MICHIGAN STATE UNIVERSITY

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.

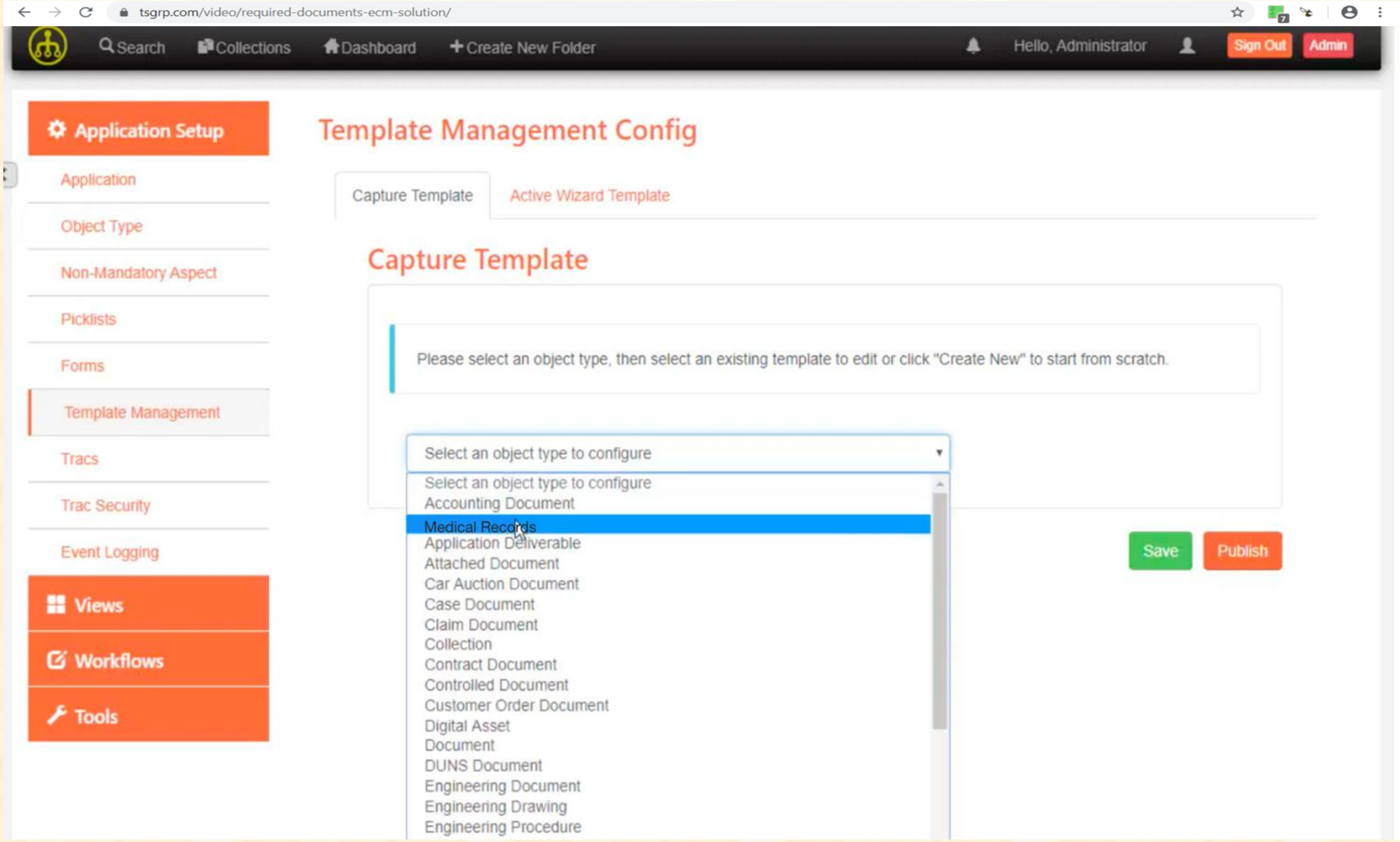
- In doing so, private information will not be viewed by who is not supposed to see it.

- 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



Template Management Config

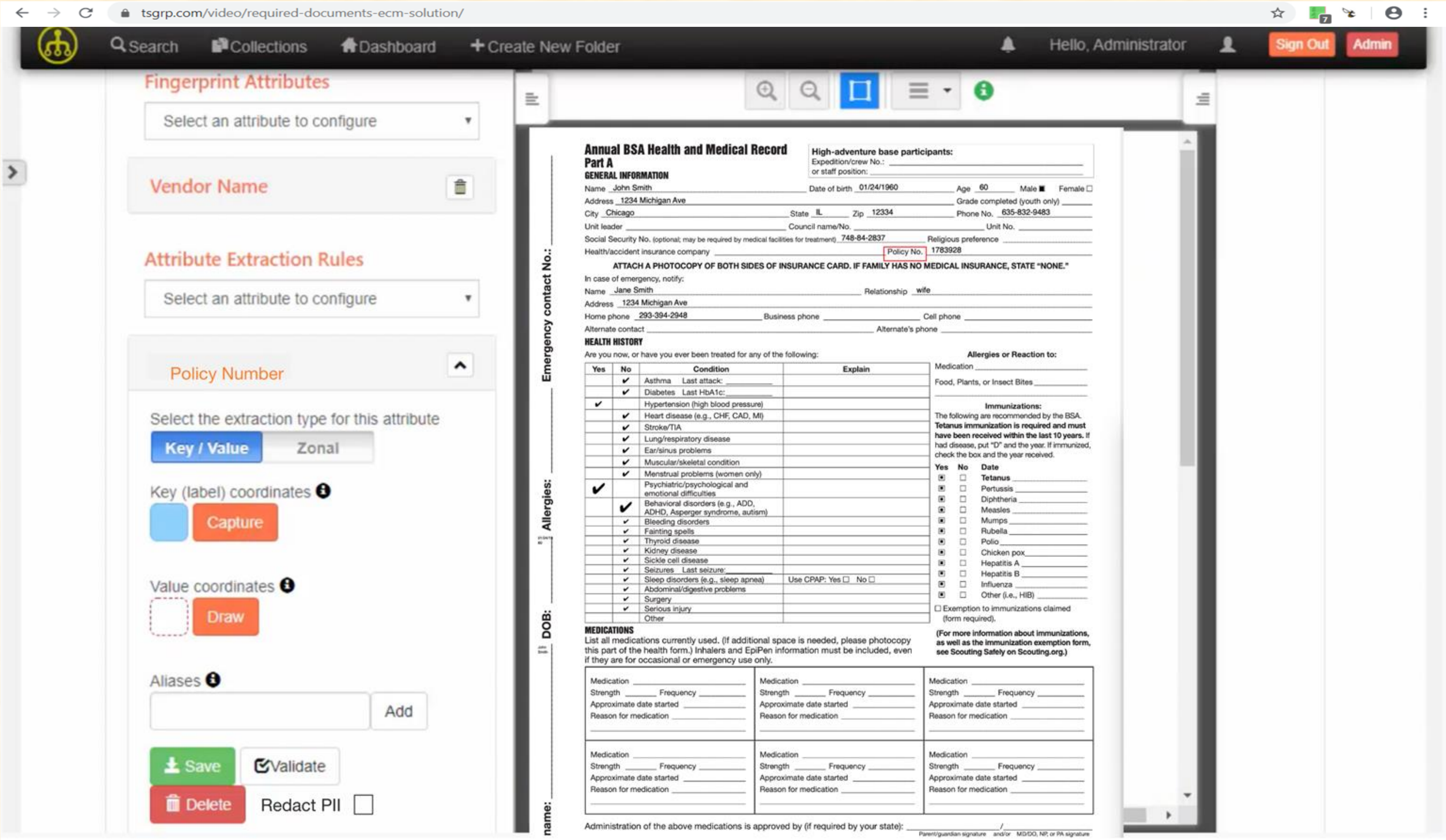
Capture Template

Please select an object type, then select an existing template to edit or click "Create New" to start from scratch.

Select an object type to configure ▾

Select an object type to configure
Accounting Document
Medical Records
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
Engineering Procedure

Screen Mockup: Identifying Values



Fingerprint Attributes

Select an attribute to configure

Vendor Name

Attribute Extraction Rules

Select an attribute to configure

Policy Number

Select the extraction type for this attribute

Key / Value | Zonal

Key (label) coordinates

Capture

Value coordinates

Draw

Aliases

Add

Save | Validate

Delete | Redact PII □

Annual BSA Health and Medical Record
Part A
GENERAL INFORMATION

High-adventure base participants:
Expedition/crew No.: ______
or staff position: ______

Name John Smith Date of birth 01/24/1960 Age 60 Male ■ Female □
Address 1234 Michigan Ave Grade completed (youth only) ______
City Chicago State IL Zip 12334 Phone No. 635-832-9483
Unit leader ______ Council name/No. ______ Unit No. ______
Social Security No. (optional; may be required by medical facilities for treatment) 748-84-2837 Religious preference ______
Health/accident insurance company ______ Policy No. 1783928

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

In case of emergency, notify:
Name Jane Smith Relationship wife
Address 1234 Michigan Ave
Home phone 293-394-2948 Business phone ______ Cell phone ______
Alternate contact ______ Alternate's phone ______

Emergency contact No.: ______
Allergies: ______
DOB: ______
name: ______

HEALTH HISTORY

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

Yes	No	Condition	Explain
	✓	Asthma Last attack: ______	
	✓	Diabetes Last HbA1c: ______	
✓		Hypertension (high blood pressure)	
	✓	Heart disease (e.g., CHF, CAD, MI)	
	✓	Stroke/TIA	
	✓	Lung/respiratory disease	
	✓	Ear/sinus problems	
	✓	Muscular/skeletal condition	
	✓	Menstrual problems (women only)	
✓		Psychiatric/psychological and emotional difficulties	
	✓	Behavioral disorders (e.g., ADD, ADHD, Asperger syndrome, autism)	
	✓	Bleeding disorders	
	✓	Fainting spells	
	✓	Thyroid disease	
	✓	Kidney disease	
	✓	Sickle cell disease	
	✓	Seizures Last seizure: ______	
	✓	Sleep disorders (e.g., sleep apnea)	Use CPAP: Yes □ No □
	✓	Abdominal/digestive problems	
	✓	Surgery	
	✓	Serious injury	
		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
☒	□	**Tetanus** ______
☒	□	Pertussis ______
☒	□	Diphtheria ______
☒	□	Measles ______
☒	□	Mumps ______
☒	□	Rubella ______
☒	□	Polio ______
☒	□	Chicken pox ______
☒	□	Hepatitis A ______
☒	□	Hepatitis B ______
☒	□	Influenza ______
☒	□	Other (i.e., HIB) ______

□ Exemption to immunizations claimed (form required).

(For more information about immunizations, as well as the immunization exemption form, see Scouting Safely on Scouting.org.)

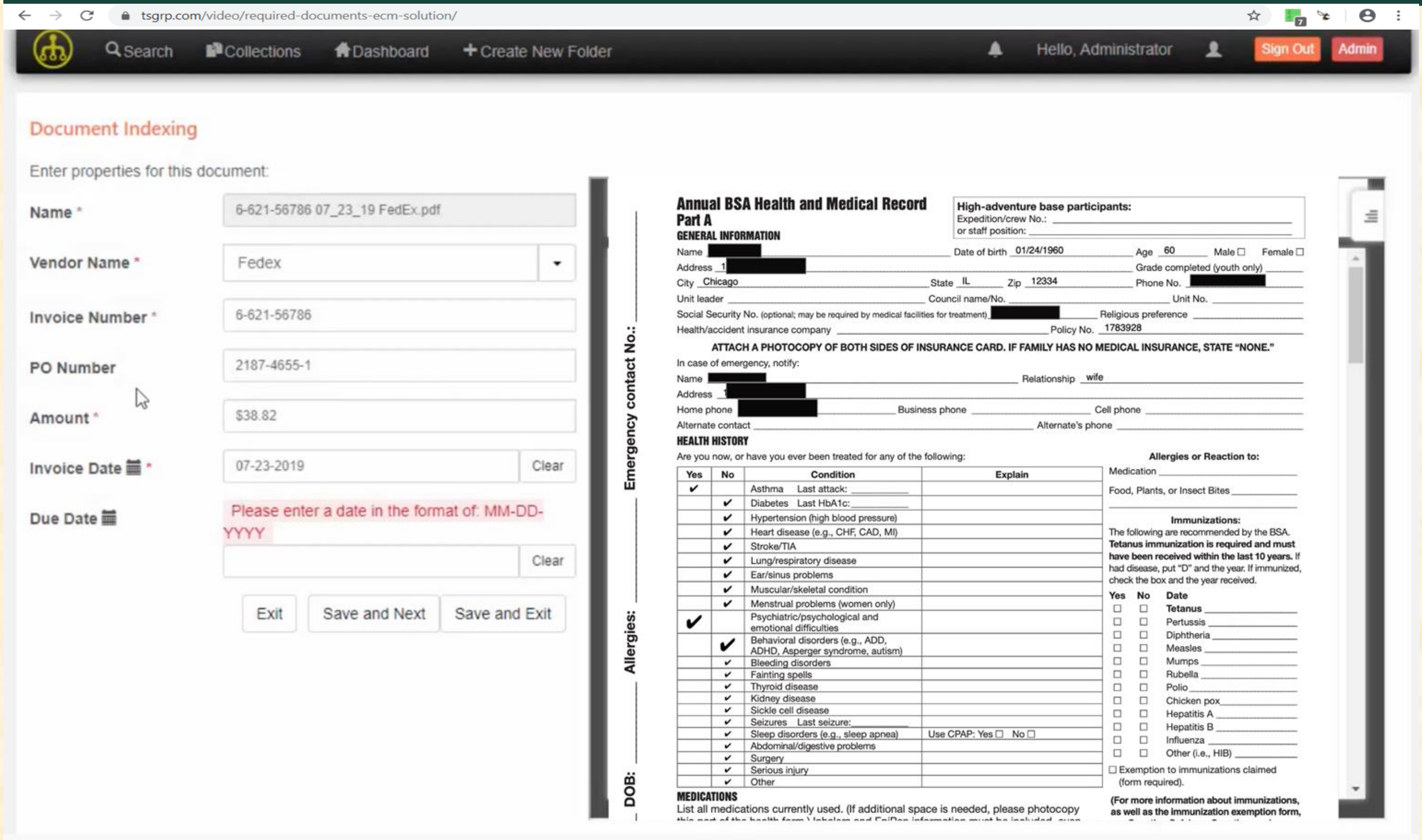
MEDICATIONS

List all medications currently used. (If additional space is needed, please photocopy this part of the health form.) Inhalers and EpiPen information must be included, even if they are for occasional or emergency use only.

Medication ______ Strength ______ Frequency ______ Approximate date started ______ Reason for medication ______	Medication ______ Strength ______ Frequency ______ Approximate date started ______ Reason for medication ______	Medication ______ Strength ______ Frequency ______ Approximate date started ______ Reason for medication ______
Medication ______ Strength ______ Frequency ______ Approximate date started ______ Reason for medication ______	Medication ______ Strength ______ Frequency ______ Approximate date started ______ Reason for medication ______	Medication ______ Strength ______ Frequency ______ Approximate date started ______ Reason for medication ______

Administration of the above medications is approved by (if required by your state): ______ / ______
Parent/guardian signature and/or MD/DO, NP, or PA signature

Screen Mockup: Redacted Indexing



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

Emergency contact No.:

Allergies:

DOB:

Annual BSA Health and Medical Record
Part A

High-adventure base participants:
Expedition/crew No.: ______
or staff position: ______

GENERAL INFORMATION

Name [redacted] Date of birth 01/24/1960 Age 60 Male □ Female □
Address [redacted] Grade completed (youth only) ______
City Chicago State IL Zip 12334 Phone No. [redacted]
Unit leader ______ Council name/No. ______ Unit No. ______
Social Security No. (optional; may be required by medical facilities for treatment) [redacted] Religious preference ______
Health/accident insurance company ______ Policy No. 1783928

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 ______ Cell phone ______
Alternate contact ______ Alternate's phone ______

HEALTH HISTORY

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

Yes	No	Condition	Explain
✓		Asthma Last attack: ______	
	✓	Diabetes Last HbA1c: ______	
	✓	Hypertension (high blood pressure)	
	✓	Heart disease (e.g., CHF, CAD, MI)	
	✓	Stroke/TIA	
	✓	Lung/respiratory disease	
	✓	Ear/sinus problems	
	✓	Muscular/skeletal condition	
	✓	Menstrual problems (women only)	
✓		Psychiatric/psychological and emotional difficulties	
	✓	Behavioral disorders (e.g., ADD, ADHD, Asperger syndrome, autism)	
	✓	Bleeding disorders	
	✓	Fainting spells	
	✓	Thyroid disease	
	✓	Kidney disease	
	✓	Sickle cell disease	
	✓	Seizures Last seizure: ______	
	✓	Sleep disorders (e.g., sleep apnea)	Use CPAP: Yes □ No □
	✓	Abdominal/digestive problems	
	✓	Surgery	
	✓	Serious injury	
	✓	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
□	□	**Tetanus** ______
□	□	Pertussis ______
□	□	Diphtheria ______
□	□	Measles ______
□	□	Mumps ______
□	□	Rubella ______
□	□	Polio ______
□	□	Chicken pox ______
□	□	Hepatitis A ______
□	□	Hepatitis B ______
□	□	Influenza ______
□	□	Other (i.e., HIB) ______

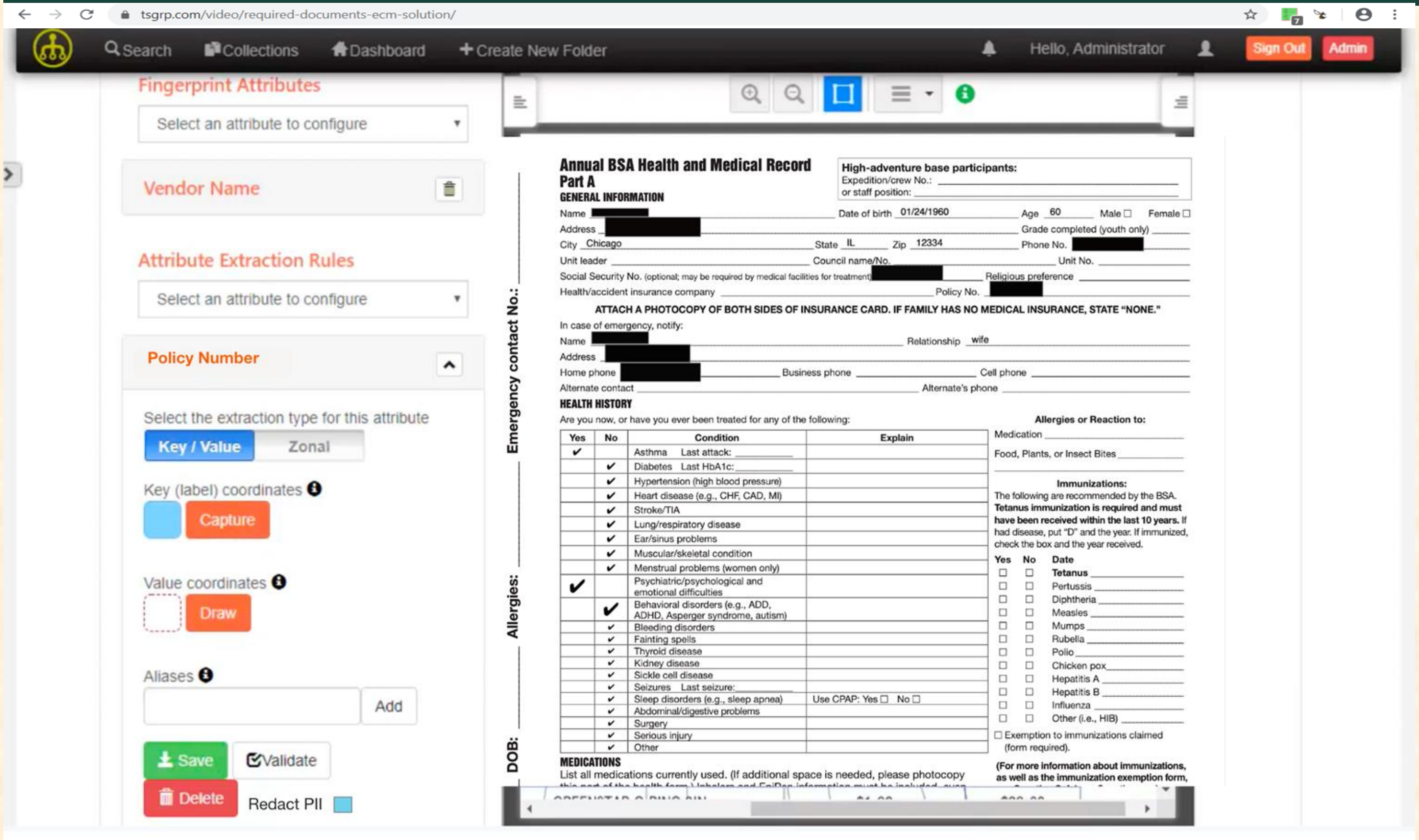
□ Exemption to immunizations claimed (form required).

(For more information about immunizations, as well as the immunization exemption form,

MEDICATIONS

List all medications currently used. (If additional space is needed, please photocopy

Screen Mockup: Redaction Editing



Annual BSA Health and Medical Record
Part A

High-adventure base participants:
Expedition/crew No.: ______
or staff position: ______

Emergency contact No.:

Allergies:

DOB:

GENERAL INFORMATION

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

Address [redacted] Grade completed (youth only) ______

City Chicago State IL Zip 12334 Phone No. [redacted]

Unit leader ______ Council name/No. ______ Unit No. ______

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

Health/accident insurance company ______ Policy No. [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 ______ Cell phone ______

Alternate contact ______ Alternate's phone ______

HEALTH HISTORY

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

Yes	No	Condition	Explain
✓		Asthma Last attack: ______	
	✓	Diabetes Last HbA1c: ______	
	✓	Hypertension (high blood pressure)	
	✓	Heart disease (e.g., CHF, CAD, MI)	
	✓	Stroke/TIA	
	✓	Lung/respiratory disease	
	✓	Ear/sinus problems	
	✓	Muscular/skeletal condition	
	✓	Menstrual problems (women only)	
✓		Psychiatric/psychological and emotional difficulties	
	✓	Behavioral disorders (e.g., ADD, ADHD, Asperger syndrome, autism)	
	✓	Bleeding disorders	
	✓	Fainting spells	
	✓	Thyroid disease	
	✓	Kidney disease	
	✓	Sickle cell disease	
	✓	Seizures Last seizure: ______	
	✓	Sleep disorders (e.g., sleep apnea)	Use CPAP: Yes □ No □
	✓	Abdominal/digestive problems	
	✓	Surgery	
	✓	Serious injury	
	✓	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
□	□	**Tetanus** ______
□	□	Pertussis ______
□	□	Diphtheria ______
□	□	Measles ______
□	□	Mumps ______
□	□	Rubella ______
□	□	Polio ______
□	□	Chicken pox ______
□	□	Hepatitis A ______
□	□	Hepatitis B ______
□	□	Influenza ______
□	□	Other (i.e., HIB) ______

□ Exemption to immunizations claimed (form required).

(For more information about immunizations, as well as the immunization exemption form,

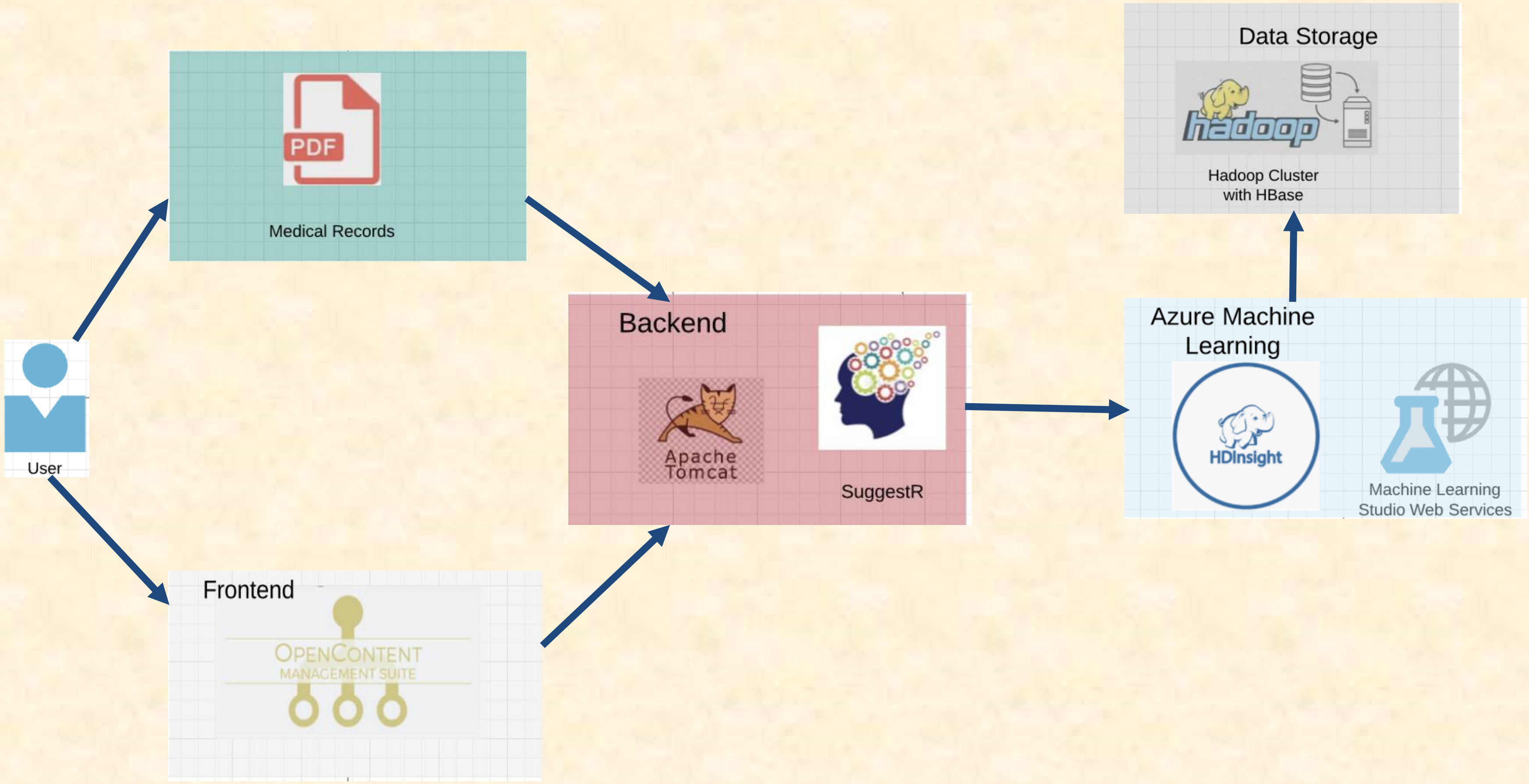
MEDICATIONS

List all medications currently used. (If additional space is needed, please photocopy

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?

? ? ? ?

? ?

? ? ?