

DIGITAL LIBRARY IN BUSINESS AND MANAGEMENT AT IIMK : PROGRESS REPORT

(Funded by the MHRD)

ABSTRACT

IIMK received an amount Rs. 15/- Lakhs from MHRD, Govt. of India, as funding for the 2-year project “Development of a Digital Library in the field of Management”, during March 2001. The essential set of Hardware (Digital Library Server, Scanner, UPS, Digital Camera etc.) and Software required for the take off of the project have been sourced and installed. World renowned open source Digital Library software “Greenstone”, has been successfully installed in the Digital Library server and the development work using Greenstone is in constant progress. The content development, interface design and digitization work etc. have been completed. Subject-wise content populating is being carried out on a war footing. A latest version of the Digital Library Software “ACADO”, developed by the IITMK, Trivandrum is being loaded on February 05, 2003. Metadata harvesting work and content aggregation from repositories outside are also in progress. The project is now operational and a detailed report is appended below for evaluation.

Principal Investigator : M.G. Sreekumar, Librarian & Head, Center for Development of Digital Libraries (CDDL)

DETAILED REPORT

IIMK Library / Information system

IIMK’s Library and Information Centre (LIC) blends digital, audio, video and print media with cutting-edge technology. The LIC aims at providing state-of-the-art information resources and value-added information services in all areas of management and related disciplines. The LIC is an outstanding learning resource centre catering to the faculty, students and research scholars. It houses around 16000 books, 360 print journals, over 30000 e-books and around 1000 electronic journals (full-text) sourced directly from publishers (which include Elsevier, Kluwer, John Wiley, Blackwell, Taylor & Francis, ACM, IEEE, and Emerald), and more than 4000 aggregated journal contents sourced through EBSCO and PROQUEST. A host of full-text/bibliographic online databases on scholarly

information coupled with a set of national and international databases on companies, industries and the markets are available to users through a local area network (LAN) powered by a high-end CD-NET Server having 2 TB storage capacity. The Web enabled Online Public Access Catalogue (OPAC) is available via the LAN. Further more, the LIC hold thousands of CD-ROM publications on contemporary issues. The audio/visual unit in the LIC hosts over 225 educational videos covering a wide range of disciplines in management. All the books in the Library are barcoded and the transactions are fully computerised. The entire set of information resources and services mentioned above are being rendered through a 'single window', the Library Web Portal, which is a unique feature of IIMK's Library & Information Centre. The portal is an integrated network of information sources. Besides serving as a web-based interface to the in-house resources, the portal also provides links to strategic sources of information outside. The search services are user friendly and hyperlinks are provided for instantaneous access to a number of strategic locations. Currently, Smart Card based access control and Smart Gate based e-security systems are being introduced in the LIC. Thanks to the excellent support of the IIM Directors and the Ministry of Human Resource Development in our endeavours to value additions, IIMK is an active member in the IIMs' Library Consortium as well as the Ministry of Human Resource Development's (MHRD, Government of India) INDEST Consortium (<http://www.library.iitb.ac.in/indest/>). A fairly good Internet connectivity (2Mbps leased line, powered by BSNL) and a state-of-art structured LAN with giga-bit speed and fibre-optic network backbone supports the information access/retrieval requirements of the users on a 24 X 7 formula.

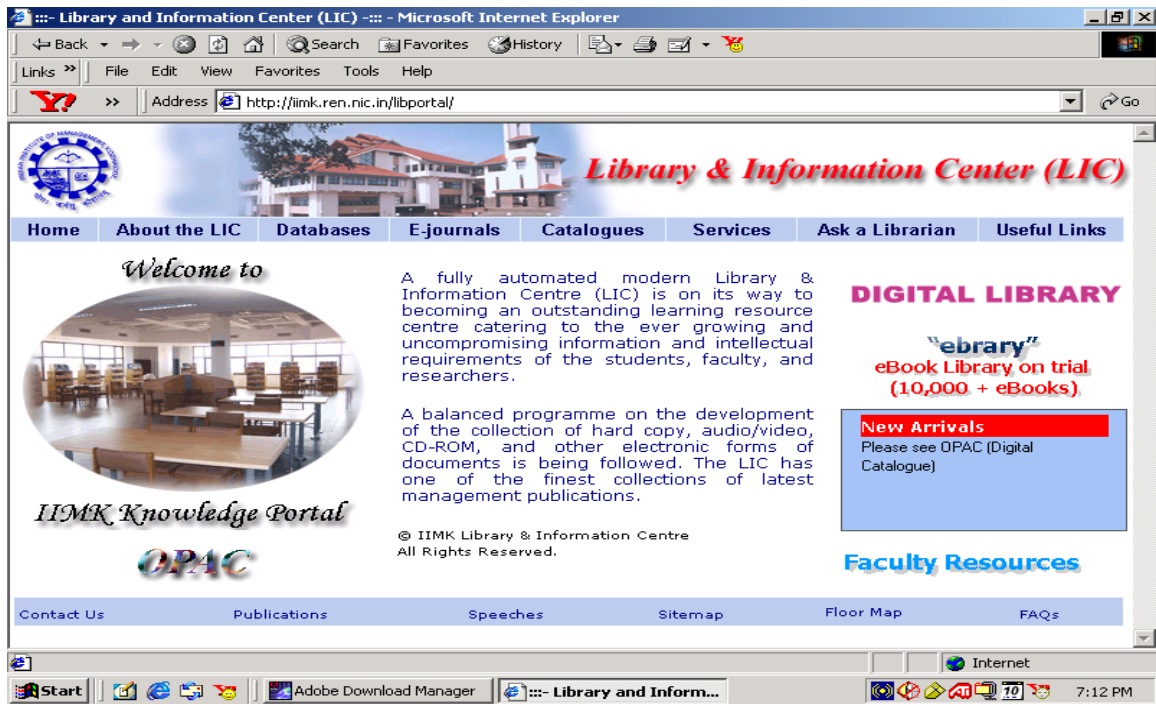


Fig. 1. IIMK Library Portal

From the above description on the environment, components, and activities relating to a digital library system it is evident that IIMK LIC is quite capable of and matured enough to embark on digital library development. It was in this context the IIMK LIC received a grant from MHRD, Government of India during 2001 for creating a digital library in the field of business and management.

DIGITAL LIBRARY Objectives and Workflow

The primary objective of the IIMK digital library was to enhance the digital collection in a substantial way, by strategically sourcing digital materials, conforming to copyright permissions, in all possible standards/formats so that scalability and flexibility is guaranteed for the future and advanced information services are assured to the user community right from beginning. The digital library should also be able to integrate and aggregate the existing collections and services mentioned above with an outstanding client interface. This implies that the digital library system should also have a strong collection interface capable of embracing almost all the popular digital standards and formats and software platforms, in line with the underlying digital library technologies in vogue. This is crucial in the case of multimedia integration, which is again important as we planned to also host a digital audio and video library as part of the core library collection . Emphasis was also given to maximise the efficiency and effectiveness of the information access and retrieval capabilities of the system by deploying cutting edge Resource Description Framework (RDF) standards for metadata and its encoding. The Internet also possesses, in addition to its mammoth proprietary information base, an invaluable wealth and a vast collection of public domain information products such as databases, books, journals, theses, technical reports, cases, standards, newsletters etc., scattered and distributed across the world. This treasure will also be explored to its maximum for collection building, based on the source and quality.

Standard workflow patterns were identified for the system which included 'content selection', 'content acquisition', 'content publishing', 'content indexing and storage', and 'content accessing and delivery'. The system will also concern about such related issues, viz., preservation, usage monitoring, access management, interoperability, administration and management etc. It was proposed to value add the digital catalogue of the library (OPAC), which acts as a stepping stone for effective information discovery in the library, by supplementing the bibliographic information with the scanned picture (.gif/.jpg) of the front cover and the contents page, so that users will be able to evaluate the book for further consultation while browsing the catalogue, right from their desktops. Dublin Core standards were identified as the desired metadata format and XML as the desired encoding scheme.

Selection of the DL Software

The software selection based on set parameters was a difficult task, as the technology itself was still emerging only. There were no strong role models too from the Indian side, to be relied upon those days. In general, we wanted a system that is flexible enough to fit the current digital information system as above and to accommodate future migration. It should be robust in technical architecture as well as the content architecture. The system should address all major digital library related issues such as 'design criteria', 'collection building', 'content organisation', 'access', 'evaluation', 'policy and legal issues' including 'intellectual property rights'. That the system should be in a position to embrace almost all predominant and emerging digital object formats and capable of supporting the standard library technology platforms, was among our major focus. It should provide two important user interfaces: a public user interface for presentation and a metadata creation interface for administration. The system should also provide a powerful search engine and the interface should be easy to navigate and there should be provision for customisation.

As we have decided not to go for a proprietary digital library software, we evaluated some of the popular Open Source Software for digital libraries, which were in use internationally. 'Dienst', 'Eprints', 'Fedora', 'Greenstone' etc. were the candidates for the preferred software. Obviously Greenstone outscored the group and we formally adopted the software for creating the IIMK digital library.

Greenstone

The Greenstone Digital Library Software (GSDL) is a top of the line and internationally renowned Open Source Software system for developing digital libraries, promoted by the New Zealand Digital Library project research group at the University of Waikato, headed by Dr. Ian H. Witten, and is sponsored by the UNESCO. The software is issued under the terms of GNU General Public License. Greenstone provides a way of building, maintaining and distributing digital library collections, opening up new possibilities for organizing information and making it available over the Internet or on CD-ROM. One of the pioneering and laudable efforts of Greenstone is its commitment to lower the bar for construction of practical digital libraries, yet at the same time leave a great deal of flexibility in the hands of the user. Even at the time of selecting Greenstone for IIMK we found that several libraries and institutions across the world were using it and with encouraging feedback. Couple of promotional articles appeared in D-Lib Magazine by the Greenstone team also were strong indicators for their continued interest and commitment to the software. Another word of inspiration came from the message of Ian Witten's team itself, that 'we should now stop talking about digital libraries and start developing them'.

GSDL: Features

The salient features of Greenstone are basically taken from two of the official publications of the software development team appeared in D-Lib Magazine during the year 2001 (Witten, 2001) and 2003 (Witten, 2003). Greenstone builds collections using almost popular and standard digital formats such as HTML, XML, Word, Post Script, PDF, RTF, and many other formats which include audio as well as video. It is provided with effective full-text searching and metadata-based browsing facilities that are attractive and easy to use. Moreover, they are easily maintained and can be augmented and rebuilt entirely automatically. The system is extensible: software "plug-ins" accommodate different document and metadata types. Greenstone incorporates an interface that makes it easy for people to create their own library collections. Collections may be built and served locally from the user's own web server, or (given appropriate permissions) remotely on a shared digital library host. End users can easily build new collections styled after existing ones from material on the Web or from their local files (or both), and collections can be updated and new ones brought on-line at any time. It runs on a wide variety of platforms such as Windows, Unix/Linux, Apple Mac etc. and provides full-text mirroring, indexing, searching, browsing and metadata extraction. Greenstone incorporates an interface that makes it easy for institutions to create their own library collections. Collections could be built and served locally from the user's own web server, or (given appropriate permissions) remotely on a shared digital library host. The other set of features include OAI plug-in (introduced since the 2.40 version) and DCMI compliance, UNICODE based multi-lingual capabilities and a user-friendly multimedia interfacing. Further more, it has a powerful search engine ('MG' PP) and metadata-based browsing facilities. A very interesting feature of Greenstone is its exhaustive set of well documented and articulated manuals (<http://www.greenstone.org/cgi-bin/library?e=p-en-docs-utfZz-8&a=p&p=docs>) such as 'Installer's Guide', 'User's Guide', 'Developer's Guide', and 'From Paper to Collection' a document describing the entire process of creating a digital library collection from paper documents. This includes the scanning and OCR process and the use of the "Organizer". There is one more interesting documentation 'Inside Greenstone Collections' which clarifies most of the trickier parts of using Greenstone, especially dealing with configuration file for the collection in question.

GSDL: Downloading, Installation, Configuration, accessing sample collections

Greenstone has been tested on Windows-x and Unix/Linux and the current version is 2.41. It can be downloaded from '<http://www.greenstone.org>' or '<http://www.sourceforge.com>'. 32 bit binaries are included in this distribution, as

well as complete source code (which is the same for Windows as well as Unix). The download is a self-extracting executable. To install, simply double-click the download file and follow the on screen instructions. The software allows installing a 'local library' or a 'Web library' setting, which could be selected at the time of installation. For a institutional library requirement, it is advised to select the 'local library' option. The software when successfully installed, provides a sample collection which could be used as a demo of the software. A thorough reading of the manuals and guides is strongly recommended before commencing the installation work. It is also advised to take help from the system administrators for want of any technical or OS/Network related doubts while installation and configuration.

GSDL: Collection building

Collection building is the second most important and continuing activity. Building up the collection using GSDL is fairly a simple process once we have read the manuals and guides properly. Basically there are three methods for collection development: the command-line, the Collector and using the GLI tool, which has been a recent introduction along with the 2.40 version. A big change with the latest release is the inclusion in the Windows download of the "Greenstone Librarian Interface", a tool for collecting, expanding and building digital libraries. It gives you access to Greenstone's functionality from an easy-to-use 'point and click' interface, and is also available for Unix systems. Other important changes include a new plug-in for importing data from perl's DBI interface, a new plug-in for parsing records from OAI servers, improvements to MGPP (Managing Gigabyte) configuration, and many bug fixes and other improvements. The collector method of collection building uses the Greenstone browser which is the most popular mode being used by many. With the introduction of the more user friendly GLI tool, it is assumed that this feature will pick up soon.

Steps in Collection Building include : (a) Providing the 'Collection information' such as ' Title for collection', 'E- mail Id of Collector', and a brief description 'About the collection'; (b) Declaring the 'Source data' such as a directory name on the Greenstone server system (beginning with "file://") , an address beginning with "http://" for files to be downloaded from the Web, or an address beginning with "ftp://" for files to be downloaded using FTP; (c) Configuring the collection. Next, the documents are converted into a standard XML form. Appropriate plug-ins to perform this operation must be specified in the collection configuration file. Then the full-text searching indexes and browsing structures specified in the collection configuration file are created. Finally, the result of the building process is moved to the area for active collections. This precaution ensures that if a version of this collection already exists, it continues

to be served right up until the new one is ready. The software assigns a global, persistent identifier to each document to ensure that the changeover is almost always invisible to users. The construction and presentation of all collections is controlled by specifications in a configuration file. Collect.cfg is configuration file, which is available in “etc” directory. It is possible to develop small collection as well as large collections as a group (batch processing), provided the documents in the collection are ready for collection building. That is, documents should be properly edited and tagged in advance as per our needs.

Each document may be *hierarchically* organized into logical *sections*, each of which comprises *paragraphs*. Metadata such as author, title, date, keywords, may be associated with documents, or with individual sections. This is the raw material for indexes. It must either be provided explicitly (for example, in an accompanying spreadsheet) or be derived automatically from the source documents. Metadata is stored with the document for internal use. Information collections built by Greenstone combine full-text search with browsing indexes based on different metadata types. There are several ways for users to find information, although they differ between collections depending on the metadata available and the collection design. *Searching* is full-text, and -- depending on the collection -- users select indexes built from different parts of the full text or the metadata. Some collections have separate indexes of full documents, sections, paragraphs, titles, and section headings, each of which can be searched for particular words or phrases. When *browsing*, users examine data structures created from metadata: lists of authors, titles, dates; hierarchical classifications; and so on. Structures for both searching and browsing are specified by instructions in the configuration file, and can be rebuilt entirely automatically. No information is inserted by hand. Details regarding collection building can be found in the Developers guide.

GSDL: Assigning metadata to a file

Metadata is simply, data about the data. That is, the descriptive information such as author, title, Keyword, and so on that is associated with the document. Metadata is stored within ‘Metadata’ tags <Metadata>. For example, <Metadata name=“Metadatatag”>Value</Metadata>, where ‘name’ is the attribute of <Metadata> tag. Metadatatag can be Title, Author, or any other element of metadata schema like Dublin Core. And the ‘Value’ is the data to which the metadata has to be assigned. The software is organized so that ‘plug-ins’ import documents and transform them into a standard XML form with metadata included. There are plug-ins for plain text documents; HTML, Word, PostScript and PDF files; email; and common bibliographic formats. New plug-ins can easily be written -- several have been specially produced for proprietary formats. If the collection contains source documents in different forms, it is just a

matter of specifying the necessary plug-ins. Plug-ins also perform metadata conversion, whether from internal sources such as HTML's *<title>* and *<meta>* tags or Word's "summary" properties, or from externally-specified XML or spreadsheet files.

GSDL: Viewing the collection

When the collection is built and installed, a *View collection* button becomes active. Clicking this button takes the user directly to the newly built collection.

GSDL: User Interface

The most significant set of features of Greenstone User Interface is its simplicity, flexibility and the high degree of customizability. The Greenstone user interface has been customized to highlight the unique features of the individual digital collections. The metadata description is presented in a standard library OPAC format with a thumbnail image. The full-size images in the digital object can be viewed with Image Viewer in another browser window. Full-text transcriptions in any formats are linked within the record and can be viewed through appropriate applications. Ian Witten and his Greenstone project suggest the interesting possibility of separating collection from interface. Greenstone opens an avenue for developing the interface iteratively, beginning with a generic facility that, over time, becomes more precisely customized. In general, Greenstone focuses the digital library user interface design effort on the portions of the interface that need to be customized, saving the effort that otherwise would be required for re-implementing parts that can be handled in a standard manner. In general Greenstone provides a Multivalent Browser, which embodies a new metaphor for information presentation and manipulation. The result allows the reader great flexibility in determining the content's visualization.

GSDL : Helpline, Archives

Greenstone's E-Mail list is a very useful and active listserv which shares and clarifies user experiences and stories dealing with real life situations. To subscribe or unsubscribe to the list via the World Wide Web, visit "<https://list.scms.waikato.ac.nz/mailman/listinfo/greenstone-users>" or, via email, send a message with subject or body 'help' to "greenstone-users-request@list.scms.waikato.ac.nz".

Greenstone Archives is a searchable collection of email messages from the Greenstone mailing list archives. The archive is available at "<http://www.sadl.uleth.ca/nz/cgi-bin/library?a=p&p=about&c=gsarch-e>". The

collection includes messages from the beginning of the mailing list in April 2000 up until fairly recently. The mailing list is used for communicating with the entire Greenstone team, therefore the content of the messages is usually global in nature. The mailing list is also a good way of getting help with problems - someone on the team will probably be able to help. This collection may be useful for finding solutions to common problems, or simply for tracking the progress of the Greenstone software.

GSDL : OAI Compliance

The OAI feature in Greenstone is a recent addition, which clearly shows the software team's persistent concern to international trends in open digital library and open archiving technologies. The OAI example is available at "<http://www.sadl.uleth.ca/nz/cgi-bin/library?a=p&p=about&c=oai-e>". This collection demonstrates Greenstone's *ImportFrom* feature. Using the Open Archive Protocol (version 1.1), it retrieves metadata from rocky.dlib.vt.edu/~jcdlpix, a collection of photographs taken at the inaugural Joint Conference on Digital Libraries. A Greenstone collection is built from the records exported from this OAI data provider. The implementation is flexible enough to cope with the minor syntax differences between OAI 1.1 and OAI 2.0.

IIMK DL Collection

The Digital Library Project at IIMK took off during September 2001, though the first four months were mainly spent on infrastructure build-up such as hardware and software, digitisation strategies, strategies regarding collection building, and finalisation of the variety of content categories etc. Since the campus intranet and the Internet connectivity were robust, there were no worries on those front. As already discussed earlier, the IIMK DL was slated to be populated with e-books, ('born digital' as well as digitised), e-journals, management cases, technical reports, working papers, faculty publications, teaching notes, CBTs, WBTs, educational videos / audios, value adding the existing OPAC, speeches, presentations etc. It took around two/three months for us to install and get proper command over the software. The DL has thus got a sizable collection of over 500 e-books covering almost all areas of business & management, literature, classics, and Information Technology. The collection comprise multiple digital formats such as HTML, PDF, PS, RTF etc. The existing collection also include over 1000 e-journals (gateway access only provided through Greenstone, but users can search on the title – free text, subject keywords, publisher, or other metadata), sourced from Elsevier, Kluwer, John Wiley, Blackwell, Emerald, Taylor & Francis, ACM, IEEE etc.; over 225 educational videos (the content development work is in progress). IIMK is planning to host the IIM Consortium digital archive and studies are underway on

the feasibility of putting these archives under Greenstone. The faculty publications are also being considered for inclusion in the collection. A massive collection development drive is being worked out with most of the popular electronic publishers specialising in business and management. One of the major bottlenecks to the speedy collection development is the of non-availability of proprietary and trade literature in Greenstone acceptable formats such as HTML, Word, or PDF. Most of these publishers put their materials in their own proprietary e-book reader formats, from which the text extraction becomes almost impossible. A set of screenshots of the IIMK DL is provided as exhibits below (Fig. 2).



Fig.2. Screenshots of IIMK DL