

CHAPTER - V

DIGITIZATION OF ARCHIVES

MATERIAL : A MODEL

CHAPTER – V

DIGITIZATION OF ARCHIVES MATERIAL: A MODEL

5.1: INTRODUCTION:

Archives are the store houses of the rare documents, preserved from hundred of years having historical value. There are many challenges for its preservation and conservation. Maximum efforts are being taken to restore such national heritage for future generation whereas there are many limitation while its restorations. The preservation of such document on microfilm was other option. Now a days there are lot of changes in technology development, and now, present technology becomes outdated, replaced with scanning and putting image in soft format can be browsed by 24 x 7 days. Now when we think about preservation, the next question comes about its hardware and software requirements. In case of software, we have to select either commercial software, in house software or open source softwares which are freely available and can be downloaded for this purpose. Customization is possible due to the proprietary nature of commercial software. Successful examples of open source initiatives in digital libraries can be seen in the extensive use of the LAMP (Linux, Apache, MySQL, Php/Perl) environment, the growing use of Fedora and Dspace digital libraries, and the development of shared data management tools (the MARC perl module and MARC Edit are two examples) within the traditional library community. These application softwares have found success useful due to a number of factors responsible including community support/development, corporate support, and non-profit contributions.

The present project of Digital Archival Center is solely built by using the “Greenstone” open source digital library tool.

5.2: SCALABILITY ASPECT:

Digitizing of rare document presents challenges around its size: how many items from any given collections will be sufficient to create added value. The popular theory of the ‘Critical mass’ is one criterion for selection that shows up in nearly all the written guidelines for selection and is commonly noted in conversation.

5.3 : THE DIGITIZATION OF ARCHIVAL MATERIAL:

The digitization comprises of the following steps:

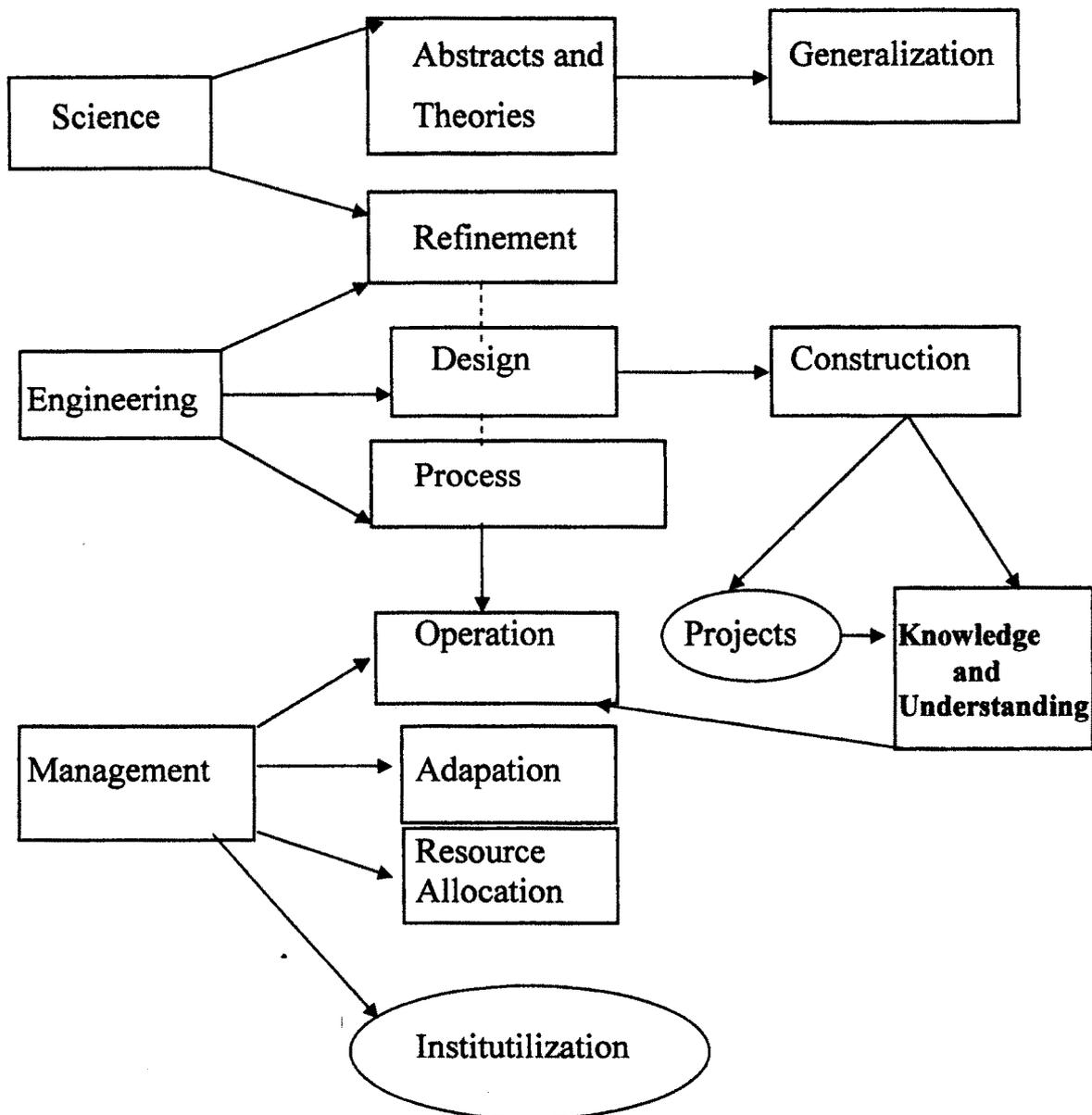
Pre Scanning process :

Planning for digitization required selection of documents from the Archival Department located at Dasara Chowk, Kolhapur and Satara Archives Collector Office, Satara.

- Handling of material: Digitization includes scanning and preservation of archives and there is risks to all material. For valuable rare documents of Chh. Shivaji III period, Queen Jijabai, Rajarshi Shahu Maharaj and Satara Archives. British Government period it requires special attention, because of the characteristics of the material and because electric power while scanning such documents the level of damage that would be high, whereas one should see that the risk acceptable is virtually zero.

5.3.1: Three (3) Facets of Developing Digital Libraries.

This table presents three facets for developing digital library as :-

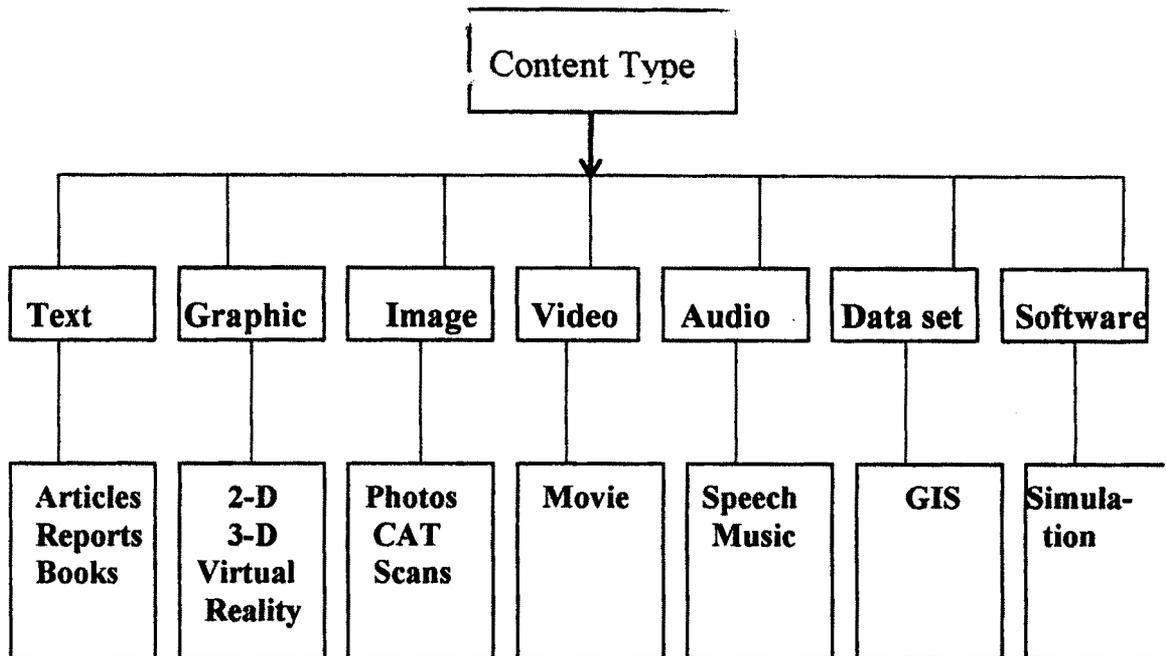


Source : ARIST V 36 (2002) P-507.

Because of increase in the knowledge and understanding, construction of numerous systems of many projects and employment of personnel in digital library laboratories and departments. Above chart is in general 3 facets of developing digital libraries and applicable in type on branch of academic environment applicable.

When we are going for digital library the concept needs to be cleared. Still there is lot of confusion among Digital Library experts. Following charts explains its concept.

Digital concept types and examples



Above chart shows that diversity of digital library contents which may be text, images, audio, video, computer programs or other forms. Newly created contents are often born digital, while older resources are typically digitized through a conversion process. Both must represent digitally, with too much attention towards its character coding, formats and file management which have dominated the discussion on digital library content.

While designing digital library for archives the issue are more complicated, such as documents are very old, brittle, difficult to scan or get image stored, and some administrative and technical issues are involved such as these documents are in the custody of Government or Government authorized body. The policy decisions are lying with both

the bodies. Therefore while designing digital library content selection its design issues such as technical and social frame work, understanding digital library concepts and its terminology, Architectural part and content storage are very important and needs to be given proper attention.

5.3.2 : Precautionary Measures:

While designing digital library following some of the precautionary measures need to be taken:

- **The size of the documents:** The size of document is very important and one has to handle safely, demanding care in handling at all stages including selection, assessment, extraction from storage, transport to the capture station, temporary storage before and after capture, mounting for capture, demounting after capture, and return to storage. Some times these documents are often found stored in rolled or folded manner which required special treatment before they can be presented flat on scanner to capture the image and they need to be returned to the same storage arrangements.
- **Digitization processes :** The document which are to be scanned with surface. Those documents are larger can be accommodated under the camera and usually requires the document to be held flat against a vertical surface, using magnetic strips, vacuum frame, removable self-adhesive strips, or something similar. All of these poses risk of damage, needing to be recognized, assessed and managed.

This is continuous process and as and when we scan documents, serving on services and providing browsable from for validation is required. For each scanned document tag has been allotted.

5.3.3 : Data Migration:

The following procedures followed during the pilot of Digital Library project to minimize risks to the rare documents:

- Xerox copies of the rare documents were used.
- Most of the selected documents were in good physical condition except for the blackened edges due to xeroxing. During data capture by digital camera, the documents were held by removable adhesive tapes at the corners.
- To minimise exposure to heat and UV associated with long exposure times, a specially made assembly was used with a lighting system consisting of four halogen lights of 1500W each with power condition to stabilize light intensities during-image capture.

5.4 Selection of Scanner:

For scanning of archive material for digitization, following points were considered for selection of scanners as :

1. Efficient use:

A planetary (overhead camera) scanner is more efficient than a face-down scanner because the rare document can be set face-up on the scanner and each page can be turned.

2. Correct ergonomics:

High-volume operators has to scan material several hours a day. A face-up scanner does not required because of constant lifting of heavy books, thus it is less likely to cause injury to the operator.

3. Preservation of rare documents:

While scanning rare documents, face-down flat bed scanner harm rare documents, because the spines of books must be pressed down upon the scan bed to obtain a clear image into the gutter of the books. Also, such scanner usually has cover that further press down upon the rare documents. With a face-up scanner the camera would make adjustments for the curvature of the book in the gutters of the spines, so that the books/rare documents need not be pressed open unnecessarily.

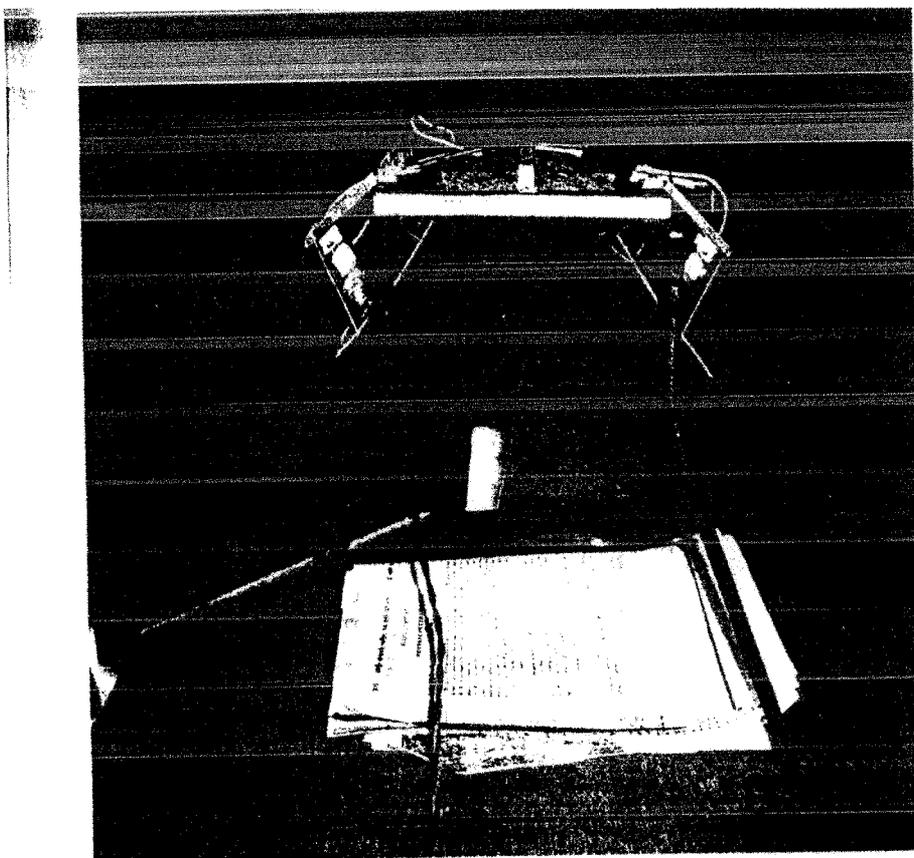
4. Scan-bed size:

Generally Libraries/archival centers have to scan a wide variety of materials from ordinary books to newspapers and rare documents, therefore it required a scan bed of at least 12" by 22 1/2" for half sheets of standard newspapers.

5.5 : Camera for capturing image of rare documents:-

Looking at the technical intricacies of focusing and clicking at normal incidence, the digital camera cant be used straight away. A specialized setup is required for camera at an appropriate distance from the documents. The focus and other parameters once fixed, need not be changed frequently with the usage of the setup.

5.1 : Figure : Image capture workstation based on digital camera.



Requirements for Digital Library Archival.

- Hardware
- Appropriate Software
- Storage – Database/File Structure.
- Accessibility – Java/Standard HTML
- Search Engines
- Security File Privilege
- Language.

This model is consists of general information pertaining towards digitization of archival material. Those who wanted to design model, this model will provide guidelines to the researchers, designers of Digital Library.

Now a days, there are lot of changes taken place in technology, software and hardware, whereas above mentioned Digital Library model is guidelines. We hope, the new study can be undertaken, based upon this model.

5.6. : Greenstone for building Digital Archival Center:

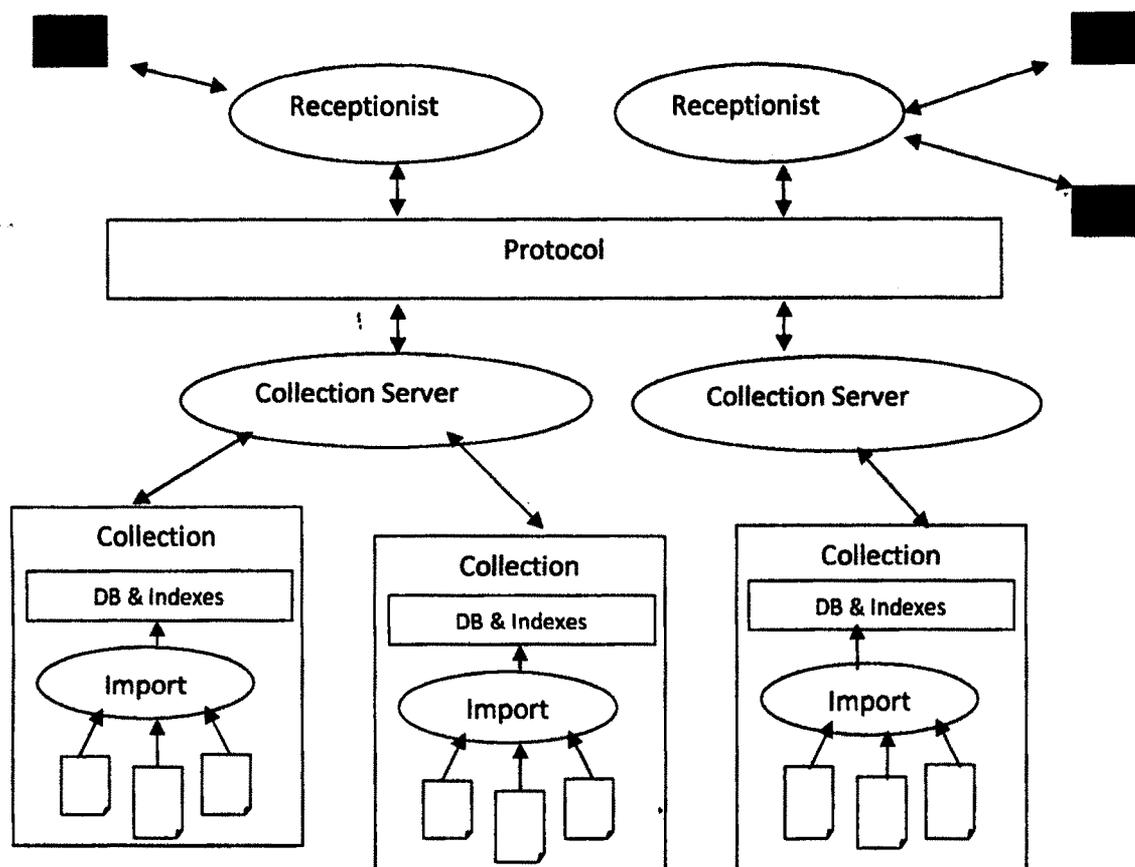
Open-source is today's freely available digital library software which has already been distributed and downloaded successfully works on the Linux operating system and the Apache web server. The present project is solely based on the Greenstone digital library software which is a suite for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM. Greenstone is open-source, multilingual software, issued under the terms of the GNU General Public License. It has gained popularity now a days because of the underlying economic advantages.

The salient features of the Greenston and relevant to the project are mentioned below:

- Greenstone: a new way of organizing information and publishing it on the Internet or on CD-ROM.
- Licensing: is open-source software, distributed under the terms of the GNU General Public License.
- Platforms required: Windows 3.1/3.11/95/98/Me/NT/2000, most distributions of GNU/Linux, Darwin (Mac OS X), Solaris. In this project it operates on the Window XP OS.
- Documents formats supported: PDF, PostScript, Word, RTF, HTML, Plain text, Latex, ZIP archives, Excel, PPT, Email (various formats), source code.

- Images and multimedia formats supported: Image formats, including GIF, JIF, JPEG, TIFF, Multimedia formats MP3 audio, Ogg Vorbis audio, and a generic plug-in that can be configured for audio formats, MPEG, MIDI, etc.
- Available reader interface: Arabic, Armenia, Bengali, Czech, Chinese (both simplified and traditional), Dutch, English, Farsi, French, Georgian, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Kannada, Kazakh, Kyrgyz, Latvian, Mongolian, Portuguese (BR and PT versions), Russian, Serbian, Spanish, Thai, Turkish, Ukrainian, Vietnamese.
- Available Librarian interface: English, French. Spanish, and Russian Architecture of the Greenstone is as shown in figure.

Figure 5.6.2 : Architecture of the Greenstone :



- There are two main parts viz. Receptionist and Collection Server:

The functions of receptionist are:

- Provides user interface
- Accept user input
- Send to appropriate collection server
- Accept results
- Dynamic page generation

5.6.2: Working with Greenstone:

Greenstone has two separate interactive interfaces, the Reader interface and the Librarian interface. End users access the digital library through the Reader interface, which operates within a web browser. The Librarian interface is a Java-based graphical user interface (also available as an applet) that makes it easy to gather material for a collection (downloading it from the web where necessary), enrich it by adding metadata, design the searching and browsing facilities that the collection will offer the user, and build and serve the collection. The collection building process is fairly straight forward and is guided by very well designed Graphical User Interface. The same is summarized in terms of steps as follows:

Creating a collection framework

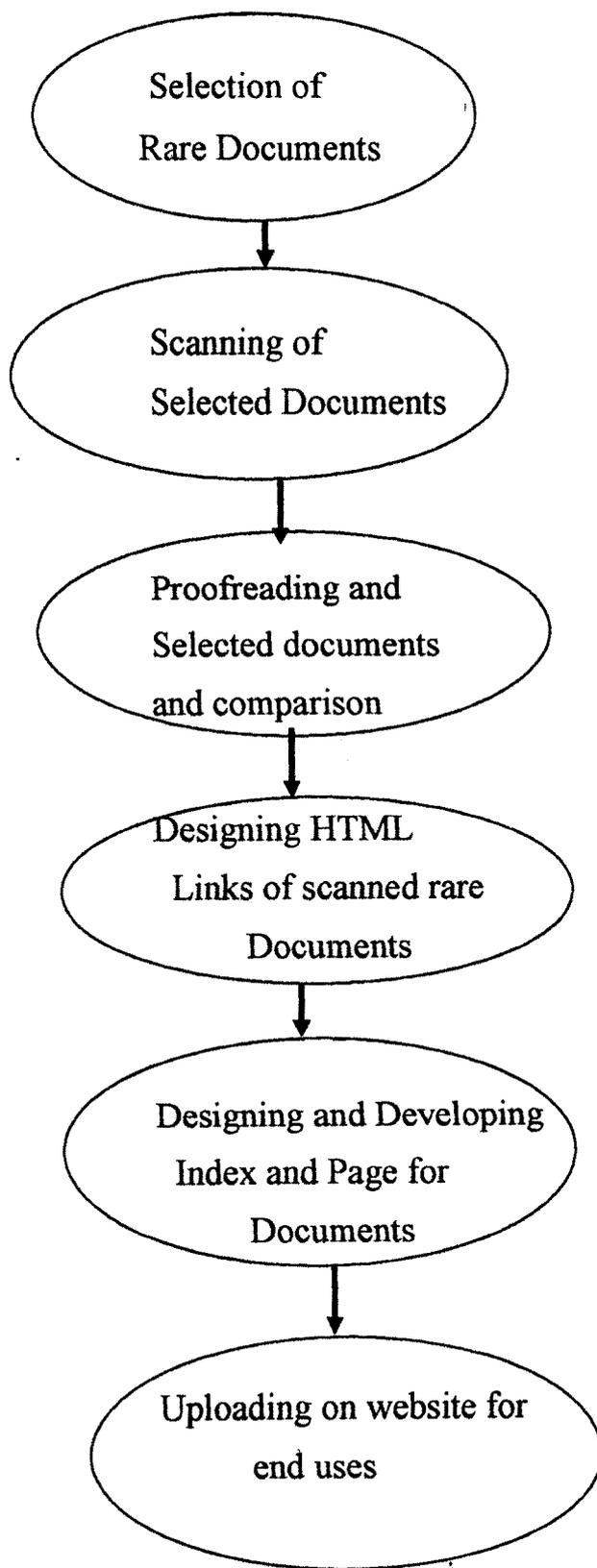
- Select documents
- Import documents
- Converts to internal XML format.
- Build collection
- Creates search indexes and browse listings

Information collections built by Greenstone combine extensive full-text search facilities with browsing indexes based on different metadata types. From the users/readers point of view, there are several ways for users to find information, although they differ between collections, depending up on the metadata available and the collection design. Typically the user can search for particular word that appear in the text, or within a section of a document, or within a title or section heading. They can browse

documents by title by just clicking on the displayed book icon to read it. Browsing documents is also possible by subject.

Subjects are represented by bookshelves by just clicking on a shelf the user can see the book. Where appropriate, documents come complete with a table of contents (constructed automatically): user can simply click on a chapter or subsection to open it, expand the full table of contents, or expand the full document. The following shows the diagram.

In digital Archival following steps are involved:



The costing of this model is not included as the cost of item is changing with item to item and now a days new developed technology available in the market with variety of model.

This model of digital library will provide general information to the libraries, information offices, digital library designers.

LIST OF ABBREVIATION

- | | | |
|----------------|---|--------------------------------------|
| 1. CAT | - | Computer Assisted Techniques |
| 2. CD-ROM | - | Compact Disk Read Only Memory. |
| 3. Email | - | Electronic Mail |
| 4. Excel | - | Executable Cell |
| 5. GIF | - | Graphic Interchange Format |
| 6. GIS | - | Geographical Information System. |
| 7. HTML | - | Hyper Text Mark up Language |
| 8. JIF | - | Jpeg Interchange Format |
| 9. JPEG | - | Joint Photographic Expert Group |
| 10. Mac OSX | - | Macintosh Operating System – 10 |
| 11. ME/NT | - | Millennium New Technology |
| 12. MIDI | - | Musical Instrument Digital Interface |
| 13. MPEG | - | Moving Picture Expert Group. |
| 14. MP3 | - | Mpeg Audio Player 3 |
| 15. MYSQU | - | My Structured Query Language |
| 16. PDF | - | Portable Document Format |
| 17. Php | - | Preprocessor |
| 18. Postscript | - | (Printer Type) |
| 19. PPT | - | Power Point Transference |
| 20. RTF | - | Rich Text Format |
| 21. SQL | - | Structural Query Language |
| 22. TIFF | - | Tagged Image File Format |
| 23. UV | - | Unique / Universal Voltage. |
| 24. Windows XP | - | Windows Xtreme Performance |
| 25. XML | - | Extensible Markup Language |
| 26. ZIP | - | Zicxac Inline Pin |