Digital Preservation and Knowledge Discovery Based on Documents from an International Health Science Program

Dharitri Misra
National Library of Medicine
Bethesda, MD, USA 20894
dmisra@mail.nih.gov

Robert H. Hall
National Institute of Allergy and Infectious Diseases
Bethesda, MD, USA 20892
robert.hall@nih.gov

Susan M. Payne
National Institute of Allergy and Infectious Diseases
Bethesda, MD, USA 20892
susan.payne@nih.gov

George R. Thoma
National Library of Medicine
Bethesda, MD, USA 20894
gthoma@mail.nih.gov

ABSTRACT

Important biomedical information is often recorded, published or archived in unstructured and semi-structured textual form. Artificial intelligence and knowledge discovery techniques may be applied to large volumes of such data to identify and extract useful metadata, not only for providing access to these documents, but also for conducting analyses and uncovering patterns and trends in a field. The System for Preservation of Electronic Resources (SPER), an information management tool developed at the U.S. National Library of Medicine, provides these capabilities by integrating machine learning, data mining and digital preservation techniques. In this paper, we present an overview of SPER and its ability to retrieve information from one such dataset. We show how SPER was applied to the semi-structured records of an international health science program, the 46-year continuous archive of conference publications and related documents from the Joint Cholera Panel of the U.S.-Japan Cooperative Medical Science Program (CMSP). We explain the techniques by which metadata was extracted automatically from the semi-structured document contents to preserve these publications, and show how such data was used to quantitatively describe the activity of a research community toward a preliminary study of a subset of its specific health science program goals.

Categories and Subject Descriptors

H.3 **[Information Storage and Retrieval]**: H3.3 Information Search and Retrieval. H3.4 Systems and Software. H3.7 Digital Libraries I.5 **[Pattern Recognition]** I.5.2 Design Methodology - Feature evaluation and selection.

General Terms

Algorithms, Management, Design, Experimentation

Keywords

Metadata, Automated Metadata Extraction, Machine Learning, Digital Preservation, Data Analysis, Knowledge Discovery

1. INTRODUCTION

Medical information, founded on a massive archive of knowledge in biomedical literature, research articles, case studies, and reviews, often carry additional information related to research policies, participants and the contemporary understanding of health science, but potentially valuable health science data are easily overlooked as the retrieval of semantic information from them is often highly labor-intensive. However, automated machine learning techniques have made it possible to identify and retrieve important metadata elements such as article titles, authors, institutions from semi-structured document sets, which can be used for archiving the documents and providing access to them. Furthermore, these metadata fields, in combination with indexed document text, can be used to discover patterns and trends, and to quantify relations among technical terms, individuals, institutions, and other concepts, making it feasible to draw inferences accessible to target communities for specific purposes such as policy formulation and peer review.

In this paper we describe how such data analysis is made possible for a historic health science collection from the Joint Cholera Panel of the U.S.-Japan Cooperative Medical Science Program (CMSP) by using an R&D system called SPER (System for the Preservation of Electronic Resources) developed at the U.S. National Library of Medicine [1].

In the following sections, we provide a brief description of SPER, discuss the CMSP collection and its analysis goals, describe the techniques used by SPER toward meeting such goals, and display some preliminary analytical results. Finally, we outline ongoing knowledge discovery work using SPER.

2. BACKGROUND AND RELATED WORK

The SPER system was originally developed to identify, locate and retrieve collection-specific metadata cost-effectively, using machine learning, from the contents of semi-structured textual documents. It was expanded to research other aspects of digital preservation, including archiving the document collections using the extracted metadata in an OAIS-compliant repository [2]. The system has been successfully used to extract domain-specific metadata for a historic medico-legal document collection from the U.S. Food and Drug Administration, and build a digital repository to access the collection [3]. Recently, we have expanded SPER to use such context-sensitive metadata to conduct semantic queries, perform quantitative analysis, and determine patterns and trends.

3. SPER SYSTEM OVERVIEW

SPER is an evolving system to research and implement digital preservation and information retrieval functions, with focus on:

1. Automated metadata extraction (AME) from textual documents using machine learning
2. Preservation of and access to documents in a digital archive
3. Support of knowledge discovery from the archived data

JCDL'12, June 10–14, 2012, Washington, DC, USA.

SPER provides an end-to-end platform for preserving and analyzing document sets by integrating these functions into its workflow. It is implemented as a Java-based, customizable framework, into which new functionalities may be added as deemed appropriate.

3.1 Function Description

a) Automated Metadata Extraction: This important functionality is used for locating and extracting domain-specific metadata from the textual contents of a document collection using supervised machine learning techniques. Textlines are classified using Support Vector Machines (SVMs) [4] as the static feature classifier - either singly, or in combination with stochastic language models such as Hidden Markov Models (HMMs) [5] when different textline classes overlap in feature space. From these classified textlines, corresponding metadata elements are extracted either using field tokenizers, or through regular expression matching.

Paper documents are usually scanned to TIFF images, prior to submission to SPER, where it is input to an optical character recognition (OCR) engine such as the ABBYY FineReader; and the output is used for the AME operations. SPER allows the extracted metadata to be reviewed and validated by an operator for quality assurance, prior to being used for ingesting the document into a digital repository.

b) Document Preservation and Access: The preservation of items in a digital collection is performed by SPER using DSpace [6] as the underlying infrastructure. The source pages, derivatives (such as PDFs), metadata for individual items, as well as the indexed text are stored in a high-volume storage device with an accompanying MySQL database.

Access to archived embedded in the documents but not directly retrievable items is provided via a Web browser using a configurable DSpace Web interface, allowing browse/search of the collection's domain-specific metadata.

c) Knowledge Discovery: Discovery of additional information, through standard search mechanisms, requires the following:

(a) Identification of additional metadata elements prior to AME operations so that they may be located and extracted from the documents during AME phase.

(b) Post-processing of metadata to transform stored elements to a suitable form for data analysis and trend discovery.

3.2 SPER System Architecture

The architecture and main dataflows within SPER are shown in Figure 1. The main system component, the SPER Server, runs as a Web application and is used to perform AME and build the archive from a collection of digitized documents. A user of the system, an operator or archivist, uses a SPER Client, installed at their site, to download the digitized documents and initiate these functions.

The AME Engine incorporates the automated metadata extraction functions, including invoking the OCR tool for recognition of text in the input documents. Once the data is ingested to the repository, the Data Analysis Engine may be used to first post-process the metadata, then to receive the semantic queries, operate on the data sets, and return the results. Access to the repository is provided by a separate DSpace-based Web application, not shown in the figure.

The AME Engine and the Data Analysis Engine are collection-specific components, whereas the rest of the SPER system is reusable and configurable for each new collection.

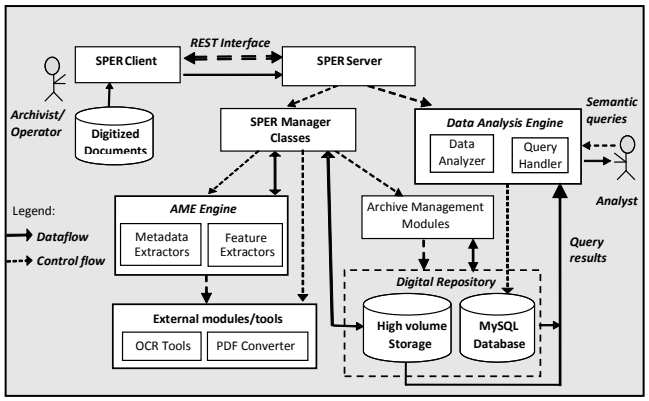


Figure 1. SPER system architecture.

4. The CMSP Program and Publications

The CMSP program [7] was founded in 1965 as a joint commitment by U.S. President Lyndon B. Johnson and Japanese Prime Minister Eisaku Sato to address health problems in South and Southeast Asia through an expanded medical research effort. Joint Panels on Cholera were founded to control cholera through vaccine development, by making effective therapy available, and by international research collaboration. Participants met annually (in meetings sponsored alternately by the U.S. and Japan), and made presentations to report progress. These presentations, published as abstract books, provide an extraordinary window into efforts against a disease that continues to cause major public health problems today.

The CMSP Joint Cholera Panel archive, held by the National Institute of Allergy and Infectious Diseases (NIAID), consists of the conference proceedings from 1965 through 2009, along with the proceedings of an earlier formative conference held in 1960. Also included are annual progress reports from a related NIAID research program within the International Centers for Medical Research and Training (ICMRT), and a list of Study Section reviewers from year 2003 to 2010.

4.1 Goal of Processing the CMSP Publications

The overall goals of processing the CMSP publications were: (a) to preserve the document collection in a digital repository, and (b) to collect statistical data, through metadata analyses, to produce metrics for evaluating the responsiveness of the research program to its strategic objectives.

The specific data for analysis were established as:

- Number of presentations on cholera research in USA, Japan, and developing countries, both individually and through international collaboration
- Levels of innovation in specific fields, such as vaccines, therapeutics, and diagnostics
- The time intervals between presentation of research discoveries in a conference forum and their subsequent appearance in formal peer-reviewed publications
- Identification of the most active and influential authors in terms of the number of presented abstracts, levels of collaboration, and overall impact

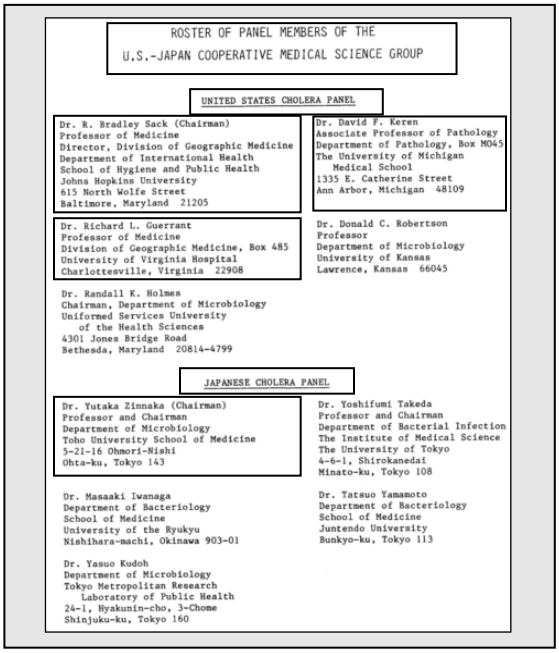
4.2 CMSP Metadata

The metadata elements required for access and/or analysis of the CMSP dataset are contained in the research articles and various rosters, within the conference proceedings and Study Section lists, and are shown in Table 1.

Table 1. Metadata fields in different CMSP documents.

Document	Type	Metadata
Conference Proceedings	Presentation	Title, Author Names, Affiliated Institutions, Keywords, Page location
	Panelist Info	Panel (U.S./Japan), Panelist Name, Affiliation, Address
	Attendee List	Attendee Name, Affiliation, Address
Study Section Rosters	Reviewer List	Reviewer Name, Affiliation, Address

An example of metadata occurrence in one sample CMSP record (containing a list of panelists in a conference proceedings page) is shown as a set of rectangular boxes in Figure 2.


ROSTER OF PANEL MEMBERS OF THE
U.S.-JAPAN COOPERATIVE MEDICAL SCIENCE GROUP

UNITED STATES CHOLERA PANEL

Dr. R. Bradley Sack (Chairman)
Professor of Medicine
Director, Division of Geographic Medicine
Department of International Health
School of Hygiene and Public Health
Johns Hopkins University
615 North Wolfe Street
Baltimore, Maryland 21205

Dr. David F. Keren
Associate Professor of Pathology
Department of Pathology, Box M045
The University of Michigan
Medical School
1335 E. Catherine Street
Ann Arbor, Michigan 48109

Dr. Richard L. Guerrant
Professor of Medicine
Division of Geographic Medicine, Box 485
University of Virginia Hospital
Charlottesville, Virginia 22908

Dr. Donald C. Robertson
Professor
Department of Microbiology
University of Kansas
Lawrence, Kansas 66045

Dr. Randall K. Holmes
Chairman, Department of Microbiology
Uniformed Services University
of the Health Sciences
4301 Jones Bridge Road
Bethesda, Maryland 20814-4799

JAPANESE CHOLERA PANEL

Dr. Yutaka Zinnaka (Chairman)
Professor and Chairman
Department of Microbiology
Toho University School of Medicine
5-21-16 Ohmori-Nishi
Ohta-ku, Tokyo 143

Dr. Yoshifumi Takeda
Professor and Chairman
Department of Bacterial Infection
The Institute of Medical Science
The University of Tokyo
4-6-1, Shirokanedai
Minato-ku, Tokyo 108

Dr. Masaaki Iwanaga
Department of Bacteriology
School of Medicine
University of the Ryukyu
Nishihara-machi, Okinawa 903-01

Dr. Tatsuo Yamamoto
Department of Bacteriology
School of Medicine
Juntendo University
Bunkyo-ku, Tokyo 113

Dr. Yasuo Kudoh
Department of Microbiology
Tokyo Metropolitan Research
Laboratory of Public Health
24-1, Hyakunin-cho, 3-Chome
Shinjuku-ku, Tokyo 160

Figure 2. Metadata location for Panelists in a sample record.

4.3 AME Technique and Feature Generation

The problem of recognizing metadata fields in the CMSP documents falls mainly under the general technique known as "named entity recognition" (NER) [8].

The recognition features selected for locating metadata elements in the CMSP documents [9] fall into six categories: dictionary (name) features, text features, punctuation features, special word features, contextual features (e.g. features from neighboring words) and weight features (e.g. clustering, ordering or frequency of specific features in a line). For our algorithms, a feature is based upon a category or a set of words representing a concept, rather than all equivalent words found for that concept. This improves the performance of the textline classifier, without significantly lowering the accuracy, by reducing the dimension of the feature space for the SVM implementation. Three separate metadata models were developed to handle AME for articles, panelist/attendees and Study Section reviewers respectively.

There were a number of challenges in locating and retrieving required metadata from the CMSP publications. Some examples are: illegibility of the text, wide variations in the format, layout and relative position of data elements (such as individual name, institution and title), occurrence of a country's non-English native language terms, and the occurrence of same words in the context of persons, institutions or addresses. These problems were reduced by developing language-specific "special word" features, field-specific tokenizers, and textline class validators.

Table 2 shows some statistics related to the CMSP collection, as derived from the extracted metadata.

Table 2. CMSP collection related statistics.

Year Span	1960-2009
Number of Conference Proceedings	55
Number of Presentations	2,556
Instances of Author	11,736
Instances of Panelist	589
Instances of Conference Attendee	4,239
Instances of Participating Institution	4,416
Instances of Study Section Reviewer	3,110
Number of Individuals (Participants)	6,923

5. ACCESS TO CMSP REPOSITORY

CMSP documents, along with the corresponding validated metadata and searchable PDFs, were ingested to the CMSP repository by the SPER server. The OCR'ed text for the entire corpus was then indexed using Apache/Lucene and stored in the archive.

SPER provides access to the archived CMSP documents, through title, author, year of publication and keyword searches, by customization of the standard DSpace-based Web interface. Furthermore, it also supports browse and search on Participants.

6. DATA ANALYSIS

The metadata retrieved from the documents do not directly contain all required information for data analysis. The "Data Analysis Engine" for CMSP, therefore, performs post-processing in the following two areas:

(a) Mapping an institution to the corresponding country, and a "Country Group." Special dictionaries were created to do this for institutions where no country was specified.
There were three specific country groups, referenced here as: USG (U.S. + other high GDP countries), Jpn (Japan), and Dev (Developing countries)

(b) Identifying an individual (with a unique first name, last name) corresponding to a contributor, whose "published name" varied widely in the document set. Special utilities and manual authentication were used to accomplish this.

6.1 Trends and Patterns

The analysis of the CMSP data enabled a long-standing health science program to be examined against specific policy goals. A detailed evaluation of the program and its policies is outside the scope of this paper, and will be presented elsewhere.

Nevertheless, the examples in Figure 3 a-c illustrate the results produced by SPER to help assess the stated CMSP Program goals. These results include examination of the U.S. sponsored meetings only, which occurred in alternating years, except 1969 and 2001, when there were no meetings.

Figure 3a shows that one specific important discovery (glucose-based oral rehydration therapy) was presented at conferences eight years before peer-review publications in PubMed [10] in 1968. It further reveals that conference activity on glucose-based oral therapy for cholera remained at a low level from 1979.

Figure 3b shows the trends in research activities over the CMSP time span, and highlights increased activity and international collaboration from 2003 that coincided with support from stakeholders and a new translational research agenda. (These results include examination of the U.S. sponsored meetings only, which occurred in alternating years, except 1969 and 2001).

Figure 3c identifies the most active authors who collaborated heavily with others for their research, and published their findings in the peer-reviewed journal Scopus.

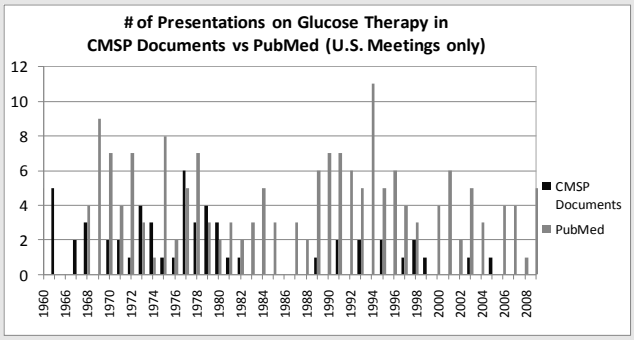


Figure 3a. Research activity on glucose in rehydration therapy reported at CMSP vs. PubMed.

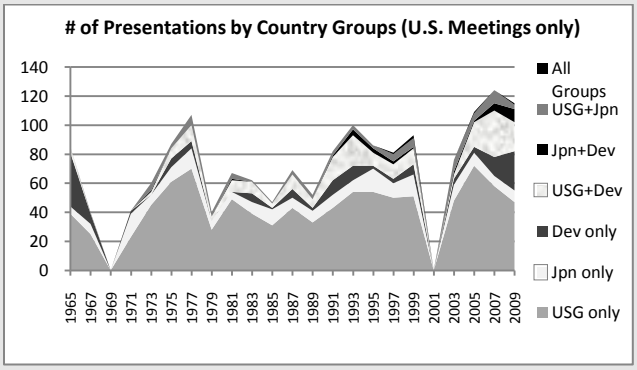


Figure 3b. Contribution of and level of collaboration by different country groups (indicated by # of presentations) in U.S. meetings (with no meetings in 1969 and 2001).

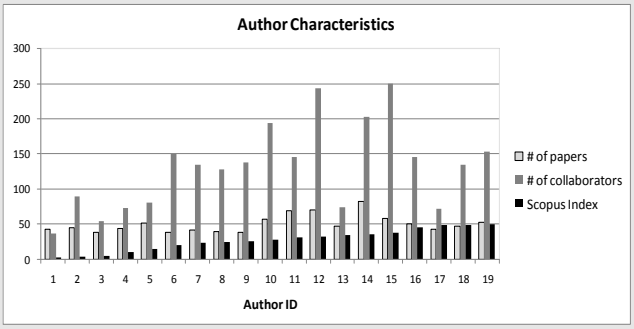


Figure 3c. Publication and collaboration by Authors.

7. CONCLUSION AND FUTURE WORK

Our work continues in further analysis of the CMSP collection in other related areas, and in providing improved semantic query capabilities by (a) building a full CMSP cholera knowledgebase, and (b) making the knowledge discovery system selectively accessible through a CMSP portal. In future, we also aim to apply the technique to other important biomedical collections, with domain-specific metadata, so as to uncover useful information for researchers and other targeted user communities.

8. ACKNOWLEDGEMENTS

The authors acknowledge Abdulmekik Shifa and Tobias T. Hall for their contributions to compiling the CMSP Joint Cholera Panels' records, and Dr. Xiaoli Zhang for her work in building the metadata models for the automated extraction.

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).

9. REFERENCES

[1] Misra, D., Mao, S., Rees, J., Thoma, G.R. 2007. Archiving a Historic Medico-legal Collection: Automation and Workflow Customization, *Proc. IS&T Archiving Conference*, Washington DC, pg 157-161. (2007).

[2] Reference Model for an Open Archival Information System (OAIS). 2002. http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf

[3] Misra, D., Chen, S., Thoma, G.R. 2009. **A** System for Automated Extraction of Metadata from Scanned Documents Using Layout Recognition and String Pattern Search Models, Proc. *IS&T Archiving Conference*. Arlin. pg 107-111. (2009).

[4] Cortes C., Vapnik V. 1995. Support-vector Network. Machine Learning. Vol. 20, pages 273-297, (1995).

[5] Rabiner, L. R., Juang, B. H. 1993. Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall. (1993).

[6] DSpace, MIT (http://www.dspace.org).

[7] The U.S.-Japan Cooperative Medical Science Program (http://www.niaid.nih.gov/topics/globalresearch/usjapan/Pages/history.aspx

[8] Named Entity Recognition (NER) and Information Extraction (IE) (http://www-nlp.stanford.edu/ner/)

[9] Zhang, X., Zou, J., Le, DX., Thoma, G.R. 2010. Investigator Name Recognition From Medical Journal Articles: A Comparative Study of SVM and Structural SVM.**,** *International Workshop on Document Analysis Systems*, pg 121-128. (2010).

[10] PubMed (http://www.ncbi.nlm.nih.gov/pubmed/).