Challenges for Large-scale Content Building for International Digital Library

Ching-chih Chen
Professor, Simmons College  U.S.A
Challenges for Large-scale Content Building for International Digital Library

Presented at
Digital Library:
IT Opportunities and Challenges in the New Millennium
July 8, 2002, Beijing

Ching-chih Chen
Professor
Graduate School of Library and Information Science
Simmons College, Boston, USA
chen@simmons.edu

Digital Libraries: Universal Access to Human Knowledge

VISION
All citizens anywhere anytime can use any Internet-connected digital device to search all of human knowledge.
Via the Internet, they can access knowledge in digital collections created by traditional libraries, museums, archives, universities, government agencies, specialized organizations, and even individuals around the world...

Some of the Challenges
1. Easily said, then done
   - Currently, only a very small fraction of information and knowledge is available online or digital—so the need for large-scale digital content-building
   - Major issues impeding creation and use: metadata, access, interoperability, intellectual property, privacy, security, preservation, applications
   - Need significant leadership to the creation and exchange of international DLs

2. Intellectual property rights need to be addressed in order to facilitate the creation of and access to digital libraries
   - Develop and deploy infrastructure (e.g., micro-payments) to support use of governmental DL materials
   - Develop and deploy methods to authenticate and verify government information
   - Develop policies and laws for managing ambiguous or unknown property rights

3. Establish large-scale international digital library testbeds.

IT Has Transformed Our Lives
Final Report of the President’s Presidential Information Technology Advisory Committee (Feb. 1999)

- Ten critical “National Challenge Transformations” identified by PITAC
- These transformations used to identify critical IT challenges

- Transforming
  - the way we communicate
  - the way we learn
  - the nature of commerce
  - the nature of work
  - our understanding of the environment
  - how we govern

Toward the Vision: Who’s responsibilities

Currently we are far from the Vision:
- All citizens anywhere anytime can use any Internet-connected digital device to search all of human knowledge.
- These new libraries offer digital versions of traditional library, museum, and archive holdings, including text, documents, video, sound, and images. The need for large-scale digital library testbeds...
- Very-high-speed networks enable groups of digital library users to work collaboratively, communicate with each other about their findings, and use simulation environments, remote scientific instruments, and streaming audio and video. No matter where the digital information resides physically, sophisticated search software can find it and present it to the user. In this vision, no classroom, group, or person is ever isolated from the world’s greatest knowledge resources.

Who should be involved to meet these challenges?
- Cross-disciplinary - Science/SCIENet, communications, library/information sciences, S&T, humanities, social sciences, etc...
- Institutional
  - Large, medium, small
  - Libraries, archives, museums
  - Libraries - academic, public, school, special libraries, etc...
  - Institutions - educational, governmental, public, etc...
- Geographical
  - Regional
  - National
  - GLOBAL (the East and the West, or the North and the South, or globally)
- Funders - Government, foundations, industrial, and private.
Essential Considerations for Content Building in Digital Libraries

- "Information-related" projects - Technology alone is not enough, we must have contents, and good contents.
- Technology - Must find ways to work synergistically.
- Content - No one has everything. Must capitalize what we have and discover how to expand to include what we don't have.
- Community building
  > Must find ways to create new communities, to COLLABORATE!
  > Find ways to overcome limitations
- Global Collaboration

Why Global Digital Libraries?

- Digital contents of various subjects and formats from different parts of the world can be shared by everyone anywhere.
- Skills and knowledge used to create the national digital libraries can be exchanged.
- Avoid competing "standards."
- Distributed systems can be integrated interoperably.
- Close existing divisions.

Barriers of International Collaboration

There are many barriers:

- Distance, time zones
- Culture
- Funding mechanisms, Increase in cost
- Ignorance
- Logistical difficulties in arranging the collaboration
- Political contexts
- General problems of collaboration
- Meeting the needs of all collaborators, Etc.

The Global Digital Library - Conceptual Diagram

(Chen presented the GDL concept at the International Conference on National Libraries - Towards the 21st Century, Tokyo, April 1993)

Benefits of International Collaboration

- Multiple types of contents - multi-subject, multi-format, multi-lingual, etc., and no one has all.
- Multiple skills needed, no one person has all.
- Multiple technological methods and tools needed, and no one has all.
- Multiple alternatives in approaching information related problems.

What Is Needed for International Collaboration

- Each partner must believe that collaboration is beneficial,
- Each partner must be willing to take some risk,
- Each partner must be willing to share,
- Targeted multi-national research programs,
- International conferences and workshops to encourage international exchange of ideas and know-how's.
Large-scale content building -- international issues

- Multilingual systems -- multilingual, translation, cross-language
- Copyright law
- Social/economics issues: technological uptake, impacts, development differences by country
- Educational systems
- Support of centralized and decentralized operations
- Multinational corporations
- Etc.

What does the future hold in the information arena?

- We can see some glimpses of the future universities without walls.
- Access to all the published works of the world anytime anywhere anyone
- Emergence of the World Bank of, not money, but knowledge

To realize this, we need large-scale digital contents!

Future Technology Considerations

- Computational power doubles every 18 months (Moore's Law)
  - 100-fold improvement every 10 years
- Disk densities double every 12 months
  - 1000-fold improvement every 10 years
- Optical bandwidth doubling every 9 months
  - 10,000-fold improvement every 10 years
- Infinite bandwidth and memory before computation
  - Cost decreasing, density increasing

All Human Recorded Knowledge

- Books, reports
- Periodicals (journals, newspapers)
- Pictures, still images of all subjects
- Music, opera, dance
- Paintings, sculptures, monuments, art objects
- Movies, videos
- Databases, software
- Etc.

All these need to be accessible via the Web

A Book vs A Digital Book?

- Collection of static content
- Linearly organized
- Selected by author, subject, etc.
- Occupying a single physical location
- Each is physically bound

- Collection dynamic multimedia content
- Non-linear, Browsable, navigable
- Selected by User as related in all possible ways
- No physical space
- No physical volume and can be instantly Transmittable
A Library vs A Digital Library

- Collection of items: Collection of digital items (potentially huge)
- Linearly organized (shelves): Non-linear -- compassing everything (someday)
- Chosen by budget constraints: Organized arbitrarily
- Occupying physical space: Occupying no physical space
- Cataloged for access: Fully content-searchable

Sizes of Digital Files

- > 1MB uncompressed - 300KB compressed
- 1 Movie = 10 GB
- > 1 petabyte = 100,000 movies
- 1 image at least 1MB
- > 1 petabyte = 1 billion painting or photos

Non-Technical Challenges (to name just a few)

- Copyright, intellectual property, licensing
- Economics (Who pays? Who gets?)
- Flat-fee subscriptions
- Metered use (electric company)
- Microcharge
- Free (paid by government or...)
- Automated permissions
- Use measured by technology
- Privacy
- Reliability and accuracy of information
- Change in the nature of use of information

Universal Library Implications

- Elimination of time, space, cost constraints
- Democratization of information
- "Knowledge is power"
- Hyperlinks to related information
- Preservation and Dissemination of Knowledge
- Faster and wider
- Larger and wider
- Interactivity
- Virtual museums

Technological Challenges (to name just a few)

- Input (scanning, digitizing, OCR)
- Non-digital media
- Conversion, scanning, correction
- Triple keyboard, uncorrected OCR
- Digital media
- PDF, HTML, XML, PDS, P6, TEL, JPEG, GIF...
- Data representation, metadata
- Navigation and search
- Output (voice, pictures, virtual reality)
- Synthetic documents - derived automatically from retrieved information: Abstracts, summaries, glossaries, etc.
- Scalability problem: what if a billion people accessing the same item of information at the same time?
- Interoperability problem

Digital Libraries - Problems and Issues

- Definition
- Roles of Digital Libraries
- Technology - infrastructure, quality of service, interoperability, scalability, sustainability, etc.
- Input method and standard - scanning, digitization, OCR...
- Data representation
- Organization - metadata, indexing
- Access - indexing and retrieval
- Navigation and search
- Content building, Collection Management and organization, preservation
- User interface and human-computer interaction
- Multilingual problems and issues
- Eeconomic, social (information rich and poor, etc...), and legal issues (copyright, privacy, security, etc...)
- Collaboration - regional, national and global
Recommendations

**Facts**
- Hardly 10% of all public information is available on the Internet
- Government needs to play a leadership role in developing digital libraries
- Significant technical and operational challenges in migrating and maintaining holdings in digital form
- Intellectual Property rights need to be addressed to facilitate creation and access digital libraries

**Recommendations** (in line with PITAC/DL Report):
- Create testbeds: million book project (creation), and other R&D testbeds
- Support research: metadata, scalability, multiple languages, security, and usability
- Place all public governmental information online
- Preserve IP rights of creators by creating tax incentives for public use of online copyrighted information

---

From Chinese Memory Net to China-US Millions Book DL to Global Digital Library

Great opportunities ahead of us!

---

International Digital Libraries Program (IDLP)

- **Goals**
  - To enable users to easily access digital collections, regardless of location, language or formats
  - To enable broad use in research, education, commerce and other purposes
- **Research will be on:**
  - Interoperable technologies
  - Technology for intellectual property protection
  - Methods and standards for ensuring long-term interoperability among distributed and separately administered databases, world data-mining etc.
- Cooperative research can help avoid duplication of effort, prevent the development of fragmented digital systems, and encourage productive interchange of knowledge and data around the world.

---

Useful references:

**GLOBAL DIGITAL LIBRARY DEVELOPMENT**

Cheng-dih Chen
Beijing: Tsinghua University Press, 2001