

AUTOMATED METADATA EXTRACTION TO PRESERVE THE DIGITAL CONTENTS OF BIOMEDICAL COLLECTIONS

George R. Thoma, Song Mao, and Dharitri Misra
U.S. National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894, USA
gthoma,smao,dmisra@mail.nih.gov

ABSTRACT

The long term preservation of digital objects, a growing problem as these are acquired by libraries and archives, requires appropriate systems, standards and institutional policies. A key requirement is the acquisition of metadata about the objects to enable future access and usage, as well as the migration of digital files from obsolete formats to newer ones. “Metadata” is data about data. It typically consists of information about the intellectual content of a digital object, the data required for appropriate digital representation and interpretation, security or rights management information, and their relation to other digital objects. The manual recording of these metadata elements is highly labor-intensive and automated means for doing this are key to successful preservation. In this paper a prototype system for digital preservation is introduced, its main functions are described highlighting the strategies adopted in designing the system to meet these functionalities in a modular and cost-effective manner, an automated metadata extraction subsystem to minimize manual entry, using string matching and machine learning techniques, is presented, and preliminary performance assessments are given.

KEY WORDS

Digital preservation, SPER, and automated metadata extraction.

1. Introduction

The long term preservation of important collections is a mandated responsibility of major national institutions. The U.S. National Library of Medicine (NLM), for example, is responsible for preserving the significant biomedical literature as well as historical material related to biomedicine. While paper-based material was traditionally preserved on microfilm, the vast and growing corpus of digital material, especially those “born digital,” is a challenge that institutions are beginning to confront.

Among the digital material considered for preservation at NLM are TIFF, PDF and HTML files of biomedical journal articles, historic photographs, laboratory notebooks and correspondence of major figures in

biomedical research, and similar documents. In addition, there are video and audio files of importance. Much of these materials are already in digital form, either as born-digital information, or converted to digital form through scanning. Their preservation involves complex administrative and technical issues, such as obtaining and storing adequate levels of metadata for each preserved resource, assuring intellectual integrity of the contents, and avoiding technical obsolescence of encoded information [1, 2].

To investigate the key technical functions required to effectively preserve NLM’s digital resources over the long term, a prototype system is being developed at NLM as part of an R&D project. This system, named the System for Preservation of Electronic Resources (SPER), is built in a modular fashion so that different strategies for, and implementations of, the necessary functions may be evaluated. Among the essential functions of such a system are: ingesting the resource to be preserved; acquiring or extracting metadata that describe the resource sufficiently to enable future access, display and migration; protecting the resource against technical obsolescence by migrating older file formats to newer ones likely to be supported; transferring files from older media to newer ones in step with advances in storage technology; and several other functions.

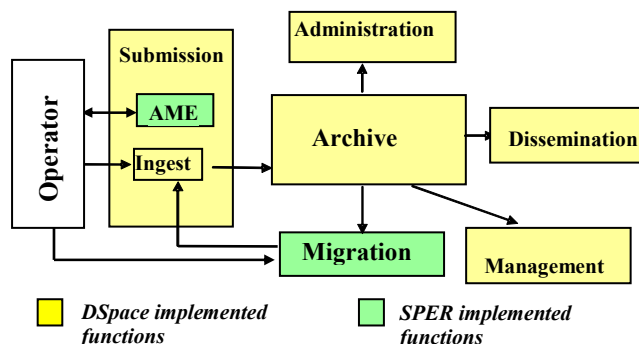


Figure 1 – SPER functional diagram.

These essential functions are shown in Figure 1: submission of resources, ingest into the archive, automated metadata extraction (AME) from the resources

to be preserved, bulk migration, and administrative tasks, among others. In our current design, SPER is developed as a Java client-server application, using 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 design is discussed in more detail elsewhere [8], but certain underlying design principles may be outlined here. First, we make use of open-source and readily available technologies. For instance, to serve as an infrastructure ‘substrate’ for the system, providing a host of low- and medium-level functionality, we use MIT’s DSpace [6], which is an OAI compliant, open-source system written in Java.

Secondly, we expect any digital preservation system to use in-house developed modules as well as those created elsewhere and offered as Web services. Consequently, our design allows the SPER server to interface with 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 make use of newer Web services and tools, as they become available, without requiring changes to the client software.

The rest of this paper is organized as follows: Section 2 describes the types of metadata useful for future access and usage; Section 3 discusses a tool for the automated extraction of technical metadata; Section 4 describes our metadata extraction tools, focusing particularly on a system for the extraction of descriptive metadata from an important type of resource to be preserved, viz., scanned medical journal articles; Section 4 also gives the results of experiments using our automated metadata extraction system for medical articles; finally, Section 5 summarizes the paper.

2. Metadata Types

An essential part of preserving a digital resource is the acquisition/generation and storage of its *preservation metadata*. A number of standards, such as METS, Dublin Core, and PREMIS [1-5] classify preservation metadata into different categories and specify the element sets comprising each category. As an example, Figure 2 shows a classification of preservation metadata drawn from the METS schema.

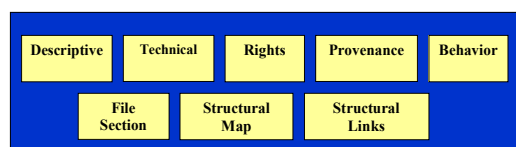


Figure 2 – SPER classification of digital resource preservation metadata.

The metadata elements used in our work are also from METS as well as Dublin Core. In SPER, these metadata elements are grouped together and stored as a record separately from the preserved item itself, using an implementation-specific schema derived from these two standards. Table 1 shows the purpose, source and usage of different types/groups of metadata elements in preserving various types of digital documents. In the following section, we discuss the extraction of descriptive and technical metadata in SPER for TIFF images, scanned journals and Web pages – which comprise the main categories of digital documents within NLM.

Table 1 Purpose and origin of preservation metadata types

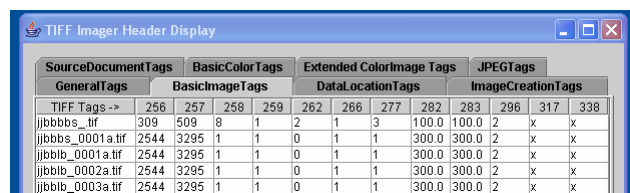
| <i>Metadata Type</i> | <i>Purpose</i> | <i>Origin</i> | <i>Comments</i> |
|--|--|--|--|
| Descriptive | For information on contents; necessary for future access | From external source, and/or in the document’s content | Required for, and unique to each document |
| Technical | To describe the technical characteristics of the resource; required for future display and usage | Mostly in document’s header; some elements from external source | Minimal set required for all documents; extensive set required for images, audio/visual data |
| Rights - copyright, access authorization | To control public access, future migration | From external source | All documents |
| Provenance (source, change history) | For tracing the origin of a preserved resource | May be added by system performing the change (during format migration) | Important for migrated documents |
| Behaviour (processing hardware/software information) | To render the object, for long term preservation | May be added programmatically at submission | Common to a document set |
| File Section, Structural Map, Structural Link | Describes the relationship of a complex resource to its subcomponents and/or external components | May be generated programmatically with information provided at resource submission | Essential for Web documents (HTML, PDF...) |

The first two categories in Table 1 are of principal interest: descriptive and technical metadata. Descriptive metadata elements, crucial for future access, are usually the ones in the bibliographic (citation) record of an item. For a journal article, for example, it consists of the article title, author names, institutional affiliations, journal name, volume, issue, page numbers, among others. The acquisition of descriptive metadata is very labor-intensive, and is usually done manually. However, when many of these elements are embedded within the document, for example in TIFF or HTML files, they may be extracted automatically as in our automated process described in Section 4.

Technical metadata, necessary for future display and usage, is drawn from the proposed NISO Standard Z39.87 [5] for *still images* (including scanned documents in TIFF format). It includes a set of mandatory elements, and rich sets of recommended and optional elements. We have tools to extract most of the mandatory elements in the TIFF header, but many of the recommended or optional elements are hard to acquire when the file is to be preserved, unless they are stored with the image itself, as an internal or external record, when created. Technical metadata, other than MIME type, file size, and checksum, have not been established for most other types of documents. Note that when a digital object is migrated from one format to another, the NISO standard requires the change history to be recorded as a part of the target object's technical metadata.

3. Automated Extraction Techniques for Technical Metadata

The TIFF file format [7] specifies a large number of metadata elements known as TIFF tags that may be stored in the file's header in the form of Image File Directories. (There is a parallel between these TIFF tags and NISO technical metadata element sets.) The mandatory tags are those which are essential to render the image by a TIFF reader; other "optional" elements are to be included to help in understanding the source and creation of the document. The mandatory tags include elements *file signature, byte order, compression scheme, color scheme, orientation, image width, image length, bits per sample, extra samples, and color map*. Some optional elements are: *source, host computer, operating system, software, scanning system*.



| SourceDocumentTags | BasicColorTags | | | | ExtendedColorImageTags | | | | JPEGTags | | | |
|--------------------|----------------|------|-----|-----|------------------------|-----|-----|-------|-------------------|-----|-----|-----|
| GeneralTags | BasicImageTags | | | | DataLocationTags | | | | ImageCreationTags | | | |
| TIFF Tags -> | 256 | 257 | 258 | 259 | 262 | 266 | 277 | 282 | 283 | 296 | 317 | 338 |
| jbbbs.tif | 309 | 509 | 8 | 1 | 2 | 1 | 3 | 100.0 | 100.0 | 2 | x | x |
| jbbbs_0001a.tif | 2544 | 3295 | 1 | 1 | 0 | 1 | 1 | 300.0 | 300.0 | 2 | x | x |
| jbbib_0001a.tif | 2544 | 3295 | 1 | 1 | 0 | 1 | 1 | 300.0 | 300.0 | 2 | x | x |
| jbbib_0002a.tif | 2544 | 3295 | 1 | 1 | 0 | 1 | 1 | 300.0 | 300.0 | 2 | x | x |
| jbbib_0003a.tif | 2544 | 3295 | 1 | 1 | 0 | 1 | 1 | 300.0 | 300.0 | 2 | x | x |

Figure 3 – Extraction and display of technical metadata in TIFF image headers. Note that tag 256 denotes image width, 257 denotes image length, 258 denotes bits per sample (1: bi-level, 8: gray level, 256 color level), 259 denotes compression (1: uncompressed), 262 denotes color space (1: black is zero, 2: RGB...), 277 denotes samples per pixel (3: RGB), 282 and 283 denote X and Y resolutions, and 296 denotes resolution unit (2: per Inch).

We have incorporated a tool, the "TIFF Header Analyzer" into SPER to automatically extract most of the mandatory elements from TIFF headers, present them visually, and to add these to the document's preservation metadata record. (These TIFF headers typically have very few optional elements, however.) An example of the technical metadata extracted by this tool is depicted in Figure 3.

The first column lists the file names, and the next eight columns (starting with Tag 256) indicate: image width, image length, bits/sample, compression, color space, samples/pixel, x- and y- resolutions.

4. Automated Extraction Techniques for Descriptive Metadata

As mentioned, resources to be preserved range widely: from paper documents and Web pages to video/audio as well as specialized biomedical imagery such as CT, MRI and other DICOM images from clinical environments. We first focus on the extraction of metadata from Web pages and then from TIFF versions of biomedical journal articles.

4.1 Descriptive Metadata from Web pages

Preservation of institutional Web pages is an ongoing activity at the National Library of Medicine. It is also an important research problem given the need to preserve various types of digital objects (text, images, audio, video, scripts, etc.) embedded in Web pages. We now describe an initial effort toward extracting descriptive metadata from the HTML text in Web pages. Important descriptive metadata items such as *title* and *keywords* are usually tagged in the HTML text and can be extracted directly. Other metadata such as *last updated date* are often inappropriately tagged, or not tagged at all. Sometimes, certain metadata items (e.g., *contact email*) cannot be found in the current page, but rather in a linked page. Figure 4 shows our approach schematically.

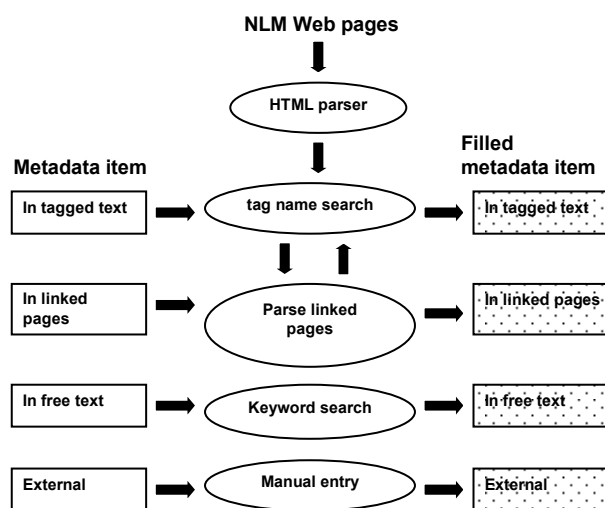


Figure 4 – Automated descriptive metadata extraction from Web pages.

We collected a set of keywords for each untagged or inappropriately tagged metadata item. These keywords along with HTML delimiters are used to search for the

corresponding metadata. For example, the keywords used to search for *last updated date* include *updated*, *published*, *revised*, *last modified*, *last updated*, *site last updated*, *page last updated*, etc. If a metadata item cannot be found in the current Web page, our system first locates links identified by keywords such as *contact us*, *contact NLM*, etc, and then searches the Web pages specified by these links.

We tested this metadata extraction method on 30 NLM Web pages. All tagged metadata (*title*, *description*, and *keywords*) are successfully extracted even though the names of the tags vary. We also extracted 20 out of a total of 22 *contact email* metadata items, 18 out of 28 *last updated date* metadata items, and 19 out of a total of 27 rights (*copyrights and permission*) metadata items. Note that some Web pages do not have certain types of metadata. Most errors are due to a) the list of keywords obtained from the training set is not sufficient, b) some metadata (e.g., *last updated date*) is generated by scripts, or c) the true metadata is in a remotely linked page at least two levels away from the current page.

In ongoing research, we plan to represent keywords, HTML delimiters, and their semantic relationships in formal statistical linguistic models such as Hidden Markov Models (HMM) [15], probabilistic context free grammar [16], and stochastic attribute grammars [17], and use associated parsing or recognition algorithms to extract untagged or inappropriately tagged metadata items in HTML documents.

4.2 Descriptive Metadata from Scanned Biomedical Journal Articles

Paper documents to be preserved include medical journal articles, historic correspondence, and lab notebooks. Each type of paper document has its own set of descriptive metadata. For example, metadata for correspondence would include the names of the sender and recipient, the date, subject of correspondence and similar items.

Here we focus on a common type of paper document in our collections: the medical journal article whose metadata includes most elements of its bibliographic record: article title, author names, authors' affiliations, abstract, journal name, issue, page numbers, and others. Such descriptive metadata are useful for indexing, as well as future search and access. To consider automated means for extraction, the first step is to scan the articles and produce TIFF images.

Unlike technical metadata that occur in file headers, descriptive metadata embedded in TIFF files are difficult to extract. In this section, we describe automated extraction of descriptive metadata from scanned medical journal article TIFF images using machine learning techniques.

An algorithm designed to work on a variety of documents with different styles usually exhibits unsatisfactory performance. On the other hand, an algorithm designed to

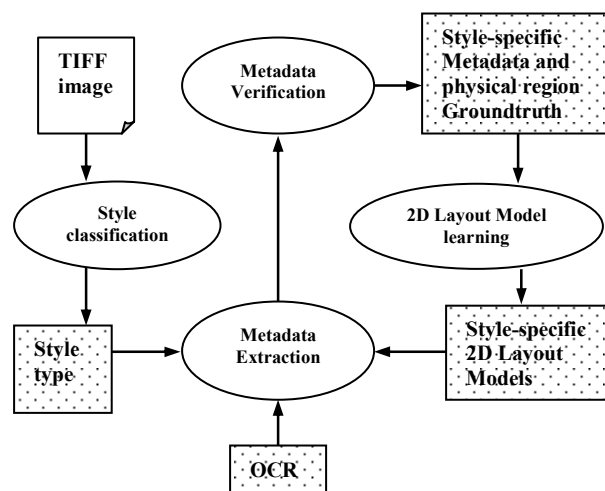


Figure 5 – Automated metadata extraction from scanned medical journal articles in a learning framework.

work on documents of a particular style usually performs well, but requires prior knowledge of that style, which in many situations has to be manually created. We propose a learning framework in which we first classify documents according to their styles. A 2-D layout model is then learned from training samples, which are verified style-specific metadata text and corresponding physical regions, of each style. In the metadata extraction phase, each learned style-specific model and an OCR engine are used to extract metadata from documents of that style. Figure 5 shows the metadata extraction system that is designed in a learning framework.

4.2.1 Classifying Document Layout Styles

The style of a document page is represented by its layout and contextual features. Examples of layouts include one-column and two-column documents. Examples of contextual features include font size, font attribute (bold, italics, underlined, etc.), keywords, and others. We classify document pages according to their styles in an unsupervised approach [10] shown in Figure 6.

We start with a set of zoned (segmented) page images with character font size information. A profile tree is then built for each page. A dynamic tree matching algorithm [11] is used to compute an edit distance between each pair of such trees. For example, the edit distance between tree 3 and tree 5 in Figure 6 is 98. Finally, a K-medoids clustering algorithm is used to group the trees into K clusters, with the number of clusters K given. Since our approach is unsupervised, it does not involve training or setting the algorithm parameters manually.

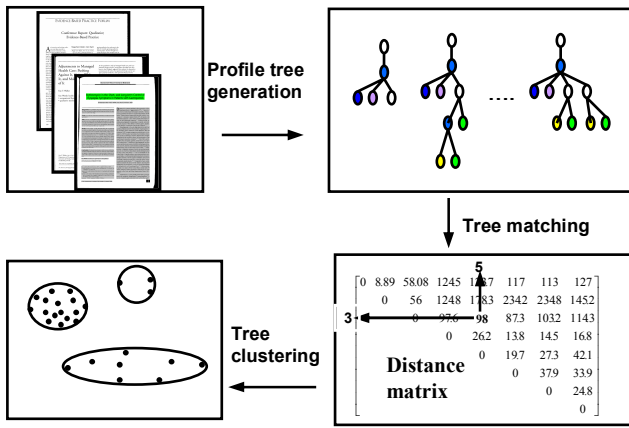


Figure 6 – An unsupervised document page style classification approach.

4.2.2 Learn a 2-D Layout Model for Each Style

A stochastic graph model [14] is used to represent the 2-D layouts of document pages in each style. The model consists of a set of attributed hidden semi-Markov models (AHSMM). Each such model is used to represent the projection profile of the zones in a document region on either the X or Y axis. They are connected by nonterminal states, where a nonterminal state represents a document region that can be partitioned in a particular direction, while a terminal state represents a document region that is not decomposable in any direction. A stochastic graph model could be learned in a Bayesian approach [14, 15] for each distinct style. The learned layout model is used in a recursive duration Viterbi algorithm [13] to segment and logically label important metadata fields simultaneously.

4.2.3 Experiment and Results

We tested our classification algorithm on the title pages of 150 medical articles, exhibiting 11 different styles, and achieved an average classification accuracy of 95.69% [10].

We then selected two of the more common styles and learn a layout model for each of them from 34 training pages. The learned model for the first style consists of only one hidden semi-Markov model since the projection of the whole document page on the Y axis is sufficient to unambiguously represent the all important metadata items (title, author, affiliation, and abstract) on one axis. The learned model for the second style [14] consists of three connected hidden semi-Markov models since projections of document regions on both X and Y axis are needed to uniquely identify those metadata items. Figure 7 shows the two styles and associated learned layout models.

The learned models are used in a duration Viterbi algorithm [13] to segment and label title, author, affiliation, and abstract metadata fields simultaneously.

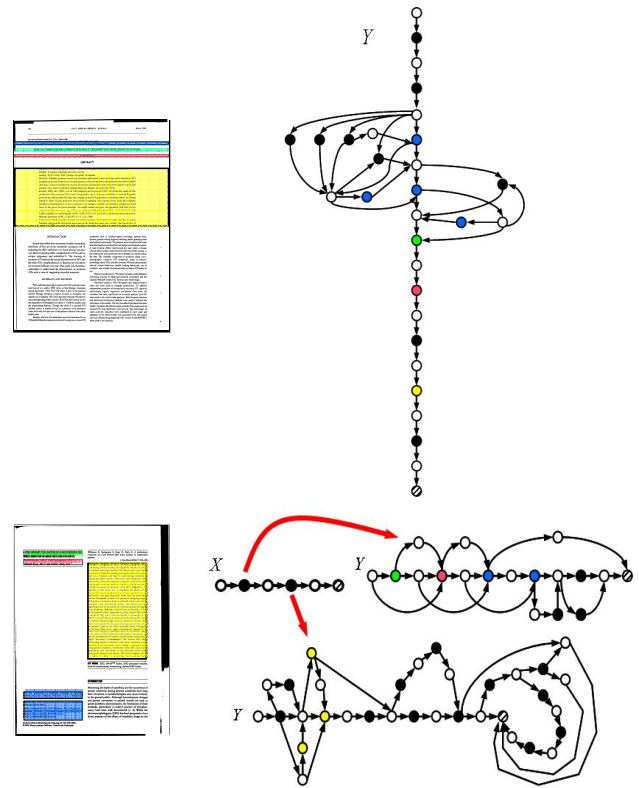


Figure 7 – The learned layout models for two document page styles. Each circle represents a document region. The white circles represent gaps or margins and the colored circles represent textual regions. The arcs in the model represent top-down order on the Y axis or left-to-right order on the X axis.

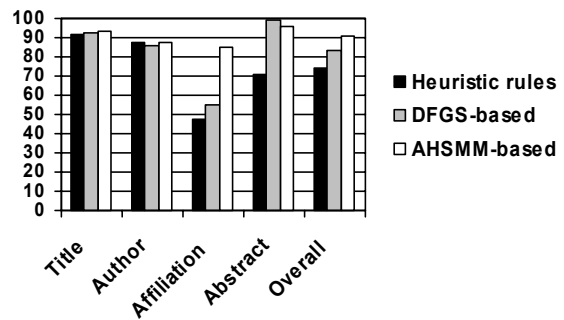


Figure 8 – The evaluation and comparison results.

We tested our algorithm on a dataset of 124 page images and compared it with two other systems developed earlier for production purposes: one called the Dynamic Feature Generation System (DFGS) for learning labeling rules [12], and a heuristic rule-based labeling algorithm [9]. These latter systems are used in production to generate bibliographic data for NLM’s MEDLINE database. Figure 8 shows the evaluation and comparison results. We can see that our current algorithm (AHSMM) performed a little better for title and author fields, far better for the affiliation field, far better than the heuristic rule-based method for abstract, though slightly poorer than the DFGS-based method for abstract. Overall, however, it achieved the best (average) segmentation and labeling accuracy.

Our current method outperforms both the alternative techniques for different reasons. In the case of the heuristic rule-based algorithm, in addition to general labeling rules (applicable to all page styles, in principle), we have to manually create style-specific rules for each new layout style, which is a laborious process. When, as often happens, the general rules do not apply to a newly encountered style, the heuristic rule-based algorithm fails badly, as is the case for labeling the affiliation and abstract fields in our experiment.

The DFSGS-based algorithm assumes that the absolute locations of fields of interest are relatively stable so that accurate bounding boxes can be estimated for them. However, when this assumption does not hold as is the case for the affiliation field in our experiment, many true affiliation zones are labeled incorrectly because they either fall outside the estimated bounding box or are confused with other zones within the bounding box.

5. Summary

The long term preservation of the biomedical literature, both historic and contemporary, requires the acquisition of descriptive and technical metadata from a variety of digital file formats. In this paper we briefly describe a system that implements digital preservation, tools for metadata extraction for Web and TIFF files, but focus particularly on a method to automatically extract descriptive metadata from scanned medical journal articles.

References:

[1] METS - Metadata Encoding and Transmission Standards, Library of Congress Network Development & MARC Standards Office, <http://www.loc.gov/standards/mets/>.

[2] Dublin Core Metadata Element Set, Version 1.1, <http://dublincore.org/documents/dces/>.

[3] Preservation metadata for digital collections, National Library of Australia (NLA), <http://www.nla.gov.au/preservation/pmeta.html>.

[4] PREMIS - Preservation metadata and the OAIS reference model. A metadata framework to support the preservation of digital objects, http://www.oclc.org/research/projects/pmwg/pm_framework.pdf.

[5] Data Dictionary – Technical metadata for digital still images (draft), National Information Standards Organization (NISO Z39.87), http://www.niso.org/standards/resources/Z39_87_trial_use.pdf.

[6] DSpace, MIT, <http://www.dspace.org>.

[7] TIFF (Tagged Image File Format) version 6.0 specification, <http://partners.adobe.com/public/developer/en/tiff/TIFF6.pdf>.

[8] S. Mao, D. Misra, J. Seamans, and G.R. Thoma, Design strategies for a prototype electronic preservation system for biomedical documents, *IS&T Archiving Conference*, Washington DC, 2005. Accepted.

[9] J.W. Kim, D.X. Le, and G.R. Thoma, Automated labeling in document images, *SPIE Conference on Document Recognition and Retrieval VIII*, San Jose, CA, 2001, 111-122.

[10] S. Mao, L. Nie, and G.R. Thoma, Unsupervised style classification of document page images, *IEEE International Conference on Image Processing*, Genova, Italy, 2005. Accepted.

[11] K. Zhang, D. Shasha, and J.T.L.Y Wang, Approximate tree matching in the presence of variable length don't cares, *Journal of Algorithms*, 16(1), 1994, 33-66.

[12] S. Mao, J.W. Kim, and G.R. Thoma, 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, 2004, 225-232.

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

[14] S. Mao and G.R. Thoma, Bayesian learning of 2D document layout models for automated preservation metadata extraction, *Proceedings of the Fourth IASTED International Conference on VISUALIZATION, IMAGING, and IMAGE PROCESSING*, Marbella, Spain, 2004, 329-34.

[15] A. Stolcke and S.M. Omohundro, Best-first model merging for hidden Markov model induction, *Technical Report, TR-94-003*, International Computer Science Institute, Berkeley, CA, 1994.

[16] A. Stolcke, An efficient probabilistic context-free parsing algorithm that computes prefix probabilities, *Computational Linguistics*, 21(2), 1995, 165-201.

[17] P.A. Chou, G.E. Kopec, A stochastic attribute grammar model of document production and its use in document image decoding, *First International Workshop on Principles of Document Processing*, Washington DC, 1992.