resource Description Framework (RDF) is essentially a scheme for annotating digital objects, so alternatively the descriptions can be stored separately from those objects.

Text Encoding Initiative (TEI) combines data with metadata.
Trends and Issues

- There are many remaining challenges in the DL field.
- The IR community must provide guidance regarding automatic indexing of marked up documents, metadata, full-text, multimedia streams so that the rich and varied content of DLs can be searched.
There are the problems of handling worldwide DLs, in the context of varying collection principles, enormous difference in response time between local and remote servers, and the needs of users for different views.

Of central concern is covering the range from personal to global DLs, the so-called ‘scaling’ problem.

At the same time, the problem of interoperability must be faced.
Conclusion

- After more than 30 years of progress in computing, the researchers still face challenges and work in the field of DLs as a segmented community, viewing DLs from one or another perspective;
- Finally, “the benefits of digital libraries will not be appreciated unless they are easy to use effectively” (C. Lynch and H. Garcia-Molina).