DIGITAL LIBRARIES

Feasibility, Features, Functionalities and the Future

Dr. M.G. Sreekumar
UNESCO Coordinator, Greenstone Support, South Asia
Indian Institute of Management Kozhikode (IIMK)
Agenda

• Organizational Knowledge – Enterprise Content Management
• Digital Library – Concepts, Principles and Technologies, Architecture…
• Web and its variant forms
• Open (Source) Digital Libraries
• Metadata – Concepts, Functions and Standards
• DL : Functional Components, Workflows & Procedures
• DL : Build up Strategies
• Hardware / Storage / Space
• Software Selection
• DL Architecture - Major Tasks - DL Hardships
Digital Libraries

- Digital Libraries are becoming important and they are gaining increasing social attention, academic and research interest.
- Demand for improved information and knowledge management solutions - universities, enterprises and institutions.
- Need for integrated access to disparate information resources.
- Key challenge - how to create online information environments facilitating internal content publishing and single point access to internal/external information sources.
- Latest DL technologies vs Traditional libraries and knowledge management.
- Fortunately we have a large number of operational digital libraries and services.
World of Digital Information: Features

- Great Potential and Dynamic
- Easy to access, disseminate, store, retrieve, archive, copy, transmit ...
- Ubiquity of the Net / Web
- Information - Any time / Anywhere / Anyone
- Access by a wide spectrum of Users
- Easiness of access - Plug & Play
- Currency of the material / information
- Increase in value
Unique Features of the Net/Web

- Reach - unprecedented
- Richness - unquestioned
- Feedback - excellent
- Content Holder
- Content Publisher
- Content Communicator
- Asynchronous
- Death of Distance / Time
The Current Environment

- Fascinating times in the history of libraries, information systems and electronic publishing
- Possibilities of building large-scale services
- Materials are stored on computers
- Network connects the computers to personal computers on the users' desks
- In a complete digital library, nothing need ever reach paper
<table>
<thead>
<tr>
<th>L 1.0</th>
<th>L 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closed stacks</td>
<td>Open stacks</td>
</tr>
<tr>
<td>Collection development</td>
<td>Library suggestion box</td>
</tr>
<tr>
<td>Preorganized ILS</td>
<td>User tagging</td>
</tr>
<tr>
<td>Walk-in services</td>
<td>Globally available services</td>
</tr>
<tr>
<td>“Read-only” catalog</td>
<td>Amazon-style comments</td>
</tr>
<tr>
<td>Print newsletter mailed out</td>
<td>Team-built blog</td>
</tr>
<tr>
<td>Easy = dumb users</td>
<td>Easy = smart systems</td>
</tr>
<tr>
<td>Limited service options</td>
<td>Broad range of options</td>
</tr>
<tr>
<td>Information as commodity</td>
<td>Information as conversation</td>
</tr>
<tr>
<td>Monolithic applications</td>
<td>Flexible, adaptive modules</td>
</tr>
<tr>
<td>Mission focus is output</td>
<td>Mission focus is outcome</td>
</tr>
<tr>
<td>Focus on bringing ‘em in</td>
<td>Focus on finding the user</td>
</tr>
<tr>
<td>ILS is core operation</td>
<td>User services are core</td>
</tr>
</tbody>
</table>
Internet / Intranet

Multimedia
Library Info System

USER @ anywhere (access to information from anywhere)
Challenges of the Day

- Proliferation of Content
- Collection Building – Acquisition, Subscriptions, Licensing…
- Diverse Datastreams - Content Categories, Publication Types
- Multimedia, Multiformats, Multilingual
- Copyright, Intellectual Property, Fair Use…
- Technology Complexities, Infrastructure Issues
- Publishers’ Stringent Policies / Monopolies
- Integration of legacy systems and the new genre
## Penetration of E-Content in Libraries

<table>
<thead>
<tr>
<th>PUBLICATION TYPES</th>
<th>DOCUMENT FORMATS</th>
</tr>
</thead>
<tbody>
<tr>
<td>• E-Books, E-Journals...</td>
<td>• ASCII, RTF, HTML, SGML, Postscript, PDF, Proprietary, Native Application Formats</td>
</tr>
<tr>
<td>• Aggregated Scholarly E-Journal Databases</td>
<td>• Images, Graphics</td>
</tr>
<tr>
<td>• Databases, CBT/ WBT</td>
<td>• Audio</td>
</tr>
<tr>
<td>• Portals, Vortals...</td>
<td>• Video</td>
</tr>
<tr>
<td>• Value added services</td>
<td>• XHTML, ASP, PHP, XML ...</td>
</tr>
<tr>
<td>• Preprints, Eprints, E-Documents...</td>
<td></td>
</tr>
</tbody>
</table>
What’s a DL?

- "Digital libraries are organized collections of digital information. They combine the structuring and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible." (Michael Lesk)

- “Is a managed collection of information, with associated services, where the information is stored in digital formats and accessible over a network. A crucial part of this definition is that the information is managed. A stream of data sent to earth from a satellite is not a library. The same data, when organized systematically, becomes a digital library collection." (William Arms)

- Digital library is "a focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenance of the collection." (Ian Witten and David Bainbridge).

- "Digital libraries are different [from traditional library automation] in that they are designed to support the creation, maintenance, management, access to, and preservation of digital content. (Bernie Hurley, the Director for Library Technologies at U.C.Berkeley. Quoted in Digital library technology trends. Sun Microsystems. August 2002)"
What is a “digital library”?

*Collection of digital objects (text, video, audio) along with methods for access and retrieval, and for selection, organization, and maintenance.*

- Traditional user/librarian distinction is blurred
- Computers make information active
- Kitchens for knowledge preparation
- WWW ≠ DL!—organization, selectivity
- Nice Web site ≠ DL!—import new documents easily

Ian Witten
Digital libraries encompass a whole range of information services related work such as

- Organization of digital information
- Information retrieval
- User interface
- Archiving and preservation
- Services and social issues
- Evaluation and applications to particular areas
Digital Libraries as ‘Collections’
NSDL is the Nation's online library for education and research in Science, Technology, Engineering, Mathematics.

Highlights

Seeing Science

Imagery has the power to communicate volumes, sometimes in a surprisingly economical way. Edward Tufte, author of *Visual Explanations* points out in the introduction, "Those who discover an explanation are often those who construct its representation." The 2009 International Science & Engineering Visualization Challenge, co-sponsored by the National Science Foundation (NSF) and the journal Science, and published by the American Association for the Advancement of Science (AAAS) aims to boost public understanding of research developments by challenging creators of photographs, illustrations, informational graphics, interactive media and non-interactive media to make new visual meaning in communicating science, engineering and technology for education and journalistic purposes.
Digital Libraries as ‘Institutions’

- Digital libraries are organizations that provide the resources, including the specialized staff, towards building and operating DLs.

- Digital libraries as a dynamic, growing organism.

- Digital libraries evolve and become the predominant mode of access to knowledge and learning, institutionalization of digital libraries appears to be an increasing possibility.
Benefits of DLs

- DLs are structured storage environments of digital data with a consistent format for index and content abstraction.
- DLs do enable the seamless integration of the scholarly electronic information, help in creating and maintaining local digital content, and strengthen the mechanisms and the capacity of the library’s information systems and services.
- DLs increase the portability, efficiency of access, flexibility, availability and preservation of digital objects.
Benefits of DLs

- Outreach - Library goes to the user
- Seamless Access - Searching and browsing
- Borderless Dissemination
- Instantaneous and Current
- Always (24X7) available
- Long term preservation
- ...
- Long term preservation
- ...
Limitations of DLs

- Technological obsolescence
  - Hardware
  - Software
- Quite Tender and hence Fragile too
- Security Issues – Being rigorously addressed
- Highly sensitive to Commands – Even a small ignorance or carelessness could be very fatal at times
- Resources, Cost, Manpower
- Bandwidth
- Rights Management
- …
Functional Components

Selection and Acquisition
- Organizaiton

Indexing and Storage
- Repository

Search and Retrieval
- Digital Library Website

Network Connectivity
- User

Content Selection
- Digitization Conversion
- Assigning Metadata

Indexing of Metadata & Content
- Document Objects
- Metadata Indexes

Browse Search Retrieval View Content
- Front End Home Page

Internet Intranet
Creation of DLs

1. Selection of Source Material
   - Identification of Material

2. Planning
   - Roadmap for implementation

3. Implementation & Project Management
   - Staff, Infrastructure and digitization of collection

4. Marketing & User Support
   - Promotion and Support Services

5. Dissemination & Provision of Services
   - Integration of library with other online libraries & services

6. Administration & Maintenance
   - Long time maintenance & administration
Digital Objects

- Digital Document (data)
  - Digital representation of the document/entity

- Metadata
  - Description of the document/entity

- Identifier
Digital Objects

- Digital objects of analogue/physical equivalents: pictures, video clips, music, publications, maps, artifacts (e.g. museum objects), living beings (plants, animals, people), animation's, slide shows, print publications, etc. In case of some of these entities (for example, artifacts like buildings and museum objects and living beings) digital objects may only carry relevant metadata information and possibly some form of multimedia representation of the entity (e.g. photographs).

- Digital objects that do not have physical counterparts and those created dynamically and in real-time: electronic publications, software, spread sheets, databases, data gathered from remote sensors, software agents, and live capture of digital versions of speech, music and video.
Libraries - Shifts

- **Traditional / Automated**
  - Organization is physical
  - Shelving of documents - Based on Subject Cln
  - Key - Index / Catalogues / Cards / Digital Catalogs
  - Cards - Real/Virtual - Author, Title, Descriptions

- **Digital**
  - Organization in terms of digital files /objects
  - Contains material digitized form
  - Contains digital material
  - Architecture
  - Key - Metadata
Shift in Approaches

Traditional
Limited/ Rigid
AACR2
CCC
CC / LCCS
DDC / UDC
Thesauri/LCSH

Automated
Improved
AACR2
ISO 2709
CCF
MARC
Thesauri

Dig. Library
Efficient/ Flexible
Metadata
DCMI -- W3C
EAD, TEI, DTD
METS, MODS, Z39.50
MARC21
OAI-PMH
What are digital libraries for?

- Knowledge/content management
  - Manage and access internal information assets
- Scholarly communication, education, research
  - E-journals, e-prints, e-books, data sets, e-learning
- Access to cultural collections
  - Cultural, heritage, historical & special collections, museums, biodiversity
- E-governance
  - Improved access to government policies, plans, procedures, rules and regulations
- Archiving and preservation
- Many more …
Features of Digital Libraries...

- Dynamic Electronic Information Systems
- Seamless Aggregation and Integration of Scholarly Content
- Create / Maintain Local Content
- Strengthens - mechanisms and capacity - Information Systems / Services
- Increase Portability
- Efficiency of Access
- Flexibility
- Availability
- Long term preservation
What Distinguishes a DL?

- Site Neutrality (3 in 1 Access-Anytime, Anywhere by Anyone Access)
- Open Access
- Greater variety and granularity of information
- Sharing of information ‘Sharium’
- Up-to-date ness
- Always available (365*7*24)
- New forms of rendering (New Genre)
Digital Libraries: An Overview

- Digital
  - Computing
  - Networking
  - Content
  - Collections
  - Services
  - Community
- Libraries
Special Requirements

- Infrastructure
- Acceptability
- Access Restrictions
- Readability
- Standardization
- Authentication
- Preservation
- Copyright
- User Interface
Need for Content Integration / Organization

• Assuring Seamless Access to the Content
• Need for a single Info. Gateway / Access Point
• Multi - Formats, Media, Platforms (Content / Data in different formats)
• Data encoding (role of markup languages)
• Role of Metadata (role of Standards)
• Structured Metadata (role of XML)
• Need for Interoperability
• Interface / Delivery / Presentation
• Exorbitant cost of proprietary DL S/W
Digital Library Technologies

• Open architectures (Open DLs)
• Componentized vs Monolithic systems
• Interoperability (role of Z39.50, OAI etc.)
• Unified interface for heterogeneous libraries
• Metadata mapping across different libraries
• OAI-compliant data and service providers
• Multilingual digital libraries
• Scalable digital library architectures
• Publication tools
• Searching tools
Space Requirements: For 100,000 Articles (Text) having 5 pages each

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characters/Page</td>
<td>2,000</td>
</tr>
<tr>
<td>Characters/Article</td>
<td>10,000</td>
</tr>
<tr>
<td>Characters/Collection</td>
<td>1,000,000,000</td>
</tr>
<tr>
<td>Raw Data Bytes</td>
<td>1,000 MB</td>
</tr>
<tr>
<td>Database Structure Overhead</td>
<td>200 MB</td>
</tr>
<tr>
<td>Index Overhead</td>
<td>1,000 MB</td>
</tr>
<tr>
<td>Bib Records Overhead</td>
<td>150 MB</td>
</tr>
<tr>
<td>Subtotal</td>
<td>2,350 MB</td>
</tr>
<tr>
<td>Processing, RAID, etc.</td>
<td>780 MB</td>
</tr>
<tr>
<td>Total</td>
<td>3,000 MB</td>
</tr>
</tbody>
</table>
Space Requirements: For 100,000 Images (640X480 in 256 colours)

<table>
<thead>
<tr>
<th>Description</th>
<th>Value 1</th>
<th>Value 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bytes/Image</td>
<td>300,000</td>
<td></td>
</tr>
<tr>
<td>Raw Data Bytes</td>
<td>30,000 MB</td>
<td>100,000 images</td>
</tr>
<tr>
<td>Database Structure Overhead</td>
<td>100 MB</td>
<td>1 KB/image</td>
</tr>
<tr>
<td>Feature Index Overhead</td>
<td>20 MB</td>
<td>200B/image</td>
</tr>
<tr>
<td>Bib Records Overhead</td>
<td>150 MB</td>
<td>500 bytes + 200%/image</td>
</tr>
<tr>
<td>Subtotal</td>
<td>30,370 MB</td>
<td></td>
</tr>
<tr>
<td>Processing, RAID, etc.</td>
<td>10,115 MB</td>
<td>33%</td>
</tr>
<tr>
<td>Total</td>
<td>40,460 MB</td>
<td>= 40 GB</td>
</tr>
</tbody>
</table>
Space Requirements: For 100,000 Audio Recordings (Half Sound, 8 Bit 11 KHz- Mono and 16 Bit 44 KHz Stereo, 10 Mins each)

<table>
<thead>
<tr>
<th>Description</th>
<th>Bytes</th>
<th>Recording Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bytes/Sound Clip</td>
<td>600 KB</td>
<td>10 minutes @ 1 KB/sec</td>
</tr>
<tr>
<td>Bytes/Music Clip</td>
<td>10,200 KB</td>
<td>10 minutes @ 17 KB/sec</td>
</tr>
<tr>
<td>Raw Data Bytes</td>
<td>540,000 MB</td>
<td>50,000 sound and 50,000 music</td>
</tr>
<tr>
<td>Bib Records Overhead</td>
<td>150 MB</td>
<td>500 bytes + 200%/article</td>
</tr>
<tr>
<td>Subtotal</td>
<td>540,150 MB</td>
<td></td>
</tr>
<tr>
<td>Processing, RAID, etc.</td>
<td>180,000 MB</td>
<td>33%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>720,150 MB</td>
<td>= 720 GB</td>
</tr>
</tbody>
</table>
Space Requirements: For 100,000 Video Clips (320X200 and 256 colours at 15 fps)

<table>
<thead>
<tr>
<th>Bytes/Video</th>
<th>15 MB</th>
<th>0.25 MB/sec assuming 4:1 compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Data Bytes</td>
<td>1,500,000 MB</td>
<td>100,000 articles/collection</td>
</tr>
<tr>
<td>Feature Index Overhead</td>
<td>200 MB</td>
<td>10 images/clip and 200B/image</td>
</tr>
<tr>
<td>Bib Records Overhead</td>
<td>150 MB</td>
<td>500 bytes + 200%/article</td>
</tr>
<tr>
<td>Subtotal</td>
<td>1,500,350 MB</td>
<td></td>
</tr>
<tr>
<td>Processing, RAID, etc.</td>
<td>500,000 MB</td>
<td>33%</td>
</tr>
<tr>
<td>Total</td>
<td>2,001,350 MB</td>
<td>= 2,000 GB = 2 TB (Terabytes)</td>
</tr>
</tbody>
</table>
## Bandwidth Requirements

**Bandwidth**
For amounts of data of the sizes given above the bandwidth needed to deliver them becomes a serious consideration.

For the sample objects considered above in the examples the delivery requirements for one object are:

<table>
<thead>
<tr>
<th>One Text Article</th>
<th>10 KB</th>
</tr>
</thead>
<tbody>
<tr>
<td>One Image</td>
<td>300 KB</td>
</tr>
<tr>
<td>One Audio Clip</td>
<td>600 KB</td>
</tr>
<tr>
<td>One Video Clip</td>
<td>1,500 KB</td>
</tr>
</tbody>
</table>

These sizes have to be compared against the available capacities of the delivery channels:

<table>
<thead>
<tr>
<th>Internal PC Disk Channel (DMA)</th>
<th>33,000 KB/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>48x CD-ROM</td>
<td>7,200 KB/s</td>
</tr>
<tr>
<td>Ethernet LAN (100 Mbps)</td>
<td>10,000 KB/s</td>
</tr>
<tr>
<td>DSL/Cable</td>
<td>256 KB/s up; 5,000 KB/s down</td>
</tr>
<tr>
<td>ISDN Connection</td>
<td>128 KB/s</td>
</tr>
<tr>
<td>56.6 Modem</td>
<td>5.7 KB/s</td>
</tr>
<tr>
<td>28.8 Modem</td>
<td>2.9 KB/s</td>
</tr>
</tbody>
</table>
Desirable Features of DL Software

- Structures
- Accessible
- Searchable
- Extensible
- Massive
- Heterogeneous
- Persistent
Desirable Features of DL ...

- Low cost, including all hardware and software components
- Technically simple to install and manage
- Robust
- Scalable
- Open and interoperable
- Modular
- User Friendly
- Multi-user (including both searching and maintenance)
- Multimedia digital object enabled
- Platform independent (including both client and server components) interoperable
DL’s operation should be examined under...

- **Architectural design** — Modular and Open
- **Backend Database** — scalable, robust, data formats
- **Network capabilities** — web-based and seamless operations, persistent IDs, security and authentication
- **Metadata and Interoperability** — compatible with world standards such as Dublin Core and OAI-PMH
Technical Issues

- Open source software Vs Commercial OS
- Hardware and peripheral requirements
- Network Components
- Standards – data formats, metadata, network, access, interoperability, encoding
Software Selection

- Goals and Requirement Specification
- Proprietary Vs Open Source
- Fit the existing Information System
- Accommodate future migration
- Embrace all possible/predominant formats
- Support standard DL technologies/platforms
- Easy installation, population, maintenance
- Comprehensive Documentation
- Software Development Team
- Active User Groups, E-Mail Lists (Users / Developers)
DL Software: Alternatives

- What are your expectations?
- Develop local web-based application?
- Commercial DL solution?
- Adopt open source software?
  - Greenstone
  - Eprints
  - DSpace
  - Fedora…
DLs: Workflows and Processes

- Content selection
- Content acquisition
- Content publishing
  - Metadata preparation
  - Content loading
- Content indexing & storage
- Content access & delivery
- Preservation
- Access management
- Usage monitoring and evaluation
- Networking and interoperation
- Maintenance
DL Software: Key requirements

- Document types (book, journal article, lecture …)
- Document formats (text, PDF, Word, PS, …)
- Content acquisition (online and offline)
  - Metadata description, content tagging
  - Content uploading
- Indexing and retrieval
  - Structured/ full text indexing
  - Automatic metadata extraction
- Storage
  - Data compression
  - Efficient storage for metadata
  - Efficient location of metadata and documents
- Access and delivery
  - Structured search, browse, hierarchical browsing
  - CD-ROM distribution
DL Software: More requirements

- Scaling up – for large collections
- Multilingual support
- Access management and security
- Usage monitoring and reporting
- Standards compliance
  - XML, Dublin Core, Unicode
- Interoperation
  - OAI, Z39.50 compliance, MARC, CDS/ISIS, …
Metadata

General Definition

- Metadata in its broadest sense is **data about data**
- Documentation about documents and objects
- Describing (Tagging) the contents of the object
- For Information Discovery from the Resource Base

Internet context

- Data describing the attributes of an electronic resource on the net
- Dublin Core (DCMI) – WWW Consortium Standard
- XML - The tool
## Dublin Core Metadata Elements

<table>
<thead>
<tr>
<th>Metadata</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Basics:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>22 Elements</strong></td>
<td></td>
</tr>
<tr>
<td>Title</td>
<td>The name given to the resource by the creator or publisher</td>
</tr>
<tr>
<td>Creator</td>
<td>The person responsible for the intellectual content of the resource</td>
</tr>
<tr>
<td>Subject</td>
<td>The Topic of the resource</td>
</tr>
<tr>
<td>Description</td>
<td>A textual description of the content of the source</td>
</tr>
<tr>
<td>Publisher</td>
<td>The Entity responsible for making the resource available</td>
</tr>
<tr>
<td>Contributor</td>
<td>A person or organization (other than the Creator) who is responsible for making significant contributions to the intellectual content of the resource</td>
</tr>
<tr>
<td>Date</td>
<td>A date associated with the creation or availability of the resource</td>
</tr>
<tr>
<td>Type</td>
<td>The nature or genre of the content of the resource</td>
</tr>
<tr>
<td>Format</td>
<td>The physical or digital manifestation of the resource</td>
</tr>
<tr>
<td>Identifier</td>
<td>An unambiguous reference that uniquely identifies the resource within a given context</td>
</tr>
<tr>
<td>Source</td>
<td>A reference to a second resource from which the present resource is derived</td>
</tr>
<tr>
<td>Language</td>
<td>The language of the intellectual content of the resource</td>
</tr>
<tr>
<td>Relation</td>
<td>A reference to a related resource, and the nature of its relationship</td>
</tr>
<tr>
<td>Coverage</td>
<td>Spatial locations and temporal durations characteristic of the content of the resource</td>
</tr>
<tr>
<td>Rights</td>
<td>Information about rights held in the resource</td>
</tr>
</tbody>
</table>
DC-2012 will explore the global, national and regional roles of metadata in addressing global challenges such as food security, the digital divide, and sustainable development. Metadata plays a significant role globally in information systems shaping how we know, monitor and change social and governmental systems affecting everything from the environment, human rights and justice to education and peace. DC 2012 will bring together in Kuching the community of metadata scholars and practitioners to engage in the exchange of knowledge and best practices in developing languages of description to meet these global challenges.

DC-2012 will take place in Kuching, Sarawak, Malaysia (Borneo) and will be part...
DL - Hardships

- Copyright Issues
- Technology Complexities
- Infrastructure Issues
- Publications/Formats – Diverse Datastreams
- Digital Objects/Formats - Multiple
- Publishers’ Policies – Stringent, Inconsistent
Approaches to Building DL

- Digitization – retro-conversion of non-digital resources to digital
- Digitally born resources – involves inter-conversion to standard formats and storage
Major Tasks

- Content identification (internal / external)
- Content Creation
- Content Collation/Signposts
- Organisation
- Updation
- Retrieval / Dissemination
- User Training
- Archiving
DIGITAL LIBRARY ARCHITECTURE

Network

OS

Z39.50/OAI-PMH

DL Software

METS/MODS

EAD/TEI

Data/Objects

DCMI
Thank you

Questions?