

Managing Occupational Medicine Historical Data

Kazem Taghva, Julie Borsack, Tom Nartker, Steve Lumos and Allen Condit
Information Science Research Institute
University of Nevada, Las Vegas

Abstract

This paper presents a prototype for converting hard-copy “medical charts” to electronic form. It discusses the purpose and direction of our project and gives a framework for the development of the complete system, which we call MCARS.

1 Introduction

The “medical chart” has always served as the repository and the assemblage of medical information for over a century. It has been the well-accepted method for tracking a person’s medical history throughout their lifetime. Recently, with the advent of electronic means for sharing information, it has been proposed that medical data could be more accessible, more useful, and more timely to physicians and medical professionals if it were stored electronically.

There are several projects that are working to develop models that will be comprehensive, portable, and robust and at the same time, benefit both the physician and the patient for the ultimate goal of better health care. This is not a simple task. The GEHR (Good Electronic Health Record) project has been working on electronic medical record representation for nearly fifteen years.

Although very much related to other medical record research projects, our group at the Information Science Research Institute (ISRI) has a more narrow focus. Our work is specific to occupational medicine which represents a concentration within a much larger arena. What’s more, the core of our research centers around the conversion of existing medical charts to electronic form. This paper gives an overview of our project, what issues we have discovered, the progress we have made, and our continued research direction.

2 Background and Purpose

Our involvement in this project began with a government agency’s need to review past employee medical records that span over 50 years. The records are warehoused in several storage facilities across the country so the process of just locating the correct patient file can take several weeks. The next step in the process of course, is to manually review the patient file to locate the sought after information — a pain-staking and time-consuming task.

Although an automated way of accessing this medical data is required, the best solution wasn’t readily obvious. We spent several months understanding the magnitude and complexity of the problem before the direction of our research came into focus.

One partial solution that has been implemented in similar projects is scanning the pages of the medical chart and “indexing” them with certain values. The problem with this solution is two-fold. First, what values should be indexed? Invariably, the values selected will be inadequate for some types of queries. Second, this method introduces manual processing. People-time is expensive and error-prone.

We believe a more sophisticated solution is feasible using current technology. Capturing all the medical record data would solve the first issue; examining related technologies and automating the expensive data capture process would solve the second. ISRI has studied and researched optical character recognition (OCR), information retrieval (IR), and related technologies for several years[4][5]. By combining and interfacing these technologies, a comprehensive medical records capture system could be built. Accomplishing these goals increases the complexity of this project by several-fold and raises a number of issues that would not otherwise be encountered. The solution though is elegant and complete.

The goals for our project are clear. Design a system specific to occupational medicine that takes as input hard copy medical data and produces correct, queryable medical information. Interfacing the required technologies and building the intermediate

processes to produce this electronic medical information establishes our research objectives.

3 Preliminaries

3.1 Occupational Medicine

Occupational medicine (occ-med) manages employee health and safety in the workplace. Occ-med encompasses prevention, treatment, and rehabilitation of work-related disease and injury. It identifies and supports outside medical consultation when an injury or illness is outside its scope of responsibility. Specific applications in occ-med include: identifying potential on-site risks, assessing fitness for work, communicating with primary care physicians and other clinical colleagues, promoting health, responding to medical emergencies, and monitoring employees for possible side-effects caused from their work environment. The key discriminate of occ-med is its direct relationship to the workplace.

The occ-med infrastructure characterizes the kind of medical data collected in this agency's patient charts. It also sets the foundation and the boundaries for building the appropriate relationships between data elements. For example, a certain job task may carry with it possible exposure to lead inhalation. By knowing that an employee's task carried this potential for exposure, we can relate his spirometric test results (from his patient chart) to his current job assignment. One may ask,

“Did employee X develop any adverse conditions while assigned to job task T?”

Because our focus is occ-med, this relationship is inherent.

3.2 Related Techechnologies

When the objective is to convert hard copy pages to electronic form, two technologies quickly come to mind: scanning technology and OCR. These of course are crucial. But other technologies will also play a role in the conversion process of medical forms data. These include image processing, forms recognition, database, and interface display. The role of these technologies in our project is described in the paragraphs below.

Scanners have become the latest required peripheral device for the PC so most people have a good grasp of what they do: scanners digitize hard copy pages. By sensing variations in light intensity, scanners represent *patterns* on the input page as analog

signals. These signals are then digitized into matrices of 1s and 0s that replicate the hard copy. The quality of the scanned image is crucial to subsequent processing steps, in particular recognition. Close attention needs to be paid to the characteristics of the hard copy and to the scanning process so that flaws in the image are minimized.

The occ-med historical data can be summarized as a collection of hundreds of unique forms with millions of instances. Each instance contains *pre-printed* information representing labels for fields in the form. The instance also contains *user-filled* information which can be typed, hand written, or check marked. The initial processes in our system are image processing and forms classification. Each form in the patient's file belongs to a specific category (tabs) such as **Audio**, **EKG's**, etc. In addition to certain image pre-processing routines to detect and correct skews, our system extracts the forms logical layout based on techniques developed in [1][2][3]. These layouts have natural tree structures which are modeled using XML DTD's[6]. The collection of these DTD's built from unfilled forms are stored in the database for incoming form identification. After form identification, the user-filled data must be extracted for recognition by OCR.

OCR typically refers to the recognition of machine printed characters which may or may not be a component of a particular form. Forms recognition is broader. It usually includes several recognition modules: handprint recognition or intelligent character recognition (ICR), optical mark recognition (OMR), barcode recognition and possibly handwritten recognition. Some forms recognition systems may include rudimentary form identification as well. *Zones* on the form are identified and bound to the appropriate recognition module. In this way, the data, whether they be textual fill-in fields or a checkboxes, they can be recognized accordingly.

The database will be used to store the data recognized from the forms. We discovered early in our analysis that the complexity of the data's relationships required more than just a flat relational representation. Our model is object relational and embodies subclass/superclass relationships, inheritance and other features that are not easily described in a traditional entity relationship (ER) diagram. Beyond the medical attributes essential to model occ-med, our design includes *attribute meta-data*. For example, we record form labels, recognition types and edit rules. This information will be used to aid recognition and subsequent correction.

The PDF display interface uses the form process-

ing capability of *Adobe's Portable Document Format* to display, in an editable fashion, field data acquired from the patient forms and stored in the database. The generated PDF file will contain an actual image of the form (obtained by scanning a blank form), with field values placed in their correct positions. In this way, we are able to provide the experienced user with a familiar interface, as well as enable the printing of copies which are nearly indistinguishable from manually-filled forms. The ability to edit these fields will lead to a query interface where the user can type queries (known field values) directly into the form and all other related information for that form can then be searched, located and displayed.

4 Medical Capture and Retrieval System

The groundwork for the *Medical Capture and Retrieval System* (MCARS) is in place. But to grasp its intricacies, we have built a prototype system that incorporates some of the technologies and concepts presented in this paper.

Since medical information is protected under the Privacy Act of 1974, the patient data we use in the prototype is contrived. The forms though are actual pages from patient charts. The information displayed is currently stored in two related but separate databases. One database records field element data; the other, relates patients to form instances, forms to categories, and forms to other related forms.

The search interface is simple: any free-form query is accepted and the string is searched in designated database fields. Any patient with a match will be returned. This simplicity was purposeful; until we know what search interface makes the most sense, we have added no artificial constraints. The patient's ID, first and last name, and employment location is displayed in the hit list.

When a patient is selected, his "electronic medical chart" is displayed. Figure 1 shows this layout. The patient charts are color-coded and divided into specific categories as shown. Even the fact that certain categories are on the left or right is meaningful. Replicating the medical chart as closely as possible to the hard copy should help simplify finding patient data. Also, the order of patient forms in the chart is preserved — more recent forms appear at the top of each category list.

When the required form is located, it can be selected from the pull-down list. Figure 2 shows the scanned image of the the selected form. At this point,

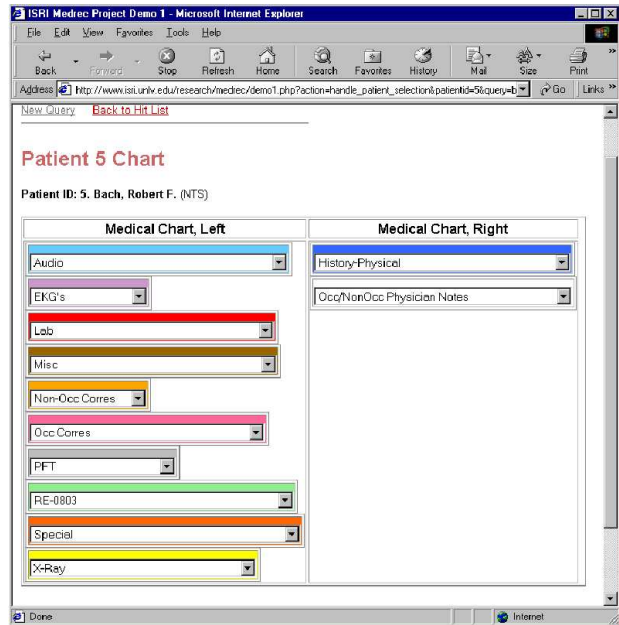


Figure 1: Electronic Medical Chart

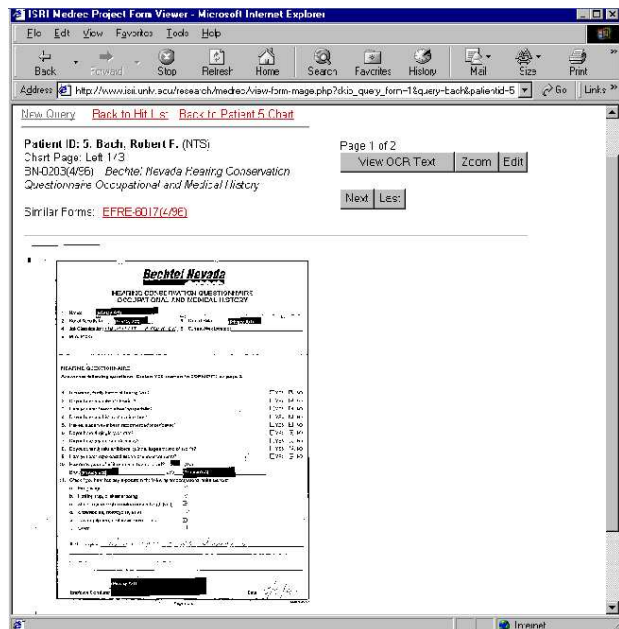


Figure 2: Patient Information and Scanned Form

The screenshot shows a web browser window with the address bar displaying a URL. The main content area contains a form titled "HEARING CONSERVATION QUESTIONNAIRE OCCUPATIONAL AND MEDICAL HISTORY". The form has several sections: a header section with fields for Name, Social Security No., Date of Birth, Job Classification, and Work Phone; a section titled "HEARING QUESTIONNAIRE" with a list of 11 questions and checkboxes for YES and NO; and a section for occupational noise sources with checkboxes for various activities like firing range, hunting, and music.

Figure 3: Editable Patient Record Form

our prototype has only a few features. They include:

1. Navigation: top left corner.
2. Patient/form information: upper left (below 1).
3. *Similar* form information: (below 2). Similar forms are forms contained in *this* patient's file that are similar to the one displayed. This concept can be important to a medical professional but we will use it to compare and correct similar forms as they are recognized.
4. Form manipulation: (right of 3). There are several buttons for form manipulation. The **Zoom**, **Next**, and **Last** are standard form navigation buttons. The **View OCR Text** displays the text generated by the OCR device. The **Edit** displays the editable version of this form based on the "captured" data. We describe this in more detail below.

Figure 3 shows what can be accomplished with MCARS. Displayed is a clean, editable version of the scanned form shown in Figure 2. What gets displayed is the data captured from the scanned form with OCR and then stored in the database. This view is an example of a *form query interface* as discussed in Section 3.2. The most obvious use of MCARS is the automated conversion of hard copy medical records to

electronic form but other valuable uses are possible. MCARS gives an efficient means for semi-automated correction; it gives a familiar interface with which to view patient data; it displays both the "historical form" (scanned image) and updated database information. As we continue to expand and improve MCARS, its versatility will become more apparent.

5 Conclusion

MCARS is in its infancy. More research and work are requisite to complete the comprehensive system we have planned. We believe though that this system has a much broader reach than a sole solution to one agency's problem. Billions of hardcopy pages exist and they may currently be the only record of essential information. Hoarding vast amounts of hardcopy data is not just one agency's problem either. One finds this behavior in nearly all government agencies and in private industry as well. But by concentrating on this agency's problem and finding a solution, we are able to identify and resolve real issues of the conversion dilemma.

References

- [1] Andrew D. Bagdanov and Marcel Worring. First order gaussian graphs for efficient structure classification. Technical Report 2001-10, Intelligent Sensory Information System Group, University of Amsterdam, August 2001.
- [2] F. Cesarini, M. Gori, S. Marinai, and G. Soda. Structured document segmentation and representation by the modified x-y tree. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 563–566, 1999.
- [3] Pinar Duygulu and Volkan Atalay. A hierarchical representation of form documents for identification and retrieval. In *Document Recognition and Retrieval, SPIE Electronics Imaging 2000*, 2000.
- [4] Kazem Taghva, Julie Borsack, and Allen Condit. Evaluation of model-based retrieval effectiveness with OCR text. *ACM Transactions on Information Systems*, 14(1):64–93, January 1996.
- [5] Kazem Taghva, Allen Condit, Julie Borsack, John Kilburg, Changshi Wu, and Jeff Gilbreth. The MANICURE document processing system. In *Proc. IS&T/SPIE 1998 Intl. Symp. on Electronic*

Imaging Science and Technology, San Jose, CA,
January 1998.

- [6] Kazem Taghva, Min Xu, Emma Regentova, and Tom Nartker. Utilizing XML schema for describing and querying still image databases. Technical Report 2002-02, Information Science Research Institute, University of Nevada, Las Vegas, April 2002.