Metadata Schemas for Long Term Preservation of Digital Documents - A Survey

Nagendra Panini Challa, Dr.R.Vasanth Kumar Mehta
Department of CSE,
SCSVMV University
Enathur, Kanchipuram-631561

Abstract—This paper gives an overview about different metadata schemas used for preserving digital documents. The resources specified here gives us a head start in understanding various metadata standards. They are used for several purposes in efficient data retrieval from database. However major standards are addressed where more research work is needed to retrieve data efficiently.

Keywords—Metadata, Digital documents, Metadata Schemas, Data Retreival.

I. INTRODUCTION

In this present world, Technology has led to the development of new generations of systems, file formats, etc., but the information often has to be interpretable over long periods of time for efficient long term preservation of data. Its main importance is giving a brief overview about the context of data. With the widespread use of ICT applications a need for conceptualization of metadata arises [1].

Any data related to our heritage can be preserved for long term with proper maintenance activities such as palm leaf manuscripts. However, they are very susceptible to degradation by nature if the conditions are not ideal. The Manuscript Library at Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya (SCSVMV University), is home to several ancient manuscripts, and served as the source for the work carried out in this study and implementation. More than 5 lakh pages have been catalogued and are in different stages of digitization. The aim of this paper is to study various metadata schemas which are used in document management systems.

Any information has 3 main features namely Content, Context and Structure. Generally, Content is "Raw information becomes content when it is given a usable form intended for one or more purposes". Context is whether the information is related to the specified content or not. Structure is the packaging of information to perform a specified task [2].

Metadata is mainly classified into 3 types namely descriptive where the data is fully described and some of the aspects associated with the object's creation and preservation are described in administrative type of metadata. In structural metadata it relates to relations among individual information objects. Here are some of the advantages of Metadata schemas like increased accessibility, retention of context, Multi-versioning, Preservation.

The rest of the paper is organized as follows: section 2 discusses the current metadata schemas which are followed for long term preservation of digital documents. Digitizing data is the only solution before researchers for preserving valuable heritage writings like palm leaf manuscripts and many more. The main challenge is retrieving the digitized data efficiently according to user requirements. For this purpose there is a need to follow some metadata standards for easy management of data. In section 3 by whom the metadata is actually created is described in brief. In section 4 after discussing various metadata schemas the paper is concluded.

II. METADATA SCHEMA

There are many schemas that have been developed for certain types of data, and if your data matches a developed schema, the use of that schema will result in the best metadata for your data.

However, many of these schemas are very complex and require a level of expertise to implement that preclude their use by many. In the following section of this paper we will discuss about different metadata schema models used for long term preserving of data. There are different types of metadata as listed below:

A. MARC:

It is generally known as Machine Readable Cataloguing. It is defined as a set of standards used to identify, store and communicate cataloguing information. Some of the cataloguing elements are

- *020 -- International standard book number*
- *037 -- Source of acquisition*
- *040 -- Original source of cataloguing*
- *041 1- Language code*
- *1xx -- Author main entry*
- *246 -- Varying form of title*
- *250 -- Edition statement*
- *500 -- General notes*
- *504 -- Bibliography, etc. note*
- *505 -- Contents note*
- *520 -- Summary, abstract, annotation, scope, etc…*

They follow machine understandable code for cataloguing information [3].

B. DUBLIN CORE

The Dublin Core Metadata Initiative (DCMI) began in 1995 with an invitational workshop in Dublin, Ohio that brought together librarians, digital library researchers, content providers, and text-mark-up experts to improve discovery standards for information resources. The original Dublin Core emerged as a small set of descriptors that quickly drew global interest from a wide variety of information providers in the arts, sciences, education, business, and government sectors. The Dublin Core is not intended to displace any other metadata standard. Rather it is intended to co-exist — often in the same resource description with metadata standards that offer other semantics. It is fully expected that descriptive records will contain a mix of elements drawn from various metadata standards, both simple and complex.

Some of the elements in Dublin core metadata are
Element Name: Title
Element Name: Creator
Element Name: Subject
Element Name: Description
Element Name: Publisher
Element Name: Contributor
Element Name: Date
Element Name: Type
Element Name: Format
Element Name: Identifier
Element Name: Source
Element Name: Language
Element Name: Relation
Element Name: Coverage
Element Name: Rights
This group of core elements are known are Dublin core metadata element set.

Darwin Core (DwC) is an extension of Dublin Core for biodiversity informatics applications. The Darwin Core is body of standards. It includes a glossary of terms (in other contexts these might be called properties, elements, fields, columns, attributes, or concepts) intended to facilitate the sharing of information about biological diversity by providing reference definitions, examples, and commentaries. The Darwin Core is primarily based on taxanomy, their occurrence in nature as documented by observations, specimens, samples, and related information. It is broader in scope and more versatile. It is meant to provide a stable standard reference for sharing information on biological diversity. As a glossary of terms, the Darwin Core is meant to provide stable semantic definitions with the goal of being maximally reusable in a variety of contexts [4].

C. MPEG-7

The Moving Picture Coding Experts Group (MPEG) is a working group standard, (International Standards Organization/International Electro-technical Committee) in charge of the development of international standards for compression, decompression, processing, and coded representation of moving pictures, audio, and a combination of the two.

It is a standard for describing features of multimedia content. It is an ISO/IEC standard being developed by MPEG, the committee that also developed the Emmy Award-winning standards known as MPEG-1 and MPEG-2, and the 1999 MPEG-4 standard. These descriptions are based on catalogue (e.g., title, creator, rights), semantic (e.g., the who, what, when, where information about objects and events) and structural (e.g., the colour histogram - measurement of the amount of colour associated with an image) features of the AV content and leverages on AV data representation defined by MPEG-1, 2 and 4 [5].MPEG-7 uses XML Schema as the language of choice for content description MPEG-7 will be interoperable with other leading standards.

D. EAD:

Encoded Archival Description (EAD) is the international metadata transmission standard for hierarchical descriptions of archival records. It is developed by the EAD Working Group of the Society of American Archivists and first published in 1998, EAD is an Extensible mark-up Language (XML) format used by archivists around the globe. It supports the general structure of finding aids used by archivists and comprises three primary groups of information:

•*Administrative Information: Repository details, how the collection was acquired, access/usage restrictions, etc.*
•*Descriptive Information: Biographical or historical note about the creator of the collection, scope note, control access terms.*
•*Folder List: A list of the materials that make up the collection, by box, folder, item, or other designation.*

AN OVERVIEW OF THE EAD TAGS IS GIVEN BELOW:

Element:
An element describes the data it contains. Elements are enclosed by angle brackets. Each use of an element must include both an opening tag and a closing tag.
For example
•<unittitle>Title Name</unittitle>
•<unitdate>5 March 1991</unitdate>
•<physdesc>Brief Summary </physdesc>
•<processinfo>Process Information</processinfo>
•<persname>Name of Person</persname>
Attribute:
Attributes provide additional information about elements. An attribute name must be followed by an equals sign (=) and the value of the attribute must be enclosed in double quotation marks (“). For example:
•<container type="Box">4</container>
•<unitdate normal="1997-03 -05">March 5, 1997</unitdate>
•<persname source="lcnaf" encoding analog="600" role="subject" normal="Adams, John Quincy, 1767-1848">John Quincy Adams</persname> [6].

E. RDF

The Resource Description Framework (RDF), developed under the auspices of the World Wide Web Consortium (W3C), is an infrastructure that enables the encoding, exchange, and reuse of structured metadata. This infrastructure enables metadata interoperability through the design of mechanisms that support common conventions of semantics, syntax, and structure. RDF uses XML (eXtensibleMarkup Language) as a common syntax for the exchange and processing of metadata. The properties associated with resources are identified by property-types, and property-types have corresponding values. Property-types express the relationships of values associated with resources. For example values such as text strings, numbers, etc. or other resources, which in turn may have their own properties.

F. MODS

It is generally known as metadata object description schema. As an XML schema, the "Metadata Object Description Schema" (MODS) is intended to be able to carry selected data from existing MARC 21 records as well as to enable the creation of original resource description records. It includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. It follows XML Schema language.
For Example: A Tag <titleInfo>is defined as a word, phrase, character, or group of characters, normally appearing in a resource, that names it or the work contained in it.
The following elements are defined for "titleInfo" as
<title>
<subTitle>
<partNumber>
<partName>
<nonSort>
A Tag <name> is defined as name of a person, organization, or event (conference, meeting, etc.) associated in some way with the resource.

The following elements are defined for “name” as
<namePart>
<displayForm>
<affiliation>
<role>
<description> [7].

G. VRA CORE

A choice of one of three elements, Work, Collection, or Image, defines a VRA 4.0 record as describing a Work (a built or created object), a Collection (an aggregate of such objects), or an Image (a visual surrogate of such objects.) In the XML schema, this differentiation is expressed as an upper-level wrapper element named work, collection, or image that contains within it the remainder of the Core element set. Each XML sub element in an element has 3 attributes namely id, refid and source.

H. MIX

This standard defines a set of metadata elements for raster digital images to enable users to develop, exchange, and interpret digital image files. The dictionary has been designed to facilitate interoperability between systems, services, and software as well as to support the long-term management of and continuing access to digital image collections.

I. DDI

Data Documentation Initiative (DDI) is an international program to produce a metadata specification for the description of social science data resources. The DDI-tree contains five main branches, or sections:

The Document Description: Citation, Guide to the documentation, Documentation Status, Documentation source. The study description: Study scope, Methodology and processing, Data access. The Data Files Description: File Description. The variable description: Variable, NCube. A consequence of this approach is, however, that the DDI do not have a level of abstraction above a concrete dataset or statistical product. There is a one-to one relationship between a DDI instance and the physical data it is meant to describe. The DDI is tied to the dataset, or put differently; the DDI abstraction ladder stops with the dataset.

J. IEEE LOM

The IEEE 1484.12.1 – 2002 Standard for Learning Object Metadata is an internationally recognized open standard (published by the Institute of Electrical and Electronics Engineers Standards Association) for the description of “learning objects”. Generally LOM is for Sharing descriptions of learning resources between resource discovery systems, tailoring of the resource descriptions to suit the specialized needs of a community and to “tag” learning resources with a description that can be associated with the resource. Generally LOM is hierarchy of elements. The data model specifies that some elements may be repeated either individually or as a group of elements. Some element datatypes simply allow a string of characters to be entered, others comprise two parts such as LangString items contain Language and String parts, allowing the same information to be recorded in multiple languages. Vocabulary items are constrained in such a way that their entries have to be chosen from a controlled list of terms. Date, Time and Duration items contain one part that allows the date or duration to be given in a machine readable format, and a second that allows a description of the date or duration. The main requirements are to understand user/community needs, to have a strategy for creating high quality metadata, to store this metadata in a form which can be exported as LOM records, to be able exchange records with other systems either as single instances or in a group [8].

III. METADATA FOR PLM’S

Palm Leaf Manuscripts are most valuable and precious writings in South Asia especially in India. There are several ways of processing palm leaves, these methods differ from region to region. Cataloguing these scripts from these PLMs is the major challenge before researchers. The metadata described for digital documents suits for these palm leaf manuscripts but additionally some elements like damage percentage, hole position, origin of manuscripts, contributors for PLMs etc… should be added in order to make manuscripts retrieved more efficiently. Many researchers proposed different schemas for metadata extraction but they are follow other document schemas. So separated approach should be implemented to auto generate all these manuscripts metadata. Some metadata schemas have proposed by different researchers but among them schema proposed by Lampang is mostly desirable. He classified the metadata lifecycle into 3 fields namely analysis of user needs, development of metadata schema and implementation and evaluation of final metadata schema [9]. In this schema firstly it consists of identification of each type of document metadata requirements and design and development of metadata elements. The design and development of metadata elements were based on the analysis of metadata elements and extraction, and metadata vocabulary development. These two important activities are done by Functional Requirements for bibliographic records model (FRBR) [12].

IV. WHO CREATES METADATA & HOW?

Many projects have found that it is more efficient to have indexers or other information professionals create the descriptive metadata, because the authors or creators of the data do not have the time or the skills.

In other cases, a combination of researcher and information professional is used. The researcher may create a skeleton, completing the elements that can be supplied most readily. Then results may be supplemented or reviewed by the information specialist for consistency and compliance with the schema syntax [10]. Many metadata project initiatives have developed tools and made them available to others, sometimes for free. A growing number of commercial software tools are also becoming available. Creation tools fall into several categories: *PreDefinedFiles(Templates)* allow a user to enter the metadata values into pre-set fields that match the element set being used. The template will then generate a formatted set of the element attributes and their corresponding values. *Mark-up tools* will structure the metadata attributes and values into the specified schema language. Most of these tools generate XML or SGML Document Type Definitions (DTD). Some templates include such a mark-up as part of their final translation of the metadata. *Aquisition tools* will automatically create metadata from an analysis of the digital resource. These tools are generally limited to textual resources. The quality of the metadata extracted can vary significantly based on the tool's algorithms as well as the content and structure of the source text. These tools should be considered as an aid to creating metadata. The resulting metadata should always be manually reviewed and edited. *Conversion tools* will translate one metadata format to another. The similarity of elements in the source and target formats will affect how much additional editing and manual input of metadata may be required. Metadata tools are generally developed to support specific metadata schemas or element sets [11].

CONCLUSION

There are many metadata standards discussed in this paper for preserving documents in their long run, but these are restricted only to documents or images. There are no specific standards specified for Heritage documents point of view. Many researchers proposed schemas for these heritage documents but there are many issues such as lack of quality in metadata. Automatic generation of metadata from digitized palm leaf manuscripts is one such field where more attention is needed towards metadata development.

REFERENCES

[1] Gonzalez, Rafael C., Richard E. Woods. : Digital Image Processing. Ed III, Pearson Education Asia, New Delhi, 2007.

[2] NISO: Understanding Metadata. National Information Standards Organisation.(2004)

[3] A. Paepcke, C.-C. K. Chang, H. Garcia-Molina, and T. Winograd. Interoperability for Digital Libraries worldwide. Communications of the ACM, 41(4):33–43, 1998.

[4] S. Weibel. The Dublin Core: A simple content description format for electronic resources. NFAIS Newsletter,40(7):117–119,1999

[5] S.M.Shafi: Digitization Perspective of Medieval Manuscripts. 2nd Convention Planner, Manipur University, Infilibnet center, Ahmedabad. 4-5 November (2004).

[6] NISO (2004). Understanding Metadata. Retrieved 12 January,2009,in http://www.niso.org/publications/press/UnderstandingMetadata.pdfS.

[7] Chen, Y.-n., Chen, S.-j., Sum, H.-c., & Lin, S. C. (2003). Functional requirements of metadata system: from user needs perspective. Retrieved 10 May, 2010.

[8] Granitzer, M., Hristakeva, M., Knight, R. and Jack, K. A Comparison of Metadata Extraction Techniques for Crowdsourced Bibliographic Metadata Management. SAC (2012) M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[9] Chamnongsri, Nisachol. (2009). Metadata development for management of a digitized palm leaf manuscript. Thesis (Ph D) Information Studies, Graduate School, Khon Kaen University, 2009.

[10] Weibel. The Dublin Core: A simple content description format for electronic resources. NFAIS Newsletter,40(7):117–119,1999.

[11] Wilhelm, A., Takhteyev, Y., Sarvas, R., Van House, N., and Davis, M. Photo Annotation on a Camera Phone. In Extended Abstracts of CHI 2004. ACM Press, New York, NY, 2004, 1403-1406.

[12] Lampang Manmart., Metadata Development for Palm Leaf Manuscripts in Thailand.,Proceedings of International Conference on Dublin Core and Metadata Applications, 2012.