WORK PLAN
ON
ESTABLISHMENT
OF

INDO – ASEAN S & T DIGITAL LIBRARY

by

Indian Institute of Information Technology
(Deemed University)
Deoghat, Jhalwa, Allahabad – 211011

CONTENTS

1. Introduction

Identification of Nodal persons from each ASEAN country and ASEAN Secretariat was completed and for “Capacity Building’, training program was held during March 1 – April 15, 2010 in the Indian Institute of Information Technology – Allahabad.

The next step is development of basic infrastructure of Content Digitization Centre for national languages of ASEAN member states. For that purpose Indian Institute of Information Technology – Allahabad, India is arranging the installation of scanner, server and computers at different locations of member states, ASEAN Secretariat H.Q. and in Indian Institute of Information Technology – Allahabad. Two major steps are required at this stage. These are:

1. Installation of scanners, servers and systems at different centers of the member states.
2. Creation and Management of Digital Library Systems

.

2. Installation of Scanner, Server and Systems

The scanner, server and systems will be provided to the specified member states. The names and addresses of different organizations in different member states are being identified by the respective member states. The training on complete digitization process was given to the participants from different member states of ASEAN countries during 1st March – 15th April 2010. The hardware and software which have been recommended are

a) 12 Servers

b) 12 Scanners

c) 24 High end Computers

d) ABBYY OCR Software

e) ScanFix software.

The servers, scanners and high end computers along with the software will be made available to each member state of ASEAN countries, the ASEAN Secretariat and Indian Institute of Information Technology – Allahabad.

The training program held earlier on digitization was given on MINOLTA PS7000 scanners. As the functionalities and features of the scanners are very similar, so the participants will have no problem to give the training to their own staff with other model.

Type of scanner	Book Scanner
Document Type	Accommodates oversized bound volumes, ledgers, artwork, maps, archival records and other large documents.
Scan area/size	A2 Landscape 2* DIN A3 Min.: 105 mm x 149 mm Max.: 432 mm x 594 mm
Publication Height	Up to 420 mm, Thickness of book: minimum 5 cm and desirable up to 10 cm
Scan modes	➢ Manual, ➢ Automatic ➢ Simplex & Duplex (two pages face up)
Scanning Speed	➢ BW 2.5 sec ➢ Color 6 sec Less than 4.5 seconds, A4 at 400 dpi with image enhancement
Scan resolution	200 dpi, 400 dpi, 600 dpi
Bit depth	24 bits (3x8 bits)
Image sensor	3 x 35 millions of pixels
Projection Lamp	Required
Exposure	Automatic exposure, can be switch to manual exposure control
Sharpness	sharpening filter required
Color Adjustment	➢ Automatic white balance adjustment on white and black sample ➢ Can be controlled manually ➢ Can also be turned to automatic white balance at each scan
Image modes	TIFF, TIFF G4, JPEG, BMP , JPG2000, PDF

Image editing/Software interface	Image processing functions included ➢ deskew, ➢ finger masking, ➢ contrast enhancement, ➢ noise reduction, ➢ curvature correction ➢ Text and photo mode ➢ Finger shadow masking ➢ Centering ➢ Centre erase Masking ➢ Character reduction correction
Hardware Interface	➢ USB2.0 ➢ Giga bit Ethernet ➢ Twain ➢ ISIS
Software Interface	Embedded
PC environment	Required with Licensed OS
Options	screen bracket (Software), shortcut USB keyboard, footswitch, book cradle manual or automatic

Table 1: Specification of color scanner

As most of the recent documents of Science and Technology are born digital and there are color images, diagrams, maps etc. so color scanners will be beneficial. Keeping the above in mind, the specifications (as given in Table 1) have been identified for the color scanners.

3. Creation and Management of Digital Library Systems

The creation and management of the digital library system requires development of software that will allow the acquisition of data for the digital library, allow their storage, indexing and subsequent retrieval based on the requests of the users and also provide services that are required for the maintenance of the digital library. The architecture of the proposed Indo-ASEAN Science and Technology Digital Library was discussed in detail during the training session held at IIIT-A from 1st March to 15th April 2010. The key functional and non-functional requirements were identified during these discussions and it was decided that a Service Oriented Architecture would suit the purpose adequately.

To build the "INDO – ASEAN Science & Technology Digital Library", we need to

1) Develop software for creation of metadata in Dublin Core

2) Convert existing metadata in the same format

3) Develop software for implementing the policies regarding access to users, based on the copyright laws of each member country.

4) Develop software for acquisition and storage of content

5) Implement basic search/query technology

6) Implement the presentation of query result

7) The processes for the primary services and secondary services.

8) Develop software for multilingual query handling

9) Development and deployment of secondary services of Digital Library.

Format of metadata:

Some of major elements for metadata are:

Title	Subject and Keyword	Description	Resource Type
Source	Relation	Coverage	Creator
Publisher	Contributor	Rights management	Date
Format	Resource Identifier	Language	Audience
Provenance	Rights Holder		

In addition, INDO – ASEAN SCIENCE & TECHNOLOGY DIGITAL LIBRARY also maintains some extra elements in the metadata that help in the management of the repository.

Architecture:

The architecture of Digital Library Workflow and the policies and Issues of Digital Repository should make the use of Data Transporter service at service layer in the architecture to provide access to digital data stored in digital repositories. Data Transporter service of Digital Library Workflow and Digital Repository Management uses a general item naming system based on metadata fields for digital data and incorporating it with network protocols to provide a fast and efficient approach for data access and depositing tasks. The architecture of Digital Library Workflow and Digital Repository Management is shown in Figure 1.1.

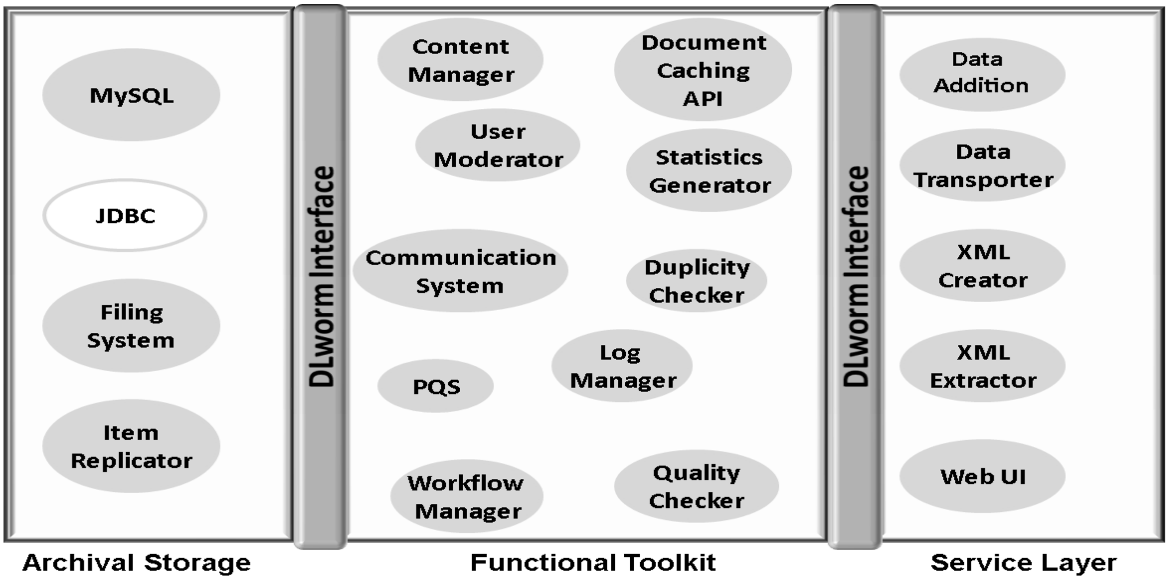

Figure 1.1 Architecture of Digital Library (Technical perspective)

The functionalities of the proposed architecture drive the workflow of a digital library are shown in Fig 1.2.

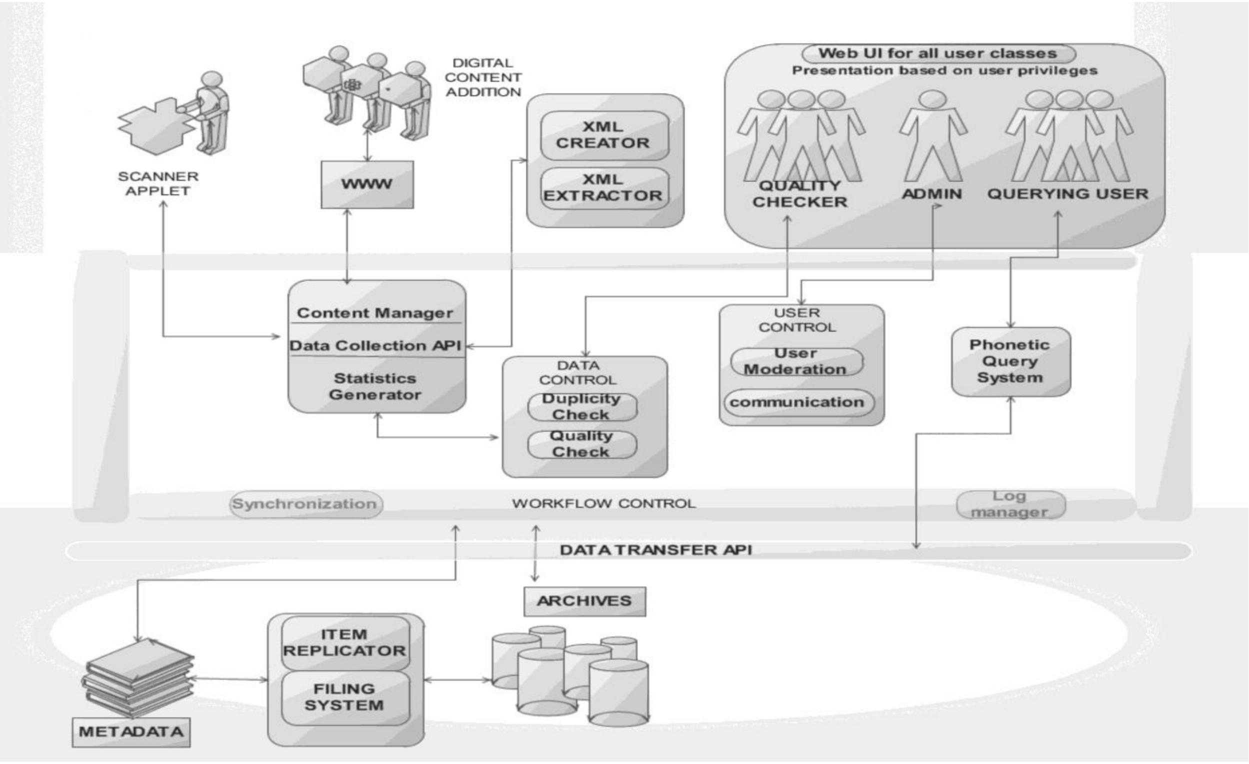

Figure 1.2 Workflow of a digital library by using the services.

The major functionalities can be summarized as:

- Central database will maintain the information, permissions and statistics for all **users**.
- Will provide a **quality checking** module for reviewing the submitted data.
- A three level hierarchy for **data organization** is defined.
- Dublin core specifications will be used for **metadata**.
- At each level metadata/ submitted data will be checked for **duplicity**.
- It provides **metadata based with phonetic matching query system** to end users for searching into the data.
- will provide **facility for communication** between users involved in the system.
- **Interface** for each user is web based.

We can divide the whole design in basic four sections:

i) INDO – ASEAN SCIENCE & TECHNOLOGY DIGITAL LIBRARY Master Server.

ii) Web-User Server

iii) Quality Checking Resource Centre and

iv) Repository Systems.

INDO – ASEAN SCIENCE & TECHNOLOGY DIGITAL LIBRARY will be developed as a modular system where new functionalities have been added in stepwise fashion taking advantage of the object oriented design of the system. The underlying code will be developed in JAVA. JAVA Applets running on Unix-like Systems such as Linux. INDO – ASEAN SCIENCE & TECHNOLOGY DIGITAL LIBRARY will fulfill most of the functionalities required for Digital Library System. Moreover, the use of Java and an Applet based interface allows the system to be platform independent.

The basic architecture of the digital library will be developed and tested in IIIT-A. After thorough testing they will deployed at the sites selected by the member states. The ASEAN-INDIA Science and Technology Digital Library system aims to become a rich repository of all science and technology related documents available in the member states. It will be a web based and distributed system which will be accessible to all ASEAN Member States as well as India. Policies regarding access to the material, available on the digital library, will be decided by the member states and India. For the project to be successful there is an urgent need of development of specific language technologies in each of the member states. This development has to be collaborative. In the following we describe each technology along with the specific linguistic resources that will be required for each technology.

4. Language Technologies Required for the Digital Library

This system will provide a multilingual query system. The expected user of this system will be the people of individual countries who need the knowledge from the rich digital repository. In order to meet the above objective, the development of the following Language Technologies were recommended along with the development of the Science and Technology digital library:

i) Optical Character Recognition
ii) Multilingual Information Retrieval Systems / Search Engines
iii) Multilingual e-dictionaries for cross language Information Retrieval
iv) Transliteration and Translation Technology

These developments will be adequate for a large section of the users, even if they do not know English. However, some physically challenged users may find it difficult to use the system. Therefore, after the completion and deployment of the above services, we can add specific modules that will address the requirements of this special category of users. A brief description of each of the above, along with the resources required for their development, is given below:

i) Optical Character Recognition

The end-to-end Optical Character Recognition (OCR) system will be developed with reference to the following functional specifications. OCR for the scripts of the main official languages of different ASEAN Member States will be developed.

The input to the scanners will be scanned pages of scientific books, documents, journals. For each script 50 books published at different times over last 50 years will be considered. Each book is expected to have on average 300 pages. These pages are expected to be representative examples of the quality of pages that the developed OCR systems will be able to handle. These pages are also expected to contain representative examples of scripts (fonts and sizes) and layout patterns (including graphics and image components). These pages will form the annotated corpus for development and testing.

Unicode text documents will be output. These will represent the pages with appropriate tags so that layout and font information, along with graphics and image component, can be retained to the maximum possible extent.

Technology Development

Technology that needs to be developed to meet the system specification can be broadly categorized into three classes. These technologies are complimentary. These technologies will be combined together to implement end-to-end script specific OCR's.

1. Script Independent Technology
2. Script Dependent Technology
3. Supporting Technology & Resource Development

The development of the OCR for each script requires resources like annotated corpus for training and testing the system. The basic OCR engine will be developed in IIIT-A.

The manpower required for the development of the linguistic resources is two professionals for each language who are familiar with their respective languages and English.

ii) Multilingual Information Retrieval Systems / Search Engines

Interactions among different languages can be facilitated through multilingual information retrieval system. We will have a rich repository of documents in several languages. It can be divided into the following sub-modules:

1. Tokenizer
2. Index of the documents stored in the digital library using the metadata and the full text (if available)
3. Query parser

4. Synset and lexicon set
5. Word sense disambiguation module
6. Named entity recognition module
7. Ranking module
8. Search module
9. Automatic key word generation tool (from full text if available)
10. Presentation module

The multilingual information retrieval system will form the heart of the digital library. This work can start only when the basic digital library is developed and deployed. It may be noted that some of the modules depend critically on the availability of the language resources (items 1 and 3 – 6 above). Each member state will have to generate these resources for their respective language to ensure the overall success of the digital library.

The manpower required for the development of these language resources is two linguists for each language who should have a thorough knowledge of their respective languages and English.

iii) <u>Multilingual e-dictionaries for cross language Information Retrieval</u>

An essential component of the Indo-ASEAN Science and Technology Digital Library is the development of multilingual e-dictionaries. It is crucial for the successful development of several secondary services like the multilingual IR system etc. The format for the e-dictionary and the software for its development will be done at IIIT-A. The software will then be in all the centers for the creation of the multilingual dictionary. The dictionary will have different entries for each sense of the word along with its possible synonyms , hypernyms and hyponyms. It will try to provide a complete ontology of the commonly used words. This will enhance the accuracy of the retrieval of documents and any machine translation system that may be developed at a later stage.

iv) Transliteration and Translation Technology

Transliteration and Translation Technology will have many modules. Each module will handle the translation of one pair of languages. The utility of these modules is that it will allow the system to translate a query in one language to all the other languages. Thus, even if the user provides the query in one language, documents relevant to that query in other languages will also be retrieved and displayed to the user. The user will have the freedom to choose the languages in which she / he wants to see the results and the system will comply with the choice.

v) Developments for the Physically Challenged

Java3D API will be used to program the structure of models and to provide visual inputs as well as the output. This API needs a lot of programming job by providing custom classes to define basic 3-dimensional primitives easily. Since Java3D is in evolution phase, there is not much help available on Internet as well as among the programming fraternity. Use of Java will assur the platform independence of the application and also easy integration with other modules developed in Java.

5. Time Estimation

Total time period for the project is 36 months.

INDO – ASEAN S & T DIGITAL LIBRARY software (English Language) – 18 Months.
To make this system available for all other languages:
Generation of Language resources: 12 – 18 months

SPECIFIC INTERMEDIATE MILESTONES

End of 6 Months

Tasks completed for English Language

1. POS Tagger
2. Stemmer

3. Named entity recognizer

4. Simple Parser

5. Word Sense Disambiguation

<u>End of 12 Months</u>

Task completed for English Language

1**.** Crawler

2. Indexing and metadata creation module

3. Ranking based on metadata

4. Document summarization

5. Identification of the character set for each script for which OCR is to be built

<u>End of 18 Months</u>

Task completed for English Language

1. S & T Digital Library System

2. Multilingual dictionary

3. Lexicon for different languages

4. Synsets for different languages

5. Annotated corpus for different languages

6. Text documents in different languages for OCR

7. Basic OCR engine and training sample generator for OCR

<u>End of 24 Months for other languages</u>

1. POS Tagger

2. Stemmer

3. Named entity recognizer

4. Simple Parser

5. Word Sense Disambiguation

6. Multilingual mapping

7. Crawler

8. Indexing

9. Ranking

10. Training and testing set for OCR in different scripts

End of 30 Months for other languages

1. Document summarization
2. Query Translation
3. Query Transliteration
4. Multiword expression
5. Query Disambiguation
6. Searching
7. Basic Optical Character Recognition system for some scripts
8. The work for System Integration and Testing will start

End of 36 Months for other languages

INDO – ASEAN S & T DIGITAL LIBRARY software for all languages of member states will be ready.

5. Manpower

The development of different modules required for the Indo-ASEAN Science and Technology Digital Library will require manpower at IIIT-A and at different centers of the member states.

The supervisors at IIIT-A will be Dr. Ratna Sanyal and Prof. Sudip Sanyal. In addition, there will be three technical persons, one support staff and twenty five students (Research Scholars, M. Tech. and B. Tech.) involved in the development and testing of the systems. These people will have experience in natural language processing, information retrieval and software engineering.

Apart from the manpower at IIIT-A, one technical person and two linguists will be required from each member state and at the ASEAN headquarters. Each linguist should have good knowledge of the official language of that country as well as English.

There should be two workshops within the first six months to finalize the standards, formats of the modules and language resources. Thereafter, a review meeting will be required at the end of each six month period to find the progress made, identify any problems / bottlenecks and find solutions for the same. These reviews are essential for the successful completion of the project.