Software for Digital Library

By
Bhupendra Ratha, Lecturer
School of Library and Information Science
Devi Ahilya University, Indore
Email: bhu261@gmail.com
Software for Digital Library

If we want to create a digital library than we require different software's like;
- Digital library software
- OCR
- DOI
- Image Editing Software's
- Other software’s
  - Operating Systems
  - Data Base Management System Software
  - Programming/ Scripting Language
  - Firewall & Protection Software
Digital library software

- **DSpace** is a digital library system to capture, store, index, preserve and redistribute the intellectual output of a university’s research faculty in digital formats. Dspace has been developed jointly by MIT Libraries and Hewlett-Packard (HP). It is now freely available to research institutions world-wide as an open source system.

- **Eprints** is generic archive software under development by the University of Southampton. It is intended to create a highly configurable web-based archive. EPrints primary goal is to be set up as an open archive for research papers, but it could be easily used for other things such as images, research data, audio archives - anything that can be stored digitally by making changes in configuration.

- **Greenstone** is a suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM. It is available for both Windows and Linux O/S. It requires Perl software to build collections.
OCR
(Optical Character Recognizer)

- OCR is also referred to as Optical Character Reader.

- It allows to scan printed, typewritten or hand written text (numerals, letters or symbols) and/or convert scanned image to a computer processable format, either in the form of a plain text or a word document or an excel spread sheet, which can be edited, used or reused in other documents.

- “A system that provides a full alphanumeronic recognition of printed or handwritten characters at electronic speed by simply scanning the documents that called OCR.”
Difference: OCR and OMR

• OCR technologies, images can be scanned, indexed and written to optical media.

• OCR can recognize all type of information.

• It is very flexible.

• OMR is a data collection technology.

• OMR cannot recognize hand-printed or machine-printed characters.

• It is not flexible.
Features of OCR

- It is a program which have recognition capabilities of characters.
- The technology provides a complete form processing and documents capture solution.
- OCR is used when recreating a document in electronic form takes more time.
- The converted text files take less space than the original image file and can be indexed.
- Bridges the gap between the paperless and the papered.
Advantages of OCR

- Savings in costs and efficiencies by not having the paper.
- Scanning and recognition allowed efficient management and planning for the rest of the processing workload.
- Reduced long term storage requirements, questionnaires could be destroyed after the initial scanning, recognition and repair.
- Quick retrieval for editing and reprocessing.
- Minimizes errors associated with physical handling of the documents.
Disadvantages of OCR

- While OCR technology can be effective in converting handwritten or typed characters, it does not give as high accuracy as OMR for reading data, where users are actually marking forms.
- Scanning speed will be determined by the quality of the scanner machines, the size of non-drop out color, paper quality, cleanness, weights.
- To compare the value of the interpreted image with the real image of the document.
- Processing can be in geographic order or in random order.
Processing of OCR

Diagram:

1. Scanner
   - Document
   - Lens
   - Detector
   - Document image

2. OCR software/hardware
   - Document Analysis
   - Character Recognition
   - Contextual Processor
   - Recognition results

3. Output interface
   - To application
Cont…

• **Scanner has 4 components:**
  – A detector, An illumination source, A scan lens and a document transport.

• **OCR hardware/software performs three operational steps:**
  – Document analysis, Character recognition, Contextual processing.

• **Output Interface**
  – Allows character recognition results to be electronically.
Types of OCRs

• Two types of OCRs
  – Task specific readers
  – General purpose readers

• Task specific readers
  – Reads only specific documents: bank cheques.

• General purpose page readers
  – High end OCR (usually for offices)
    • Speed and Accuracy are important
    • Format preservation
    • Good proof reading solutions
  – Low end OCR (usually for house use)
    • Speed is not required
    • Proof reading is done manually
Factors affecting OCR quality

• Scanner quality
• Scan resolution
• Type of printed documents, whether laser printer outputs or photocopied
• Paper quality
• Fonts used in the text
• Linguistic complexities
Evaluating OCRs

- Neat interface
- Easy-to-use wizards
- Accurate recognition
- Scan resolution setting (600 dpi is advisable)
- Time taken from scanning to deliver the final product
- Enhanced usability of the product
- Ability to modify the scan setting
DOI
(Digital Object Identifier)

• An open standard for creating an alphanumerical name that identifies digital content, mostly scholarly contents such as an e-book or journal article.

• A DOI is a unique ID number for a document that is paired with the object’s electronic address, or URL (updatable), along with other metadata.
Basic aspects of DOI

“The DOI is like the Bar Code, but for objects on the Internet.”

Two aspects:

1. **Uniquely Identifies the Object** –
2. **Provides Linking to the Object Itself** (or to any related objects, transactions or services). These links are:
   - Permanent
   - Dynamically maintainable
   - Capable of one-to-many routing
   - Capable of supporting new applications over time
Features of DOI

• Applies to any type or format of object
  – text, music, film, video, photographs, software, database record.
• Applies at any level of specificity
  – whole book/individual chapters, music collection/individual tracks, software programs/individual routines products/components…
• Compatible with every other numbering scheme (UPC, ISBN, ISSN)
• Permanent (Once assigned, never changes. “A DOI is Forever.”)
• It protect to copyright.
• A central directory provides a level of indirection between the ID and its locations or services
DOI number format

- **10.1065/abc123defg** = the whole DOI
- **10.1065** = Publisher Prefix
- **abc123defg** = Suffix
  - item identifier
  - any format
  - naming authority (publisher)

- in use, a DOI is an opaque string (a “dumb number” - a **good** thing)
IDF

- **International DOI Foundation (IDF)** established October 1997
- Offices in Washington & Geneva
- Non-profit: supported primarily by membership fees
- Develops policies and governance procedures (“policy infrastructure”)
- Liaises with standards organizations internationally
- Manages the relationship with CNRI (as technology provider) via service contract
1. **Send DOI Query**
   - User PC
   - DOI Directory Server
   - DOI = Where to go next

2. **Forward Query to Publisher**
   - User PC
   - Browser
   - Publisher/Gateway
   - DOI = Where to go next

3. **Receive Object Information**
   - User PC
   - Browser
   - Publisher/Gateway
   - Object Information

**Working Process of DOI**
Image Editing Software

- Photoshop
- Coral draw
- Paint brush
Other Software for DL

- **Operating Systems:**
  - MS-Word, excel, power point etc.

- **Data Base Management System Software**
  - Oracle/SQL, MS-Access, Fox-Pro etc

- **Programming/ Scripting Language**
  - Html, Java, .Net, VB. C++ etc.

- **Firewall & Protection Software**
  - Firewall, Anti-virus, bioinformatics programs etc