### Block

2

<table>
<thead>
<tr>
<th>UNIT 5</th>
<th>Introduction to Digital Library</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>UNIT 6</td>
<td>Digitisation Process</td>
<td>24</td>
</tr>
<tr>
<td>UNIT 7</td>
<td>Creating Digital Libraries Using DSpace</td>
<td>42</td>
</tr>
<tr>
<td>UNIT 8</td>
<td>Creating Digital Libraries Using GSDL</td>
<td>60</td>
</tr>
</tbody>
</table>
Programme Design Committee

Prof. Uma Kanjilal (Chairperson)
Faculty of LIS, SOSS, IGNOU

Prof. B.K. Sen, Retired Scientist NISCAIR,
New Delhi

Prof. K.S. Raghavan, DRTC
Indian Statistical Institute, Bangalore

Prof. Krishan Kumar, Retired Professor
Dept. of LIS, University of Delhi, Delhi

Prof. M.M. Kashyap, Retired Professor
Dept. of LIS, University of Delhi, Delhi

Prof. R.Satyanarayana
Retired Professor, Faculty of LIS, SOSS
IGNOU

Dr. R. Sevukan
(Former Faculty Member) Faculty of LIS
SOSS, IGNOU

Prof. S.B. Ghosh, Retired Professor
Faculty of LIS, SOSS, IGNOU

Prof. T. Viswanathan
Retired Director, NISCAIR, New Delhi

Dr. Zuchamo Yanthan
Faculty of LIS, SOSS, IGNOU

Prof. K.S. Raghavan, Retired Scientist
NISCAIR, New Delhi

Prof. Neena Talwar Kanungo
Faculty of LIS, SOSS, IGNOU

Conveners:

Dr. Jaideep Sharma
Faculty of LIS, SOSS, IGNOU

Prof. Neena Talwar Kanungo
Faculty of LIS, SOSS, IGNOU

Programme Coordinators

Prof. Jaideep Sharma and Prof. Neena Talwar Kanungo

Course Coordinator

Prof. Uma Kanjilal

Course Preparation Team

<table>
<thead>
<tr>
<th>Unit No</th>
<th>Unit Writer</th>
<th>Course Editor</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-8</td>
<td>Prof. Uma Kanjilal</td>
<td>Prof. Uma Kanjilal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Print Production</th>
<th>Secretarial Assistance</th>
<th>Cover Design</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr. Manjit Singh</td>
<td>Ms. Sunita Soni</td>
<td>Ms. Ruchi Sethi</td>
</tr>
<tr>
<td>Section Officer (Pub.)</td>
<td>Mr. Vijendra</td>
<td>Web Designer</td>
</tr>
<tr>
<td>SOSS, IGNOU, New Delhi</td>
<td>SOSS, IGNOU</td>
<td>E Gyankosh, IGNOU</td>
</tr>
</tbody>
</table>

November, 2014 (Second Revised Edition)

© Indira Gandhi National Open University, 2014

ISBN-978-81-266-0112-9

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

“The University does not warrant or assume any legal liability or responsibility for the academic content of this course provided by the authors as far as the copyright issues are concerned.”

Further information on Indira Gandhi National Open University courses may be obtained from the University’s office at Maidan Garhi, New Delhi-110 068 or visit University’s website http://www.ignou.ac.in

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at:
Introduction

The automation of the library during past few decades have been mainly focusing on creation of surrogate records of printed documents available in a library or for providing services through secondary databases held locally on CD ROM or magnetic tapes. The scope and functions of integrated library packages, till recently, were essentially restricted to providing access to documents at bibliographic level. The new versions of, integrated library packages, however, tend to provide additional features and functionalities akin to digital libraries. However, since the automated systems till recently provided only bibliographic information, users had to depend heavily on physical collection available either in their institutional library or on inter-library loan from other libraries for references retrieved from the secondary services.

Digitisation is the process of converting the content of the physical media (text, audio, video) into digital media. For printed material an image of the physical object is captured using a scanner or digital camera and converted into a digital format that can be stored electronically and accessed via computer or mobile devices. For audio and video material encoders are used for digitisation.

Once document and media content are digitised, these need to be archived and made accessible to the users. For this, tools for organising digital collection are needed. DSpace and Greenstone Digital Library Software are two major application being used by libraries world over to organising digital collection and building digital libraries.

This block has four Units. **Unit 5** on Introduction to Digital Library provides an overview on the concept of digital library and major worldwide initiatives. **Unit 6** discusses the Digitisation Process. **Units 7 and 8** deal with Creating Digital Libraries Using D-Space and GSDL respectively.
Digitisation and Digital Libraries – DSpace and GSDL
UNIT 5  INTRODUCTION TO DIGITAL LIBRARY

Structure
5.0  Objectives
5.1  Introduction
5.2  Concept
5.3  Types of Digital Libraries
5.4  Major Digital Library Initiatives
5.5  Future Trends
5.6  Summary
5.7  Answers to Self Check Exercises
5.8  Keywords
5.9  References and Further Reading

5.0  OBJECTIVES

After going through this Unit, you will be able to:

• understand the basic concept, and need for digital libraries;
• explain different types of digitisation; and
• discuss future trends of digital libraries.

5.1  INTRODUCTION

Digital age has brought a tremendous change in the way information is stored and accessed. It is marked by three distinct features: abundance, currency and easy access of information. This has brought about a change in the concept of libraries, their collection and services. Many new terms viz., ‘digital libraries’, ‘libraries without walls’, ‘virtual libraries’ are emerging to describe the libraries of present day age.

The term ‘digital library’ is a shift from the earlier term electronic library which was used for the last two decades to describe the book-less library which relies on telecommunication and computers to provide users with whatever information they need. A digital library is popularly viewed as an electronic version of a library where storage is in digital form, allowing direct communication to obtain material and copying it from a master version. It combines technology and information resources to allow remote access, breaking down the physical barrier between resources. In Wilensky’s view “the digital library will be a collection of distributed information services, producers will make it available, and consumers will find it through the automated agents”. In this model it appears that the traditional libraries will have no role to play. How far this will be true only time can tell.

In the early stages of development of digital libraries the main focus was on providing dial up access to Online Public Access Catalogues (OPAC). The term however evokes different meaning for different people. To some it may simply mean computerisation of the traditional library system. To those with library science
Digitisation and Digital Libraries – DSpace and GSDL

background it means doing things in a new way, using new type of information resources, new approach to acquisition, new methods of storage and preservation, new approaches to classification and cataloguing, new ways of interaction with the patrons with more reliance on electronic system and networks. As it stands today, most libraries in the developed countries have their own homepages providing links to local information, electronic databases, bibliographic as well as full text, apart from its own online system of collection and services.

Digital libraries in future will not be a standalone version. The explosive growth in networked connectivity and rapid advances in computing power are replacing the older notions of standalone information utilities with newer notions of integrated digital libraries. The integrated digital library creates a shared environment linking everything from personal collection, collection of conventional libraries and large databases spread all over the world.

In the recent years the term ‘virtual library’ is becoming more popular. It is being used to describe libraries that provide access to digital information using variety of networks, specifically the internet and the World Wide Web, irrespective of place and time. According to Gilbert “it is an aggregate of libraries or literature bases, the catalogue or bibliographies of which are accessible electronically (e.g. with a personal computer) and of which some may offer document ordering and delivery services. The center of the virtual library is by definition the individual user, or his/her work station”. Thus in the present day context virtual library is the convergence of a number of concepts: electronic browsers, online catalogues and literature bases, and empowerment of the end users.

In Toren and Czech’s view, libraries in future will become icons on the screen and library buildings will function as book warehouses. The future implication of such a situation needs to be contemplated seriously.

5.2 CONCEPT

Defining Digital Libraries

The term “digital library” is the most recent in a long series of names for a concept that has been written about nearly as long as the development of the first computer: a computerised “library” that would supplement, adds functionality, and even replaces traditional libraries.

In comparison to traditional libraries, digital libraries provide efficient and qualitative services by collecting, organizing, storing, disseminating, retrieving and preserving the information. Digital libraries support preservation besides making information retrieval and delivery more comfortable. It provides online access to historical and cultural documents whose existence is endangered due to physical decay. The major areas which offer digital libraries great exploitation are: Information retrieval, multimedia database, data mining, data warehouse, on-line information repositories, image processing, hypertext, World Wide Web and Wide Area Information Services (WAIS).

Digital libraries necessarily include a strong focus on the management of digital content, just as traditional libraries have focused for long on the management of content in physical forms. Most of the digital content that is being managed includes human language, either in the form of character-coded electronic text, scanned versions of
printed or handwritten text, or digital representations of human speech. Language technology therefore plays a major role in managing digital content. This comes as no surprise, of course. Digital libraries today make good use of what we know about searching large collections, and techniques such as machine-assisted indexing are employed increasingly often as we strive to extend our reach to progressively larger collections. But we are on the verge of a new era, one in which our machines will learn from what we do and then apply those capabilities to enable the management of digital content at a far larger scale than we could ever hope to do ourselves.

Few advantages of digital libraries according to Haddouti are:

- User can access the information anywhere
- Reduces bureaucracy by providing access to the information
- The information is not necessarily located in same place
- Understanding the catalogue structure is not necessary
- Cross references to other documents speed up the work of users
- Full text search
- Protected information source
- Wide exploration and exploitation of the information

The knowledge dissemination is an integral part of success story of popularity of creating digital libraries. The aim is to provide universal access to human knowledge, and given the advancement of digital storage and communications this goal is now achievable.

**Distributed Models**

Libraries are increasingly adopting distributed models for information access and management, and more often use open and collaborative models for developing library content and services. With the incorporation of open models and distributed technologies, the libraries have the potential to get more involved in knowledge creation, dissemination, and use. In reference to libraries, the creation and dissemination of knowledge—in ways that represent the library’s contributions more broadly and that intertwine the library with the other stakeholders in these activities. The library becomes a collaborator within the academy, yet retains its distinct identity.

**Open Paradigms and Models**

There is new trend emerging as Open Source movement— the concept of collaborative software development with developers sharing the source code — reflects a fundamental shift away from proprietary software and systems. These open models are appearing in new applications areas such as the Open Knowledge Initiative to share learning technologies. The increasing interest in open models is leading towards more generalized acceptance of collaborative development and sharing of intellectual goods and services. Cyber law experts suggest that the creation of a “commons,” wherein the free exchange of ideas and collaboration prevail, is fundamental to an open society. Themes of openness and collaborative exchange have also emerged in the context of publishing, particularly with respect to the relationship between authors and commercial publishers. As information becomes more distributed and open models of exchange become more common, the library’s relationship with content creators, publishers, and consumers will change. In these open trends there is evidence of a shift from publication as product to publication as process. When content is available
Digitisation and Digital Libraries – DSpace and GSDL

In such a shape that can be enhanced or supplemented over time, it becomes more dynamic and the “versions” become more cumulative. Few people forecast this shift as the ultimate challenge to current copyright law. Such a shift will have significant impact on organisations whose current role is to manage publications in both traditional and digital forms. As this shift continues, there are likely to be further changes in the library’s information management functions.

In this second phase in the evolution of library roles, the library starts to engage in collaboration as a strategy to address its core mission of building collections, maintaining access, and providing service. As responsibilities for content and services become more distributed, models of central control give way to new mechanisms for coordination and collaboration. Ultimately, the processes of scholarly communication become as critical as traditional publication products.

**Digital Collections Vs Digital Library**

In the last decade substantial progress has been made in creating large-scale digital collections. It is extremely important to distinguish digital collections from digital libraries. There is no clear definition about what exactly constitutes a digital library. Digital collections are “raw content,” while “digital libraries [are] the systems that make digital collections come alive, make it usefully accessible, useful for accomplishing work, and connect them with communities.” The collections gain value only when these are surrounded by a matrix of content and interpretation that makes them useful. Therefore it should be ascertained that we develop digital libraries, not just digital collections.

Care should be taken to surround collections with appropriate metadata supplying context and interpretation, to develop synergy. It is the time to “build massive, comprehensive digital collections that scholars, students, and other researchers can use with more ease than they use the book-based collections.”

Three general characteristics of the digital library of the future are:

- A comprehensive collection of resources important for Scholarship, teaching, and learning;
- Readily accessible to all types of users
- Managed and maintained by professionals

The information explosion, the wide bandwidth data networks and the potential of Internet-based technologies - such as the Web - make digital libraries one of the important application areas of computer science.

**Self Check Exercise**

**Note:** i) Write your answers in the space given below.

 ii) Check your answers with the answers given at the end of this Unit.

1) Discuss three general characteristics of the digital library of the future.

........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
5.3 TYPES OF DIGITAL LIBRARIES

Digital libraries can be grouped in different ways. They can be classified by origin, such as digital libraries developed in the USA as part of DLI 1 and DLI 2 (the Digital Library Initiatives), digital libraries developed in the course of the eLib (Electronic Libraries) programme in the UK, digital libraries built by individual institutions, digital libraries that are part of national libraries, digital libraries that are part of universities; or by period, by country of origin, and so on.

- early digital libraries, e.g. ELINOR, Gutenberg
- digital libraries of institutional publications, e.g. ACM, IEL
- digital library developments at national libraries, e.g. the British Library, Library of Congress (THOMAS), Digital Library of Canada
- digital libraries at universities, e.g. Berkeley Digital Library SunSITE Bodleian Library Digital Library Projects, California Digital Library, DIGILIB, iGEMS and SETIS
- digital libraries of special materials, e.g. Alexandria, Informedia, Grainger Engineering Library
- digital libraries as research projects, e.g. GDL, NCSTRL, NDLTD
- digital libraries as hybrid library projects, e.g., HeadLine.

Self Check Exercise

Note: i) Write your answers in the space given below.
   ii) Check your answers with the answers given at the end of this Unit.

2) Classify different types of digital libraries.

5.4 MAJOR DIGITAL LIBRARY INITIATIVES

- The British Library’s Digital Libraries Programme
  (http://www.bl.uk/aboutus/stratpolprog/digi/dom/index.html)

  The Digital Libraries Research Programme at British Library Research and Innovation Centre (BLRIC) is establishing a digital library information service based on the British library collections.
• THOMAS - Library of Congress Digital Library (http://thomas.loc.gov/)
The Library of Congress Digital Library, Thomas was launched in January 1995, at the inception of the 104th Congress to make federal legislative information freely available to the public.

• California Digital Library (http://www.cdlib.org/)
The California Digital Library was established in 1997 at the University of California. It supports the University of California libraries in their mission of providing access to the world’s knowledge for the UC campuses and the communities they serve. The CDL also maintains its own distinctive programs emphasizing the development and management of digital collections, innovation in scholarly publishing, and the long-term preservation of digital information.
• **Google Digital Library of Alexandria**
  
  Google announced the library scanning project in December 2004. It has four library partners viz. Stanford University, Oxford University, New York Public Library and University of Michigan. The major publishing houses like McGraw-Hill and Penguin Group have sued Google for scanning books without permission.


• **Gutenberg** ([http://promo.net/pg/](http://promo.net/pg/))
  
  The project Gutenberg began in 1971 at the Materials Research Lab, the University of Illinois. The prime objective of this project was to facilitate the world’s great literature to electronic versions for the public access.
• **The IEEE Electronic Library**
(http://www.ieee.org/portal/innovate/products/research/ieee_iel.html)

The IEEE digital library is the gateway to valuable, cutting-edge research, standards and educational courses with more than two million articles. It offers 100% full-text searchable content with full-page PDF images of all IEEE articles, papers and standards.

• **International Children’s Digital Library (ICDL)** (http://en.childrenslibrary.org/)

The ICDL was created by an interdisciplinary research team at the University of Maryland in cooperation with the Internet Archives. This was established to create a collection of more than 10,000 books in at least 100 languages that is freely available to children, teachers, librarians, parents, and scholars throughout the world via the Internet.
• **The New Zealand Digital Library Project** ([http://nzdl.sadl.uleth.ca/cgi-bin/library.cgi](http://nzdl.sadl.uleth.ca/cgi-bin/library.cgi))

The New Zealand Digital Library Project is a research programme at the University of Waikato. The main objective of this project is to develop the underlying technology for digital libraries and make it available publicly.

• **Digital Library of the Commons** ([http://dlc.dlib.indiana.edu/](http://dlc.dlib.indiana.edu/))

The Digital Library of the Commons (DLC) is running on Eprints2, which provides free access to an archive of international literature on the commons, common-pool resources and common property. Features for authors and readers include advanced searching; browsing by region, sector, and author name; an author submission portal for uploading a variety of document formats; and a service that uses email to alert subscribers to new documents in their area of interest.
• **Perseus Digital Library** ([http://www.perseus.tufts.edu/hopper/](http://www.perseus.tufts.edu/hopper/))

Perseus is an evolving digital library, to bring a wide range of source materials to as large an audience as possible.

• **The German Digital Library Programme GLOBAL INFO**

The German Digital Library Programme GLOBAL INFO is funded by the federal ministry for education and research from 1998. The main objective of this initiative is to provide optimal access to the world-wide electronic and multimedia information on full texts, literature references, factual databases and software.


• **The Sydney Electronic Text and Image Service (SETIS)** ([http://setis.library.usyd.edu.au/](http://setis.library.usyd.edu.au/))

SETIS was launched in 1995 at the University of Sydney. It provides access to a large number of networked and in-house full text databases. It also engaged in a number of text and image creation projects.
- **The Berkeley Digital Library** ([http://sunsite.berkeley.edu/](http://sunsite.berkeley.edu/))

The Berkeley Digital Library project began as an inter-agency, academic teaming to research collaboration techniques. It continues and in currently developing the tools and technologies to support highly improved models of the “scholarly information life cycle”. The goal is to facilitate the move from the current centralized, discrete publishing model, to a distributed continuous, and self-publishing model. It provide access to a large variety of scholarly publications.

- **Informedia Digital Video Library** ([http://www.informedia.cs.cmu.edu/](http://www.informedia.cs.cmu.edu/))

This is a project at Carnegie Mellon University and the overarching goal of the Informedia initiative is to achieve machine understanding of video and film media, including all aspects of search, retrieval, visualization and summarization in both contemporaneous and archival content collections. The Informedia-II seeks to improve the dynamic extraction, summarization, visualization and presentation of distributed video.

The Networked Digital Library of Theses and Dissertations is an international organisation dedicated to promoting the adoption, creation, use, dissemination and preservation of electronic analogues to the traditional paper-based theses and dissertations. This contains information about the initiative, how to set up Electronic Thesis and Dissertation (ETD) programmes, how to create and locate ETDs, and current research in digital libraries related to NDLTD and ETDs.


This digital library was created to give world wide access to collection of memorabilia devoted to Sir Don Bradman and held by the Mortlock State Library of South Australia. It contains biographical information about Bradman, a digital exhibition of artifacts, and a series of scrapbooks covering the years 1925-26 to 1948-49, containing press cuttings, notes and photographs.
• The University of Adelaide Digital Library (http://digital.library.adelaide.edu.au/)
  The Digital Library undertakes projects aimed at enhancing online access to information for their members. This provides access to exam papers available online, Australian digital theses collection and e-books available at Adelaide.

• National Science Foundation Digital Library (http://nsdl.org/)
  The National Science Foundation Digital Library at the University of Texas at Austin is a dynamic archive of information on digital morphology and high-resolution X-ray computed tomography of biological specimens.
  The Cuneiform Digital Library initiative represents the efforts of an international group of Assyriologists, museum curators and historians of science to make available through the internet the form and content of cuneiform tablets dating from the beginning of writing until the end of the pre-Christian era.

- **UQ eSpace** ([http://espace.library.uq.edu.au/](http://espace.library.uq.edu.au/))
  UQ eSpace is the University of Queensland’s institutional digital repository for publications, research, and teaching materials. Deposited material covers a very wide range of subjects and disciplines. This also holds the electronic full text of many peer-reviewed published articles and conference papers, book chapters, theses and other forms of written research from UQ academic staff and students.
• **Traditional Knowledge Digital Library**
  
  (http://www.tkdl.res.in/tkdl/langdefault/common/home.asp?GL=Eng)

  The Traditional Knowledge Digital Library is a well known Indian digital library initiative being implemented by the National Institute of Science Communication and Information Resources (NISCAIR). The major objective is to provide information on the Indian system of medicine such as Ayurveda, Unani, Siddha, Yoga, Naturopathy and Tribal Medicine.

• **The Digital Library of India (DLI)** (http://dli.iit.ac.in/)

  The Digital Library of India is the greatest digital library initiative in the country. DLI is a part of Universal Digital Library (UDL) and Million Books Projects, coordinated by the Carnegie Mellon University, USA.
The Archives of Indian Labour (http://www.indialabourarchives.org/)

The Archives of Indian Labour is a collaborative project of V.V.Giri National Labour Institute and the Association of Indian Labour Historians. The main objective is to preserve and make accessible archival documents on the working class of India.

5.5 FUTURE TRENDS

Although the term digital library is used widely in the literature, a new term, ‘hybrid library’, appeared in the course of digital library research in the UK. A hybrid library has been defined as a library where digital and printed information resources co-exist and are brought together in an integrated information service accessible locally as well as remotely (HyLife, 2002a). A number of researchers believe that for the foreseeable future we shall live in the world of hybrid libraries that will integrate traditional libraries with the emerging digital ones (for example, Oppenheim and Smithson, 1999; Pinfield et al., 1998; Rusbridge, 1998). Pinfield et al. (1998) comment that the hybrid library is on the continuum between the conventional and digital library, where electronic and paper-based information sources are used alongside each other. Rusbridge (1998) suggests that a hybrid library brings a range of technologies from different sources together, and integrates systems and services in both the electronic and print environments. He further argues that ‘the name hybrid library is intended to reflect the transitional state of the library, which today can neither be fully print not fully digital’.

There are numerous areas of research related to the historic interests of the digital library community that are at the crossroads of technology and social science and which will demand investment and attention in the coming years; many of these are natural extensions and elaborations of the collaborations initiated by the past decade of digital library research programs. Below mentioned are some of the driving force areas for future of digitisation.
• Personal information management. As more and more of the activities in our lives are captured, represented and stored in digital form, the questions of how we organize, manage, share, and preserve these digital representations will become increasingly crucial. Among the trends lending urgency to this research area are the development of digital medical records (in the broadest sense), e-portfolios in the education environment, the overall shift of communications to email, and the amassing of very large personal collections of digital content (text, images, video, sound recordings, etc.)

• Long term relationships between humans and information collections and systems. This is related to personal information management, but also considers evolutionary characteristics of behaviour, systems that learn, personalization, system to system migration across generations of technologies, and similar questions. This is connected to human-computer interface studies and also to studies of how individuals and groups seek, discovers, use and share information, but goes beyond the typical concerns of both to take a very long time horizon perspective.

• Role of digital libraries, digital collections and other information services in supporting teaching, learning, and human development. The analysis here needs to be done not on a relatively transactional basis (i.e. how can a given system support achievement of a specific curricular goal in seventh grade mathematics) but how information resources and services can be partners over development and learning that spans an entire human lifetime, from early childhood to old age.

• Active environments for computer supported collaborative work offer the starting point for another research program. These environments are called for, under the term “colaboratories”, by the various cyber infrastructure and e-science programs, but have much more general applicability for collaboration and social interactions. From one perspective, these environments are natural extensions of digital library environments, but at least some sectors of the digital library community have always found active work environments to be an uncomfortable fit with the rather passive tradition of libraries; perhaps here the baggage of “digital libraries” as the disciplinary frame is less than helpful. But there is a rich research agenda that connects literatures and evidence with authoring, analysis and re-use in a much more comprehensive way than we have done to date; this would consider, for example, the interactions between the practices of scholarly authoring and communication on one hand, and on the other, the shifting practices of scholarship that are being recognized and accelerated by investments in e-science and e-research.

5.6 SUMMARY

Libraries have always played a significant role in society, and digital libraries with the promise of breaking the barriers of geographical distance, language and culture, have a potentially even more significant social role. Digital libraries will not only change our reading and information use habits, they are also going to bring major changes in the economic models of information generation, distribution and management functions.

A tremendous amount of research and development activity has gone into the study of digital libraries. Many issues have been addressed and problems have been partly or fully resolved. Researchers from a variety of disciplines, such as library and
information science, computer science and engineering, social sciences and humanities are working closely together to look into the myriad of unresolved issues.

For exploiting the benefits of Digital Library in Indian languages there is urgent need of tools and applications such as OCRs and Machine Translation systems so that user can take benefit of reading rare classics published in any language and researchers are able to use these tools for their linguistic research. This parallel aligned corpus development is first attempt in context of Indian languages. This is the initiation of several efforts which will follow the trend of enhancing the research in the field of Computational Linguistics. The parallel corpus as a Translation Memory (TM) will be a valuable source in improving the translation system and translators’ efficiency.

It will boost the development of Lexical and Terminology databases with the combination of Quantitative and Qualitative Analysis of Text. Text Analyzer is a new kind of tool which is helpful in lexicography, knowledge acquisition, language and writing variation studies. Digital libraries creation have been a good test bed for OCR’s and now that the world is moving towards speech to speech translation all these tools together will help building one for Indian languages.

5.7 ANSWERS TO SELF CHECK EXERCISES

1) Three general characteristics of the digital library of the future are:
   • A comprehensive collection of resources important for Scholarship, teaching, and learning
   • Readily accessible to all types of users
   • Managed and maintained by professionals.

2) Digital libraries can be classified broadly into:
   • early digital libraries, e.g. ELINOR, Gutenberg
   • digital libraries of institutional publications, e.g. ACM, IEL
   • digital library developments at national libraries, e.g. the British Library, Library of Congress (THOMAS), Digital Library of Canada
   • digital libraries at universities, e.g. Berkeley Digital Library SunSITE Bodleian Library Digital Library Projects, California Digital Library, DIGILIB, iGEMS and SETIS
   • digital libraries of special materials, e.g. Alexandria, Informedia, Grainger Engineering Library
   • digital libraries as research projects, e.g. GDL, NCSTRL, NDLTD
   • digital libraries as hybrid library projects, e.g., HeadLine.

5.8 KEYWORDS

Hybrid library : Libraries containing a mix of traditional print library resources and the growing number of electronic resources.

OCR : Optical Character Recognition, or OCR, is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data.
**Open Knowledge Initiative**: The Open Knowledge Initiative (O.K.I.) is an open and extensible architecture for learning technology specifically targeted to the needs of the higher education community.

**Open Source Movement**: A broad-reaching movement of individuals who support the use of open source licences for some or all software. Open source software is made available for anybody to use or modify, as its source code is made available.

### 5.9 REFERENCES AND FURTHER READING


http://dspace.iimk.ac.in/bitstream/2259/252/1/05-mgs-ps-paper.pdf

http://www.dlib.org/dlib/july05/lynch/07lynch.html


6.0 OBJECTIVES

After going through this Unit, you will be able to:

• Understand the digitisation process of text, audio and video;
• Know different types of file formats; and
• Explain the file compression process.

6.1 INTRODUCTION

A digital library may contain materials that are born digital, such as e-journals and e-books, or may contain materials that were originally produced in another form but subsequently digitised. The process of digitising materials involves different steps depending upon material, technology and requirement. Various technical issues, like hardware and software, file formats and file compression and then the post processing requirements for making the digitised file accessible to end-user will be discussed.

6.2 DIGITISATION OF PRINT BASED DOCUMENTS

Once you have taken decision as to what needs to be digitised, the first step is to capture the documents available in print or analogue form for conversion into digital form. In the case of print based material, it is the hard copy of the document which needs to be scanned and digitised. The hard copy can be a paper based document, microforms or projection slides. For audio/video media conversion is done from the analogue form to digital formats. Capturing devices for print based material include scanners and digital cameras attached with a computer. For audio/video material
appropriate players like music system or VHS players attached with a computer will be required. The computer that you use must have appropriate audio/video capture cards in it.

6.2.1 Capturing Print Based Document

For converting hard copies into machine readable form there are three options available for a library:

1) Keying in the text
2) Scanning and capturing them as image files
3) OCR the files

Fresh keying in costs ten times more than scanning and saving as image files. However, if you are converting them into OCR, then some costs will be involved in error correction and editing.

Scanning technology has improved considerably over the years in terms of speed and resolution. There are several types of scanning devices available in the market now. Scanners come in three broad price ranges: i) low cost flatbed scanners or hand held devices, ii) low end sheet feeder type, iii) high end professional or book scanners. Scanning machines are generally based on Charge Couple Device (CCD) technology. In low end devices Contact Image sensor (CIS) technology is used generally whereas in some high end devices Photo Multiplier Tube (PMT) technology is used. PMT based drum scanners produce very high quality images which come at a high cost. CMOS (Complementary Metal Oxide Semiconductor) is another sensing technology that is used in hand held digital cameras.

The scanners operate by shining light on the document and directing the reflected light through a series of mirrors and lenses onto photo sensitive element. The photo sensitive element could be CCD, CIS or PMT based technology depending on the type of the scanners. Light sensitive photosites arrayed along the photosensitive element are converted into electronic signals which finally processed into digital image.

Self Check Exercise

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

1) Enumerate three options for converting hard copies into machine readable form.

........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
........................................................................................................................................
The **steps for scanning** a document

**Step 1: Place the document on the scanner bed**

Here we show the process using the Konica Minolta PS 7000 book scanner, which is a superior system for scanning large-sized books, artwork, ledgers and other bound materials. It is a face-up scanning system.

![Fig. 6.1: Book Scanner](image)

**Step 2: Open the Adobe Acrobat**

Click on **File>Import>Scan...**

Fill in the information for device, format and destination in the dialogue box that appears.
Digitisation Process

To scan the documents click on the **Scan All** option. From the **Minolta PS7000 Scanner Setup Dialog Box** that appears.

Click on **Done** option from the **Minolta PS7000 Scanner Setup Dialog Box** which shows the file like this:
Save the file as PDF version giving .pdf extension. To change the resolution, Click on Scan Setting >> Resolution (DPI) from the Minolta PS7000 Scanner Setup Dialog Box. To change the Scan Area click on, Scan Setting >> Scan Area. You can also change the Brightness and Contrast of the scanned file by using the drag button from the right panel. If you want to change the Image Type then click on Scan Setting >> Image Type. You can also change the Brightness and Contrast of the scanned file by using the drag button from the right panel. Scanned pages can be saved as individual files or as a complete document by appending them to the current document while scanning.

6.2.2 Digitising

The process of digitisation involves capturing the physical or analogue object through devices like scanners, digital camera, recorder etc., converting them into numerical values in bits and bytes which enables them to be read electronically.

Digitisation of text is possible either through text transcription or using optical character recognition method. Text transcription can be through keying in the text using a keyboard or by voice recognition software. Keyed in text are saved in ASCII format which do not replicate the structure and format of the original text.

OCR software converts image of text captured by a scanner into computer editable text which a word processor can read. The software tries to match the image of each letter against the pattern it recognizes making use of the stored knowledge about the shapes of individual characters. The OCR software has options for either storing the text and graphics in their original layout or converting them into ASCII or word processing format. Omnipage Pro and ABBYY Fine Reader are two commonly used OCR software.

After OCR, you can export the resulting text to a variety of word-processing, page layout, and spreadsheet applications. It also provides the option to save it directly as a PDF file.

Self Check Exercise

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

2) Name two commonly used OCR software.

................................................................................................................................................
................................................................................................................................................
................................................................................................................................................
................................................................................................................................................
................................................................................................................................................

To perform OCR with automatic processing the following steps are to be followed:

1) Select all settings needed to process pages. Do this in the following:

- Get Pages drop-down list
- Layout Description drop-down list
• Export Results drop-down list
• Options dialog box panels (Tools menu)

2) Click the button or Click on the shortcut icon

Start button with 1-2-3 selected in the Workflow drop-down list. Your pages will be acquired, auto-zoned and recognized one after the other. Proofing will start if you requested it. When proofing and/or recognition are finished, an export dialog box appears. Select the destination, file type and file name to save the file.

To manually perform the OCR, follow the steps given below.

1) Scan the document as an image
   • Launch Omni Page Pro. Start>Programs>Scansoft Omnipage Pro
   • The Program will open with the toolbar shown below.

   - Place the document to be scanned in the scanner.
   - Click the icon above the Scan Color menu. The Program will scan the document.

2) Select for Recognition
   • Once the document opens up in Omni Page, draw a box around the text you want.
   • You can categorize the objects on the scanned image into text, table or image by selecting the appropriate option on the side toolbar.
You can skip this step if you want OmniPage to automatically perform the OCR and select the regions.

- Then choose the file to convert OCR by clicking on the 123 option (Start Button).

- It will give this type of screen to browse the file from any location.

3) **Perform OCR and Proof Read**

- Select the third icon above the automatic menu. This begins the proof reading. In this step you can easily proofread recognized text by comparing it to the original image and using the built-in spellchecker as shown below.

It also gives suggestions from its built-in spell checker. If OmniPage Pro does not recognize some words in the document, the OCR Proofreader window will appear. Choose the appropriate response to each unrecognized word.
4) **Select page layout**

Once the proofreading is complete the document is exported to the text editor in **OmniPage**. Here you can edit the text and change the page layout.

5) **Save as a file**

- To do this click on the icon above the **Save to File menu**.
- Choose the location to save it at and give it a file name and select the file type to save it as. Now you can save the file in the available format you want. The typical formats available are MS-Word document *.doc, PDF *.pdf, HTML *.html, Text *.txt
- Enter your desired file name in the **File Name** text field.
- You can choose a document format from the Files of type pull-down menu. The default selection of RTF Word (*.rtf) is highly recommended, as it can be opened by most of the word processing programs.
- Click **OK** to save the file.
6.3 VIDEO DIGITISATION

Analogue mediums such as vinyl, VHS cassettes, and TVs have now been replaced by superior digital medium, such as CDs, DVDs, and HDTVs. The digital medium provides higher quality content. It also allows exact reproduction from copy to copy, barring any encryption technology implemented to stop copying.

Digital video refers to video being viewed or manipulated in the digital system (for instance on a computer), or sometimes simply video stored in a digital tape format. The video may have originally been analogue source material digitised into a computer, or it may have been stored directly to a digital tape format. Traditionally, digital tape formats were only available at the professional level (D-1, Digital Betacam, etc.), but now that some digital tape formats (DV) have emerged on the consumer scene, there is even more confusion about the generic term “digital video.”

DV (and related DVCAM and DVCPRO) is a digital tape format developed by a consortium of 10 companies as a “consumer” digital video format. There are now over 60 companies in the DVC consortium, including Sony, Panasonic, JVC, Philips, and other similar names you’ve heard before.

6.3.1 Video Capturing

In the simplest terms multimedia capturing can be stated as the process of storing or displaying the video/audio from the devices like Camcorders, Digital Cameras etc to some digital form like that of Monitor or in the binary forms (files).

As we have moved into the 21st Century, traditional analogue mediums such as vinyl, VHS cassettes, and TVs are being replaced by superior digital ones, such as CDs, DVDs, and HDTVs. Not only does digital formats allow for higher quality content, but also allows exact reproduction from copy to copy, barring any encryption technology implemented to stop copying. As computers become faster and disk storage space becomes larger, users are able to more deftly manipulate their digital data taken from analogue mediums and frequently “improve” the original analogue content using various techniques in the digital world.

System Requirements for a beginner multimedia processing system:

- x86-based PC @ 800+Mhz
- 256+MB RAM
- 40+GB of Free HD space (7200 rpm drive)
- Microsoft Windows98/ME/2000/XP
- Sound card with Line-in
- Video Capture card

These are the minimum requirements to perform reliable video capture. It is entirely possible to do video capture with less than this configuration, but good results cannot be guaranteed. Obviously, a faster CPU, more RAM, and more HD space are nothing but a good thing. Windows 9x/ME users should be aware that the FAT32 file system has a limitation preventing files from being larger than 4GB.
Digitisation Process

Choosing the Right Device to Capture the Video/Audio

One can purchase a video card with video-in support built right onto the card. We require the device which has a built-in “Analogue-to-Digital Conversion with Pass-Through” ability. This feature is quite useful since it will allow us to attach any analogue device (VCR, 8mm camcorder, etc.) to our Handy cam and then stream the digital data over FireWire to our computer.

6.3.2 Video Digitisation Process

Video digitisation is the next step used where the captured data from the analogue/digital device like cam coder is processed and saved in various file formats understandable by Media Players (both hardware and software based).

Software for video digitisation:

1) **VideoLAN**

VLC Player is one of the open source technologies that we are using to do the following things:

- Digitisation of content in various formats
- Re-Digitisation of multimedia video/audio content on LIVE and VOD.

![VideoLAN Streaming Solution](image)

Fig. 6.2: VideoLan Streaming

2) **Virtual DUB**

Virtual Dub is an open source video capture/processing utility for 32-bit Windows platforms, licensed under the GNU General Public License (GPL). It lacks the editing power of a general-purpose editor such as Adobe Premiere, but is streamlined for fast linear operations over video.
It has batch-processing capabilities for processing large numbers of files and can be extended with third-party video filters. VirtualDub is mainly geared toward processing AVI files, although it can read (not write) MPEG-1 and also handle sets of BMP images.

3) **FFmpeg**

It is a complete Open Source, cross-platform solution to record, convert and stream audio and video. It includes **libavcodec** - the leading audio/video codec library.

4) **Adobe Flash Media Encoder**

Adobe® Flash® Media Live Encoder 3 software is designed to enable us to capture live audio and video while streaming it in real time to RED 5 (Open Source) or Flash Media Server software or Flash Video Streaming Service (FVSS).
When high-quality streaming along with a very low bandwidth is our priority, Flash Media Live Encoder 3 can help you broadcast live events and around-the-clock broadcasting such as:

- Sporting events
- Concerts
- Webcasts
- News
- Educational events

### 6.4 AUDIO DIGITISATION

Analogue audio tapes are available in two formats: open reels and cassettes. They are available in various playing speeds and recording formats such as mono aural, stereophonic, and quadraphonic with tracking configurations like 2 track or 4 track. To digitise analogue audio data a player needs to be attached with a computer system through audio capture card. This process of analogue to digital conversion of audio data is known as sampling. The process involves sampling the original sound many times per second. The frequency of this sample is measured in Hertz (Hz) and the range of each sample is measured in bits. When digitising sound, the frequency range in kHz determines the sampling rate and the dynamic range i.e., the ratio between lowest and highest sound determines the number of bits per sample.

Various open source products are used for the audio digitisation. Here we are basically using Open Source and Free encoders.

#### 6.4.1 Audio Capturing

Audio can be captured using microphone. For better quality audio capture and storage of audio data via USB and Portable modes one can use voice recorders like shown in the figure below:

---

Fig. 6.17: Audio Capturing Devices
**LAME Audio Encoder**

LAME is a high quality MPEG Audio Layer III (MP3) encoder licensed under the LGPL. Currently LAME is considered the best MP3 encoder at mid-high bitrates and at VBR.

**VLC Media Player**

As already seen in the Video Processing the VideoLAN can be also used for the audio processing as well.
Self Check Exercise

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

3) What is LAME encoder?

........................................................................................................................................................
........................................................................................................................................................
........................................................................................................................................................
........................................................................................................................................................
........................................................................................................................................................
........................................................................................................................................................
........................................................................................................................................................
........................................................................................................................................................
........................................................................................................................................................

6.5 AUDIO/VIDEO COMPRESSION

Audio compression algorithms are implemented in computer software as audio codecs. A codec is a device or program capable of performing encoding and decoding on a digital data stream or signal. Generic data compression algorithms perform poorly with audio data, seldom reducing file sizes much below 87% of the original, and are not designed for use in real time. Consequently, specific audio "lossless" and "lossy" algorithms have been created. Lossy algorithms provide far greater compression ratios and are used in mainstream consumer audio devices. In addition to the direct applications (mp3 players or computers), digitally compressed audio streams are used in most video DVDs; digital television; streaming media on the internet; satellite and cable radio; and increasingly in terrestrial radio broadcasts.

There are five MPEG standards designed with a specific application and bit rate in mind for video compression. They include:

**MPEG-1**: for Video CD designed for up to 1.5 Mbit/sec application transmitted as .mpg files.

**MPEG-2** for the compression and transmission of digital broadcast television between 1.5 and 15 Mbit/sec rate of transmission. Digital Television set top boxes and DVD compression is based on this standard.

**MPEG-4** for multimedia and Web compression based on object-based compression technique.

**MPEG-7** also called the Multimedia Content Description Interface provides a framework for multimedia content that will include information on content manipulation, filtering and personalization, as well as the integrity and security of the content.

**MPEG-21** also called the Multimedia Framework attempts to describe the elements needed to build an infrastructure for the delivery and consumption of multimedia content, and how they will relate to each other. The work on this standard is still on.

Other video compressions are:

**DV** is a high-resolution digital video format used with video cameras and camcorders. DV images are compressed with a similar but superior technique to motion-JPEG allowing for higher-quality 5:1 compression. DV video information
is a constant data-rate of about 36 Mbps. The resulting video stream is transferred from the recording device via FireWire (IEEE 1394). IEEE-1394 (“FireWire”) is a communications protocol for high-speed, short-distance data transfer.

H.261 is an ITU standard designed for two-way communication over ISDN lines (video conferencing) and supports data rates which are multiples of 64Kbit/s.

H.263 is based on H.261 with enhancements that improve video quality over modems.

DivX is a software application that uses the MPEG-4 standard to compress digital video, so it can be downloaded over a DSL/cable modem connection in a relatively short time with no reduced visual quality.

6.6 AUDIO/VIDEO STREAMING

With the advent of high end streaming media technology, the concept of doing live/on-demand webcast has gained popularity like never before. Webcasting allows us to extend the reach of audio/video programmes to all corners of the world, with no limitations of physical or geographical boundaries.

Web casting can be either live or on demand. The modalities of these two types of delivery are explained below:

- **Live Webcast**: The transmission of live or pre-recorded audio or video to personal computers that are connected to the Internet. A user who clicks a link to a live clip joins the live event in progress. Because the event is happening in real time, fast-forward, rewind, and pause capabilities are not available. Live Web casts are most suitable for high demand live presentations to large geographically dispersed audiences. Participants can attend these virtual presentations from their desktop by visiting a web site. Interaction between instructor and learners occurs in real-time. Participants can use a chat window to type in questions to the presenter during the session. Web casts simulate the look and feel of a live event and can even be recorded for later viewing for those who missed the original web cast. This method is also less expensive than satellite broadcasting.

- **On-Demand Webcast**: Pre-recorded clips are delivered, or streamed, to users upon request. A user who clicks a link to an on-demand clip watches the clip from the beginning. The user can fast-forward, rewind, or pause the clip. Therefore on demand streams can be created from archived live events or recorded clips.

6.7 FILE FORMATS AND CONTENT CREATION

As large amount of document are being digitised and made available online through digital libraries throughout the world, it is pertinent that while archiving documents, physical survival, interpretability, and usability of the data is given importance. For this it is important to give due consideration to encoding standards, file formats and also ensure that the formats are usable and accessible in future. An ideal format for the purpose of archiving would be the one that is a representation rather than a presentation. The most common formats for text archiving are native formats (mostly MS Word), pdf, pdf-a, tex/latex, and xml applications. Other formats that are also prevalent are html, sgml, xhtml. Document formats may be broadly grouped into three types: text based formats, image formats, audio and video formats.
<table>
<thead>
<tr>
<th>Category</th>
<th>Type</th>
<th>Formats</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text formats</td>
<td>Plain text</td>
<td>Text Files (*.txt)</td>
<td>ASCII text files viewed with an editor (such as Edit or Notepad) or with a Word Processor (such as MS Word). Do not contain any kind of formatting on the document (such as bold, italics, font colour, images, etc.).</td>
</tr>
<tr>
<td>Formatted text</td>
<td>1. doc or odf</td>
<td>2. pdf files</td>
<td>Document files created, viewed and edited using programs such as MS Word or OpenOffice Writer. Formatting features such as bold, italics, justification, adding bullets and numbering, etc., is possible in such formats. Portable Document Format (pdf) was developed by Adobe Systems to transfer formatted documents over the net so that they gave a ‘printed document’ look and feel. This file typerequires Adobe Acrobat Reader which is freely downloadable from the net.</td>
</tr>
<tr>
<td>Image formats</td>
<td>Tagged Image File Format (TIFF)</td>
<td>• standard for describing and storing raster image data from scanners, faxes and digital photography applications. It is capable of describing bilevel, grayscale, palette-colour, and full-colour images in several colour spaces. TIFF is extensible, portable and does not favour a particular computer operating system, compiler or processor.</td>
<td></td>
</tr>
<tr>
<td>Audio/video formats</td>
<td>Graphics Interchange Format (GIF)</td>
<td>• free and open specification for the storage of raster imagery and to facilitate the exchange of digital imagery between different computer platforms and operating systems</td>
<td></td>
</tr>
<tr>
<td>Audio/video Video</td>
<td>Joint Photographic Experts Group (JPEG)</td>
<td>• JPEG is a standardized lossy image compression mechanism that is designed for compressing full-colour and grayscale images.</td>
<td></td>
</tr>
<tr>
<td>Audio Video Interleave</td>
<td>Audio Video Interleave Format (AVI)</td>
<td>• for storing and playing audio and video data on a PC. The format is limited to a 320 x 240 video resolution and playback rate of 30 fps.</td>
<td></td>
</tr>
<tr>
<td>Video formats</td>
<td>MPEG-4</td>
<td>• MPEG-4 is built on the MPEG-1, MPEG-2 and Quicktime MOV standards. These files are designed for transmission over a narrow Internet bandwidth,</td>
<td></td>
</tr>
<tr>
<td>Video formats</td>
<td>Quicktime (MOV)</td>
<td>• The MOV file format was developed by Apple Computer to create, play and stream high-quality audio and video files on both Macintosh and Windows computers using the Quicktime software application</td>
<td></td>
</tr>
<tr>
<td>Real Networks’ RealVideo</td>
<td>Real Networks’ RealVideo (RM)</td>
<td>• RealVideo was the first streaming video format available on the World Wide Web. A RealVideo clip consists of two parts, a visual track that is encoded with RealVideo codecs (COMpression/DECompression) and an audio track encoded using RealAudio codecs</td>
<td></td>
</tr>
</tbody>
</table>
### Table 6.2: Common Formats

<table>
<thead>
<tr>
<th>Format</th>
<th>File Extension</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>XML</td>
<td>.xml</td>
<td>An XML file, validated with DTD or schema specified, is a format suitable for preservation.</td>
</tr>
<tr>
<td>SGML</td>
<td>.sgml</td>
<td>A SGML file, validated, with DTD specified, is suitable for preservation.</td>
</tr>
<tr>
<td>HTML</td>
<td>.htm, .html</td>
<td>Hypertext markup language file, which may in principle be validated against a DTD. In practice invalid documents are often produced and used.</td>
</tr>
<tr>
<td>XHTML</td>
<td>.xhtml, .htm</td>
<td>XML-conformant HTML file, is required to be well-formed and valid.</td>
</tr>
<tr>
<td>DTD</td>
<td>.dtd</td>
<td>Document Type Definition. Defines the rules and syntax applied to a document. To be supplied with an SGML or XML document.</td>
</tr>
<tr>
<td>Pseudo-SGML</td>
<td>.sgm, .sgml, .txt or other</td>
<td>A text file employing some SGML-like formalisms for inserting markup, but not valid SGML. Suitability depends on whether tagging is consistently applied and well-documented, sufficient for later migration.</td>
</tr>
<tr>
<td>Various non-SGML encodings in text files</td>
<td>.txt or other</td>
<td>Suitability depends on acceptance as de facto standard in an academic community, plus an assessment of its likely future viability and level of documentation</td>
</tr>
</tbody>
</table>

### 6.8 SUMMARY

The conversion of analogue sources into digital form and their appropriate storage and processing form an important part of building a digital library. Digitisation is a complex process requiring managerial and technical skills. Proper planning and management help in keeping the cost down, and they also lead to the successful completion of a digitisation project. Digitisation can be carried out in-house or outsourced.

Various technical issues need to be considered in a digitisation project ranging from hardware to software and standards for file formats, file compression and post-processing. Selection of metadata format depends on the nature of the documents as well as the nature and needs of the users.

### 6.9 ANSWERS TO SELF CHECK EXERCISES

1) For converting hard copies into machine readable form three options available are:
   1) Keying in the text
   2) Scanning and capturing them as image files
   3) OCR the files
2) Omnipage Pro and ABBYY Fine Reader are two commonly used OCR software.

3) LAME is a high quality MPEG Audio Layer III (MP3) encoder licensed under the LGPL. Currently LAME is considered the best MP3 encoder at mid-high bitrates and at VBR.

6.10 KEYWORDS

Charge-coupled device (CCD) : A device for the movement of electrical charge, usually from within the device to an area where the charge can be manipulated, for example conversion into a digital value.

Contact Image Sensors (CIS) : Relatively recent technological innovation in the field of optical flatbed scanners that are rapidly replacing CCDs in low power and portable applications.

Photomultiplier Tubes (PMT) : Members of the class of vacuum tubes, and more specifically vacuum phototubes, are extremely sensitive detectors of light in the ultraviolet, visible, and near-infrared ranges of the electromagnetic spectrum.

6.11 REFERENCES AND FURTHER READING

http://www.librarydigitisation.com/
http://www.jiscdigitalmedia.ac.uk/digitisation
http://www.tape-online.net/Short_Guidelines_Video_Digitisation.pdf
UNIT 7  CREATING DIGITAL LIBRARIES USING DSPACE

Structure
7.0  Objectives
7.1  Introduction
7.2  Functional Features of Dspace
7.3  Installing Dspace on Windows
7.4  Working with Dspace
7.5  Summary
7.6  Answers to Self Check Exercises
7.7  Keywords
7.8  References and Further Reading

7.0  OBJECTIVES

After going through this Unit, you will be able to:
•  Describe the functional features of DSpace;
•  Install Windows version of Dspace; and
•  Create digital library using DSpace.

7.1  INTRODUCTION

DSpace is open source software, a turnkey repository application used by more than 1000+ organisations and institutions worldwide to provide durable access to digital resources. In India more than 140 institutions are using DSpace for building digital repositories.

DSpace is a software platform that enables organisations to:
•  capture and describe digital material using a submission workflow module, or a variety of programmatic ingest options.
•  distribute an organisation’s digital assets over the web through a search and retrieval system.
•  preserve digital assets over the long term.

The DSpace project was initiated in July 2000 as part of the HP-MIT alliance. The project was given $1.8 million USD by HP over two years to build a digital archive for MIT that would handle the 10,000 articles produced by MIT authors annually. DSpace has gone through several versions and the current stable release available is version 4.2.

The DSpace Foundation was formed in 2007 as a non-profit organisation to provide support to the growing community of institutions that use DSpace. The foundation’s mission is to lead the collaborative development of open source software to enable permanent access to digital works.
The code for DSpace is kept within a source code control system (http://dspace.svn.sourceforge.net/viewvc/dspace/) that allows code to be added or modified over time, whilst maintaining a track of all changes and a note of why the change was made and who made it. The Control of the source code repository is delegated to a small group of ‘committers’ who have the ability to change the code and release new versions. The committers work with the wider community of DSpace users to fix bugs and improve the software with new features.

In this we will guide you through the process of installation of DSpace (on a window platform) and familiarise you with the process of using and building collection in Dspace.

The Unit has been adapted from the DSpace official documentation and the Courseware developed by Aberystwyth University. Both the documents are available under the terms of either the GNU General Public License (http://www.gnu.org/licenses/gpl.html) and the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), for distribution and modification. The documents used are listed in the References and Further Readings section for further reference and you may refer them for further details.

### 7.2 FUNCTIONAL FEATURES OF DSPACE

The digital content in DSpace is presented in an organised tree structure of Community and Collections. Individual items can be accessed either through browsing the tree structure or searching with the Java freeware search engine Lucene built within. Each item gets a metadata description together with files available for download.

**Full-text search**: DSpace can process uploaded text based contents for full-text searching. Users may search for specific keywords that only appear in the actual content and not in the provided description.

**Navigation**: Users in DSpace find their way to relevant content through:

- **Searching** for one or more keywords in metadata or extracted full-text
- **Faceted browsing** through any field provided in the item description.
- Through **external reference**, such as a Handle
- **Browse** is another important mechanism for discovery in DSpace, whereby the user views a particular index, such as the title index, and navigates around it in search of interesting items.

**Supported file types**: While DSpace is most known for hosting text based materials including scholarly communication and electronic theses and dissertations (ETDs), it can accommodate any type of uploaded file. Files uploaded on DSpace are referred to as “Bitstreams” as after ingestion, files in DSpace are stored on the file system as a stream of bits without the file extension.

**Optimized for Google Indexing**: For the Google Scholar indexing, DSpace has added specific metadata in the page head tags that facilitates indexing in Scholar. Popular DSpace repositories often generate over 60% of their visits from Google pages.
OpenURL Support
DSpace supports the OpenURL protocol through linking server software called SFX server. DSpace will display an OpenURL link on every item page, automatically using the Dublin Core metadata if SFX server is implemented.

Self Check Exercise
Note: i) Write your answers in the space given below.
    ii) Check your answers with the answers given at the end of this Unit.

1) Enumerate functional features of DSpace.

Metadata Management
DSpace holds three types of metadata about archived content:

- **Descriptive Metadata**: A qualified Dublin Core metadata schema loosely based on the Library Application Profile set of elements and qualifiers is provided by default. However, one can configure multiple schemas and select metadata fields from a mix of configured schemas to describe items.

- **Administrative Metadata**: This includes preservation metadata, provenance and authorization policy data.

- **Structural Metadata**: This includes information about how to present an item, or bitstreams within an item, to an end-user, and the relationships between constituent parts of the item.

Choice Management and Authority Control
This is a configurable framework that lets you define plug-in classes to control the choice of values for a given DSpace metadata fields. It also lets you configure fields to include “authority” values along with the textual metadata value. The choice-control system includes a user interface in both the Configurable Submission UI and the Admin UI (edit Item pages) that assists the user in choosing metadata value.

Licensing
DSpace offers support for licenses on different levels:

- Collection and Community Licenses
- License granted by the submitter to the repository
- Creative Commons Support for DSpace Items

Persistent URLs and Identifiers
Researchers require a stable point of reference for their works. To help solve this problem, a core DSpace feature is the creation of a persistent identifier for every
item, collection and community stored in DSpace. To persist identifiers, DSpace requires a storage- and location- independent mechanism for creating and maintaining identifiers. DSpace uses the CNRI Handle System for creating these identifiers.

Similar to handles for DSpace items, bitstreams also have ‘Persistent’ identifiers. They are more volatile than Handles, since if the content is moved to a different server or organisation, they will no longer work (hence the quotes around ‘persistent’). However, they are more easily persisted than the simple URLs based on database primary key previously used. This means that external systems can more reliably refer to specific bitstreams stored in a DSpace instance.

**Getting content into DSpace**

Rather than being a single subsystem, ingesting is a process that spans several. Below is a simple illustration of the current ingesting process in DSpace.

![Fig. 7.1: DSpace Ingest Process](https://wiki.duraspace.org/display/DSDOC4x/Functional+Overview)

The batch item importer is an application, which turns an external SIP (an XML metadata document with some content files) into an “in progress submission” object. The Web submission UI is similarly used by an end-user to assemble an “in progress submission” object.

When the Batch Ingester or Web Submit UI completes the In Progress Submission object, and invokes the next stage of ingest (be that workflow or item installation), a provenance message is added to the Dublin Core which includes the filenames and checksums of the content of the submission. Likewise, each time a workflow changes state (e.g. a reviewer accepts the submission), a similar provenance statement is added. This allows us to track how the item has changed since a user submitted it.

Once any workflow process is successfully and positively completed, the In Progress Submission object is consumed by an “item installer”, that converts the In Progress Submission into a fully blown archived item in DSpace. The item installer:

- Assigns an accession date
- Adds a “date.available” value to the Dublin Core metadata record of the item
• Adds an issue date if none already present
• Adds a provenance message (including bitstream checksums)
• Assigns a Handle persistent identifier
• Adds the item to the target collection, and adds appropriate authorization policies
• Adds the new item to the search and browse index.

Workflow Steps
A collection’s workflow can have up to three steps. Each collection may have an associated e-person group for performing each step; if no group is associated with a certain step, that step is skipped. If a collection has no e-person groups associated with any step, submissions to that collection are installed straight into the main archive.

In other words, the sequence is this: The collection receives a submission. If the collection has a group assigned for workflow step 1, that step is invoked, and the group is notified. Otherwise, workflow step 1 is skipped. Likewise, workflow steps 2 and 3 are performed if and only if the collection has a group assigned to those steps.

When a step is invoked, the submission is put into the ‘task pool’ of the step’s associated group. One member of that group takes the task from the pool, and it is then removed from the task pool, to avoid the situation where several people in the group may be performing the same task without realizing it.

The member of the group who has taken the task from the pool may then perform one of three actions:

<table>
<thead>
<tr>
<th>Workflow Step</th>
<th>Possible actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Can accept submission for inclusion, or reject submission.</td>
</tr>
<tr>
<td>2</td>
<td>Can edit metadata provided by the user with the submission, but cannot change the submitted files. Can accept submission for inclusion, or reject submission.</td>
</tr>
<tr>
<td>3</td>
<td>Can edit metadata provided by the user with the submission, but cannot change the submitted files. Must then commit to archive; may not reject submission.</td>
</tr>
</tbody>
</table>

Fig. 7.2: Submission Workflow in DSpace
(Source: https://wiki.duraspace.org/display/DSDOC4x/Functional+Overview)
If a submission is rejected, the reason (entered by the workflow participant) is e-mailed to the submitter, and it is returned to the submitter’s ‘My DSpace’ page. The submitter can then make any necessary modifications and re-submit, whereupon the process starts again.

If a submission is ‘accepted’, it is passed to the next step in the workflow. If there are no more workflow steps with associated groups, the submission is installed in the main archive.

One last possibility is that a workflow can be ‘aborted’ by a DSpace site administrator. This is accomplished using the administration UI.

The reason for this apparently arbitrary design is that it was the simplest case that covered the needs of the early adopter communities at MIT. The functionality of the workflow system will no doubt be extended in the future.

**Command line import facilities**

DSpace includes batch tools to import items in a simple directory structure, where the Dublin Core metadata is stored in an XML file. This may be used as the basis for moving content between DSpace and other systems.

**Registration for externally hosted files**

Registration is an alternate means of incorporating items, their metadata, and their bitstreams into DSpace by taking advantage of the bitstreams already being in accessible computer storage.

**Getting content out of DSpace**

- **OAI Support**

  The Open Archives Initiative has developed a protocol for metadata harvesting. This allows sites to programatically retrieve or ‘harvest’ the metadata from several sources, and offer services using that metadata, such as indexing or linking services. Such a service could allow users to access information from a large number of sites from one place.

- **SWORD Support**

  SWORD (Simple Web-service Offering Repository Deposit) is a protocol that allows the remote deposit of items into repositories.

- **Command Line Export Facilities**

  DSpace includes batch tools to export items in a simple directory structure, where the Dublin Core metadata is stored in an XML file.

- **Packager Plugins**

  *Packagers* are software modules that translate between DSpace Item objects and a self-contained external representation, or “package”. A *Package Ingester* interprets, or *ingests*, the package and creates an Item. A *Package Disseminator* writes out the contents of an Item in the package format.
Crosswalk Plugins

*Crosswalks* are software modules that translate between DSpace object metadata and a specific external representation. An *Ingestion Crosswalk* interprets the external format and crosswalks it to DSpace’s internal data structure, while a *Dissemination Crosswalk* does the opposite.

The Packager plugins and OAH-PMH server make use of crosswalk plugins.

Supervision and Collaboration

In order to facilitate, as a primary objective, the opportunity for thesis authors to be supervised in the preparation of their e-theses, a supervision order system exists to bind groups of other users (thesis supervisors) to an item in someone’s pre-submission workspace. The bound group can have system policies associated with it that allow different levels of interaction with the student’s item; a small set of default policy groups are provided:

- Full editorial control
- View item contents
- No policies

User Management

E-People and Groups are the way DSpace identifies application users for the purpose of granting privileges. Both E-People and Groups are granted privileges by the authorization system described below.

- **User Accounts (E-Person)**

DSpace holds the following information about each e-person:

- E-mail address.
- First and last names.
- Whether the user is able to log in to the system via the Web UI, and whether they must use an X509 certificate to do so.
- A password (encrypted), if appropriate.
- A list of collections for which the e-person wishes to be notified of new items.
- Whether the e-person ‘self-registered’ with the system; that is, whether the system created the e-person record automatically as a result of the end-user independently registering with the system, as opposed to the e-person record being generated from the institution’s personnel database, for example.
- The network ID for the corresponding LDAP record, if LDAP authentication is used for this E-Person.
– **Subscriptions**

As noted above, end-users (e-people) may ‘subscribe’ to collections in order to be alerted when new items appear in those collections.

– **Groups**

Groups are another kind of entity that can be granted permissions in the authorization system. A group is usually an explicit list of E-People; anyone identified as one of those E-People also gains the privileges granted to the group.

Administrators can also use groups as “roles” to manage the granting of privileges more efficiently.

**Access Control**

**Authentication**

*Authentication* is when an application session positively identifies itself as belonging to an E-Person and/or Group.

**Authorization**

DSpace’s authorization system is based on associating actions with objects and the lists of EPeople who can perform them. The associations are called Resource Policies, and the lists of EPeople are called Groups. There are two built-in groups: ‘Administrators’, who can do anything in a site, and ‘Anonymous’, which is a list that contains all users. Assigning a policy for an action on an object to anonymous means giving everyone permission to do that action. The following actions are possible:

**Usage Metrics**

DSpace is equipped with SOLR based infrastructure to log and display page views and file downloads.

- **Item, Collection and Community Usage Statistics**

  Usage statistics can be retrieved from individual item, collection and community pages.

- **System Statistics**

  Various statistical reports about the contents and use of your system can be automatically generated by the system. These are generated by analyzing DSpace’s log files.

**Digital Preservation**

- **Checksum Checker**

  The purpose of the checker is to verify that the content in a DSpace repository has not become corrupted or been tampered with.
Each DSpace site is divided into **communities**, which can be further divided into **sub-communities** reflecting the typical university structure of college, department, research center, or laboratory.

Communities contain **collections**, which are groupings of related content. A collection may appear in more than one community.

Each collection is composed of **items**, which are the basic archival elements of the archive. Each item is owned by one collection. Additionally, an item may appear in additional collections; however every item has one and only one owning collection.

Items are further subdivided into named **bundles** of **bitstreams**. Bitstreams are, as the name suggests, streams of bits, usually ordinary computer files. Bitstreams that
are somehow closely related, for example HTML files and images that compose a single HTML document, are organized into bundles.

**Storage Resource Broker (SRB) Support**

DSpace offers two means for storing bitstreams. The first is in the file system on the server. The second is using SRB (Storage Resource Broker). Both are achieved using a simple, lightweight API.

SRB is purely an option but may be used in lieu of the server’s file system or in addition to the file system. Without going into a full description, SRB is a very robust, sophisticated storage manager that offers essentially unlimited storage and straightforward means to replicate (in simple terms, backup) the content on other local or remote storage resources.

### 7.3 INSTALLING DSPACE ON WINDOWS

Running DSpace on Windows is actually rather similar to running it on any other operating system. For the most part, you should be able to follow the normal DSpace Installation Documentation. However, this page provides you with some hints that are specific to Windows.

**Pre-requisite Software**

You’ll need to install this pre-requisite software (for DSpace 1.5.x and higher). Check the “Windows Installation” section of the System Documentation for the most recent pre-requisites, as they sometimes differ based on the version of DSpace you are running.

- **Java SDK (jdk-6u14-javafx-1_2-windows-i586)**: JDK is a development environment for building applications, applets, and components using the Java programming language. Download it from [http://java.sun.com/javase/downloads/widget/jdk6.jsp](http://java.sun.com/javase/downloads/widget/jdk6.jsp). For Ant to work properly, you should ensure that JAVA_HOME is set.

- **PostgreSQL (8.x for Windows)**: PostgreSQL is a powerful, open source object-relational database system. It has native programming interfaces for C/ C++, Java, .Net, Perl, Python, Ruby, Tcl, ODBC, among others. This comes with a Windows installer app. Make sure the ODBC + JDBC options are selected, as well as the pgAdmin III tool. We will be using it for storing the database of our repository. You can download it from: [http://www.postgresql.org/download/windows](http://www.postgresql.org/download/windows).

- **Apache Tomcat (apache-tomcat-5.5.28)**: An open source software implementation of the Java Servlets to serve as a Web server. You can download it from: [http://tomcat.apache.org/download-60.cgi](http://tomcat.apache.org/download-60.cgi).

- **Apache Maven (2.2.1-bin)**: Apache Maven is a software project management and comprehension tool. Just unzip it wherever you want it installed, and add [path-to-apache-maven]bin to your system PATH.

- **Apache Ant 1.7.x.** is a Java-based build tool. Just unzip it wherever you want it installed, and add [path-to-apache-ant]bin to your system PATH.
Self Check Exercise

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

2) What are the prerequisite software required for DSpace?

General Installation Steps

1) Download the DSpace software from SourceForge (http://sourceforge.net/projects/dspace/) and other prerequisite software. Untar or unzip it and save it in a folder.

2) **Install JDK**: Double click and execute the installer file of Java that you have downloaded. Finish JDK installation by clicking Finish. Another installer will start automatically for installing JRE. Click next (or you may cancel it also) > Click finish to close the installer. Next is to set up the Environmental variables and JAVA HOME.

   Right click on My computer > go to Properties > go to Advanced TAB > click on Environmental variables > select PATH in system variables section > click EDIT button. Open your program files directories in C drive and locate JAVA in Programme Files Directory > JAVA\JDK x.x.x.x\bin (C:\Program Files\Java\jdk1.6.0_14\bin) folder. Copy the file path from the address bar of windows explorer. Paste this path in system variable window (opened earlier) but use as a separator ‘semicolons’ (;) before it > Click ok. In User Variable segment > Click on NEW to set up a new user variable of JAVA_HOME. Variable name will : JAVA_HOME and Variable Value: C:\Program Files\Java\jdk1.6.0_14 or according to your installed version. Give the path of your java home directory located in program files. Click ok, and apply the settings.

3) **Install Apache Maven and Apache Ant**: Extract the files of Apache Maven and Apache Ant into C drive. Then give path for apache maven in system variables: Right click my computer > properties > Advanced > Environmental variables > Click on path and edit it > Add path “C:\apache-maven-2.2.1\bin” Now define path variable for apache ant in the same way. Open the extracted folder of apache ant in C drive, copy the folder path from windows explorer address bar and paste it in system path. Click ok. This will complete the task of defining all system paths [C:\Program Files\Java\jkl.6.0\_14\bin; C:\apache-maven-2.2.1\bin; C:\apache-ant-1.8.0\bin]. Now define ANT HOME in user variables. Variable name: ANT HOME Variable value: C:\apache-ant-1.8.0 > Click ok and apply the settings. All system paths and user variables are defined. We can also check, what we have done till now. Open command prompt and run the following command to see the java version C:/> ‘java –version’.
you can check ‘ant –version’ and ‘mvn –version’ and the command prompt will show relevant information regarding the respective software. If it appears all right then we may conclude that all packages java, maven and ant are successfully installed and paths are appropriately defined.

4) **Install Apache Tomcat**: Double click on Apache tomcat installer file and
> Now, tick mark all the components in order to **do full installation** and then click next. > In this window give your **username and password**, that will give you access to monitor and control your tomcat server web interface. Then click next. > **Make sure that your java virtual machine path is appropriate with your JRE installation folder**: Click install. > Click finish… this will start tomcat service automatically. You will see Apache icon in Notification area of Taskbar.

5) **Install PostgreSQL**: Before installing PostgreSql check the file system of your local disc. It needs be NTFS. To identify this, right click the local disc > select “Properties”, see the “File system”. If all the drives in your system are FAT, then convert a convenient disc to NTFS. For converting, go to command prompt and type `C:\>CONVERT C:/fs:ntfs` this command will convert your c drive into NTFS file system. If you already have C drive with NTFS file system partition, you may simply proceed to install PostgresSql by double clicking the installer file of postgresql. You must provide the database password to administrate your DATABASE. Click next. > Check DATABASE port number. **The port number should be 5432**. After installation of Postgres SQL is over, the next task is to create database and login rolls. For this open pgAdminIII. Connect to the database (provide password). Database will start… and then create login roll. > Right click on Login roles icon and click New Login Role. > Fill up the fields of role name with dspace and your password is also dspace. Then open role privileges tab. > Tick mark on icons named: Can create database objects, and can create roles. And then click ok. Login role is created. Now create Database. Right click on Database icon and click New Database. Fill up **Database name: dspace** and select **database owner dspace**. Click ok. Dspace database is created.

6) **Install DSpace**: Ensure the PostgreSQL service is running, and then run pgAdmin III (Start -> PostgreSQL 8.x -> pgAdmin III). Create the directory for the DSpace installation (e.g. C:\DSpace).

Build DSpace in the normal fashion. From [dspace-source]\dspace run:

```
mvn package
```

Then install DSpace to your specified location. From [dspace-source]\dspace\target\dspace-[version].dir run:

```
ant fresh_install
```

Create an administrator account, e.g. assuming C:\dspace is where your DSpace installation is:

```
C:\dspace\bin\dsrun org.dspace.administer.CreateAdministrator
```

(then enter the required info)
Copy the .war Web application files from C:\dspace\webapps to Tomcat’s webapps dir, which should be somewhere like C:\Program Files\Apache Software Foundation\Tomcat\webapps

Start the Tomcat service

Browse http://localhost:8080/dspace. You should see the DSpace home page!

### 7.4 WORKING WITH DSPACE

1) **Creating Communities**

| Sign in as an administrator  
| • Select ‘Community & Collection’ from the browse menu  
| • Select ‘Create Top-Level Community’ from the Admin Tools menu  
| • Complete the descriptive metadata for the Community  
| • Click ‘Create’ to complete the Community |

2) **Creating Collection**

| Navigate to the parent Community of the collection to be created  
| • Select ‘Create Collection’ from the Admin Tools menu  
| • Select the appropriate statements that apply to this collection |

3) **Descriptive Metadata for the Collection**

| Provide Descriptive Metadata for the collection  
| • Select the users who can submit to the Collection and the ‘Next’  
| • Click ‘Update’ to complete the collection creation process |
4) **Creating a user and groups**

Users require accounts to be able to log in and submit or edit items. Logical collections of users can be placed in groups to make administration easier.

DSpace has the facility User Self creation of account for which the following steps are to be followed:

- Click on My DSpace link
- Click on ‘New user? Click here to register.’
- Enter an email address and press ‘Register’
- Follow the link in the email that is sent for verification
- Provide name, telephone number, and a password
- New users have no privileges.

Users may be combined into logical groups for managing users and assigning privileges. Two special groups are possible on DSpace: i) Anonymous group in which there are no users in this group. Anyone can view the content without being logged, ii) Administrator group contains users who have full administrator access.

Administrator needs to be created directly on the DSpace server ([dspace]/bin/create-administrator) with the email address, first name, last name, and password details.

5) **Metadata in DSpace**

DSpace uses Dublin Core by default. Dublin core is made up of elements, and qualifiers. There are 15 base elements:

<table>
<thead>
<tr>
<th>Title</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>Identifier</td>
</tr>
<tr>
<td>Subject</td>
<td>Source</td>
</tr>
<tr>
<td>Description</td>
<td>Language</td>
</tr>
<tr>
<td>Publisher</td>
<td>Relation</td>
</tr>
<tr>
<td>Contributor</td>
<td>Coverage</td>
</tr>
<tr>
<td>Date</td>
<td>Rights</td>
</tr>
<tr>
<td>Type</td>
<td></td>
</tr>
</tbody>
</table>

The elements can be further refined through the use of qualifiers as shown below in the case of the base DC element Title:

**Schema** = ‘dc’

**Elements** viz. Title / Creator / Subject / Description

**Qualifiers** e.g. Title.main / Title.subtitle / Title.series.

Multiple schemas can be held in the metadata registry of DSpace and the access for which is through Administer menu -> Metadata Registry.
A schema can be edited and submitted using the ‘Update’ button, deleted using the ‘Delete’ button next to an element and new elements can be added using the ‘Add Metadata Field’ section at the bottom of the page.

6) Item submission Workflow

In the ‘Describe your Collection’ step while creating a new collection, one can select different workflow steps. During the process of creating the collection you will then be asked to select users and groups to assign to the workflow stages you have selected.
There are three options available for decision on the workflow:

- **Accept/Reject Step** – allows a user to simply accept an item, or reject it (with proper justification).

- **Accept/Reject/Edit Metadata Step** – allows a user to either accept or reject and item, and edit its metadata.

- **Edit Metadata Step** – allows the user to edit the metadata. This might be done to correct the metadata, or to improve it.

Any or all of the steps may be used. Workflow steps are worked through in order. If step 1 and 3 are selected, step 1 must be completed before step 3 will be initiated.

For an existing collection you may create the workflow through the following steps:

Log in as an administrator; go to the collection where you wish to create a workflow for. Click on the button ‘Edit’ in the ‘Admin Tools’ box.

Find the ‘Submission Workflow’ section, and click on whichever step you wish to create.

Edit the list of user and groups who can participate in the workflow as shown below:
When you have finished, press ‘Update Group’.

Use the same process to edit and delete workflow in a collection.

Once an item has entered into a workflow, the concerned users and group members will receive an email alert that there is a task awaiting attention. When a user visits their ‘My DSpace’ page they will see any tasks in the pool.

On clicking on ‘Take Task’ the user gets an overview of the item take a decision whether they wish to take the task.

Clicking ‘Accept This Task’ will take the user into the workflow task page where they have several option for action such as, Approve, Reject, Edit Metadata, Do Later and Return Task to Pool.

7.5 SUMMARY

DSpace is a platform that allows you to capture items in any format – text, video, audio, and data and distribute it over the web. It indexes all the collection so that users can search and retrieve your items. It is best suited for preservation of digital work over the long term.

The Web-based interface of DSpace makes it easy for a submitter to create an archival item by depositing files. Data files, also called bitstreams, are organized
Creating Digital Libraries Using DSpace

together into related sets. Each bitstream has a technical format and other technical information. This technical information is kept with bitstreams to assist with preservation over time. An item in DSpace is an “archival atom” consisting of grouped, related content and associated descriptions (metadata). An item’s exposed metadata is indexed for browsing and searching. Items are organised into collections of logically-related material.

In this Unit we have discussed in detail the technical features of DSpace along with the process of installation on your system and also using it for developing digital library.

### 7.6 ANSWERS TO SELF CHECK EXERCISES

1) The functional features of DSpace are:

- Full-text search
- Navigation
- Supported file types
- Optimized for Google Indexing
- OpenURL Support

2) The prerequisite applications required for installation of DSpace are:

- Java SDK (jdk-6u14-javafx-1_2-windows-i586)
- PostgreSQL (8.x for Windows)
- Apache Tomcat (apache-tomcat-5.5.28)
- Apache Maven (2.2.1-bin)
- Apache Ant 1.7.x.

### 7.7 KEYWORDS

**Bitstream** : a stream of data in binary form.

**Checksum Checker** : A checksum is a count of the number of bits in a transmission unit that is included with the unit so that the receiver can check to see whether the same number of bits arrived.

**OpenURL** : A standardised format of Uniform Resource Locator (URL) intended to enable Internet users to more easily find a copy of a resource that they are allowed to access.

### 7.8 REFERENCES AND FURTHER READING

The DSpace Course < http://cadair.aber.ac.uk/dspace/handle/2160/615>

DSpace Documentation <https://wiki.duraspace.org/display/DSDOC4x/DSpace+4.x+Documentation>
UNIT 8  CREATING DIGITAL LIBRARIES USING GSDL

Structure
8.0  Objectives
8.1  Introduction
8.2  Technical Features
8.3  Installation of GSDL on Windows
8.4  Greenstone Interfaces
8.5  Collection Building In Greenstone
8.6  Summary
8.7  Answers to Self Check Exercises
8.8  Keywords
8.9  References and Further Reading

8.0  OBJECTIVES

After going through this Unit, you will be able to:
• explain the technical features of Greenstone Digital Library (GSDL) Software;
• install GSDL on your system; and
• build a digital collection for the web as well as CD-ROM for your library.

8.1  INTRODUCTION

Greenstone is an open-source, multilingual software, issued under the terms of the GNU General Public License for building and distributing digital library collections. The aim of the Greenstone software is to empower users, particularly in universities, libraries, and other public service institutions, to build their own digital libraries. It provides a new way of organizing information and publishing it on the Internet or on CD-ROM in the form of a fully-searchable, metadata-driven digital library.

Greenstone has been produced by the New Zealand Digital Library Project at the University of Waikato, and is now being further developed and distributed in cooperation with UNESCO and the Human Info NGO in Belgium.

The exact user base for Greenstone is unknown. However, since it is being distributed on SourceForge, since November 2000, it has been found that the average downloads per month since then is around 4500.

The advantages of GSDL are:
• It is based on FOSS platform and has active community supporting it.
• It is Multi-platform application and can run on various operating system platforms, including Windows (any version), Linux, Sun Solaris, and Mac OSX. It is available in both binary (executable) and source code form for the Windows (all versions), Linux, and Mac OS X operating systems and in source code form for other operating systems (Unix).
• A Greenstone Collection can be served on the World Wide Web or it can be exported to a CD-ROM and accessed from the CD-ROM or local hard disc without the need for Internet connectivity.

• Greenstone can build indexes from full text documents and also metadata associated with these documents. It supports creation of indexes for various metadata fields, either automatically extracted or manually assigned.

• It uses Perl-scripting, MG(PP) or Lucene for indexing, Apache (or built-in webserver), XML, which are proven technologies.

• Greenstone lets you build collections of multimedia documents such as audio, video, and pictures accompanied by textual description or metadata to allow searching and browsing.

• UNICODE compliant facilitating building, searching and browsing documents in any Unicode-compliant language.

• Separate modules are available for different uses:
  – JAVA-based interface for management
  – Web-browser based access to collections
  – CLI client : remote collection building

• Multi-metadata (with editor)

• Practical GLI interface for editing/managing GSDL

• Plug-ins for most document formats also available as well as for crosswalks for ISIS, Dspace, e-mails, MARC, MARCXML.

The Unit has been adapted from the Greenstone official documentation and the IMARK tutorial developed by FAO. Both the documents are available under the terms of either the GNU General Public License (http://www.gnu.org/licenses/gpl.html) and the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), for distribution and modification. The documents used are listed in the References and Further Readings section for further reference and you may refer them for further details.

8.2 TECHNICAL FEATURES

Multiplatform user friendly application

Greenstone runs on all versions of Windows, Unix/Linux, and Mac OS-X. The process of installation is quite simple. The default Windows installation does not require any configuration. End users routinely install Greenstone on their personal laptops or workstations. The Institutional users, however, generally run it on their main web server, where it interoperates with standard web server software i.e. Apache.

Interoperability

It is highly interoperable, based on contemporary standards. Greenstone can harvest documents over OAI-PMH and include them in a collection. Greenstone can ingest documents in METS (Metadata Encoding and Transmission Standard) form. This facilitates export and import of any collection to and from DSpace through DSpace batch import program.
Interfaces

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.

Metadata formats

Users define metadata interactively within the Librarian interface. Unlike DSpace, Greenstone allows several sets of metadata, including locally produced ones to be merged. The metadata sets are predefined:

- Dublin Core (qualified and unqualified)
- RFC 1807
- NZGLS (New Zealand Government Locator Service)
- AGLS (Australian Government Locator Service)

All metadata are stored in XML-format with the documents. Metadata can also be extracted from XML-statements within the documents. It can be assigned easily through the GSDL Librarian interface using Greenstone’s Metadata Set Editor. “Plug-ins” are used to ingest externally-prepared metadata in different forms, and plug-ins exist for: XML, MARC, CDS/ISIS, ProCite, BibTex, Refer, OAI, DSpace and METS.

Document formats

Plug-ins are also used to ingest documents. For textual documents, there are plug-ins for: PDF, PostScript, Word, RTF, HTML, Plain text, Latex, ZIP archives, Excel, PPT, Email (various formats), source code. For multimedia documents, there are plug-ins for: Images (any format, including GIF, JIF, JPEG, TIFF), MP3 audio, Ogg Vorbis audio, and a generic plug-in that can be configured for audio formats, MPEG, MIDI, etc.

Languages

One of Greenstone’s unique strengths is its multilingual nature. The reader’s interface is available in the following languages: Arabic, Armenian, Bengali, Catalan, Croatian, Czech, Chinese (both simplified and traditional), Dutch, English, Farsi, Finnish, French, Galician, Georgian, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Kannada, Kazakh, Kyrgyz, Latvian, Maori, Mongolian, Portuguese (BR and PT versions), Russian, Serbian, Spanish, Thai, Turkish, Ukrainian, Vietnamese

The Librarian interface and the full Greenstone documentation (which is extensive) is in: English, French, Spanish, and Russian.

In GSDL the server (library.exe) uses PERL-scripts to create web-pages and forms to deal with the library of documents and its indexes. The documents are stored in their native format as such (PDF, DOC, HTML, XML etc.) which are converted (‘imported’) as XML in a collection with their text-only content. ‘Plug-ins’ for each type of content extract words from the documents and pass them
onto the indexing engine. Metadata are also stored in XML. A web-interface allows searching, browsing results and opening full-text documents either in original or converted format.

There are three indexers available in GSDL:

- MG (‘Managing Gigabytes’) : at section level (=~field), Boolean or ranked
- MGPP : word level indexing (field, phrase + proximity) with Boolean+ranking
- Lucene (from the Apache SF) : field+proximity indexing but either on whole document or section, Boolean+ranking plus : single-character wildcards and range-searching; allows incremental collection building (not possible with MG(PP))

Unlike DSpace, GSDL allows several sets of metadata, including locally produced ones, even merged. Dublin Core (v.1.1) is provided together with RFC 1807, Development Library Subset, as well as LOM required for indexing learning objects. All metadata are stored in XML-format with the documents and can also be extracted from XML-statements within the documents. Metadata can be assigned easily through the GSDL Librarian interface. One limitation is that since GSDL does not use a DB for handling its XML-data, this imposes real limitations on speed.

Self Check Exercise

Note:  
i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

1) Enumerate technical features of GSDL.

...................................................................................................................

...................................................................................................................

...................................................................................................................

...................................................................................................................

...................................................................................................................

...................................................................................................................

...................................................................................................................


8.3 INSTALLATION OF GSDL ON WINDOWS

Before installing the software, be sure you have all the hardware and software requirements!

Hardware and software requirements

Storage requirements:
- 50MB for a binary installation
- 155MB for compiling Greenstone from source code
- 200MB for optional Greenstone demonstration collections
- 5MB for documentation
- 24MB for Greenstone’s “CD exporting” function
Software:
- Java Run-time Environment (JRE) version 1.4 or above (Install JRE before installing GSDL) - JRE is required for GLI
- [Not required for default Windows installation] Web Server (Apache Recommended)
- PERL - gets installed automatically
- C++ compiler, if you wish to compile the source code (Visual Studio or GCC)
- A Web Browser

There are different options for getting the GSDL software:

1) UNESCO CD-ROM (version 2.70) or FAO IMARK CD-ROM, (but this is an earlier version 2.51) which contain the Greenstone software, plus documented example collections, four language interfaces (English French Spanish Russian), the Export to CD-ROM package, the ImageMagick graphics package, the Java runtime environment, and an installer that installs all of these.

2) IITE Digital Libraries in Education CD-ROM, or a Greenstone workshop CD-ROM. This CD-ROMs contains the tutorial exercises and a set of sample files to be used for these exercises apart from the requisite software listed above.

3) Download directly from http://www.greenstone.org that contains the latest version of Greenstone.

You will need Java to run Greenstone. You might already have it installed on your system otherwise, download it from http://java.sun.com. To work with image collections, you need ImageMagick (from http://www.imagemagick.org).

Most Greenstone CD-ROMs have AutoPlay feature and start the installation process as soon as they are inserted into the drive. If installation does not begin by itself, locate the file setup.exe and double click it to start the installation process.

If you download Greenstone over the web then just double-click installer.

If Greenstone is already installed on your system then completely remove the old version before installing a new one. You need not remove any pre-packaged collections that you may have installed for this.

The following steps need to be carried out to install Greenstone:

1) Install the Java 2 Runtime Environment (latest version).
2) After installing J2RE, go for GSDL folder choose setup gsdl 2.70.
3) Choose setup Language. English (US) is the default. We choose English
4) Welcome to the InstallShield Wizard for the Greenstone Digital Library Software. Click <Next>
5) License Agreement. Accept the agreement and then click <Next>
6) Choose location to install Greenstone. Leave at the default and click <Next>
7) Setup Type. Leave at the default (Local Library) and click <Next>
8) (For older installers you must now select collections. Leave at the default, Documented Example Collections, and click <Next>.)

9) Set admin password. Choose a suitable password and click <Next> (If your computer will not be serving collections online, the password doesn’t matter)

10) Click <Install> to complete the installation

11) Files are copied across and installation is complete.

If you are installing from a CD-ROM, the installer will offer to install ImageMagick, and Java, if necessary.

To invoke the Greenstone Reader’s interface, go to the Greenstone Digital Library Software item under Programs on the Windows Start menu and select Greenstone Digital Library. To invoke the Librarian interface, go to the same item and select Greenstone Librarian Interface.

**Installing ImageMagick on a Windows system**

Once Greenstone has been installed, ensure that ImageMagick is installed on your system, if you wish to build any image collections. If you are installing from a Greenstone CD-ROM, you will be asked whether you want to install ImageMagick: say Yes. If you are not, you will need to download ImageMagick (from [http://www.imagemagick.org](http://www.imagemagick.org)). To install this program you must have Windows “Administrator” privileges.

The remaining steps are straightforward, and, as before, it is recommend that you use the default settings. Here is what you need to do for installing ImageMagick:

1) “This will install ImageMagick 5.5.7 Q8. Do you wish to continue?” Yes
2) “Welcome to the ImageMagick Setup Wizard” Click <Next>
3) “Information: Please read the following ...” Click <Next>
4) “Select Destination Directory ...” Leave at default and click <Next>
5) “Select Start Menu Folder ...” Leave at default and click <Next>
6) “Select Additional Tasks ...” Leave at default and click <Next>
7) “Ready to Install”. Click <Install>
8) Files are copied across
9) “You have now installed ...” Click <Next>
10) “Setup has finished ...”. Deselect “View index.html” and click <Finish>.

### 8.4 GREENSTONE INTERFACES

GSDL comprises two interfaces, the Librarians Interface and the Website which serves as the user interface.

The “librarian’s interface” in GSDL is for creation, management and updating collections. It is programmed in JAVA highly based on creation of the necessary commands.

The website is served by internal www-server or Apache. Webpages are created by Perl and Java Servlets which is customisable via CSS and text-files.
A) Librarian’s Interface

A JAVA-PERL applet (gliserver.pl) provides an interactive graphical interface for the Greenstone Librarian Interface with the following main functions:

1) Gathering - documents into a Selecting files from ‘local file space’ or Local Network or downloading using protocols viz. WWW, OAI (Open Archives Initiative), Z39.50, SRW (Search and Retrieve Web service), MediaWiki.

![Fig. 8.1: Librarian’s Interface- Collection Building](image1)

2) Enriching - cataloguing with metadata, i.e. assign values to metadata-fields - Dublin Core and/or others or local sets. Metadata editor allows creating/changing sets and assigning values- automatic inheriting for lower levels, multiple values, picklists or hierarchical at level1|level2|level3

![Fig. 8.2: Librarian’s Interface- Metadata Input](image2)
3) **Design** – this involves selection of plugins (e.g. GA, TEXT, PPT, Word, PDF, RTF, e-mail, XLS, Fox, DB, as well as ISIS, DSpace, MARC, ProCite…), defining Search index, Partitioning of sub-collections and setting Browsing classifiers, hierarchical or A-Z.

4) ‘plug-ins’ (filters), Indexing the documents and providing preview facility for direct access to webpage with search-interface produced by GLI is done at this stage. Once build is successful then the collection needs to be linked to previewing.
Self Check Exercise

Note: i) Write your answers in the space given below.

ii) Check your answers with the answers given at the end of this Unit.

2) What functions are available in the Librarian’s Interface?

B) Greenstone User Interface

Although the user interface of different Greenstone collections may appear remarkably similar, each one can provide varying search, browse and display features, depending on access requirements, nature of documents comprising the collection and metadata associated with these documents. As a digital library developer you can define the desired end-user interface features for your collection at the designing stage.

Collection Searching

Greenstone supports different ways of searching collections. They can be grouped in two main categories: “plain search” (through Google-like single search box) and “form-based search”.

- **Plain search:**
  
  **Simple** - Users can search for words or phrases in the full text of the document or limit the search to a specific index (e.g. document title or author) by selecting the available index from the drop-down box.

  **Advanced** - Boolean queries.

- **Form-based search**
  
  **Simple** - Users can search for words or phrases across different fields.

  **Advanced** - Users can search for words or phrases across different fields, with support for Boolean query combination, case folding and stemming.

Document Browsing

Greenstone supports browsing of documents in a collection by specific metadata fields.

Available browse elements for a collection are shown on the navigation bar in the collection home page. Hierarchical browsing of classification-like structures (e.g. a subject classification) with different levels is possible.
Presentation of Search Results

The web pages the users see when using Greenstone are not pre-stored but are generated “on the fly” as they are needed. This includes the way the browse and search results appear and individual documents are presented. After obtaining a document (selected from results of browse/search), a user can:

- view complete content or contract it (in a full-text tagged document);
- highlight matching search terms or not; and
- detach the document for viewing in a different window.
Greenstone supports multilingual interface. Through the preferences setting, the user can change the language of the Greenstone interface. It can also support indexing and searching of document collections in non-Latin scripts.

8.5 COLLECTION BUILDING IN GREENSTONE

You will need some source files like those in the sample_files\Word_and_PDF folder to work on the collection building.

1) Start a new collection called reports, fill out appropriate fields for it, and choose Dublin Core as the metadata set.

2) Copy the 12 files from sample_files 'Word_and_PDF' Documents into the collection. You can select multiple files by clicking on the first one and shift-clicking on the last one, and drag them all across together. (This is the normal technique of multiple selection.)

3) Switch to the Create panel, and build and preview the collection.

4) Again, this collection contains no manually assigned metadata. All the information that appears—title and filename—is extracted automatically from the documents themselves. Because of this the quality of some of the title metadata is suspect.

5) Back in the Librarian Interface, click the Enrich tab to view the automatically extracted metadata. You will need to scroll down to see the extracted metadata, which begins with “ex.”. The PostScript documents (cluster.ps and langmodl.ps do not have extracted titles: what appears in the titles a-z list is just the first few characters of the document).

6) Manually adding metadata to documents in a collection

In the Enrich panel, manually add Dublin Core dc.Title metadata to one of these documents. Select word03.doc and double-click to open it. Copy the title of this document (“Greenstone: A comprehensive open-source digital library software system”) and return to the Librarian Interface. Scroll up or down in the metadata table until you can see dc.Title. Click in the value box, paste in the metadata and press Enter.

7) Now add dc.Creator information for the same document. You can add more than one value for the same field: when you press Enter in a metadata value field, a new empty field of the same type will be generated.

8) Close the document when you have finished copying metadata from it. External programs opened when viewing documents must be closed before building the collection, otherwise errors can occur.

9) Next add title and creator metadata for a few of the other documents.

If you build and preview your collection at this point, you will find that nothing has changed. You need to alter the collection design to use the new Dublin Core metadata instead of the original extracted metadata.
10) **Collection design; branding a collection with an image**

Change to the Design panel, which is split into several sections. The first section General appears. This allows you to modify the values you provided when defining the collection, if desired. You can also brand the collection using a suitable image.

11) Click on the `<Browse...>` button associated with URL to about page icon, and browse to the image sample_files`'/Word_and_PDF'/wrpdf.gif` on your computer. When you select this image, Greenstone automatically generates an appropriate URL for the image. Preview the collection.

If you are on the web, you can easily make your own Greenstone-style icon by going to and following the instructions there.

http://www.greenstone.org/make-images.html

**Document plugins**

12) Now look at the Document Plugins section, by clicking on this in the list to the left. Here you can add, configure or remove plugins to be used in the collection. There is no need to remove any plugins, but it will speed up processing a little. In this case we have only Word, PDF, RTF, and PostScript documents, and can remove the ZIPPlug, TEXTPlug, HTMLPlug, EMAILPlug, ImagePlug, ISISPlug and NULPlug plugins. To delete a plugin, select it and click `<Remove Plugin>`. GAPlug is required for any type of source collection and should not be removed.

13) **Search types and fielded searching**

Go to the Search Types section. This specifies what kind of search interface and what search indexes will be provided for the collection. Let’s add a form search option. Click `<Enable Advanced Searches>`; this allows form searching to be added to the collection.

14) To include “form search” as well as the default “plain search”, pull down the Search Types menu and select form; then click `<Add Search Type>`. Plain search will be the default search type as it is first in the list.

**Search indexes**

15) The next step in the Design panel is Search Indexes. These specify what parts of the collection are searchable (e.g. searching by title and author). Delete the ex.Title and ex.Source indexes, which are not particularly useful, by selecting them one at a time and clicking `<Remove Index>`. Only the text index remains.

16) Now add a Title index based on dc.Title by providing an Index Name (e.g. “Document Title”) and selecting dc.Title from the Index Source box. Then click `<Add Index>`. 

17) You can add indexes based on any metadata. Add an index called “Authors” based on dc.Creator metadata.

*The next two sections are Partition Indexes and Cross-Collection Search. In this exercise, we will not make any changes to these.*
Digitisation and Digital Libraries – DSpace and GSDL

18) The **Browsing Classifiers** section adds “classifiers,” which provide the collection with browsing functions. Go to this section and observe that Greenstone has provided two classifiers, *AZLists* based on *ex. Title* and *ex. Source* metadata. Remove both of these by selecting them in turn and clicking `<Remove Classifier>`.

19) Now we add an *AZList* classifier for *dc.Title* metadata. Select *AZList* from the **Select classifier to add** drop-down list and click `<Add Classifier>`.

20) A popup window **Configuring Arguments** appears. Select *dc.Title* from the **metadata** drop-down list and click `<OK>`.

21) Now add an *AZCompactList* classifier. Click `<Add Classifier>` and configure it to use *dc.Creator* metadata, with button name “Creator”. Click `<OK>`.

*The last three sections are Format Features, Translate Text and Metadata Sets. In this exercise, we will not make any changes to these.*

22) Switch to the **Create** panel, and **build** and **preview** the collection.

23) Check that all the facilities work properly. There should be three full-text indexes, called *text*, *Document Title*, and *Authors*. In the *titles a-z* list should appear all the documents to which you have assigned *dc.Title* metadata (and only those documents). In the *authors a-z* list should appear one bookshelf for each author you have assigned as *dc.Creator*, and clicking on that bookshelf should take you to all the documents they authored.

In the similar fashion you can build up collection for other types of file formats. For details visit the tutorial site of Greenstone.

### 8.6 SUMMARY

Greenstone is a freely available open source software for building and distributing digital library collections through Internet or Multiplatform availability, the capability of providing access in different ways and managing different file formats, media and languages are some of the major advantages of Greenstone. The Librarian Interface provides the most advanced and at the same time a very user friendly approach to collection building and also metadata management.

In this Unit we discussed the technical features of Greenstone, installation process and building a digital library.

### 8.7 ANSWERS TO SELF CHECK EXERCISES

1) Technical features of GSDL are:
   - Multiplatform user friendly application
   - Interoperability
   - Independent librarian and user interfaces
   - Supports variety of Metadata formats
   - Supports variety of Document formats
   - Supports multiple Languages
2) Following functions are available in the Librarian’s Interface:

- Creation of New Collection
- Selection Metadata
- Gathering
- Enrich
- Design
- Create

8.8 KEYWORDS

Lucene : Open source search engine.
Perl : A script programming language that is similar in syntax to the C language and that includes a number of popular UNIX facilities.
UNICODE : An international encoding standard for use with different languages and scripts, by which each letter, digit, or symbol is assigned a unique numeric value that applies across different platforms and programs.
XML : Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format which is both human-readable and machine-readable.

8.9 REFERENCES AND FURTHER READING