A tool for Entering Structural Metadata in Digital Libraries

Lavanya Prahallad, Indira Thammishetty, E.Veera Raghavendra, Vamshi Ambati

MSIT Division, International Institute of Information Technology, Gachibowli, Hyderabad 500032, India
Email: lavanyap@cmu.edu; indirat@andrew.cmu.edu; raghavendra@iiit.net; vamshi@andrew.cmu.edu

***Abstract* - Digital libraries harbour thousands of scanned books as Tiff images. Given a book, a user would like to quickly see the main details like Author, Title, no of chapters, start and end of the chapters in the book. However, the scanning process of a book results in a set of digital images and OCR-ed text (if OCR is available in that language). Extraction of details of structural meta-data from OCR-ed text is not that easy, and it is even more difficult to extract the information from the images. To overcome this issue, the structural meta-data is entered manually for each book, and a tool is required to key-in the structural meta-data. In this paper, we describe a tool developed to key-in the structural meta data for different language including Indian languages using QWERTY key-board.**

***Index Terms* — Structural Metadata, Metadata, IT3, Dublin Core, Universal digital library.**

I. INTRODUCTION

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information. In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital. Other metadata schemes have been developed to describe various types of textual and non-textual objects including published books, electronic documents, archival finding aids, art objects, educational and training materials, and scientific datasets.

There are three main types of metadata:

- ***Descriptive metadata*** describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
- ***Structural metadata*** indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- ***Administrative metadata*** provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

Metadata is a key to ensure that resources will survive and continue to be accessible into the future.

What Does Metadata Do?

Metadata is used to speed up and enrich searching for resources. In general, search queries using metadata can save users from performing more complex filter operations manually. Using metadata we will be able to search or identify the book uniquely.

Descriptive metadata can be used to automize workflows. For example, if a tool knows content and structure of data it can convert them automatically and give them to another tool as input. By that, users could save many copy-and-paste actions that are necessary when analyzing data with different tools.

II. STRUCTURAL METADATA

Information that ties each object to others to make up logical units (e.g., information that relates individual images of pages from a book to the others that make up the book) is known as structural Metadata. Metadata schemes (also called schema) are sets of metadata elements

designed for a specific purpose, such as describing a particular type of information resource. The definition or meaning of the elements themselves is known as the semantics of the scheme. The values given to metadata elements are the content. Metadata schemes generally specify names of elements and their semantics. Optionally, they may specify content rules for how content must be formulated (for example, how to identify the main title, page no, image no etc.)

A metadata scheme with no prescribed syntax rules is called syntax independent. Metadata can be encoded in any definable syntax.

Many current metadata schemes use SGML (Standard Generalized Mark-up Language) or XML (Extensible Mark-up Language). XML, developed by the World Wide Web Consortium (W3C), is an extended form of HTML that allows for locally defined tag sets and the easy exchange of structured information. SGML is a superset of both HTML and XML and allows for the richest mark-up of a document. Useful XML tools are becoming widely available as XML plays an increasingly crucial role in the exchange of a variety of data on the Web.

Types of structural metadata:

Many different metadata types have been developed in a variety of user environments and disciplines. Some of the most common ones are listed below:

1. Dublin core
2. MODS (Metadata object description schema)
3. METS(Metadata encoding and transmission standard)

Most common schema followed in UDL (Universal digital library) is Dublin core as shown in Fig. 1 and Fig. 2. Fig. 1 shows the example of structural meta-data in Dublin core and Fig. 2 shows implementation of Dublin core in XML format.

Dublin core:

The Dublin Core Metadata Element Set arose from discussions at a 1995 workshop sponsored by OCLC and the National Center for Supercomputing Applications (NCSA). As the workshop was held in Dublin, Ohio, the element set was named the Dublin Core. The original objective of the Dublin Core was to define a set of elements that could be used by authors to describe their own Web resources. Faced with a proliferation of electronic resources and the inability of the library profession to catalog all these resources, the goal was to define a few elements and some simple rules that could be applied by non-catalogers. The original 13 core elements were later increased to 15: *Title*, *Creator*, *Subject*, *Description*, *Publisher*, *Contributor*, *Date*, *Type*, *Format*, *Identifier*, *Source*, *Language*, *Relation*, *Coverage*, and *Rights*. The Dublin Core was developed to be simple and concise, and to describe Web-based documents. Find below an example of the Dublin Core:

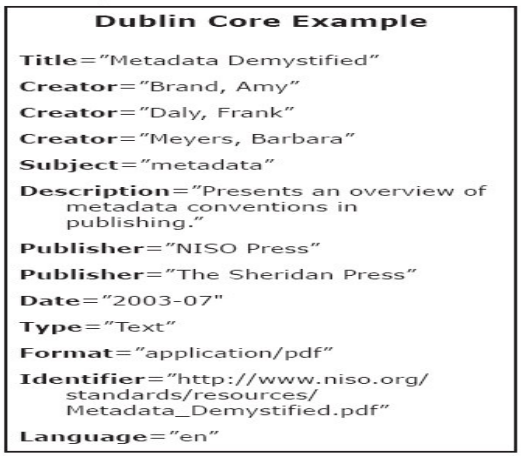

Fig 1: Dublin Core Example

Fig 2 Sample Metadata in Dublin Core in XML format

All Dublin Core elements are optional and all are repeatable. The elements may be presented in any order. Because of its simplicity, the Dublin Core element set is now used by many outside the library community— researchers, museum curators, and music collectors to name only a few. There are hundreds of projects worldwide that use the Dublin Core either for cataloging or to collect data from the Internet. The subjects range from cultural heritage and art to math and physics. Meanwhile the Dublin Core Metadata Initiative has expanded beyond simply maintaining the Dublin Core Metadata Element Set into an organization that describes itself as "dedicated to promoting the widespread adoption of interoperable metadata standards and developing specialized metadata vocabularies for discovery systems."

III. TOOL FOR ENTERING STRUCTURAL METADATA

Digital libraries harbour thousands of scanned books as Tiff images. Given a book, a user would like to quickly see the main details like Author, Title, no of chapters, start and end of the chapters in the book. However, the scanning process of a book results in a set of digital images and OCR-ed text (if OCR is available in that language). Extraction of details of structural meta-data from OCR-ed text is not that easy, and it is even more difficult to extract the information from the images. To overcome this issue, the structural meta-data is entered manually for each book, and a tool is required to key-in the structural meta-data.

The issues or requirements of building such tool are:

1. Tool should be able to handle images (obtained from scanning) in different formats such as Tiff.
2. It should allow to key-in the text in any language typically non-English languages using QWERTY keyboards
3. It should allow the display the keyed-in text in the native script using Unicode
4. An easy and user-friendly interface to browse through the pages, to key-in/edit the meta-data.

Our tool is based on the windows acquisition library where the operator (who keys-in the meta-data) needs to browse for the book from their local drives and select the Tiff images folder. All the tiff images in that folder are displayed, and there could be browsed one after the other easily with the click of a button or a short-cut key. The user can also keep going to any desired page. There would be an edit button where he can enter the structural Metadata. The interface of our application is user friendly as the operator can see the scanned page as tiff image clearly, so that he would be able to see where the particular chapter starts or ends and accordingly he can enter the Metadata. The fields provided in the edit page are: Title of the book, Barcode no, Image no, Page no, structure type and structure description.

Title of the book: Refers to the name of the book. This can be automatically taken from the already entered metadata page which is stored as meta.xml.

Barcode no: Each book is being identified by its unique barcode number. It is used to locate the book.

Image no: Every page in a book is scanned and saved as a tiff image. The number corresponding to each tiff image is the Image no and the convention followed to name the book is as shown below: 00000001.tif.

Page No: Page number is the number given inside the image. This is helpful to navigate between the pages in Book reader.

Structure Type: This is also used for easy navigation. The structure type is used to classify the whole book in terms of Chapter, Introduction, Title Page, Appendix, Glossary etc.

Structure Description: This is a detailed description of a structure type. For Example: If we select chapter 1 as structure type, the name of the chapter and the page it starts or ends can be the structure description. This data can be in the form of IT3 for Indian Languages.

The operator needs to enter the fields of Page no, Structure type and Structure description manually, where as Title of the book, Barcode no and the Image no are automatically displayed with the help of the metadata for that book which is saved as an XML file. After entering the structural metadata for a page, we need to just use the return key or press the Save button, the entered data will automatically gets saved into an XML file (E.g. structure.xml). The same procedure is followed for the remaining pages in the book, and the entered structural metadata will get appended to the already saved XML file.

Fig 3 and Fig 4 show the screen shots of the tool and interface to enter the structural metadata.

Fig 5 shows the structural metadata which is saved as an XML document.

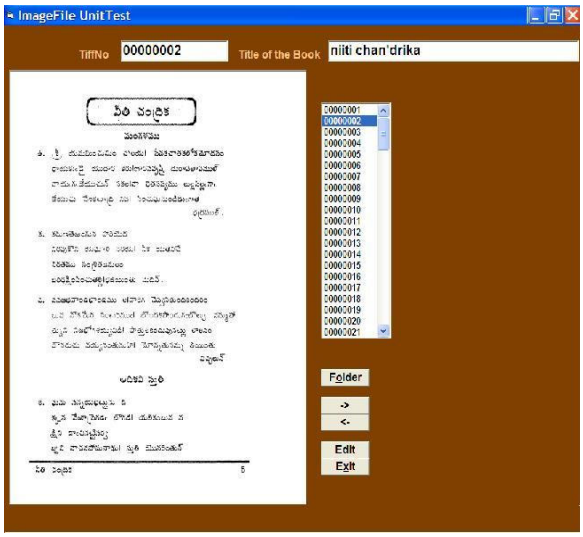

Fig 3 Screen shot of the tool

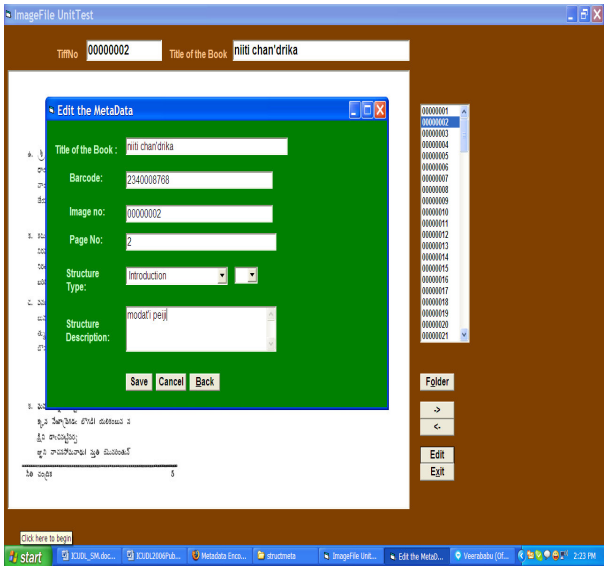

Fig 4 Structural Metadata entry form

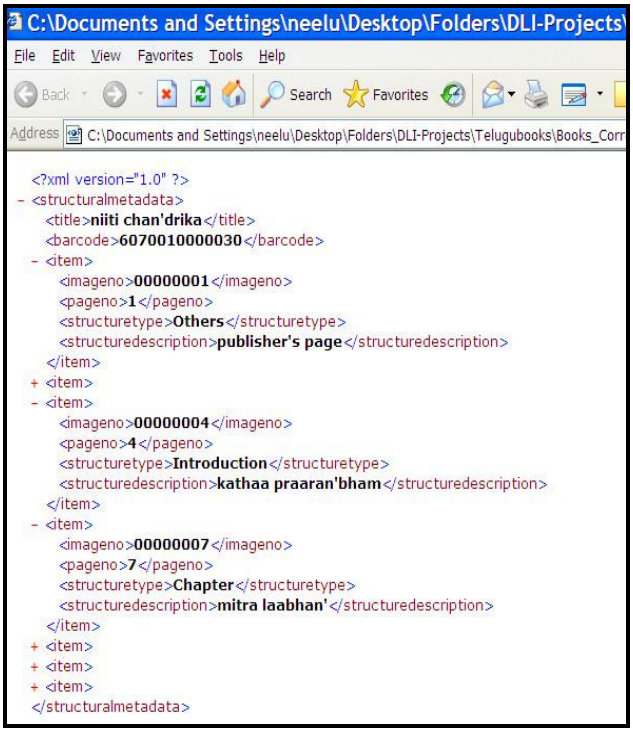

Fig 5 Sample Structural Metadata

Technology Used:

1. Our tool is built on the windows platform and works only on windows XP with service pack 2 installed.
2. We are using the windows acquisition library and its DLLs are used to display the tiff image.
3. We are using the XML parser to store the data in the xml format as it plays crucial role in the exchange of a variety of data on the Web.
4. Out tool supports storage of the IT3 characters in XML files.
5. This software is a stand alone application developed in Microsoft Visual Basic 6.0

Features:

Requirement	Features of our tool
Handles Tiff images	Uses windows acquisition library to display the tiff images
IT3 display	Our tool enables to enter the data in IT3 which is used to key in for the non English languages.
User Friendly	Interface is simple and easy for the operators to enter the structural metadata
Save the values to XML	Our tool uses the XML parser and saves the entered structural metadata (even IT3) into an XML document.

IV. ADVANTAGES OF OUR TOOL

1. Easy Navigation of the Book
2. Key board shortcuts – To reduce number of mouse clicks and to save time.
3. Gives the Broader description of the books.
4. Well structured XML format files are generated.
5. Special characters are recognized.
6. User Friendly.

V. CONCLUSIONS

In this paper, we have discussed the requirements of entering structural meta-data for a digital book. We briefly described different types of structural meta-data and Dublin-core format that is being adapted in Universal Digital Library (UDL). Finally, we described the tool developed which aid in faster key-in of the structural meta-data.

ACKNOWLEDGEMENTS.

We express our sincere gratitude to Prof. Raj Reddy, Carnegie Mellon University, for his motivation and guidance. We would like to thank Mr. S.P Kishore for his valuable comments and feedback. We would like to thank Vamsi Krishna for his help during the development of the tool.

REFERENCES

[1] http://www.ifla.org/II/metadata.htm
[2] http://en.wikipedia.org/wiki/Metadata
[3] http://dublincore.org/
[4] http://niso.org/standards/resources/UnderstandingMetadata.pdf
[5] Vamshi Ambati, N.Balakrishnan, Raj Reddy, Lakshmi Pratha, C V Jawahar: The Digital Library of India Project: Process, Policies and Architecture, In the Proceedings of 2nd International Conference on Digital Libraries(ICDL), 2006
[6] http://www.ulib.org