## 3. Reader's Feedback

- ".....  I liked the INSROT scheme proposed in your newsletter VishwaBharat@tdil January 2002 issue received recently. It is very nice. Make effective use of 'h' as an operator. It is also reversible. My congratulations…"

  *-Prof. Rajeev Sangal, Director IIIT, Hyderabad*
  *E-mail : sangal@iiit.net*

- ".....  Thankyou for sending me a copy of the 5th issue of the newsletter "VishwaBharat@tdil". I received the previous 4th issue also some months ago. These issues are excellently produced and carry a lot of useful information. I congratulate you & your colleagues engaged in this project for coming out with these issues.

  *-Prof. R. Narasimhan, E-mail : rn@ncb.ernet.in*

- ".....  It is very informative and may be useful for developing our programme Initiative B@bel. This programme is foreseen to start in 2002 and we hope it will contribute to the implementation of the Recommendation on the promotion and use of multilingualism in cyberspace, which will be submitted for approval to the forthcoming session of our General Conference."

  *-Victor Montviloff, UNESCO, Paris*
  *E-mail : V.MontViloff@unesco.org*

- ".....We read with interest about your initiatives in India, especially the TDIL (Technology Development for Indian Language) and 'Digital Unite and knowledge for All' projects and the possibility of setting up a Technological Interchange mechanism. We in ILCA (Instituto de Lenguay Cultura Aymara) are very interested in the advances in language technology that you are making, and wish to know more about them, to see how we can apply similar measures here."

  *- Juan de Dios Yapita,*
  *Instituto de Lenguay Cultura Aymara,*
  *Bolivia, Sudamerica E-mail : ilca@mail.megalink.com*

- ".....You brought so much useful and important information it was a great pleasure to me to find out that so much had been achieved over the two years since my previous visit to India. I would particularly appreciate something that described and catalogued the technologies now available, as described in the excellent booklet that was distributed."

  *-Prof. Pat Hall, Open University, Milton Keynes*
  *E-mail : p.a.v.hall@open.ac.uk*

- ".....Thanks for the fourth issue of the TDIL Newsletter. Nowadays there is so much work under way on various aspects of the computerisation of Indian languages that no one person can hope to keep up to date with it all. Your Newsletter acts as a very valuable clearing house for this information — it lets users know what software is under production, and it gives producers a chance to give advance publicity to the projects they are working on. This is a very valuable service : no more duplicated effort, no more multiple "standards", no more working in the dark. Keep up the good work!"

  *-Dr. J.D. Smith, Faculty of Oriental Studies, Sidgwick Avenue, Cambridge CB3 9DA*
  *E-mail : jds10@cam.ac.uk*

## 4. Universal Digital Library

### Workshop on Universal Digital Library (UDL) at Carnegie Mellon University, Pittsburgh, USA on 26 - 31 May 2002

#### Universal Digital Library (UDL)

#### Vision

- Digitally preserve literary, artistic and scientific information.
- Make the works of man permanently accessible to the billions of people all over the world.
- A Library through free access on the Internet, will improve the global society in ways beyond measurement.
- The Internet will house a Universal Library that is free to the people.

#### Mission

- Create the Universal Library with a free-to-read, searchable collection of one million books.
- Make available 10 million books in 10 years accessible to everyone

| | |
|---|---|
| High School Library | < 30,000 Books. |
| Most Libraries in world | < 1 Millions Books |
| Total no. of Books indexed in OCLC database | 48 Million. |

One Million books on web will be bigger than what first rank university Libraries hold.

- One Million books gives us an excellent tested material for language processing research in
  - Machine translation
  - Summarization
  - Intelligent indexing
  - Information Retrieval

#### Goals

##### Primary Objective

- Digitizing 1 million books (less than 1% of all books in all languages ever published) by 2005.

##### Secondary Objective

- A test bed for other researchers working on improved optical character recognition, and improved indexing.

- The corpus this project creates will be one to three orders of magnitude larger than any existing free resource.

### Goals Specific to Indian languages

#### Primary Objective

- Digitizing 10,000 books as technology demonstrator.
- Indian OCR Research
- Train other centres in India.
- To be a DL Referral agency in India.
- Focus on Indian Collections and Indian Heritage as well as Science and Technology.

#### Secondary Objective

- A test bed for other researchers working on improved optical character recognition, and improved indexing in Indian languages.
- The corpus this project creates will be one to three orders of magnitude larger than any existing free resource in Indian languages.

### Technical Details

| | |
|---|---|
| Scanning / day (on 3-shift basis) | 10,000 pages |
| Images stored in | TIF Format |
| OCR (for English using Abbyy fine Reader 6.0) | Stored in HTM, TXT & RTF formats for searching purpose. |
| Metadata | In XML (Dublin Core format) |
| Scanned Book on Server takes | 50-60 MB |

### Benefits

- To supplement the format education system by making knowledge available to anyone who can read and has access.
- Free to everyone around the world, will enhance the learning process.
- This digital library would be open all the 168 hours of the week on a 24x7x365 basis.
- It is hoped that at least 10,000 books among the million will be available in more than one language.

### Content Selection

- The Million book project will adhere to the copy right law.
- Materials published before 1920 are in the public domain and may be scanned for this project.

### Partners

#### India

- Anna University
- Arulmigu Kalasalingam College of Engineering
- Goa University
- Indian Institute of Information Technology - Allahabad
- International Institute of Information Technology- Hyderabad
- Shanmugha Arts, Science, Technology & Research Academy
- Tirumala Tirupati Devasthanams.
- Maharashtra Industrial Development Corporation
- University of Pune.
- University of Mysore.

#### USA

- Carnegie Mellon University

#### China

- Beijing University.
- Chinese Academy of Science.
- Fudan University.
- Ministry of Education of China.
- Nanjing University.
- State Planning Commission of China.
- Tsinghua University.
- Zhejiang University.



## At CMU 26-30 May, 2002

Shri Y.S. Bhave, JS&FA, Dr. Om Vikas, Senior Director and Dr. P.K. Chaturvedi, Director of the Department of IT Participated in the Indo-US Meeting on *Universal Digital Library (UDL)* project.

On 26th May, following presentations were made:

| | |
|---|---|
| Eric Burns: | Technical Challenges & Solutions |
| Gabrielle Michalek: | Meta Data issues & demo |
| Madhavi Ganapathiraju : | Book to Web-process |

Subsequently demonstrations of scanning and proofing tools were made.

### Madhavi's Presentation : Taking a book to the Web : Processes

#### Digital Library Manual - Contents

- Methodology
- Process
- Trouble Shooting
- Research Areas

#### Process

- Process involved
- Identification of books
- Pre-Scanning process
- Scanning Process
- Image Processing
- Conversion Process

#### Identification of Books

- No copy right law is violated.
- They are of relevance to the selected user group.
- The books are available for scanning.

#### File Formats

- Image File Formats used to store scanned images
  - Tagged Image Format (TIFF or TIF) - Recommended
  - Windows Bitmap (BMP)
  - Joint Photographic Experts Group (JPEG or JPG)
- Text file format used to store the OCR output
  - Rich text format (RTF) - Recommended
  - Hypertext Markup Language (HTML)
  - ASCII text (TXT)
  - Portable Document format (PDF)

#### Pre-Scanning Process

- Create a folder on your local hard disk and name it BOOKS. This will be the main folder where you will store all the Books
- In BOOKS folder Create a folder With the Book Name that You are about to Scan, Say Book1.
- Within the Book1, create 5 new folders. They should be named as follows,
  - 1AUTHOR
  - 2TITLE
  - 4RTF
  - 5TIF
  - 9HTM

#### About Quick Scan Scanning Software

- Pix Tools/Quick Scan is a high-performance Microsoft Windows utility application that provides an integrated image acquisition environment that allows you to scan, view, print, annotate, store, and perform image processing on documents.

#### Scanning

- Quick Scan uses Pixel Translations ISIS (Image and Scanner Interface Specification) libraries to support more than 125 scanners from many manufacturers.

- ISIS drivers enable scanning at the full rated speed of your scanner.
- Quick Scan also incorporates support for full control of your scanner's capabilities, allowing you to adjust brightness, contrast, scan resolution, scan mode, dithering, and any other settings available in your scanner.

Viewing

- Quick Scan is also a high-performance image viewer that includes many features to make it easy to display and manipulate images.
- Features include a main viewer, a thumbnail viewer, fast scaling and rotation, background preloading, annotations, scale-to-gray conversion for binary images, and a pan window.

Printing

- Quick Scan prints images using any standard Windows-supported printer. Options for convenient image printing include Fit Page, Actual Pixels, and Actual Size.

Saving

- Quick Scan saves acquired images in a variety of popular image file formats and compression schemes.
- By using TIFF Group 4, you can achieve compression rations of 35:1 to 50:1, depending on the type of image and the quality of the scan.
- Color and gray scale capabilities vary according to your scanner and the image file format in use.
- Quick Scan supports color and gray scale file storage, scanning, viewing, and printing.

## Scanner Setting

Make sure that all the options in the Minolta PS 7000 Special Features window are selected correctly and according to the Book.

## Tips during Scanning

- Proper care has to be taken while placing the book.
- The book should not go beyond the stopper.
- Steel mirror has to be kept clean and wipe it with a soft cloth.
- Before placing the books clean the books so that dust does not enter from books to scanner bed.
- Place the scanner away from direct lighting source.
- Do not place any magnetic material or liquid near the scanner.
- Do not try look at the light source when scanning.
- Always make sure you switch off the scanner when you have completed scanning.
- Take a break for 5 minutes after every hour of scanning.

## Image Processing

- Enhances the quality of the scanned Images by removing noise.
- Reduces File size.

Tools used

- Despeckle - Removes isolated black pixels.
- Deskew - Detects and removes the skew
- Crop - Removes the extra white spaces

Execution Procedure

- Install JDK version 1.2.1 or above version
- Set the path C:\crop\debug
- Run the command for single Book.
- java cropper1 "D:\Books\Book1"
- For multiple book
- java cropperm "D:\Books\Book1"
- Execute the batch file Cropper.bat to process the TIF files
- After completion, execute the batch file Clearup.bat to remove the OLD TIF files



### Conversion Process

- RTF2HTM Conversion
- RTF2XML Conversion

### Research Areas

- Enhancements in Image Processing tool.
- Error free character recognizer in Indian Languages.
- Software to identify various fonts of different languages and to create meta files that can be used for accessing the information.
- Most of the Indian languages are not available in printed form. In such case Speech Technologies to convert speech to text in Indian Languages and Photographic scanner technologies may have to be developed.

### Trouble Shooting

- During Scanning
- During Recognition

### The setting for scanning Palm Leaves

The Brightness and contrast were varied and the best results were obtained. The various variation are shown.

- Brightness - Normal (5 in scale) and contrast - 0 (in scale)
- Brightness - Lite (10 in scale) and contrast - 0 (in scale)
- Brightness - Normal (5 in scale) and contrast - 9 (in scale)
- Brightness - Dark (1 in scale) and contrast - 9 (in scale)
- Brightness - Lite (7 in scale) and contrast - 9 (in scale)

### Network and Multimedia Technology or Digital Library

- Delivery Technologies
- Content Generation
- Security
- Scalability

### The Strategy for Scanning of books

- A planetary Scanner like the Minolta PS 7000
- Takes about two hours to scan a 500 page book, crop, OCR and convert it to TIFF, HTML and XML files
- Storage per book is around ~ 50MB
- Distributed data bases

| (Mean) Average Times Spent on Reformatting | | | |
|---|---|---|---|
| medium for reformatting | pre-conversion | conversion | post-conversion |
| Photocopy | 17.00% | 74.70% | 8.30% |
| Microfilm (RLG median times) | 28.10% | 58.90% | 13.00% |
| Digital images (CLASS) | 23.30% | 56.10% | 20.60% |
| Digital images (COM Project) | 23.10% | 57.20% | 19.70% |
| Digital images (Project Open Book) | 16.30% | 32.10% | 51.60% |

### Average Size of a Book

- Average book size ~ 500 Pages
- Size of Page as Image ~ 50-150 KB
- Size of Page as text file (rtf /htm) ~ 8 - 15 KB
- Average size of Digitized book ~ 60MB

### Scanner choice for scanning the palm leaves

- 17 inch flatbed-scanner with curvature correction
- A smaller scanner with graphics software support is also possible to join two parts into one.
- A digital camera with a resolution of 3450*1050 pixels and 24-bit color has also been used for larger palm leaves

- Image enhancement
- File format: TIFF, HTML and XML
- Scanner settings should be 24-bit, 200 to 300 dpi
- expected file sizes: around 10 MB.
- Lossless compression used by TIFF is preferred

## Delivery Technologies in Education

- Preloaded Lectures
- Live Lectures
  - – Multicasting and broadcasting
- Static Courseware
  - – Delivery and Development
  - – E Journals
  - – Books born Digital and archived by scanning
- Streaming
- Video Conferencing
- Multicast
- Broadcast - Using IP and Satellite
- Secure Multicast
- Traditional Search Engines
- Distributed

## Formats

- Formats for Video
  - – MPEG-1, MPEG-2, MPEG-4, AVI, QT, H.263
- Formats for Audio
  - – WAV, MP3, RA
- Formats for books
  - – TIFF, JPEG, RTF, PDF, HTML, XML

## Methodology

*The Steps involved in the project*

- Identification of Digital Items or books to be scanned.
- Scanning the Books using Minolta PS 7000 which is a Planetary scanner.
- Enhancing the scanned Image .
- Run the Optical character Recognizer.
- Generate the HTML files.
- Append to the Digital Library data base.
- Create a master file for searching and display.

## On 27th May, 2002
## Prof. Raj Reddy's Presentation : Information Technology and Digital Libraries

Prof. Raj Reddy presented scenario of exponential growth trends of computing performance; doubling between every 15 months to 24 months; memory doubling every 12 months; 50 GB disk (in $100 in 2001) to 50 TB disk (in $100 by 2010) to 50 Peta Byte disk (in $100 by 2002).

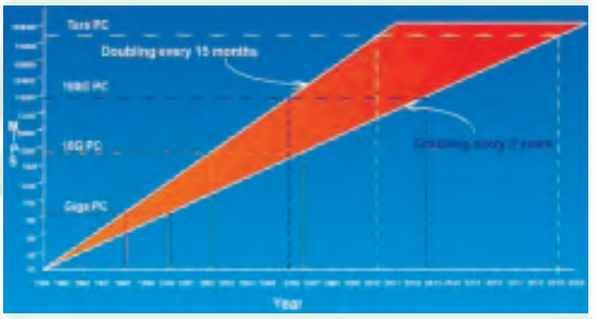
Every 10 years, Processing power improves 100 fold, memory 1000 fold and Bandwidth 10,000 fold.

'Digital Library' term was coined in 1995. There is need to launch Y3K project, i.e. any thing that works today, should be readable 1000 years later.

## How will Technology Impact Our Lives?

- Exponential advances in Information and Communication Technologies will result in
  - – *innovations* that will *transform* the way we live, learn and work.
  - – In retrospect, these transformations will be seen as *revolutionary* by the future generations

## Exponential Growth Trends in Computer Performance




## Technology Trends

- A Giga-PC in 2002
  - – Billion operations per second,
  - – Billion bits of memory
  - – Billion bits per second Network bandwidth
- Less than $2 k
- A Tera-PC by the year 2015
- A Peta-PC by the year 2030

What do we do with all this power?

- Social systems not affected:
  - Food we eat
  - Clothes we wear
  - Mating rituals
- The processing will transform the way, we
  - Live
  - Learn
  - Work, and
  - Provide health care

## Trends in Magnetic Disk Memory

- Densities doubling every 12 months
- Thousandfold improvement every 10 years
- 50GB disk memory costs ~ $100 (2002)
- 50 GB can be used to store
  - 15 hrs of video, a 150 paintings, 1500 hours of MP3 music and 15000 books
  - larger than most of our personal collections at home
- By 2010, 50 Tera Bytes cost ~ $100
  - A personal Library of several million books, a lifetime collection of music and videos- all on our home PC
- By 2020, 50 Peta Bytes ~ $100
  - Infinite amount of memory for all practical purposes

## What do we do with a Peta Byte?

- Capture everything you ever said
  - From the moment of birth
  - To the moment you die
  - Takes less than 1% of a Peta Byte !!
- Everything you did or experienced
  - can be captured in living color
  - with only a few Peta bytes

## Advances in Fiber Optic Technology

- 1.6 Tera bits per second on a single fiber
  - 160 wavelengths each at 10 Gbps
  - Dense Wavelength Division Multiplexing (DWDM)

What can you do with 1.6 Tera bits per second ?

- 10 HDTV movies
- 40 regular full-length movies
- 20000 hours of MP3 music
- In one second on a Single fiber !
- 50 seconds to transmit ALL books in the Library of Congress !
- ALL phone calls on a single fiber with room to spare !

## Bottlenecks in Data Transmission

Main bottleneck is not fiber bandwidth

- It is :
  - Bus bandwidth
  - Router capacity and speed
  - Speed of light!
  - Round-trip delay times in TCP/IP
  - At Tera bit rates with RT times of about 30 ms across the US, 30 billion bits would have been transmitted before an acknowledgment is received

## Technology Trends

- Exponential doubling of memory and band width will continue for 10 to 20 years
  - Leading to the availability of
    - Peta-byte disks
    - Peta-bytes per second bandwidth
    - At a cost of pennies per day
- Leading to Changes in Computer Science and Theory of Algorithms

Compensate for scarcity of computation, memory, and/or bandwidth by using another

- Memory compensates for Lack of Computation
- Bandwidth compensates for Lack of Memory
- Computation compensates for Lack of Bandwidth
- And vice versa

## Access to Information in the 21st Century

- Maxim: Access to all human knowledge anytime anywhere
- Access, query, and print any book, magazine, newspaper, video, data item, or reference document

  - regardless of language
  - using speech, touch screen, or gestures
  - *Universal Library video*
- Challenges in data access
  - High bandwidth networking for multimedia access
  - Intellectual property protection while facilitating access
  - Intelligent information retrieval
  - Delivery and protection of critical information

## Universal Library

- All published works online
- Instantly available
- In any language
- Anywhere in the world
- Searchable, browsable, navigable
- By humans and machines

## Universal Library Goals

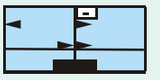
- Research, technology transfer
- Democratization of information
  - Knowledge is power
- Education, distance learning
- Promotion of understanding
- Preservation of human culture

## Research Challenges

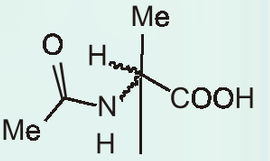
- Input (scanning, digitizing, OCR)
- Metadata creation
- Data representation
- Navigation and search
- Multilingual issues
- Output (voice, pictures, virtual reality)
- Synthetic documents, data mining

## Input Issues

- Low-cost scanning
  - Want $25 per book. Scan WorldCat ~ $1 billion
- Digital media
  - Formats, conversions, color representation
  - ASCII, HTML, SGML, XML, PDF, PS, TEX?
  - Graphics: JPEG, TIFF, GIF?
  - Archiving, persistence
- OCR
  - Unavailable for most languages, e.g. Hindi, Korean
- Structured matter



  - Musical notation, Laban
  - Chemistry
- 3D Items
- Duplication of effort (no registry)

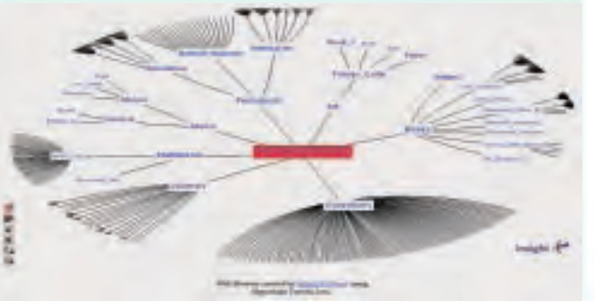



- Web documents

## Navigation

- Keyword searching does not scale. 106 hits?
- Browsing, finding, searching, flying, zooming
  - View whole collections or one glyph
- Fractal view
  - Keys are granularity and connectivity
  - Hyperbolic trees, virtual reality, discovered similarities, user-defined catalogs

## Hyperbolic Tree Navigation



## Searching Mathematics

$$\int_0^\infty e^{-x^2} \sin x^2 \, dx$$

Has this integral ever been evaluated?

$$\int_0^\infty e^{-x^2} \sin x^2 \, dx = \frac{\sqrt{\pi}\sqrt{2-\sqrt{2}}}{2^{9/4}}$$

MATHEMATICA C.F.:

Integrate[Times[Power[E,Times[-1,Power[V1,2]]], Sin[Power[V1,2]]], {V1,0,Infinity}]

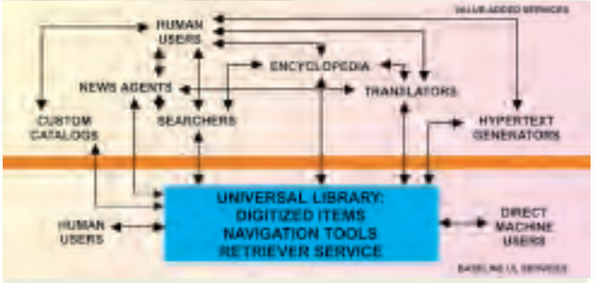
## Multilingual Issues

- Character sets (UNICODE?)
- Representations
- Multilingual Navigation
- Translation Assistance

## Synthetic Documents

- Documents derived automatically from retrieved information via intelligent agents
- Abstracts, summaries, glossaries
  - maximum marginal relevance
- Translations
- Encyclopedia-on-demand
- Critical reviews

## Layered UL Model




## Policy Challenges

- Convenience displaces quality (Gresham)
- What to digitize first?
- Use of copyrighted material
- Economics (Who pays? Who gets?)
- Privacy
- Reliability of information
- Change in the nature of teaching

## Use of © Content

- Philosophy: must pay for use
  - Authors, publishers must not lose
- Implied license
- Bulk licensing
- Compulsory licensing
  - Owner CAN'T refuse; user MUST pay
  - Music industry (1.35¢/min, 7.1¢/song)

## Economic Models

(Determined by the Marketplace)

- "Buy" button
- Metered use (electric company)
- Microcharge (Tobias "clickl")
- Flat-fee subscriptions (e.g. HBO)
- Free (paid by government)
- Automated permissions
- Use measured by technology

## Action Items

Initial goal of 1 Million Books to be digitized within 5 years

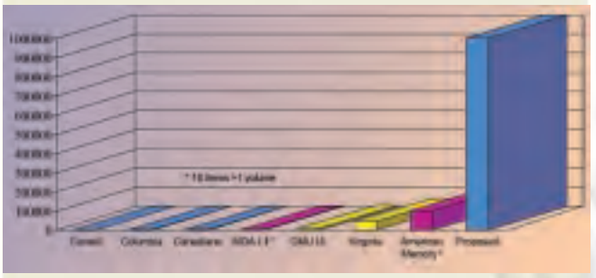
- Expected cost around $25M
- The collection may include books, magazines, newspapers, journals, art, music, and video

## Dr. Michael Shamos Presentation : Why online? Will people read them?

Dr. Michael Shamos Director, Universal Library, www.ulib.org <http://www.ulib.org> Language Technologies Institute Carnegie Mellon University elaborated the Universal Library model - any language, anywhere searchable, browsable, navigable by humans and machines.

There will be UDL Server at CMU and 3 Mirror sites in India, China and Australia. (US Congress should spend $ 25 Billion for UDL project). He discussed the Economic models and the policy challenges.

### World Digital Content



### How Are Books Used?

- Reading
  - pleasure
  - learning, support for distance education
- Reference
- Machines can use books too
  - lookup
  - question-answering
  - textual analysis
  - parallel corpora
  - knowledge representation

### Will People Read Online?

- No. Not until there is an e-Book that imitates a paperback
- UL then becomes source of e-Books

### High-Level Functions

- Hyperlinked documents
- Reliability data
  - Information about authors, institutions, critiques
- Synthetic documents

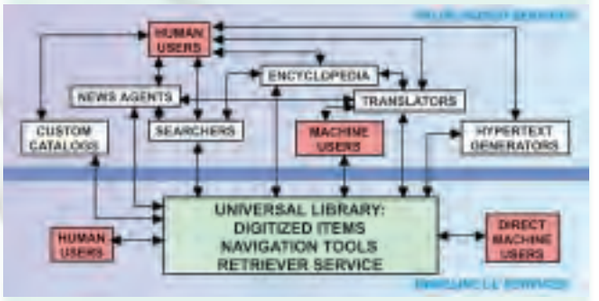
### Synthetic Documents

- Documents derived automatically from retrieved information via intelligent agents
- Abstracts, summaries, glossaries

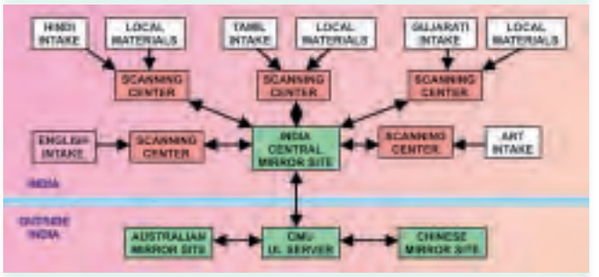
– maximum marginal relevance

- Translations
- Encyclopedia-on-demand
- Critical reviews

### Layered UL Model




### Possible Intake Model

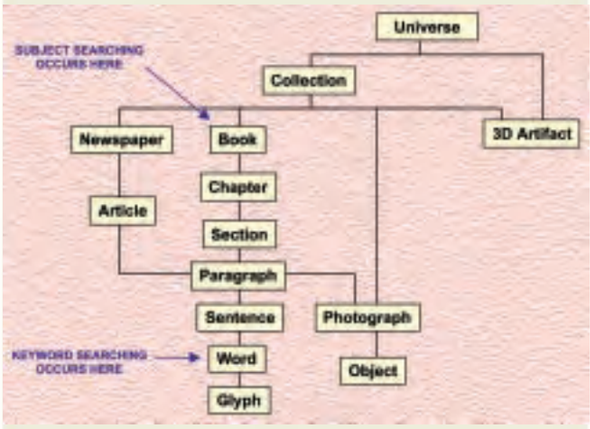



### Hierarchical Nature of Aboutness

- What does it mean to say that a book is "about" chemistry? Can a word be about chemistry?
- If one paragraph is about chemistry, is the book about chemistry?
- If the book is about chemistry, is every sentence in it about chemistry?

- Aboutness is central to cataloging and retrieval

## Aboutness Hierarchy




## Set Theory of Aboutness

- Given a finite universe W of objects (e.g. all words)
- Define a topic T ⊆ W to be a subset of W (a wordlist)
- Topic inclusion (defines the hierarchy):
  - Topic T includes topic S iff S ⊆ T
- Definition of aboutness:
  - A subset P ⊆ W of the universe (e.g., a book) is about topic T iff P ∩ T ≠ ∅ (intersection is nonempty)
- Hierarchical nature of aboutness:
  - If P is about S and T includes S, then P is also about T

## Thesauri and Aboutness

- A set of numbered thesaurus entries defines a topic
- Thesaurus is topic-hierarchical
- 1011 Hindrance
  - 1011.5 barrier, bar, gate, fence, wall, rampart, dam, moat …
- A word is "about" any topic to which it belongs Dam:
  - 241.1 lake
  - 293.7 close (v.)
  - 560.11 mother
  - 757.2 horse
  - 856.11 put a stop to (v.)
  - 1011.5 barrier

Thesaurus + aboutness hierarchy can be used to disambiguate meanings without "understanding"

Note: topic numbers are language independent

## Economic Models (Determined by the marketplace)

- "Buy" button
- Metered use (electric company)
- Microcharge (Tobias "clickl")
- Flat-fee subscriptions (e.g. HBO)
- Free (paid by government)
- Automated permissions
- Use measured by technology

## Use of © Content

- Philosophy: must pay for use
  - Authors, publishers must not lose
- Implied license
- Bulk licensing
- Compulsory licensing
  - Owner CAN'T refuse; user MUST pay
  - Music industry (1.55¢/min, 8.0¢/song)

## Policy Challenges

- Convenience displaces quality (Gresham)
- What to digitize first?
- Use of copyrighted material
- Economics (Who pays? Who gets?)
- Privacy
- Reliability of information
- Change in the nature of teaching

### Dr. Gloriana St. Clair discussed copyright issues, categorized books for scanning

(i) Copyright cleared books for college library (60,000 best books published by 1988),

(ii) Technical reports,

(iii) University Press Publications (MIT, National Academy Press),

(iv) Govt. documents, and

(v) Pre-1923 books.

There is need for US $ 80,000 for copyright clearance project.

Selection of books may follow Gresham's law "convenience displaces quality".

#### Copyright is the Biggest Reality

- U. S. copyright is now 95 years.
- Many books are out of print but restricted by copyright for ~93 years.
- Copyright can be cleared by asking permission to scan.

#### A Collection of Collections

- Books will represent a variety of languages including material originating in China and India, our partners.
- Partners will select and include collections of cultural importance.
- Copyright-cleared *Books for College Libraries.*
- Technical reports.
- University press publications.
- Government documents.
- Pre-1923 books, now in depositories.

#### Best Books Feature

- *Books for College Libraries* 60,000 best books, published in 1988.
- $80,000 needed for a copyright clearance project.

#### Subject Collections

- Sending proposals to small foundations for money to support copyright clearance for resources in subjects such as history and environment.
- Discussed putting up a segment of materials to support literacy.

#### University Press Negotiations

- National Academy Press

  *We will scan early materials and exchange for some 2,500 they are scanning.*

- MIT Press

*Discussed scanning some of their backlog.*

#### Government Documents

- U. S. government produces about 100,000 documents per year; mostly in the public domain.
- Have 30 boxes of Dept of Education materials.
- Negotiating with other agencies
- Some highly desired collections
- British Parliamentary Papers, thousands, 1950 back copyright o.k.

#### Conclusions

- The collection must be composed of many sub-collections.
- Copyright is a serious barrier to an effective effort.

Librarians need to be brought into the picture to ensure solid selection criteria.

### Dr. Robert Thibadeau's Presentation

Dr. Robert Thibadeau described Seagate's Historical NewYork Times Project. He cited phenomenon of free & fee. Many books are available free and some on fee. He cited a number of e-books



websites, e.g. www.getwellbooks.com, www.octavo.com, antiquebooks.net, www.nap.edu, nyt.ulib.org.

During the discussion, it was felt necessary that MCIT in collaboration with MHRD pilots "Indian Digital Library Act". Issues to tackle may include compulsory Licensing, digital pack book (incentive: 10% tax deduction on lifetime revenue); deemed out of print (donate electronic rights); concept shift in Royalty per copy to per preview; public lending rights (as in Japan); 4Cs (Consortium for Compensation for Creative Content), formula to respect content creator and pay compensation, (min. Rs. 100/- to max Rs. 1 lakh), inclusion of books, music and movie with higher & higher privacy value.

### Presentation by Indian delegates

In the afternoon of 27th May, 2002, there were presentations by the Indian delegates, namely, Prof. MGK Menon, Sh. .Y.S. Bhave, Dr. G.V. Subbarao, Dr. I.V. Subbarao, Sh. Ajay Sawhney, Dr. Om Vikas, Dr. Ashok Kolaskar (PU), Dr. Krishniah, Dr. Vaidhya (Sastra), Dr. Thangraj (TN) and Sh. Surendra Baga... (MIDC).

Australian perspective of UDL was presented by Prof. Arun K. Sharma (NSW Univ. Australia)

Prof. N. Balakrishnan presented the national focus with achievable goals of Operative OCR for all Indian languages and the example based translation. India's initiative of Parallel corpora was greatly appreciated.

Task forces were formed for;

i. Standards for Medadata,
ii. Content Selection
iii. Hosting of DL contents
iv. Transport for books & bytes
v. PetaByte Server Initiatives
vi. OCR for Indian languages.

### 29-30 May, 2002

Visits to e-learning centre, Speech technology research groups were arranged

### 31st May, 2002

Speech Technology Research at T J Watson IBM Research Center, New York

Dr. Om Vikas & Dr. P.K. Chaturvedi visited T.J. Watson, IBM Research centre at Yorktown Heights, New York; and interacted with Speech Technology experts : David Nahamoo (incharge, Human language Technologies Research), Michael Picheny (TTS/ Super Human project), Mahesh Vishwanathan (TTS), Ponani Gopalakrishnan (Speech Technologies for Mobile computing), Gregg Daggett (iPaq), David Lubensky (Conversational services), Salim Roukos (MT and crosslingual Information Retrieval, Q&A), Yuqing Gao (Speech-to-Speech Translation), Chalapathy Neti (Audio-visual speech technologies)

Now over 50 % Websites are non-English ones. French appears the second most frequently spoken language.

Speech Technology projects at IBM Research Centre include Speech Recognition, Wireless Roust Speech, 8K/wire telephone based, recognition, Hand held iPaq, Speaker Recognition biometrics, Broadcast news transcription, Audio-visual, Text-to-Speech, Speech-to-Speech MT. About 60 researchers are working in the field of speech technology alone. Major projects on Speech Technology are funded by DARPA (US Defense).

Conventional approach of S2S Translation by cascading S2T, T2T and T2S units will yield limited accuracy upto 30-40% and hence it is not a workable model. Dr. Yuqing Gao (IBM) follows another approach based on Natural Language Understanding - understand meaning and translate concepts. English-to-Chinese S2S translation was demonstrated.

Speech recognition improves on considering Audio model, Language model and Visual model together. There is 10-db gain upon using visual model in addition.