World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:8, 2015

Tool for Metadata Extraction and Content Packaging as Endorsed in OAIS Framework

Payal Abichandani, Rishi Prakash, Paras Nath Barwal, B. K. Murthy

***Abstract*—**Information generated from various computerization processes is a potential rich source of knowledge for its designated community. To pass this information from generation to generation without modifying the meaning is a challenging activity. To preserve and archive the data for future generations it's very essential to prove the authenticity of the data. It can be achieved by extracting the metadata from the data which can prove the authenticity and create trust on the archived data. Subsequent challenge is the technology obsolescence. Metadata extraction and standardization can be effectively used to resolve and tackle this problem. Metadata can be categorized at two levels i.e. Technical and Domain level broadly. Technical metadata will provide the information that can be used to understand and interpret the data record, but only this level of metadata isn't sufficient to create trustworthiness. We have developed a tool which will extract and standardize the technical as well as domain level metadata. This paper is about the different features of the tool and how we have developed this.

***Keyword*—**Digital Preservation, Metadata, OAIS, PDI, XML.

I. Introduction

THIS paper is about that how the critical assets of the Indian legal system can be organized for long term digital preservation with the help of software tool called *Disposed Case Portfolio Manager (DCPM)*. This software is based on Open Archival Information System (OAIS: ISO 14721 standard) framework [1] developed by NASA's Consultative Committee for Space Data Systems (CCSDS) [2].

There are about millions of cases disposed by Indian courts per year and each court is maintaining their data independently in different ways. The first step is to collect the data from distributed architecture [3], [4]. After that, the metadata extraction and standardization along with its data is another critical and essential activity. DCPM tool will collect the data from NIC system followed by the process of metadata extraction and standardization [5], [6].

One of the key features of this tool is that, it will classify the metadata into two categories, that is, *Descriptive metadata and Representation Information.*

II. DCPM Architecture

This tool will be deployed in Delhi courts to collect the Disposed Case records.

Payal Abichandani, Senior Technical Officer, Rishi Prakash, Principle Technical Officer, Paras Nath Barwal, Joint director and Group coordinator, and B. K Murthy, Executive Director, are with the Center for Development of Advance Computing-Noida (phone: +91-120-306-3311; fax: +91-120--306-3317; e-mail: payalabichandani@cdac.in, pvrishi@cdac.in, pnbarwal@cdac.in, bkm@cdac.in).

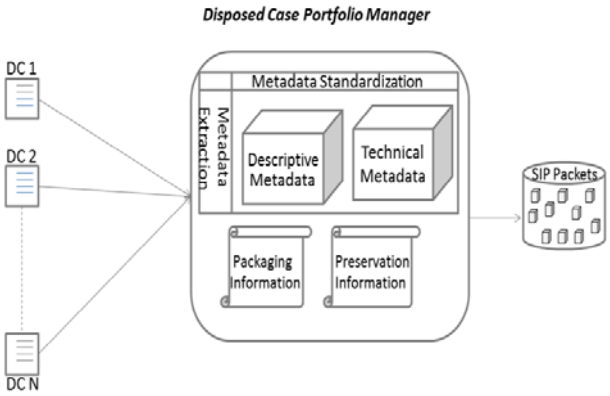

Fig. 1 Disposed Case Portfolio manager Architecture (DCPM)

There are three major processes of DCPM:

1. Data Collection
2. Metadata Generation and Standardization
3. Packet Creation for Long Term Preservation.

III. Data Collection

Data collection will be done at producer level. Data collection can be a fully/semi-automatic procedure depends upon the situation. In this pilot project CDAC has developed a procedure to fetch the data from the NIC database and will also have the option to upload the disposed cases files into the DCPM software. During the upload activity of individual documents it's very essential to maintain the indexing of these documents. The indexing feature in DCPM is dynamic i.e. each respective court can maintain the order of placement of documents as per their standard. This tool is also maintaining the type of documents which needs to be uploaded.

Descriptive Metadata - In this direction, CDAC Noida had developed a Metadata Standards for Long Term Digital preservation of Disposed Case Records. This standard has been made mandatory for Delhi courts under National Digital Preservation Program. Presently each court is maintaining their records in different manners and there are about millions of cases which need to be standardized per court. This challenge has been catered by DCPM tool. It will collect the data from NIC database for all Delhi courts as per the respective configuration and after collecting the data it will extract and standardize the metadata as per the above mentioned standard.

Representation Information is used to interpret and understand the data. In this evolving era, technology is changing every day and the data with specific technology can be interpreted with the help of its technical metadata only e.g.

if record is available in .pdf format then pdf version number, software configuration etc. is required to render and interpret the stored record.

IV. Metadata Generation

In the metadata extraction phase DCPM will separate the domain and technical metadata. The domain metadata will be segregated into different sections as per the standard. For technical metadata the Representation Information will be extracted from each record. All the extracted metadata will be stored in the open source format. In case of domain metadata, there will be provision in the software to manually enter the missing metadata as per the access rights.

Once the data collection and metadata generation is complete, next step is to standardize the metadata.

V. Packet Creation

CDAC has developed a standard folder structure for keeping the packets termed as Submission Information Package. This packet will consist of data and its standardized metadata along with packing information.

Following are the components of the packet:

1. Data- It will contain the digitized document and media files.
2. Metadata- It will have the domain and technical metadata for each and every file.
3. Manifest File- It will contain the packaging information about the data and its metadata. This information will be useful for secure packet travelling and packet validation.
4. PDI file

VI. SIP Architecture

Submission Information Package is one of the components of OAIS information. SIP is a collection of data and it's metadata along with preservation information. The main goal of DCPM is to produce SIP packets at producer level which can be later on transferred to the central location for long term digital preservation. Therefore, we have to standardize a SIP structure, so that later on it can be converted into an Archival Information Package for long term storage.

- XML

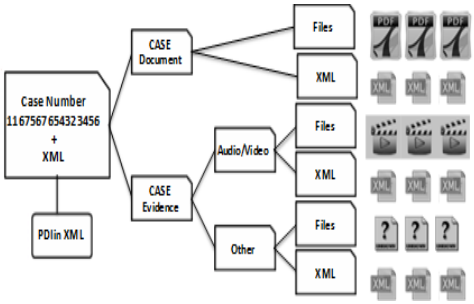

Fig. 2 SIP Packet Architecture

SIP is basically a combination of data and its metadata. Therefore, to preserve the case record it is essential to extract the metadata from every document and encode it into an open source format. For each case record there will be single domain file, which will contain the domain level metadata filled/finalized by expertise. To make SIP more secure and to maintain the transaction details, packet will enclosed by the manifest file.

VII. Result

A SIP packet in the above defined structure architecture will be the result for each and every disposed off case. There must be four types of XML file in each SIP packet.

A. Manifest File

Manifest file [7] will contain the transactions details of SIP packet, such as, sender and receiver's details, path and checksum of all file folders.

Sample File is:

Fig. 3 Sample Manifest File

B. Provenance Description Information (PDI) File

PDI [8], [9] XML file will contain provenance information of the packet, as from where this comes, from which machine this packet is generated. Access right information will also goes in this file.

Sample PDI file is:

Fig. 4 Sample PDI File

C. Case Descriptive Information

This file in SIP packet will contain complete information related to a case in xml format.

Sample file is:

Fig. 5 Sample Case Description File

D. Technical Information (File Level)

SIP packet will have technical XML files for each uploaded files respectively. For an uploaded pdf file format version, total no of pages, language and mime type of file is required. Whereas this metadata will change with file if an image will be uploaded then color depth height, width and pixel depth etc will be important information.

Technical XML File for uploaded PDF File:

Fig. 6 Sample Technical Information File

VIII. Conclusion

The solution developed by CDAC Noida will help the Indian Judiciary System to preserve their records for longer period of time. These records are assets of the Indian Judicial fertility. Current and future Law students, lawyers and judges can refer these old records for various reasons. Even after, decades these records can be rendered because this solution is storing the records in the object form along with their metadata.

This is a solution to the threat of technology obsolescence. Besides long-term storage, this solution will also gather and standardize the scattered records and its associated metadata among all district court in centralized location. The records in the form of packets can be rendered for various statistical analyses also. Standardization of metadata among all Delhi district courts will solve the following concerns:

1) Maintenance of case records will become easy.
2) Fast and efficient searching.
3) Reduction in redundancy
4) No scattered information

Acknowledgment

This paper is the result of research effort at C-DAC, Noida for the development of Digital preservation repository for courts. The project work was done under the constructive guidance and active participation of Dr. B.K Murthy, Executive Director CDAC Noida. This research project was supported by DeitY, MCIT, Govt. of India under their flagship National Digital Preservation Program. We sincerely thank Member-Convener Shri. S. A. Kumar, Scientist ‘F’, DeitY for keeping a close eye on the project from ministry point of view. We thank PRSG (Project Review and Steering Group) chaired by Dr. Mukul K Sinha, Director, ESC Ltd, Noida for overall guidance and steering the project to meet out the project deliverables. We are heartily thankful to Hon'ble Justice B.D. Ahmed Delhi High Court, Hon'ble Dr. Justice S Muralidhar Delhi High Court whose initiative, guidance and support from the initial to the final level enabled us to develop an understanding of the subject and resulted in this valuable deliverable and implementation of pilot project at Karkardooma Court Delhi. We are also grateful to eCommittee Supreme Court of India for acknowledging the work done and incorporating it phase II policy.

References

[1] Lavoie, Brian F. "The open archival information system reference model: Introductory guide." Microform & imaging review 33.2 (2004): 68-81.

[2] Consultative Committee for Space Data Systems *public.ccsds.org/ (Accessed on 2nd July 2015)*

[3] Gasser, Morrie, and Ellen McDermott. "An architecture for practical delegation in a distributed system." Research in Security and Privacy, 1990. Proceedings, 1990 IEEE Computer Society Symposium on. IEEE, 1990.

[4] d'Inverno, Mark, et al. "The dMARS architecture: A specification of the distributed multi-agent reasoning system." Autonomous Agents and Multi-Agent Systems 9.1-2 (2004): 5-53.

[5] Giuffrida, Giovanni, Eddie C. Shek, and Jihoon Yang. "Knowledge-based metadata extraction from PostScript files." Proceedings of the fifth ACM conference on Digital libraries. ACM, 2000.
[6] Han, Hui, et al. "Automatic document metadata extraction using support vector machines." Digital Libraries, 2003. Proceedings. 2003 Joint Conference on. IEEE, 2003.
[7] Jung, Kil-soo, and Kwang-Min Kim. "Manifest file structure, method of downloading contents usng the same, and apparatus for reproducing the contents." U.S. Patent Application 11/322,354.
[8] Hartig, Olaf. "Provenance Information in the Web of Data." LDOW 538 (2009).
[9] Tansley, Robert, Mick Bass, and MacKenzie Smith. "DSpace as an open archival information system: Current status and future directions." Research and advanced technology for digital libraries. Springer Berlin Heidelberg, 2003. 446-460.

Payal Abichandani is working as a Sr. Technical Officer in C-DAC NOIDA. She has done her engineering in Computer Science and MBA in Operations Management. She is leading the development activity of the Project called as "Center of Excellence for Digital Preservation" under National Digital Preservation Program.

She is also leading other projects as well which is under DeitY-"Department of Electronics and Information Technology", such as, M-SIPS and ICMR. She has worked on metadata standards required for the long-term digital preservation of disposed case records. In the Conference "e-Court Present and Future" held on July 2012, she presented the metadata standards required for the digital preservation of Judiciary records. She joined CDAC in 2011. Previously she was associated with AVL India Software Pvt. Ltd. from Automobile Industry Her total experience is 8yrs. The Project handled by her in AVL was used to test the engine performance of automobiles called as EDACS. Her expertise is System Analysis, Architecture Design and Implementation.

Rishi Prakash received his B.E in Instrumentation and Control from NSIT (Netaji Subhash Institute of Technology) Delhi University. He has joined C-DAC Noida in early 2000 and has worked with embedded system and e-Governance division.

He has completed the "e-Court" pilot project which established a "proof of concept" of development of software for search, archive of records and writing depositions in local language along with setting up video recording and video conferencing inside the courtroom.

This is very novel and prestigious project implemented at City civil and Sessions Court, Ahmadabad, under the guidance of High Court of Gujarat and has the honor of having implemented for the first time India. The same project was awarded with "Gold Medal" in 13th National e-Governance conference held in 18-19th Feb 2010 Jaipur. Currently he is working as a chief-investigator for Digital Preservation of Case Records project under National Digital Preservation program funded by DeitY, MCIT, Govt. of India.

Paras Nath Barwal received his M.Tech in Computer Science from BIT (Birla Institute of Technology, Mesra, Ranchi). He has joined CDAC, Noida in Feb 2001. Currently he is Joint Director and Group-coordinator of e-Governance group in CDAC-Noida.

He was Project Manager for Design Development of various e-Governance projects like executing setting up of RTC (Rural Tele Centers) in Lao PDR under Lao PDR and India ICT Bilateral Co-operation ,ITAS for Northern Railways, BIS, DeitY, MCD, TRAI, LMIS for DDA, DTL, IPGCL, Forensic Science Laboratory (FSL), Decision Support System (DSS), IPO.

Dr. B.K Murthy is currently holding the charge of Executive Director, C-DAC, Noida Centre, on deputation from DeitY, MCIT, Govt of India. As the Head of the Noida Centre of C-DAC, he is responsible for policy formulation, planning, implementation and deployment of various R&D Projects.

In his capacity as Senior Director, DeitY he was responsible for the National Knowledge Network, Human Resources Development, Productivity Enhancement & Employment Generation Divisions.

He was awarded Ph.D. degree by IIT Delhi. He is instrumental in the initiation of e-learning activities and contributed for the incorporation of Media Lab Asia. He also served as Executive Director, NIELIT (erstwhile DOEACC). He was a member of Technical Committee & Project Approval Board of National Mission on Education through ICT (NMEICT), MHRD; Academic Advisory Committee, School of Computer Sciences, IGNOU.

Dr. Murthy's research interests include Artificial Intelligence, Natural Language Processing, Knowledge Engineering, Object Oriented Design and Semantic Web. He has published and presented more than 45 papers in various journals and conferences.