Design Strategies for a Prototype Electronic Preservation System for Biomedical Documents

Song Mao, Dharitri Misra, James Seamans, and George R. Thoma
National Library of Medicine
Bethesda, Maryland, USA

Abstract

Among the digital material considered for preservation at the National Library of Medicine (NLM) are TIFF, PDF and HTML files of biomedical journals, laboratory notebooks, correspondence of major figures in biomedical research, and similar documents. Although most of these materials are already in digital form (either as born-digital information, or converted to digital form through scanning), preservation of these materials involve complex administrative and technical issues, such as obtaining and storing adequate levels of metadata for a preserved resource, assuring intellectual integrity of the contents, and avoiding technical obsolescence of encoded information. [1,2]. An R&D project has been initiated at NLM to develop a prototype system that would help investigate the key technical functions required to effectively preserve NLM's digital resources over the long term. This system, named the System for Preservation of Electronic Resources (SPER) has had its initial design and implementation phase completed. Here we describe the main functions of SPER, and the strategies adopted in designing the system to meet these functionalities in a modular and cost-effective manner. In particular, an automated metadata extraction subsystem is designed to minimize manual entry, using string matching and machine learning techniques. Also given are preliminary performance assessments of the subsystems in this prototype. We discuss the overall system architecture, automated metadata extraction techniques, and file migration in the SPER system.

SPER System

SPER, in its initial implementation, has focused upon two critical functions that are essential for digital preservation and in addition to a digital archive's main-stream responsibilities such as ingest, archiving and dissemination [3]. These functions are: (a) to assemble and store sufficient metadata describing a resource: SPER attempts to retrieve the metadata for a document in a number of different ways, including automated extraction from the document's body and header sections (depending upon the document's format), and stores it in a METS-compliant format [4]. A SPER operator may review this data and edit in changes or missing information, if appropriate; (b) to protect the resource against technical obsolescence. Currently, SPER uses the migration approach to transform a document from a to-be-obsolete format to one that is more reliable.

System Architecture

SPER uses the following strategies to implement the system in a modular and cost-effective way:

- The archive infrastructure, upon which the SPER functionalities may be built, is acquired through DSpace [5], which is an OAIS compliant, open-source, preservation system written in Java.
- The SPER system itself is developed as a Java Client-Server application, using the Java Remote Method Invocation (RMI) procedures to communicate between the client and the server processes, and a Swing-based GUI. (To implement SPER as a Web application using a Java Server Faces-based GUI is a future goal.)
- The SPER Server, in turn, interfaces with a number of remote applications, both internal to the SPER system and external Web servers, to carry out necessary functions and receive the required data. This data may then be formatted and sent to the SPER Client to be presented to the operator. This allows SPER to plug in newer Web services and tools, as they become available, to get the required information, with minimal effort, with no change to the client software.

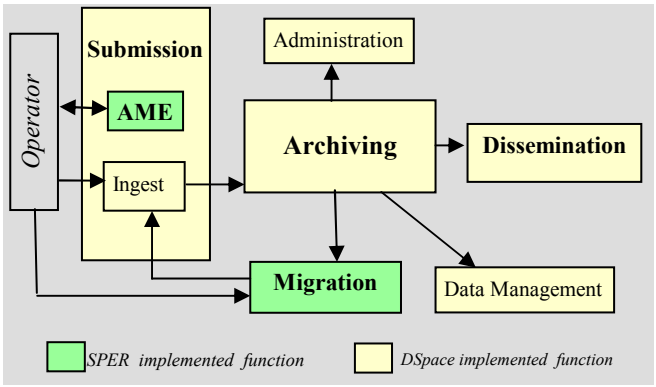

Figure 1 - SPER Functional Overview. AME denotes Automated Metadata Extraction.

The functional overview of SPER, within a digital archive, is shown in Figure 1. The shaded functions are provided by SPER, and the rest are obtained through DSpace, with certain decomposition of its layers, and NLM-specific customization.

Interfaces to External Services

As mentioned earlier, SPER leverages the services provided by other preservation related systems, wherever feasible, to accomplish its objectives, and augments these with its own software components as necessary. The interfaces to these external services are based upon standard, open protocols such as SOAP, HTTP and RPC, as required by the service providers. The general architecture of SPER along with its external interfaces, is shown in Figure 2.

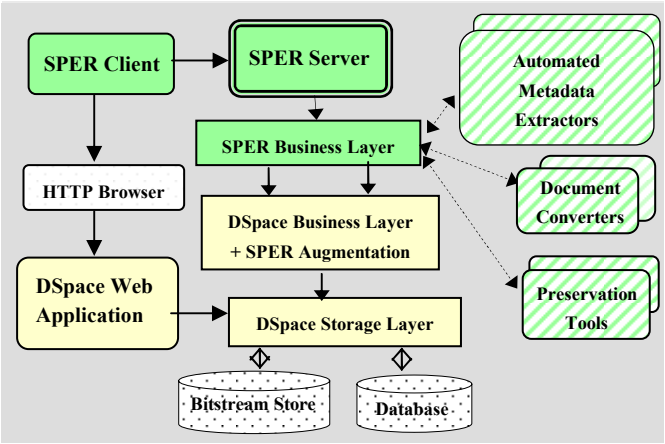

Figure 2 – SPER System Architecture

All Administrative functions such as collection and authorization management, as well document search and retrieval operations are performed using DSpace services directly through a Web browser, which can be invoked from the SPER Client GUI itself for convenience.

Metadata Extraction and Ingest Workflow

Using the SPER Client, an operator can request the extraction and display of metadata from a selected document. Upon receiving the request, the SPER Server initiates metadata generation workflow as follows:

- If descriptive metadata already exists for the document at NLM, such as for the Profiles in Science collection [6] and WebMARS collection [7] the SPER Server retrieves the information programmatically using HTTP and SOAP protocols.
- For scanned journals, the Server invokes, using SOAP protocols, an in-house rule-based tool which analyses the layout of the first page of the journal articles and extracts fields such as title, abstract, author, and affiliation (all elements of the 'descriptive metadata' for the document). The rules are learned using dynamically generated geometric and contextual features. This is discussed in more detail in the automated metadata extraction subsystem design section later in this paper.
- For HTML-based Web pages at NLM, again the SPER Server receives the descriptive metadata, embedded in the source document, through a WebPage Metadata Extractor tool developed in-house for this purpose.
- For TIFF documents, available technical metadata [8] is extracted by the Server using TIFF Image I/O classes provided by Java Advanced Imaging packages.

After all available metadata elements are gathered and packaged into a subset of METS-defined elements [4], it is sent to the Client and displayed for operator review. Editing is allowed for certain automatically generated terms and missing fields. Once that is performed and verified, the operator may ingest the document to the archive – which is performed by the Server by invoking lower level DSpace layers. All available metadata are stored as a separate XML bitstream along with the document bitstream in the archive's data store.

Automated Metadata Extraction

In most current archives it appears that metadata is manually entered and hence is labor-intensive. In SPER, an automated metadata extraction subsystem has been designed to minimize manual entry. It consists of three parts: document page style classification, style-specific rule or model learning, and metadata extraction using learned rules or models. The metadata extraction subsystem design is shown in Figure 3.

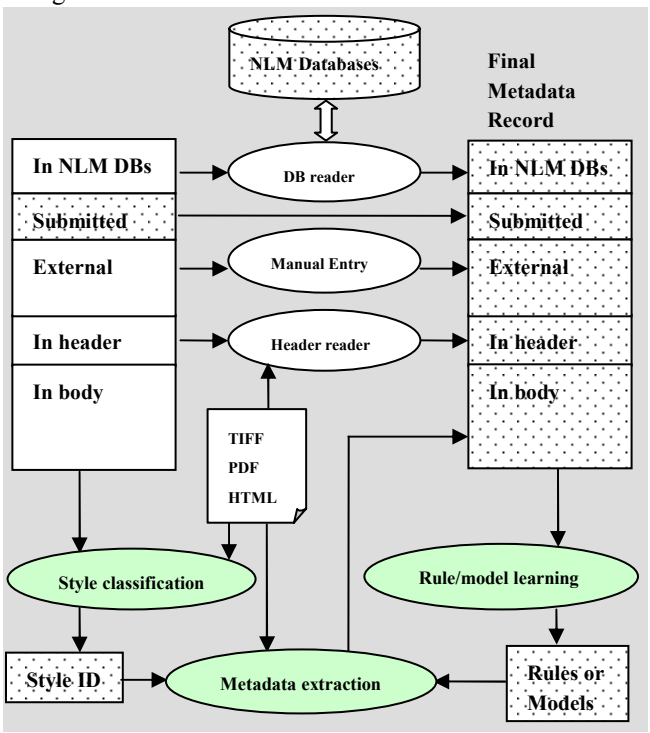

Figure 3 – The Automated Metadata Extraction subsystem. Dotted boxes signify that their content is available.

If some metadata elements are provided upon submission, they can be directly copied into the final metadata record after verification. Existing metadata records in NLM databases, e.g., PubMed [9] and Profiles in Science [6] are obtained programmatically from the databases. External metadata that is not otherwise available at submission time have to be manually entered. Technical metadata that exist explicitly in the header portion of a digital document can be directly extracted by SPER using header reader programs. For descriptive metadata that are embedded in the bitmap of TIFF images or in the ASCII text in PDF and HTML documents, machine learning algorithms are used for their extraction.

Automated extraction of descriptive metadata consists of three steps: zoning, logical labeling, and text extraction. Zoning and logical labeling are critical not only for scanned journals, but also for online journals due to inadequate or inappropriate HTML tags. In addition, OCR is required for text extraction from scanned journals. In order to achieve these tasks, the automated metadata extraction subsystem is designed to consist of three parts: document page style classification module, metadata extraction module, and a module for learning rules or models to be used in the metadata extraction module.

Document Page Style Classification

Document page style classification is necessary for effective logical structure analysis of heterogeneous collections of document pages. Logical structure analysis is a process of assigning logical labels to physical zones and arranging them in a logically meaningful structure.

The style of a document page is represented by the physical layout and contextual features of its physical zones, which can be concisely summarized by an ordered labeled tree [10]. Figure 4 shows a document page and its tree representation.

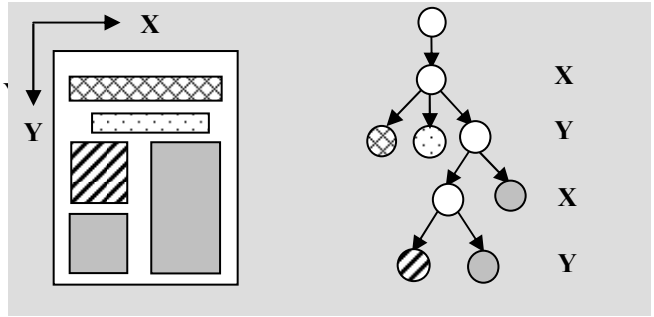

Figure 4 – A document page and its tree representation. Zones with different filling patterns denote different logical entities. X and Y denote projection direction at each tree level.

A document page is first partitioned into zones in an initial zoning process [11]. An ordered labeled tree is then generated as follows: the zones are first projected onto the X axis, each non-overlapping zone forms a leaf and overlapping zones form an internal node. The overlapping zones of an internal node are further projected onto the Y axis at the next tree level. This process is repeated until all leaf nodes denote singular zones. Each node of the tree is associated with a label represented by a set of contextual features such as font size and attribute. The children of a parent node are ordered in read order.

Style dissimilarity between two document pages is computed as the edit distance between their ordered labeled tree representations. We compute such a distance for each pair of document pages and then use the *K*-medoids algorithm to classify them into *K* clusters, each of which corresponds to a distinct style. While the number of clusters is assumed given, it can be estimated using the Gap statistic [12].

Learning Style-specific Rules or Models

The knowledge about logical entities in documents can be summarized as rules or formal models based on geometric and contextual features. A rule-based algorithm [13] has been used to label scanned biomedical documents in MARS (Medical Article Records System) for automated bibliographic data extraction to populate NLM's MEDLINE® database. The rules can be tailored to work well for documents of a particular style. But for documents of a different style, new rules have to be manually created. Manual creation of a set of rules or models for each document style is very expensive. We automate this process by learning style-specific rules using dynamically generated features [14].

In MARS, a Dynamic Feature Generation System (DFGS), shown in Figure 5, is designed to generate style-specific empirical probability distributions of geometric and contextual features using string-matching technique.

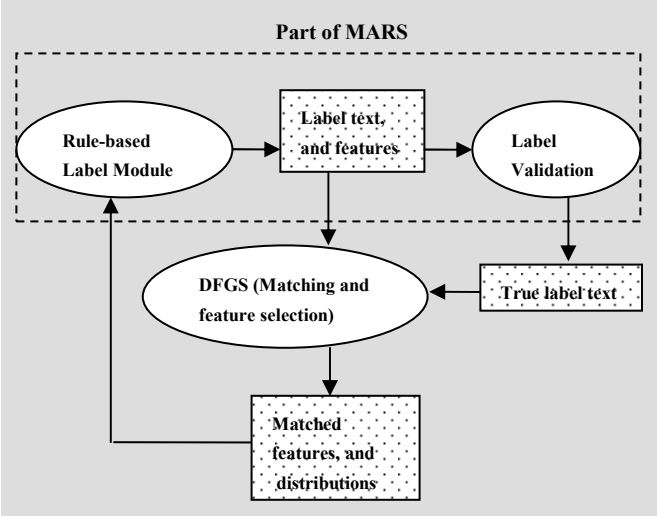

Figure 5 – Dynamic Feature Generation System (DFGS) in MARS.

The features include zone bounding box coordinates, font size, and font attribute for each of the title, author, affiliation, and abstract fields. They are obtained by matching the validated text result with the noisy text in the

zoning results. The features associated with the noisy text are then used to learn appropriate labeling rules for each of significant fields. DFGS is currently implemented in the MARS production system.

A stochastic tree-like 2D graph model [15] is proposed to formally represent logical entities of document pages in a probabilistic framework. This model represents not only the absolute physical size and location of logical entities, but also their relative placement and gaps. A Bayesian-based method [16] is designed to automatically learn such models from unstructured and groundtruthed zoning and labeling results of document pages of a particular style.

The learning process starts with an initial trivial model M_0 where each path in M_0 represents zones of a document page projected onto the X axis in the order from left to right. The number of paths equals to the number of training document pages. Each node in a path represents a non-overlapping zone or a set of overlapping zones. The nodes in M_0 are merged subsequently until the posterior probability $P(M|X)$ of the current model M, given the observation X on zones, is maximized. The observations include the width and height of zones, the size of the gaps between two adjacent zones, and character count in the zones. If a node in M represents more than one overlapping zones, the same learning scheme is repeated on the zones corresponding to this node on the Y axis. When all leaf nodes in the model represent singular zones, the learning process stops. The learning process can be considered as a top-down process at each tree level as shown in Figure 4, but a real model at each level is a finite state automaton.

Metadata Extraction Based on Learned Rules or Models

The distributions of features generated by the DFGS system are used by a rule-based labeling algorithm [13] to create appropriate rules for scanned articles from each journal. Zone location rules are modified by the bounding box coordinate distributions such that candidate zones for a logical entity should significantly overlap with the bounding boxes that have high probabilities. Contextual rules are modified such that the font size and attribute distributions are appropriate description of the characters in the candidate zones. For example, only the zones that significantly overlap with the title bounding box with significant probability can be considered as title candidates. The true title zone can then be found by filtering out other zones in which the average font size has insignificant probability. Currently, SPER is designed to employ this MARS module as a remote source of metadata for scanned biomedical journals.

In the metadata extraction algorithm where formal models are used, the learned stochastic graph model M and the observation X are used in a recursive duration Viterbi algorithm [16]. In this algorithm, an optimal path is searched at each level of the learned model by maximizing the probability $P(E_i, X_i|M_i)$, where i is the level number, E_i is an optimal path at level i, and M_i is the model component at the level i of M. Finally, the obtained optimal paths E_is are hierarchically connected to constitute a logically labeled physical structure S. We plan to improve this algorithm by incorporating more features.

File Migration

File migration is implemented by both an external service as well as modules within SPER. Conversion of a TIFF image to a PDF file is provided by an NLM Web service named DocMorph, with a SOAP-driven interface protocol [17,18]. Conversion from TIFF image to JPEG2000 (JP2) and vice versa is performed by the SPER system itself through in-house software using Java JPEG2000 libraries, SPER software also stores the metadata information present in the TIFF file header in the target JP2 file header [19].

Using the SPER Client, an operator selects a specific document from the archive, and sends the migration request to the server, along with the chosen target format. This document, along with its metadata may also be viewed by the operator. The workflow is shown in Figure 6, and described below:

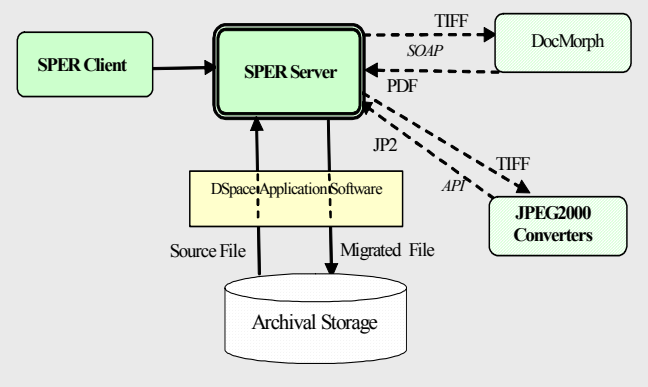

Figure 6 – File migration work flow.

- Upon receiving the request, the SPER Server retrieves the document from the archive, and sends it to the DocMorph server using SOAP protocols, or invokes the internal JPEG2000 converter, as the case may be.
- Upon completion of format transformation, the server generates a metadata record for the target document from that of the source document and adds a "Change History" section to the document, reflecting the migration process.
- The operator may then view the migrated resource and perform a comparison of the source and target images (for JP2 conversion) to assure that they are identical.
- He/she may then authorize the replacement of the source document by the target one. The target document, along with its metadata is then added to the archive as a *new* item, and the original item is "withdrawn" from circulation, but not deleted.

Note that PDF file format, although widely used, is recognized as a non-preservable format, giving way to the PDF/A format in the future. When that happens and DocMorph or other publicly accessible servers support

PDF/A conversion, SPER will support TIFF to PDF/A migration.

Preliminary Performance Assessment of SPER

In analyzing *automated metadata extraction* results, the classification algorithm achieved an average accuracy of 95.69% over six datasets consisting of 150 scanned document pages of 11 different styles. Experimental results also show that the classification algorithm is robust to over-segmentation problem in initial zoning results and variations in the width and height of zones. A large experiment on several thousand title pages in 580 medical journal issues shows that the learning module [14] improves the bibliographic metadata labeling accuracy of a previous algorithm [13] by 50.98%. An initial experiment shows that the recognition module using the 2D document layout models [16] learned from 19 training pages achieved a labeling accuracy of 90.53% on 69 test pages. Currently, we can also extract the following metadata items from Web pages using HTML parsers: URL, title, description/abstract, keywords, last updated date, contact email, copyright information, publisher, permanence level, and language. When a metadata item is not in the current Web page, its linked pages are searched. An automatic character encoding detection algorithm is used to detect language for those encodings that can be mapped to a unique language.

In analyzing *migration results*, it was observed that in migrating TIFF files to PDF (V 1.4), the original document contents remain unchanged, but it is difficult to recover the original metadata in the TIFF image header. In TIFF to JPEG2000 conversion the "lossless" conversion is preferred, since this yields transformation with no loss of data. Furthermore, JPEG2000 provides a much better metadata recording capability in its file header – making it a good choice as a preservation format for raster images.

References

1. An Introduction to Digital Preservation Technical Advisory Services for Images (TASI) http://www.tasi.ac.uk/index.html.

2. Comparing Preservation Strategies and Practices for Electronic Records. http://www.rlg.org/events/pres-2000/cloonan.html.

3. Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems. http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf.

4. Metadata Encoding and Transmission Standards (METS). http://www.loc.gov/standards/mets/.

5. DSpace at MIT. http://www/dspace.org.

6. Profiles in Science at http://profiles.nlm.nih.gov/.

7. Le DX, Tran LQ, Chow J, Kim J, Hauser SE, Moon CW, Thoma GR, Automated medical citation records creation for Web-based online journals, Proc. 14th IEEE Symposium on Computer-based medical systems, Los Alamitos CA. (2001).

8. NISO Z39.87 -2002 Data Dictionary - Technical Metadata for Digital Still Images.

9. PubMed at http://www.ncbi.nlm.nih.gov/entrez/query.fcgi.

10. Krishnamoorthy M, Nagy G, Seth S, and Viswanathan M, Syntactic segmentation and labeling of digitized pages from technical journals, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 15, pg. 737-47. (1993).

11. Hauser SE, Le DX, Thoma GR, Automated zone correction in bitmapped document images, Proc. SPIE: Document Recognition and Retrieval VII, San Jose, CA, pg. 248-58. (2001).

12. Hastie, T, Tibshirani R, and Walther G, Estimating the number of data clusters via the Gap statistic, Technical Report, Stanford University, Stanford, (2000).

13. Kim J, Le DX, and Thoma GR, Automated labeling in document images, SPIE Conference on Document Recognition and Retrieval VIII, San Jose, CA, pg. 111-122. (2001).

14. Mao S, Kim J, and Thoma GR, A Dynamic Feature Generation System for Automated Metadata Extraction in Preservation of Digital Materials, the First International Workshop on Document Image Analysis for Libraries, Palo Alto, CA, pg. 225-232. (2004).

15. Mao S, Rosenfeld A, and Kanungo T, Stochastic attributed K-D tree modeling of technical paper title pages, IEEE International Conference on Image Processing, Barcelona, Spain, pg. 533-536. (2003).

16. Mao S and Thoma GR, Bayesian Learning of 2D Document Layout Models for Automated Preservation Metadata Extraction, Proceedings of the Fourth IASTED International Conference on VISULIZATION, IMAGING, and IMAGE PROCESSING Marbella, Spain, pg. 329-34. (2004).

17. NLM DocMorph at http://docmorph.nlm.nih.gov/ docmorph/.

18. Walker FL and Thoma GR. A SOAP-Based Tool for User Feedback and Analysis. Proc. of InfoToday 2003. Medford N.J.: Information Today, pg. 313-322. (2003).

19. JPEG 2000 (Joint Photographic Experts Group) http://www.jpeg.org/JPEG2000.html

Biography

Song Mao is a Postdoctoral Fellow at an R&D division of the U.S. National Library of Medicine. His conducts research in document image analysis, machine learning, and pattern recognition. He is currently working on algorithms in the digital preservation project and Medical Article Record System (MARS) for automated metadata extraction from biomedical documents. He earned the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Maryland. He is a member of the IEEE.

Dharitri Misra is a Senior Consultant at Aquilent, Inc., currently working on the digital preservation project at the U.S. National Library of Medicine. Her work involves developing experiments and tools to help in the long term preservation of digital resources. She earned M.S. and Ph. D. degrees in Physics from the University of Maryland.

James Seamans is a Senior System Scientist with MSD, Inc. Mr. Seamans has worked on many medical research and development computer imaging projects. Since 1996 Mr. Seamans has been participating in and developing programs for the Visible Human Project. Mr. Seamans received his B.S. degree in mathematics from Ricker College and A.S. degree in Electronics and Computer Technology from DeVry University.

George R. Thoma is a Branch Chief at an R&D division of the U.S. National Library of Medicine. He directs R&D programs in document image analysis, biomedical image processing, animated virtual books, and related areas. He earned a B.S. from Swarthmore College, and the M.S. and Ph.D. from the University of Pennsylvania, all in electrical engineering. Dr. Thoma is a Fellow of the SPIE, the International Society for Optical Engineering.