

MICHIGAN STATE

U N I V E R S I T Y

Project Plan

Machine Learning Document Classification and Redaction

The Capstone Experience

Team Technology Services Group

Lazaro Cruz

Genya Dobrev

Will Giger

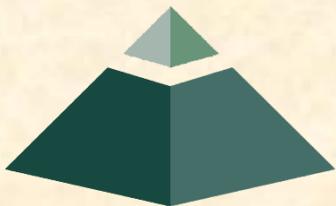
Jacob Harris

Xiaokuan Zhang

Department of Computer Science and Engineering

Michigan State University

Spring 2020



*From Students...
...to Professionals*

Functional Specifications

- Removes sensitive personal information from documents.
- In doing so, private information will not be viewed by who is not supposed to see it.
- This is important in medical records especially.



Design Specifications

- A person should be able to upload a document through a computer.
- PII can be redacted during template configuration and fields can be un-redacted as needed.
- User is displayed redacted version during indexing of the document.



Screen Mockup: Selecting Template

The screenshot shows a web browser window with the URL `tsgroup.com/video/required-documents-ecm-solution/`. The page title is "Template Management Config". The left sidebar contains a navigation menu with the following items: Application Setup (selected), Application, Object Type, Non-Mandatory Aspect, Picklists, Forms, Template Management (highlighted), Tracs, Trac Security, and Event Logging. Below the sidebar are three orange buttons: Views, Workflows, and Tools. The main content area has two tabs: "Capture Template" (selected) and "Active Wizard Template". Under the "Capture Template" tab, there is a text box with the instruction: "Please select an object type, then select an existing template to edit or click 'Create New' to start from scratch." Below this is a dropdown menu with the following options: "Select an object type to configure" (selected), "Accounting Document", "Medical Records" (highlighted), "Application Deliverable", "Attached Document", "Car Auction Document", "Case Document", "Claim Document", "Collection", "Contract Document", "Controlled Document", "Customer Order Document", "Digital Asset", "Document", "DUNS Document", "Engineering Document", "Engineering Drawing", and "Engineering Procedure". To the right of the dropdown are two buttons: "Save" (green) and "Publish" (orange).



Screen Mockup: Identifying Values

The screenshot displays a web application interface with a navigation bar at the top containing a search icon, 'Collections', 'Dashboard', and 'Create New Folder' buttons. The user is logged in as 'Hello, Administrator' with 'Sign Out' and 'Admin' buttons. The main content area is split into two panels. The left panel, titled 'Fingerprint Attributes', has a dropdown menu 'Select an attribute to configure'. Below it is a 'Vendor Name' field with a trash icon. The 'Attribute Extraction Rules' section also has a dropdown 'Select an attribute to configure'. The 'Policy Number' section features a dropdown 'Select the extraction type for this attribute' with 'Key / Value' and 'Zonal' options, a 'Capture' button, and a 'Draw' button. At the bottom of this panel are 'Aliases', 'Save', 'Validate', and 'Delete' buttons, along with a 'Redact PII' checkbox. The right panel shows a form titled 'Annual BSA Health and Medical Record Part A'. It includes 'GENERAL INFORMATION' with fields for Name (John Smith), Date of birth (01/24/1960), Age (60), Sex (Male), Address (1234 Michigan Ave, Chicago, IL 12334), Phone No. (635-832-9483), Unit leader, Council name/No., Unit No., Social Security No. (748-84-2837), Religious preference, and Health/accident insurance company (Policy No. 1783928). A note asks for a photocopy of insurance card sides. The 'HEALTH HISTORY' section has a table with columns 'Yes', 'No', 'Condition', and 'Explain'. The 'Allergies or Reaction to:' section includes a table for 'Medication' and 'Immunizations'. The 'MEDICATIONS' section has a table for listing current medications. The form also includes fields for 'Emergency contact No.', 'Allergies:', 'DOB:', and 'name:'. At the bottom, it asks for 'Administration of the above medications is approved by' with a signature line.



Screen Mockup: Redacted Indexing

tsgrp.com/video/required-documents-ecm-solution/

Search Collections Dashboard Create New Folder Hello, Administrator Sign Out Admin

Document Indexing

Enter properties for this document:

Name * 6-621-56786 07_23_19 FedEx.pdf

Vendor Name * Fedex

Invoice Number * 6-621-56786

PO Number 2187-4655-1

Amount * \$38.82

Invoice Date * 07-23-2019 Clear

Due Date * Please enter a date in the format of: MM-DD-YYYY
Clear

Exit Save and Next Save and Exit

Emergency contact No.:

DOB:

Allergies:

Annual BSA Health and Medical Record Part A

GENERAL INFORMATION

Name [REDACTED] Date of birth 01/24/1960 Age 60 Male Female

Address 1 [REDACTED] City Chicago State IL Zip 12334 Phone No. [REDACTED]

Unit leader [REDACTED] Council name/No. [REDACTED] Unit No. [REDACTED]

Social Security No. (optional; may be required by medical facilities for treatment) [REDACTED] Religious preference [REDACTED]

Health/accident insurance company [REDACTED] Policy No. 1783928

High-adventure base participants:
Expedition/crew No.: [REDACTED]
or staff position: [REDACTED]

ATTACH A PHOTOCOPY OF BOTH SIDES OF INSURANCE CARD. IF FAMILY HAS NO MEDICAL INSURANCE, STATE "NONE."

In case of emergency, notify:
Name [REDACTED] Relationship wife
Address [REDACTED]
Home phone [REDACTED] Business phone [REDACTED] Cell phone [REDACTED]
Alternate contact [REDACTED] Alternate's phone [REDACTED]

HEALTH HISTORY

Are you now, or have you ever been treated for any of the following:

Yes	No	Condition	Explain
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Asthma Last attack: _____	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Diabetes Last HbA1c: _____	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Hypertension (high blood pressure)	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Heart disease (e.g., CHF, CAD, MI)	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Stroke/TIA	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Lung/respiratory disease	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Ear/sinus problems	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Muscular/skeletal condition	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Menstrual problems (women only)	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Psychiatric/psychological and emotional difficulties	
<input checked="" type="checkbox"/>	<input type="checkbox"/>	Behavioral disorders (e.g., ADD, ADHD, Asperger syndrome, autism)	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Bleeding disorders	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Fainting spells	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Thyroid disease	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Kidney disease	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Sickle cell disease	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Seizures Last seizure: _____	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Sleep disorders (e.g., sleep apnea) Use CPAP: Yes <input type="checkbox"/> No <input type="checkbox"/>	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Abdominal/digestive problems	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Surgery	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Serious injury	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	Other	

ALLERGIES OR REACTION TO:

Medication _____

Food, Plants, or Insect Bites _____

IMMUNIZATIONS:

The following are recommended by the BSA. Tetanus immunization is required and must have been received within the last 10 years. If had disease, put "D" and the year. If immunized, check the box and the year received.

Yes	No	Date
<input type="checkbox"/>	<input type="checkbox"/>	Tetanus _____
<input type="checkbox"/>	<input type="checkbox"/>	Pertussis _____
<input type="checkbox"/>	<input type="checkbox"/>	Diphtheria _____
<input type="checkbox"/>	<input type="checkbox"/>	Measles _____
<input type="checkbox"/>	<input type="checkbox"/>	Mumps _____
<input type="checkbox"/>	<input type="checkbox"/>	Rubella _____
<input type="checkbox"/>	<input type="checkbox"/>	Polio _____
<input type="checkbox"/>	<input type="checkbox"/>	Chicken pox _____
<input type="checkbox"/>	<input type="checkbox"/>	Hepatitis A _____
<input type="checkbox"/>	<input type="checkbox"/>	Hepatitis B _____
<input type="checkbox"/>	<input type="checkbox"/>	Influenza _____
<input type="checkbox"/>	<input type="checkbox"/>	Other (i.e., Hib) _____

Exemption to immunizations claimed (form required).

MEDICATIONS

List all medications currently used. (If additional space is needed, please photocopy this part of the health form.)

(For more information about immunizations, as well as the immunization exemption form, see the back of this form.)



Screen Mockup: Redaction Editing

The screenshot displays a web application interface for document redaction. On the left, a sidebar contains configuration options for different document attributes: 'Fingerprint Attributes', 'Vendor Name', 'Attribute Extraction Rules', and 'Policy Number'. Each option includes a dropdown menu to 'Select an attribute to configure'. Below these, there are sections for 'Key (label) coordinates' and 'Value coordinates', each with a 'Capture' or 'Draw' button. At the bottom of the sidebar are 'Save', 'Validate', and 'Delete' buttons, along with a 'Redact PII' checkbox.

The main content area shows a 'Annual BSA Health and Medical Record Part A' form. The form is divided into several sections:

- GENERAL INFORMATION:** Fields for Name, Date of birth (01/24/1960), Age (60), Sex (Male), Address, City (Chicago), State (IL), Zip (12334), Phone No., Unit leader, Council name/No., and Unit No.
- ATTACH A PHOTOGRAPH OF BOTH SIDES OF INSURANCE CARD. IF FAMILY HAS NO MEDICAL INSURANCE, STATE "NONE.":** A section for emergency contact information, including Name, Relationship (wife), Address, Home phone, Business phone, Cell phone, and Alternate contact/phone.
- HEALTH HISTORY:** A table with columns 'Yes', 'No', 'Condition', and 'Explain'. Checked conditions include Asthma, Diabetes, Hypertension, Heart disease, Stroke/TIA, Lung/respiratory disease, Ear/sinus problems, Muscular/skeletal condition, Menstrual problems, Psychiatric/psychological and emotional difficulties, Behavioral disorders (ADD, ADHD, Asperger syndrome, autism), Bleeding disorders, Fainting spells, Thyroid disease, Kidney disease, Sickle cell disease, Seizures, Sleep disorders, Abdominal/digestive problems, Surgery, and Serious injury.
- Allergies or Reaction to:** A section for listing allergies and reactions to medication, food, plants, or insect bites.
- IMMUNIZATIONS:** A section for recording immunizations, including Tetanus, Pertussis, Diphtheria, Measles, Mumps, Rubella, Polio, Chicken pox, Hepatitis A, Hepatitis B, and Influenza.
- MEDICATIONS:** A section for listing currently used medications.

Vertical labels on the left side of the form indicate redaction targets: 'Emergency contact No.:', 'Allergies:', and 'DOB:'. The browser's address bar at the top shows the URL 'tsgroup.com/video/required-documents-ecm-solution/'.

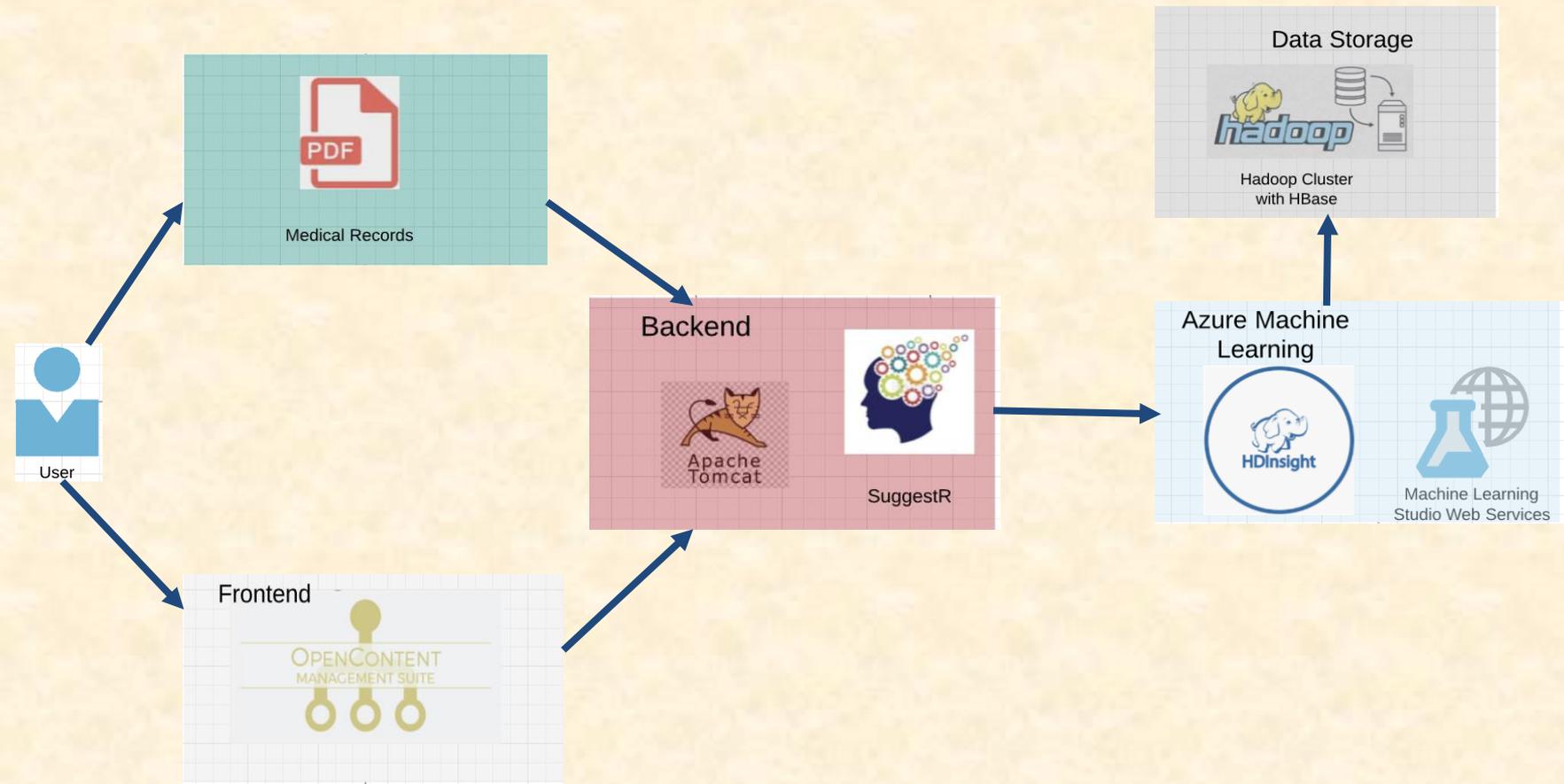


Technical Specifications

- Using Apache TomCat server to host the program.
- Uploading documents through front-end with a JavaScript application (OCMS).
- Managing back-end with Java and HBase database.
- Using Azure machine learning to recognize information and redact



System Architecture



System Components

- Hardware Platforms
 - Linux
 - Ubuntu
- Software Platforms / Technologies
 - Apache Tomcat Server
 - Hadoop cluster
 - OpenContent from client
 - Azure Machine Learning



Risks

- Which Azure Machine Learning environment would work best, if any would?
 - Time consuming process of testing each environment to find the best solution.
 - Mitigation: Upfront research for the most integratable environment.
- Redaction confidence level
 - Current client software can find metadata with strong confidence. In dealing with PII redaction the confidence level will need to be much higher.
 - Mitigation: Making sure to benchmark our Machine Learning continuously throughout development.
- What is PII exactly and how to measure it accurately
 - What information is PII exactly and how to train model to recognize it?
 - Mitigation: Worst case is doing it manually by going through documents and running it by the person redacting manually in the client's company.
- Client storage platform is still unclear
 - In last call with client it was unclear which storage platform the client would like to use.
 - Mitigation: Upfront research on Azure to fall back on.



Questions?

?

?

?

?

?

?

?

?

?

