2002年7月8日 中国 北京
July 8, 2002 Beijing China

“数字图书馆 —新世纪信息技术的机遇与挑战”国际研讨会

Digital Library

— IT Opportunities and Challenges in the New Millennium

专题报告

Tutorials/Panels

http://www.nlc.gov.cn/dloc/

适应国际数字图书馆合作的大型资

陈刘钦智（Ching-chih Chen）
西蒙斯学院图书馆及信息学系研究生

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

PITAC's Panel on Digital Libraries Report, Feb. 2001

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 on-line 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., micropayments) to support use of governmental DL materials
   - develop and deploy methods to authenticate and verify government information
   - explore safe harbor policies for DLs supporting research and scholarship
   - develop practical and fair policies 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 deal with information
  - the way we learn
  - the practice of health care
  - the nature of commerce
  - the nature of work
  - how we design and build things
  - how we conduct research
  - our understanding of the environment
  - how we govern

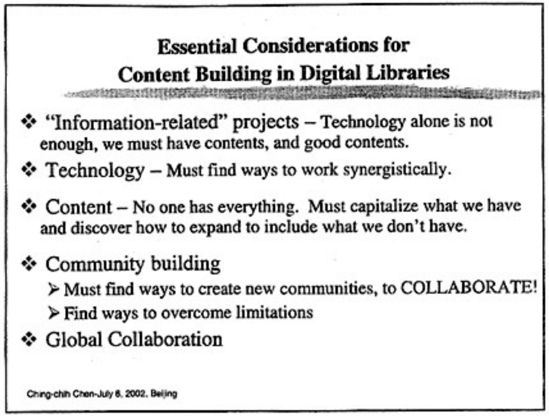
Toward the Vision: Who's responsibilities

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?

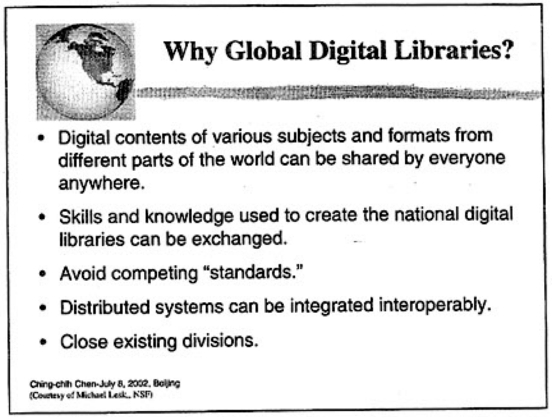
- **Cross-disciplinary** – computer science, network, communications, library/information science, S&T, humanities, social sciences, etc…
- **Institutional**
  - large, medium, small,
  - Libraries, archieves, museums
  - Libraries -- academic, public, school, special libraries, etc…
  - Institutions – educational, governmental, public, etc…
- **Geographical**
  - Regional,
  - national, and
  - GLOBAL (the East and the West, or the North and the South, or globally) collaboration
- **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

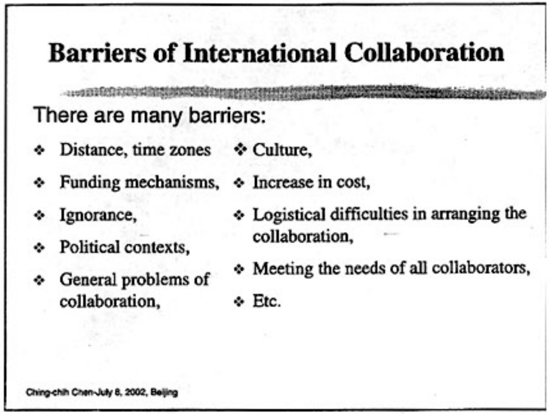
Ching-chih Chen-July 8, 2002, Beijing



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.

Ching-chih Chen-July 8, 2002, Beijing
(Courtesy of Michael Lesk, NSF)

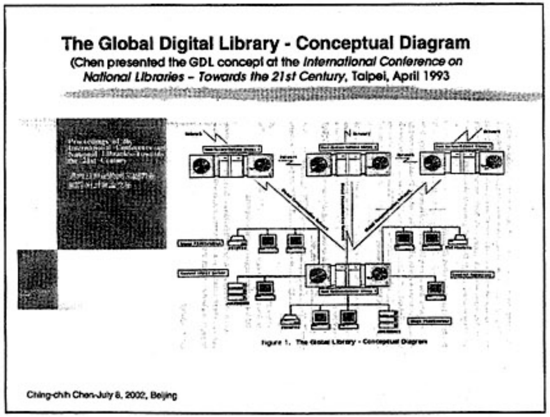


Barriers of International Collaboration

There are many barriers:

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

Ching-chih Chen-July 8, 2002, Beijing

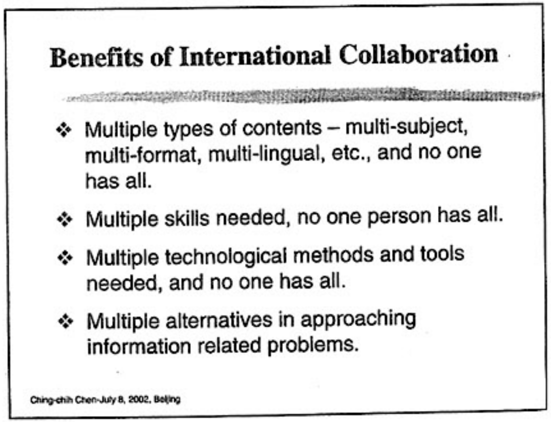


The Global Digital Library - Conceptual Diagram

(Chen presented the GDL concept at the *International Conference on National Libraries – Towards the 21st Century*, Taipei, April 1993

Figure 1. The Global Library - Conceptual Diagram

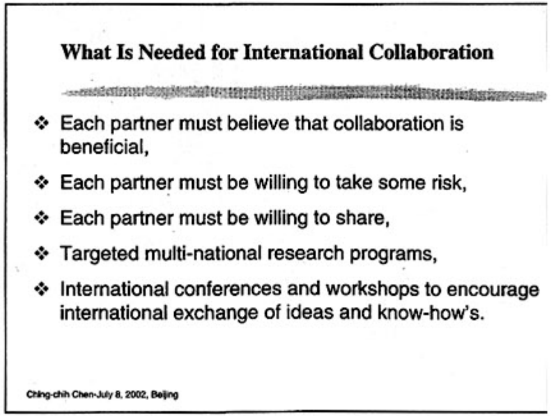
Ching-chih Chen-July 8, 2002, Beijing



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.

Ching-chih Chen-July 8, 2002, Beijing



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.

Ching-chih Chen-July 8, 2002, Beijing

Large-scale content building -- International issues

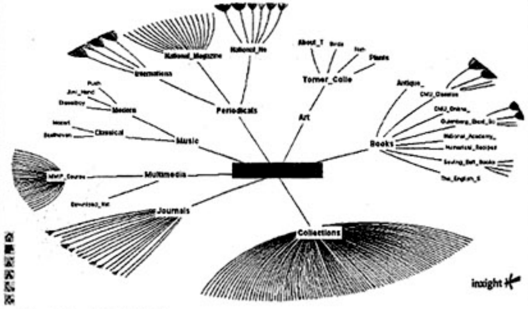
- ❖ Multilingual systems -- multilingual, translation, cross-language
- ❖ Copyright law
- ❖ Social/economics issues: technological uptake, impacts, development differences by country
- ❖ Educational systems
- ❖ Support of centralized and decentralized operations
- ❖ Multinational corporations
- ❖ Etc.

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!

Hyperbolic Tree Navigation



From Raj Reddy

Web Browser created by Inxight Software using Hyperbolic Tree for Java.

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
  - ➢ 10000-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.

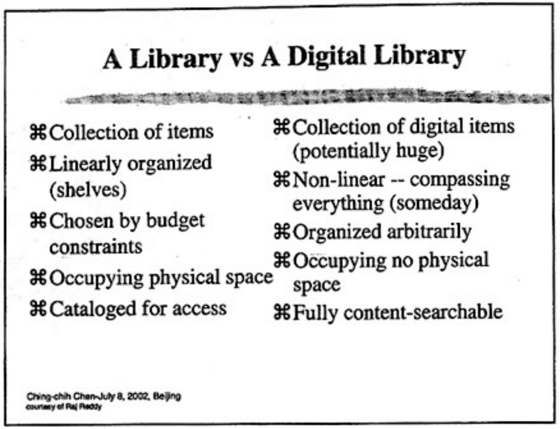
All these need to be accessible via the Web

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 i*nstantly Transmittable*

courtesy of Raj Reddy

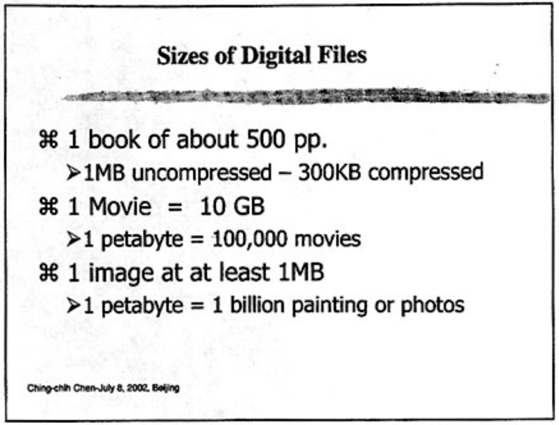


A Library vs A Digital Library

- ⌘Collection of items
- ⌘Linearly organized (shelves)
- ⌘Chosen by budget constraints
- ⌘Occupying physical space
- ⌘Cataloged for access

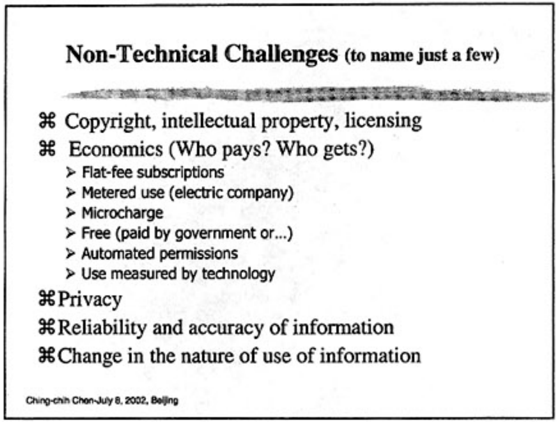
- ⌘Collection of digital items (potentially huge)
- ⌘Non-linear -- compassing everything (someday)
- ⌘Organized arbitrarily
- ⌘Occupying no physical space
- ⌘Fully content-searchable

courtesy of Raj Reddy



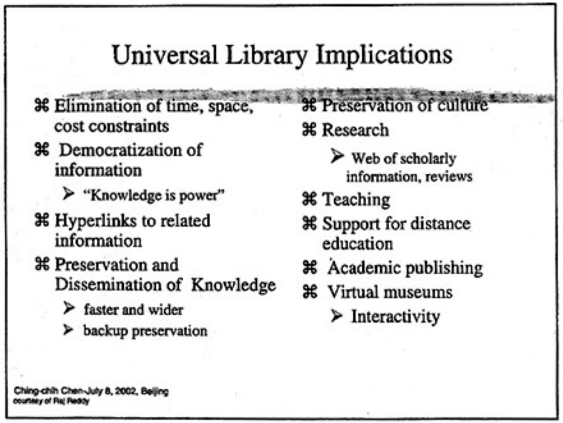
Sizes of Digital Files

- ⌘ 1 book of about 500 pp.
  - ➢1MB uncompressed – 300KB compressed
- ⌘ 1 Movie = 10 GB
  - ➢1 petabyte = 100,000 movies
- ⌘ 1 image at 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

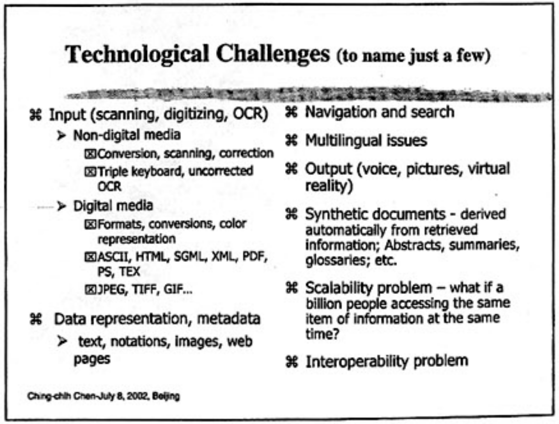


Universal Library Implications

- ⌘ Elimination of time, space, cost constraints
- ⌘ Democratization of information
  - ➢ "Knowledge is power"
- ⌘ Hyperlinks to related information
- ⌘ Preservation and Dissemination of Knowledge
  - ➢ faster and wider
  - ➢ backup preservation

- ⌘ Preservation of culture
- ⌘ Research
  - ➢ Web of scholarly information, reviews
- ⌘ Teaching
- ⌘ Support for distance education
- ⌘ Academic publishing
- ⌘ Virtual museums
  - ➢ Interactivity

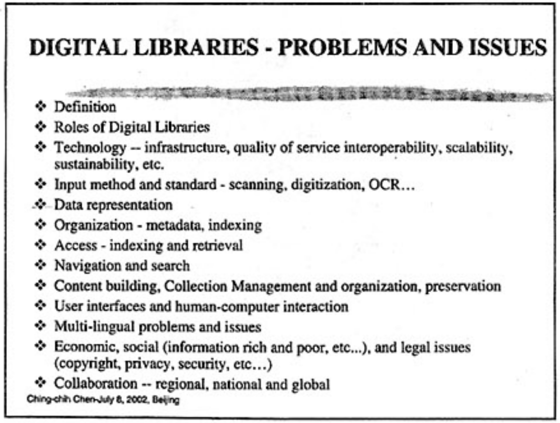
courtesy of Raj Reddy



Technological Challenges (to name just a few)

- ⌘ Input (scanning, digitizing, OCR)
  - ➢ Non-digital media
    - ☒Conversion, scanning, correction
    - ☒Triple keyboard, uncorrected OCR
  - ➢ Digital media
    - ☒Formats, conversions, color representation
    - ☒ASCII, HTML, SGML, XML, PDF, PS, TEX
    - ☒JPEG, TIFF, GIF...
- ⌘ Data representation, metadata
  - ➢ text, notations, images, web pages

- ⌘ Navigation and search
- ⌘ Multilingual issues
- ⌘ 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 interfaces and human-computer interaction
- ❖ Multi-lingual problems and issues
- ❖ Economic, social (information rich and poor, etc...), and legal issues (copyright, privacy, security, etc…)
- ❖ Collaboration -- regional, national and global

Recommendations

⌘ Facts

- ➢ Barely 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

NSF's IDLP (International Digital Library Program)

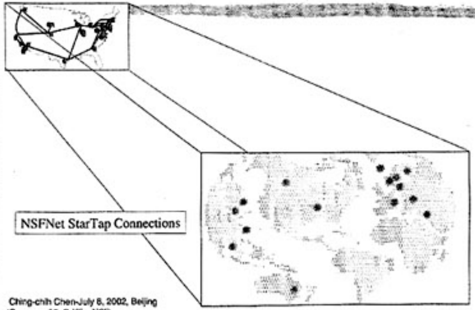
❖ Introduced in 1999, the new IDLP program (NSF 99-6) is intended to contribute to the fundamental knowledge required to create information systems that can operate in **multiple languages, formats, media, and social and organizational contexts.**

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

Great opportunities ahead of us!

More presentations on July 10, 2002!

Changing Scales and Contexts of Interaction and Collaboration since late 1990's

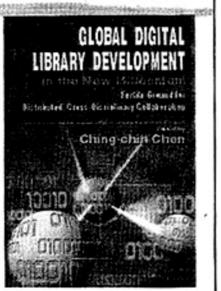


(Courtesy of S. Griffin, NSF)

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:



Beijing: Tsinghua University Press, 2001