---@@@---

TOOLS FOR ADDING AND EXTRACTING MARC METADATA TO PDF DOCUMENTS

M.R. Ramesh
Scientific Assistant,
Indira Gandhi Centre for Atomic Research,
Kalpakkam-60302
Email: gascramesh@gmail.com

ABSTRACT

The proliferation of information and communication technologies with electronic publishing resulted in the production of abundant digital documents. These digital documents are in the form of any of the well established file formats specifications such as TIFF(Tagged Image File Format), JPEG(Joint Photographic Expert Group), PDF(Portable Document Format), HTML(Hypertext Markup Language), etc. often the metadata of digital documents is not properly added to the document properties of the files. It is also for the digital documents pertaining to e-books whose metadata is generally stored and access to metadata of printed documents were well established practices by the library community for a reasonable period using various cataloguing, classification schemes, indexing and retrieval mechanisms. The metadata for the digital documents generally maintained in an external database for storage and retrieval purpose. In order to make the digital documents more meaningful and describing, it is $necessary to store the metadata inside the files for easy identification and retrieval. This paper describes the various tools for adding and extracting MARC metadata to PDF files of storing and retrieving metadata within the digital documents.

Keywords: Portable Document File, MARC metadata, Extraction tools, digital documents

INTRODUCTION

It is an easy task for an individual to archive and retrieve a particular document when the collection is small. When the collection increases, the process of archiving and retrieving of documents becomes tedious and time consuming. Many libraries have collected, archived, catalogued and classified printed documents using internationally accepted cataloguing codes and classification schemes. Due to the changing technological innovations, the paper media is becoming obsolete and is unable to support the features and formats that computer technology is supporting i.e. audio, video and hypertext. It is increasingly visible in the society that people have generated and accumulated large number of files which stores their pictures-mails and documents in digital format. However, file formats i.e. PDF, JPEG, GIF, TIFF, etc., for storing data or information also increased and varied in their functionality. Even today, we could see the manuscripts and paintings, which are thousand of years ago in the same manner and same sense that of an author or artist. But the scenario is totally different with digital documents. It is necessary to employ frequently changing computer systems and software to read digital documents. In addition to computer systems, storage media such as floppy disks, CD-ROMs, and hard disks are becoming obsolete and fail to store data. It is observed that the computer systems

and their components are becoming obsolete even faster than that of digital media. Individual and organization world wide realized and addressed the importance of long term preservation and access to digital documents. Archival and preservation of digital document and associated metadata is a challenging task which draws upon several technical issues and considerations. If digital documents are preserved in any media and any form along with metadata, it becomes easier to categorize the document collection, search and retrieve relevant documents based on certain criteria. The methodology adopted in this paper alleviates the problem of loss of metadata which generally stored in external databases and facilitates quickly develop the index/metadata database.

PORTABLE DOCUMENT FORMAT (PDF)

Portable Document Format (PDF) - an introduction

Abode turned PDF as a free standard in 1993 with an aim to establish PDF as a successor to the postscript format. PDF documents may contain images, interactive elements and referencing via hypertext links. PDF perhaps allow storing metadata such as page numbering, page formats, and simple comments about contents of document. Several software tools were developed for creating & processing PDF documents and PDF has become the de-facto standard for electronic publishing. Search options were incorporated in PDF but are less powerful compared to HTML/XML. In an effort towards digital preservation, ISO has published a variant of PDF 1.4 called PDF/A (ISO 19005-1:2005) for long term preservation purposes.

Describing and finding portable document format files

Popular publishers world wide like Springer, Taylor & Francis, Elsevier, etc., have already started publishing scientific books in form of e-books mostly as PDF files. The trend is going to replace print versions completely into electronic books in the near future. As a result librarians are compelled to subscribe or purchase e-books and catalogue the documents with metadata similar to the card catalog in the library. In order to describe and find the PDF files, it is necessary to accurately describe the electronic document with appropriate metadata so that the electronic documents are easily and quickly findable with the help of search engine. In the basic form the PDF support four attributes to describe and search the electronic document i.e. title, author, subject and keywords. PDF also supports to define metadata with name and value pair and allows adding sufficient number of metadata elements to describe any electronic document.

Ways of adding metadata to PDF files

Metadata can be added to PDF by (1) allowing the author or document creator to input metadata using original document creation software tool and transfer during PDF conversion. (2) Manually inputting the metadata by the author or document creator or indexer after the electronic document has been converted to PDF format with the help of document properties. (3) Adding metadata through program tools using already indexed and stored in an external database which corresponds to PDF files. (4) Automatically adding metadata through software tools which electronically scans the document content and identify relevant content by employing intelligent algorithms.

TOOLS FOR ADDING METADATA TO PDF FILES

Several software tools have been developed for adding and retrieving metadata to and from PDF files. Some of the tools are:

Advanced PDF fools[1] from verypdf.com, Inc.: advanced PDF tools is a fast and easy to use utility to edit or add data into the document information fields of single or multiple PDF files, to set open action, page layout, page orientation, metadata, optimize for the web(linearism) compression and others into the PDF files.

A-PDF preview and Rename[2] from a-pdf.com: A-PDF preview and rename is a simple and fast desktop utility program that allows renaming multiple PDF document while previewing. The tool support edit or add metadata to PDF files.

A-PDF scan paper[3] from a-pdf.com: A-PDF scan paper is a desktop program that allows scanning and organizing paper in PDF format. A-PDF scan paper uses thumbnails and metadata to organize, filter, secure, send and retrieval scanned documents.

Calibre[4] by Kovind Goyal: Calibre is free and open source e-book software that organizes saves and manages e-books, supporting a variety of formats. It allows the user to sort and group e-books by metadata fields. Metadata can be extracted from many different sources (ISBNdb.com, Google books, Amazon, Library thing).

Catalogue[5] from informer technologies, Inc: catalogue is a file metadata miner utility which enables quick viewing, management and updating of metadata PDF documents associated with Microsoft office, star office or open office.org documents

JabRef[6] from sourceforge.com: JabRef is an open source bibliography reference manager, which uses the standard LaTeX bibliography format BibTeX for formatting the references. JabRef runs on the java VM, automatically find the right metadata when a PDF file is placed in jabRef and creates a new BibTeX entry which is linked to the PDF file.

Pdftk[7] from PDF labs: Pdftk is an electronic staple remover, hole-punch, binder, secret decoder ring and x-ray glasses if pdf is an electronic paper. Pdftk is a simple command line tool for manipulating PDF documents such as merging, splitting, rotating, decrypting/encrypting, water marking, reporting metrics, updating metadata, etc.

Pdfmark[8] from CrossRef: Pdfmark is an experimental open source tool that enables adding crossref metadata to a PDF file. Metadata can be added to pdf file by passing a pre generated XMP file or by applying CrossRef bibliographic metadata passing as an argument at command line. If CrossRef DOI is passed, the tools will automatically lookup the metadata using CrossRef OpenURL API, generate XMP and applies to pdf file

Pdflib TET[9] from pdflib.com: PDFlib TET (text extraction toolkit) reliably extracts text, images and metadata from PDF documents. TET contains advanced content analysis algorithms for determining word boundaries, grouping text into columns and removing redundant text.

Metadata extraction tool[10] by the national library of New Zealand: this tool

programmatically extract preservation metadata from a range of file formats like PDF documents, image files (BMP, GIF, TIFF and JPEG), audio and video files (WAV, MP3), Microsoft office documents, markup languages (HTML, XML) and many others and output the metadata in a standard format (XML) for use in preservation activities. The tool is written in java and XML and is distributed under the Apache public license.

CONCLUSION

Rapid advances in hardware, software and publishing technologies resulted in the publication of more number of digital documents. Today, preservation of digital documents became an imperative because of obsolescence of hardware, software and standard file formats. Number of digitalization projects is increasing and are more visible in the society. Libraries and archivists need to study thoroughly various issues and concerns about “digital preservation” for long term preservation because more number of print collections is shifting towards digital collections. Libraries need to concentrate on selection of reliable hardware, software, storage media, standard file formats, metadata, physical care and handling, safer environment, disaster recovery and legality of copying/reproduction of digital documents for long term preservation and access. In addition to preservation of digital documents, it is more important to preserve metadata of digital documents.

References

1. Advanced PDF tools. Computer software. Verypdf.com, Inc. Web. http://www.verypdf.com/pdfinfoeditor/index.html (accessed on 05.09.2012)
2. A-PDF preview and rename. Computer software.A-pdf.com. Web. http://www.a-pdf.com/preview-rename/index.htm (accessed on 05.09.2012)
3. A-PDF scan paper. Computer software. A-pdf.com. Web. http://www.a-pdf.com/scan-paper/index.htm (accessed on 05.09.2012)
4. Calibre. Computer software. Kovid Goyal. Web. http://calibre-ebook.com (accessed on 05.09.2012)
5. Catalogue. Computer software. Informer technologies, Inc. Web. http://catalogue9.software.informer.com (accessed on 05.09.2012)
6. Jabref. Computer software. Sourcforge.net. Web. http://jabref.sourceforge.net (accessed on 05.09.2012)
7. Pdftk. Computer software. Web. http://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/ (accessed on 05.09.2012)
8. Pdfmark. Computer software. Crossref.org. Web. http://labs.crossref.org/styled-6/pdf_extract.html (accessed on 05.09.2012)
9. PDFlibTET. Computer software. Pdflib GmBH. Web. http://www.pdflib.com/download/tet/ (accessed on 05.09.2012)
10. Metadata extraction tool. Computer software. National library of New Zealand. Web. http://www.meta-extractor.sourceforge.net (accessed on 05.09.2012)

---@@@---