Archiving a Historic Medico-legal Collection: Automation and Workflow Customization

Dharitri Misra, Song Mao, John Rees, George R. Thoma, National Library of Medicine, Bethesda, Maryland, USA

Abstract

The U.S. National Library of Medicine (NLM) has acquired a historical collection of documents, released by the Food and Drug Administration, specifying the Notices of Judgment (NJs) against manufacturers of adulterated or misbranded food, drugs and cosmetics. These documents, consisting of 70,000+ pages containing more than 65,000 NJs, are to be preserved and made accessible over the long term due to their legal and historical value.

We developed a preservation system, named SPER (System for Preservation of Electronic Resources), based on DSpace infrastructure, for archiving and disseminating NJs contained in these documents. For efficiency and cost-effectiveness, we developed algorithms to automatically identify the NJs and extract metadata from their contents, and then have an archivist review and edit the metadata, and ingest the NJs into the archive. Contents of the documents are also captured as text streams to provide full-text search capability for the NJs.

These functionalities required a number of changes to the open source DSpace software, including changing the ingest interface and workflow, handling metadata schema that does not map to Dublin Core, and enhancing the database schema.

This paper describes the overall SPER system, customized workflow for automated metadata extraction, the automated metadata extraction process, and an estimate of labor savings through automation.

Overview

The Collection

The 1906 Federal Food and Drug Act [1] established mechanisms for the federal government to seize, adjudicate, and punish manufacturers of adulterated or misbranded food, drugs and cosmetics. These federal activities were carried out by U.S. Food and Drug Administration (FDA). The legal proceedings associated with each case resulting from these activities were documented as Notices of Judgments (NJs), published synopses created monthly.

The U.S. National Library of Medicine (NLM) has acquired a collection of these documents (70,000+ pages) containing more than 65,000 NJs dating between 1906 and 1964. Moreover, the NJs include keys to the original evidence files created for each NJ, containing correspondence, lab results, and seized product samples and labeling. To preserve these NJs and make them accessible, our goal is to create a digital archive of both page images and metadata for each NJ. This archive will offer insight into U.S. legal and governmental history, but also into the evolution of clinical trial science and the social impact of medicine on health. The legal history of some of our best-known consumer items of today, such as Coca Cola, can be traced in the collection.

Design Approach

The creation of this archive requires a system for identifying and capturing metadata for each NJ from the document text, ingesting the scanned FDA documents and captured metadata to the archive, and indexing the contents of each NJ to allow access through a Web server.

Prior to designing the system, we surveyed well-known open-source digital archiving systems such as DSpace and Fedora, but they did not meet our specific needs – especially with respect to batch-based operations and inclusion of automated metadata extraction (AME) tools. This led to the design of the *System for Preservation of Electronic Resources* (SPER), to allow automated metadata extraction, review and ingest of a batch of documents in the FDA NJ and similar collections, in an integrated fashion with archiving operations. SPER implements AME and ingest functionalities through in-house developed software, but leverages the powerful archiving infrastructure and access mechanisms provided by DSpace for storage and dissemination.

A prototype version of SPER was implemented and reported earlier [2]. This paper describes the architecture of the *operational* SPER system, details of automated metadata extraction and overview of system workflow supporting AME and metadata quality assurance, as applicable to the NJ documents. We also quantify labor savings as a result of automation. Although this specific collection is our first operational example, SPER design is flexible enough to allow ingest of other collections requiring different types of metadata extraction or acquisition mechanisms.

SPER System Description

SPER is an evolving Java-based system to research digital preservation functions and capabilities, including automated metadata extraction for documents, retrieval of available metadata from databases, document archiving, and ensuring long term use through bulk file format migration. (The last feature is not covered in this paper). SPER is built upon MIT's DSpace software (Version 1.4) [3], with some modifications and enhancements mentioned below. It runs on Windows platforms, using MySQL (Version 5.02 or higher) database, but may also run on Solaris systems.

The metadata acquisition/review and data ingest component of SPER is separated from the data access part. The latter simply uses the tailored JSP-based Web interface provided by DSpace for browsing, searching and downloading documents by an end-user through a standard Web browser. The first component involving AME and ingest, and implemented as a Java RMI-based [4] Client-Server system, is the subject of this paper. Note that although an RMI-based system restricts access to the server from supported client facilities only, it is more efficient and better suited to our AME and ingest needs than a Web-based interface.

System Architecture

The basic architecture of SPER is shown in Figure 1 and described below:

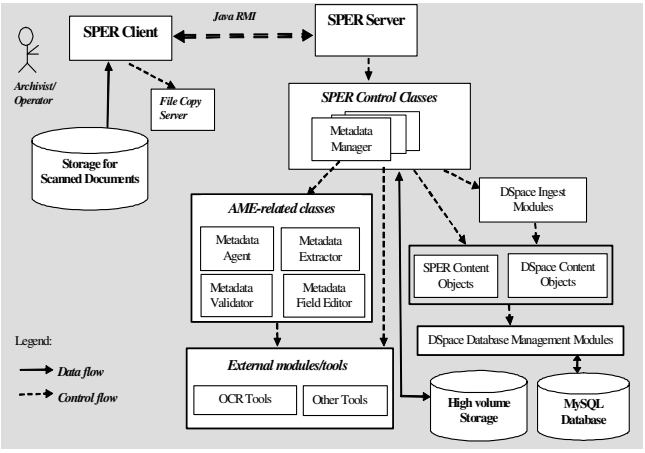

Figure 1. *SPER system architecture*

SPER Client is a Java application that runs on the client's machine and interacts with the SPER Server using RMI protocols. It provides the user (the archivist or an operator) with a Java/Swing-based graphical user interface for batch creation, AME initiation, and metadata review and ingest. The Client uses a FileCopyServer module on a dedicated RMI port to copy scanned documents from its local file system to the Server for OCR and storage.

SPER Server is a dedicated server module, running at the SPER server facility and serving multiple clients simultaneously, each on a separate thread. Actual processing corresponding to user request is performed by the lower level Control classes.

SPER Control classes include *Batch Manager, Metadata Manager, Ingest Manager* and *Property Manager* to handle specific SPER functions. For example: Batch Manager handles creation, deletion and queries for document batches, and Property Manager provides information on configurable properties of each collection, defined in a collection-specific property file.

Metadata Manager is the top-level class that controls the generation/acquisition of item specific metadata, and their formatting, storage, retrieval and modification. It uses collection-specific ***AME-related classes*** (*Metadata Agent*, *Metadata Extractor*, *Metadata Validator* and *Metadata Field Editor*) that may be developed and plugged in to support new collections. The Metadata Extractor class uses ***External modules*** such as *OCR Tools* to access textual information in a collection's documents.

AME Workflow

Automated metadata extraction from digitized documents, shown in Figure 2, is a two-step process: (a) model building, and (b) AME operations. Prior to these steps, the FDA NJ documents (either the originals, or, more frequently, their reproduction copies) are digitized into TIFF image format at an external scanning facility, after which they are stored at the site of the archivist in charge of the collection.

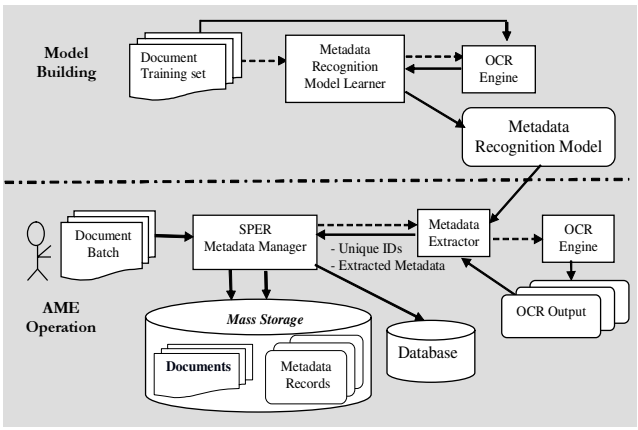

Figure 2. *Workflow involving automated metadata extraction*

Next, in the model-building phase, the archivist works with the SPER researchers to provide a small subset of document samples of each layout style for building collection-specific *Metadata Recognition Models*. These models are built by the *Metadata Recognition Model Learner* (implemented using machine-learning techniques) by sending each image through an optical character recognition (OCR) engine and interpreting the resulting output. The layout-specific classifiers, keywords identifying metadata fields and other attributes are stored as model data. Further AME details specific to the FDA NJ collection are provided in the next section.

During operations, the operator/archivist submits a sequence of digitized TIFF documents as a Batch for AME operations, either in synchronous or asynchronous mode. These images are OCRed by the SPER Server using the FineReader OCR engine 8.0 [5]. The Metadata Extractor, by referencing the Recognition model created earlier, identifies NJ boundaries and embedded metadata fields in the OCRed text. The NJ case number is used as the unique ID of each case in a collection, and all relevant metadata, along with their bounding box pixel coordinates in the TIFF image, are written by the Metadata Manager to an XML-formatted file and stored in the server. Review and editing of these automatically extracted metadata is described under "Metadata Verification."

SPER-specific Modifications to DSpace

To support batch-based operations, DSpace Content classes and database tables are augmented by SPER with new classes, namely: Batch, Page and MetaItem. Other classes are added to tailor the system for handling collection of documents with different attributes such as AME requirements, metadata acquisition methods, and descriptive metadata schemas.

The relationship between a Batch and DSpace archived Items is shown in Figure 3. A Page corresponds to a scanned image file in SPER. A MetaItem corresponds to an item (NJ) with an associated metadata record in the SPER *staging area*, before it is submitted for ingest. It is converted to a DSpace Item during ingest, where it loses its correlation with Page or Batch.

Another noteworthy enhancement to the DSpace infrastructure is the porting of its database schema and RDBMS classes to use the MySQL (V 5.02) database management system instead of the PostgreSQL or Oracle DBMS. This allowed SPER

the flexibility of having its database either on Windows or on Solaris platforms.

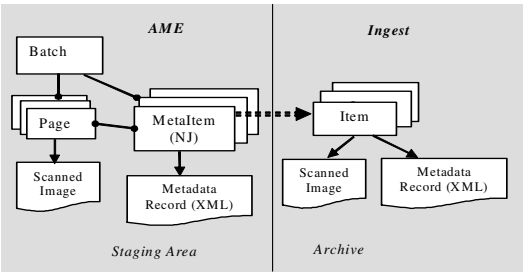

Figure 3. *Relationship between Batch and archived Items*

Automated Metadata Extraction for FDA NJ Collection

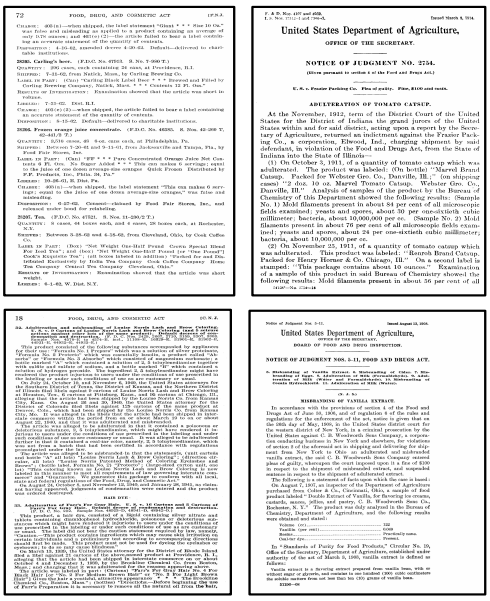

Figure 4*. Four typical layout styles in FDA NJ documents.*

Automated metadata extraction from scanned FDA NJ documents is a challenging problem since: (a) NJs have different layout styles as shown in Figure 4, (b) each NJ could start and end anywhere in a page, (c) each NJ has variable length that varies from part of a page to several pages (d) metadata items of interest are located in both labeled fields as well as in free text, and (e) there are OCR errors in extracted metadata items due to suboptimal photocopying, scanning quality, and small and old-style fonts in the documents.

The AME system we developed for this collection of FDA documents [6] is designed as follows: We first segment each document page into a sequence of textlines. From each textline, we extract a set of fourteen image and textual features, which are used to train static classifiers called Support Vector Machines (SVMs). We then model the class syntax in the sequence of logical labels of the sequence of textlines using Hidden Markov Models (HMMs). Finally, we combine the static classifiers (SVMs) with the class syntax models (HMMs) in our unified algorithm for optimal textline classification.

The AME system is used to automatically detect the boundary of each NJ, and extract twelve metadata fields for each NJ. The extracted fields are: *issue date, evidence number, NJ number, unique identifier (UI), title, keywords, defendant names, adjudicating court jurisdiction, seizure date, seizure location, shipped from city/state,* and *shipped into city/state*. In the training (or learning) phase, a recognition model is learned from randomly selected training samples for each distinct layout style. In the recognition phase, the layout style of an input document page is detected and the appropriate recognition model is used to classify each textline of the page into a logical entity. The first six metadata fields are directly extracted from logically labeled textlines and the remaining six are extracted from the NJ body free text via regular expressions used to model field-specific metadata string patterns.

To correct OCR errors in extracted metadata items, we built a user-defined dictionary and integrated it into the OCR engine to correct common errors at the word level, and built regular expressions to correct common OCR errors at the character level.

Metadata Verification

After NJs are identified, and the corresponding metadata is extracted and stored for a batch of documents, their manual review and validation is an essential step prior to ingest. It is especially necessary for the FDA NJ collection, as its older paper stock and outdated fonts produce frequent OCR errors, the most important being misidentification or omission of NJ numbers and failed recognition of metadata fields.

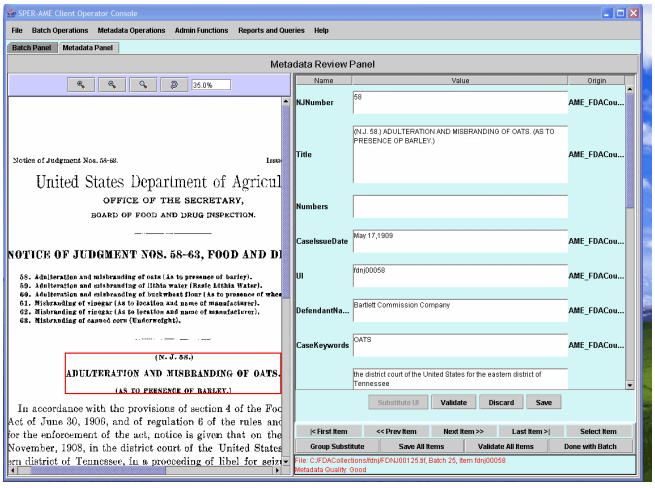

Figure 5. *Metadata review screen in SPER client*

The Metadata review screen in SPER client, shown in Figure 5, allows the following steps:

- The operator requests metadata for a set of NJs extracted from a submitted batch. The NJs are displayed one at a time, but the operator may navigate easily through the set by using buttons such as First, Next, Select etc.
- As the extracted metadata for each NJ is displayed on the screen, the corresponding document image is displayed along

with it for a comparison of the document text with the extracted values. Any corrections may then be made and saved.

- To help the operator refer to the metadata text in the document, a bounding box enclosing the text is displayed on the image, by using coordinates extracted and stored in the metadata file during AME.
- If an NJ was not included in a set or an extraneous NJ is added to it (due to errors in recognizing the NJ case number), the operator may insert or delete a record using special editing functions. To help insert a new record, SPER displays a metadata record template with field titles for operator to fill in.
- Group Editing: Since NJ numbers are sequential (and increasing) within a set of ascending order documents, and certain fields such as the NJ Issue Date are common to a group of NJs in a batch, SPER exploits this attribute and provides a number of functions to conveniently edit a set of metadata records as a group rather than individually.

Archiving and Dissemination of NJs

Once a batch of NJs is validated by the operator, it may be ingested directly to the SPER archive. During ingest, SPER server identifies each NJ within the batch, and retrieves the corresponding MetaItem from its database. Then it acquires the associated TIFF images and the metadata record from the staging area and transfers the MetaItem to a DSpace *Item*; and stores both the source files and the XML-based metadata record in the archive as bitstreams for the NJ.

For full-text searching of NJs contents, the corresponding text from the OCRed files of the source images is also stored as a separate text bitstream for each NJ, and is indexed using the Lucene search engine provided with DSpace.

For Web access to NJ records in the database, DSpace-provided JSPs are customized to use SPER-specific logo and terminologies, and access is provided via a Tomcat Web server. All search and retrieval functions are then carried out by DSpace modules from the SPER archive in an independent manner.

AME Performance Assessment

A short time-motion test was conducted to estimate the potential savings accruing from automatic data extraction versus human data entry. The time taken for an individual to read NJ records and manually enter correct metadata elements was measured. The test used 23 randomly chosen NJ records representing the three heavily narrative layout styles. The text ranges from three paragraphs to several pages in length. The fourth style was not chosen because each data element is preceded by an element label and the text itself is very brief (akin to a bulleted list), making identification and selection quite easy. The human operator had no previous experience with the source content, although he was an experienced staff member with an advanced degree. He was briefed on the content and given instructions about the text strings and other clues about where to find the metadata. This was the same information used to train the AME algorithms. This test represents the best-case scenario for human extraction, and results would likely be different using student workers or other inexperienced contract staff.

Using a blank form from the AME interface, the twelve metadata elements mentioned earlier (which the AME extracts from the free text), were manually identified by the operator. The AME creates five additional constant data elements that were not tested for manual entry: Metadata Creation Date; Publisher; MIME type; Language; Copyright Status.

The average time to manually enter metadata was 2.39 minutes per NJ (55 minutes total for all 23 records). The time includes reading through the text for the data elements plus input time. Alternatively, the AME process took 33 seconds per NJ to OCR the TIFF image (9 seconds), extract the same data (2 seconds) on a standard Pentium 4, 3.6 GHz/512 RAM, and for human QC to manually correct OCR errors (22 seconds). This difference is not insignificant over the course of the entire project. Extrapolating over the 65,000 NJs, manual data entry would take 2,590 hours—more than a year of full-time equivalent (FTE) staff time. Alternatively the AME process would take only 596 hours (3.4 months FTE).

Summary

This paper describes a system for the preservation of a historic collection at the National Library of Medicine. We customize MIT's open source DSpace software to provide the archiving infrastructure and access mechanisms, but also incorporate in the system special modules developed for automated metadata extraction (based on machine learning techniques), ingest and operator validation functions. A preliminary time-motion study indicates a four-fold improvement in throughput as a result of automation.

Acknowledgment

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM), and Lister Hill National Center for Biomedical Communications (LHNCBC).

References

[1] Public Law 59-384, repealed in 1938 by 21 U.S.C. Sec 329 (a). And U.S Food and Drug Administration, "Federal Food and Drugs Act of 1906 (The "Wiley Act")," http://www.fda.gov/opacom/laws/wileyact.htm (3 Feb. 2006).

[2] Mao S, Misra D, Seamans J, Thoma, G. R.: Design Strategies for a Prototype Electronic Preservation System for Biomedical Documents, Proc. IS&T Archiving Conference, Washington DC, pg 48. (2005).

[3] DSpace at MIT, http://www.dspace.org.

[4] Remote Method Invocation, http://java.sun.com/products/jdk/rmi

[5] ABBYY FineReader OCR Engine 8.0. http://www.abbyy.com

[6] Thoma GR, Mao S, Misra D, Rees J. Design of a Digital Library for Early 20th century Medico-legal Documents Proc. ECDL 2006. Eds: Gonzalo J et al. Berlin: Springer-Verlag; LNCS 4172: 147-57.

Author Biography

Dharitri Misra is a researcher at the U.S. National Library of Medicine, working on digital preservation topics. She earned her M.S. and Ph.D. degrees in Physics from the University of Maryland.

Song Mao is a Staff Scientist at the U.S. National Library of Medicine. He conducts research in document image analysis, information extraction, machine learning, and pattern recognition. He earned his M.S. and Ph.D. degrees in Electrical and Computer Engineering from the University of Maryland. He is a member of the IEEE and IEEE Computer Society.

John Rees is curator of the Archives and Modern Manuscripts Program at NLM's History of Medicine Division. He earned his MA at the University of Mississippi and MLIS at the University of Texas-Austin.

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.