

DISSERTATION

APPLICATION OF SYSTEMS ENGINEERING PRINCIPLES IN THE ANALYSIS,
MODELING, AND DEVELOPMENT OF A DOD DATA PROCESSING SYSTEM

Submitted by

Kevin P. Fenton

Department of Systems Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2023

Doctoral Committee:

Advisor: Steven J. Simske

Thomas Bradley

Ken Carlson

Rebecca Atadero

Copyright by Kevin P. Fenton 2023

All Rights Reserved

ABSTRACT

APPLICATION OF SYSTEMS ENGINEERING PRINCIPLES IN THE ANALYSIS, MODELING, AND DEVELOPMENT OF A DOD DATA PROCESSING SYSTEM

In support of over 1000 military installations worldwide, the Department of Defense (DoD) has procured contracts with thousands of vendors that supply the military with hazardous materials constituting billions of dollars of defense expenses in support of facility and asset maintenance. These materials are used for a variety of purposes ranging from weapon system maintenance to industrial and facility operations. In order to comply with environmental, health, and safety (EHS) regulations, the vendors are contractually obligated to provide Safety Data Sheets (SDSs) listing EHS concerns compliant with the requirements set forth by the United Nations Globally Harmonized System of Classification and Labeling of Chemicals (GHS).

Each year chemical vendors provide over 100 thousand SDSs in a PDF or hard copy format. These SDSs are then entered manually by data stewards into the DoD centralized SDS repository – the Hazardous Material Management Information System (HMIRS). In addition, the majority of these SDS are also loaded separately by separate data stewards into downstream environmental compliance systems that support specific military branches. The association between the vendor-provided SDSs and the materials themselves was then lost until the material reaches an installation at which point personnel must select the SDS associated to the hazardous material within the service-specific hazardous material tracking system.

This research applied systems engineering principles in the analysis, modeling, and development of a DoD data processing system that could be used to increase efficiency, reduce

costs, and provide an automated solution not only to data entry reduction but in transitioning and modernizing the hazard communication and data transfer towards a standardized approach. Research for the processing system covered a spectrum of modern analytics and data extraction techniques including optical character recognition, artificial neural networks, and meta-algorithmic processes. Additionally, the research covered potential integration into existing DoD framework and optimization to solve many long-standing chemical management problems. While the long-term focus was for chemical manufacturers to provide SDS data in a standardized machine-encoded format, this system is designed to act as a transitional tool to reduce manual data entry and costs of over \$3 million each year while also enhancing system features to address other major obstacles in the hazard communication process. Complexities involved with the data processing of SDSs included multi-lingual translation needs, image and text recognition, periodic use of tables, and while SDSs are structured with 16 distinct sections – a general lack of standardization on how these sections were formatted. These complexities have been addressed using a patent-pending meta-algorithmic approach to produce higher data extraction yields than what an artificial neural network can produce alone while also providing SDS-specific data validation and calculation of SDS-derived data points.

As the research progressed, this system functionality was communicated throughout the DoD and became part of a larger conceptual digital hazard communication transformation effort currently underway by the Office of the Secretary of Defense and the Defense Logistics Agency. This research led to five publications, a pending patent, an award for \$280,000 for prototype development, and a project for the development of this system to be used as one of the potential systems in a larger DoD effort for full chemical disclosure and proactive management of not only hazardous chemicals but potentially all DoD-procured products.

ACKNOWLEDGEMENTS

I would first like thank my wife who has been at my side for over 20 years. She has been there for me to help in low moments and inspire me to go further and achieve more than I could have dreamed. I would not have been able to achieve this without her support. I would like to thank my parents, sister, family, and close friends for their love and patience through this process and for always being there for me. I would like to thank my colleagues at CEMML and AFCEC whom I've learned so much from over the years. I would like to thank Jonathan Luu and USAFSAM for allowing me to formulate my dissertation around the needs of their organization and in providing me a conduit to impact change in the Air Force. I would like to thank my committee members – Dr. Thomas Bradley, Dr. Ken Carlson, and Dr. Rebecca Atadero for their time and guidance to help progress my research. Lastly, I want to thank Dr. Steven Simske for all the guidance I received from him, the lessons I learned, the time he dedicated to my research, and his incredible accessibility and kindness. For most of us, we're lucky if we get a teacher that makes such a big impact in our lives. I've had the good fortune of two – Mr. Henry, my freshman year of high school science teacher who brought forth my first interest, fascination, and love science; and Dr. Simske who reinvigorated that love for science and engineering and whom I've learned so much from and has inspired me. He has been an excellent teacher, mentor, and friend throughout this process.

TABLE OF CONTENTS

ABSTRACT.....	ii
LIST OF FIGURES	viii
LIST OF ACRONYMS	ix
1 CHAPTER 1 - INTRODUCTION AND RESEARCH MOTIVATIONS.....	1
1.1 INTRODUCTION.....	1
1.2 HAZARD COMMUNICATION BREAKDOWN	3
1.3 LACK OF SDS AVAILABILITY, PRECISION, AND DATA QUALITY	6
1.4 LACK OF IMMEDIATE SDS RETRIEVAL ABILITIES	8
1.5 LOSS IN INVENTORY ACCOUNTABILITY	9
2 CHAPTER 2 - APPLICABLE TECHNOLOGIES REVIEW	10
2.1 OPTICAL CHARACTER RECOGNITION (OCR).....	10
2.2 ARTIFICIAL INTELLIGENCE	15
SYSTEM TRAINING	17
2.3 “NATURAL LANGUAGE PROCESSING” (NLP)	19
2.4 COMPUTER VISION.....	21
3 CHAPTER 3 - APPLICABLE REGULATORY DRIVERS REVIEW	22
3.1 EMERGENCY PLANNING AND COMMUNITY RIGHT-TO-KNOW ACT (EPCRA) 22	
3.2 COMPREHENSIVE ENVIRONMENTAL RESPONSE, COMPENSATION, AND LIABILITY ACT (CERCLA).....	23
3.3 RESOURCE CONSERVATION AND RECOVERY ACT (RCRA)	24
3.4 CLEAN AIR ACT (CAA).....	24
3.5 OCCUPATIONAL SAFETY AND HEALTH ADMINISTRATION HAZARD COMMUNICATION (OSHA HAZCOM)	25
3.6 GLOBALLY HARMONIZED SYSTEM (GHS)	25
3.7 EUROPEAN UNION REGISTRATION, EVALUATION, AUTHORISATION, AND RESTRICTION OF CHEMICALS (REACH).....	29
3.8 SDS FORMATTING AND POTENTIAL FOR OPTICAL CHARACTER RECOGNITION AND TEXT PARSING.....	30
3.9 PRECISION OF CALCULATED VALUES.....	31
3.10 HAZARDOUS MATERIALS INFORMATION RESOURE SYSTEM (HMIRS)	32
3.11 HMIRS XML STANDARD.....	32
4 CHAPTER 4 - SYSTEMS ENGINEERING METHODOLOGIES	33

4.1	SDS PROCESSING AS A “SYSTEM”	33
4.2	STAKEHOLDER INFORMATION	33
4.3	NEEDS ANALYSIS	34
4.4	CONCEPT OF OPERATIONS (CONOPS)	35
4.5	PROTOTYPE RESOURCES AND FUNDING	41
4.6	INTELLECTUAL PROPERTY	41
5	CHAPTER 5 - CONCEPTUAL DESIGN	44
5.1	PARALLEL PROCESSING VIA META-ALGORITHMS	44
5.2	NORMALIZED CROSS-CORRELATION	46
5.3	CONVOLUTIONAL NEURAL NETWORK	47
5.4	MACHINE LEARNING KEY-VALUE PATTERN ARRAYS	47
5.5	TESSELLATION AND RECOMBINATION	49
5.6	DATA VALIDATION AND CLEAN UP	49
5.7	RESULTS	50
5.8	DATA ANALYTICS, HAZARD CLASSIFICATION CALCULATION, AND META-ALGORITHMIC VALIDATION	51
5.9	ENTITY RELATIONSHIP DIAGRAM	54
5.10	GRAPHICAL USER INTERFACE DESIGN	55
5.11	REQUIREMENTS	57
5.12	VENDOR CONTRIBUTION QUERY RESULTS	60
5.13	DOCUMENT PROCESSING FLOW	62
5.14	SYSTEM DELIVERABLES	64
6	CHAPTER 6 - SYSTEM OPTIMIZATION	64
6.1	UNIVERSAL SDS DATABASE AND REPOSITORY	64
6.2	Universal SDS Database and Repository	64
6.3	LABEL ADVANCEMENTS	67
6.4	BASIC SDS QUICK RESPONSE (QR) INTEGRATION	67
6.5	QR CODE APPLICATION EXPANSION	68
6.6	RADIO FREQUENCY IDENTIFICATION (RFID) INTEGRATION	71
6.7	ENHANCED COMPUTER VISION USING AUTOMATED NEURAL NETWORK IMAGE PRE-PROCESSING	74
7	FUTURE DEVELOPMENT AND FURTHER SYSTEM INTEGRATION	79
7.1	CURRENT DEVELOPMENT	79
7.2	FULL CHEMICAL DISCLOSURE	79
8	SUMMARY	81

8.1	MILESTONES	81
8.2	CONCLUSION	83
	BIBLIOGRAPHY	85
	APPENDIX A – PATENT APPLICATION	90
	APPENDIX B: CSU AWARD	112

LIST OF FIGURES

Figure 1 2019 TOXIC RELEASE INVENTORY DATA AND LOCATIONS.....	5
Figure 2 CHEMICAL USAGE CALCULATION	6
Figure 3 PYTHON BINARIZATION USING GAUSSIAN ADAPTIVE THRESHOLD	11
Figure 4 GLOBAL THRESHOLD RESULTS FROM PYTHON BINARIZATION.....	11
Figure 5 OTSU BINARIZATION METHOD.....	12
Figure 6 MAXIMA MINIMA BINARIZATION.....	13
Figure 7 PYTHON OPEN CV OCR TEST	14
Figure 8 BINARIZATION & NEURAL NETWORK	16
Figure 9 EXAMPLE OF ML APPLICATION FOR CHEMICAL/PRODUCT SUBSTITUTION	18
Figure 10 GHS PICTOGRAM BINARIZATION	21
Figure 11 GHS SDS SECTIONS.....	28
Figure 12 SDS PHYSICAL & CHEMICAL PROPERTY SECTIONS FORMATTING	31
Figure 13 SDS UPLOAD METHODS	38
Figure 14 PATENT GRAPHIC 1	42
Figure 15 PATENT GRAPHIC 2	43
Figure 16 SDS OCR AND AI PROCESSING SYSTEM	45
Figure 17 NORMALIZED CROSS CORRELATION EQUATION	46
Figure 18 AI SDS META-ALGORITHMIC PROCESSING RESULTS	50
Figure 19 CONCEPTUAL ERD.....	54
Figure 20 GUI DESIGN 1	55
Figure 21 FIGURE 20 GUI DESIGN 2.....	55
Figure 22 GUI DESIGN 3	56
Figure 23 GUI DESIGN 4	56
Figure 24 ADVANCED DEVELOPMENT PHASE	62
Figure 25 CENTRALIZED SDS REPOSITORY	66
Figure 26 EXAMPLES OF VARIOUS EXPANDED QR CODES CAPABLE OF PRODUCT AUTHENTICATION AND SDS RETRIEVAL	69
Figure 27 QR CODE SDS RETRIEVAL AND AI PROCESSING METHODOLOGY.....	70
Figure 28 RFID HAZARDOUS MATERIAL/WASTE TRACKING	73
Figure 29 ENHANCED NEURAL NETWORK PROCESS.....	75
Figure 30 OPTIMIZED PRE-PROCESSING EQUATION.....	76
Figure 31 OPTIMIZED IMAGE	77
Figure 32 PATENT IMAGE 1.....	109
Figure 33 PATENT IMAGE 2.....	110
Figure 34 PATENT IMAGE 3.....	111

LIST OF ACRONYMS

ANN – Artificial Neural Network
CAA – Clean Air Act
CERCLA – Comprehensive Environmental Response, Compensation, and Liability Act
CNN – Convolutional Neural Network
DLA – Defense Logistics Agency
DOD – Department of Defense
DODI – Department of Defense Instruction
EESOH-MIS – Enterprise, Environmental, Safety, and Occupational Health Management Information System
ECHA - European Chemicals Agency
EHS – Environmental, Health, and Safety
EMS – Environmental Management System
EPCRA – Emergency Planning Community Right-to-know Act
EU – European Union
GHS – Globally Harmonized System
HCC – Hazard Characteristic Code
HMIRS – Hazardous Material Information Resource System
MSDS – Material Safety Data Sheet
NFPA - National Fire Protection Association
NSN – National Stock Number
OCONUS - Outside Continental United States
OCR – Optical Character Recognition
OSHA – Occupational Safety, and Health Administration
RCRA – Resource Conservation and Recovery Act
REACH – Registration, Evaluation, Authorisation and restriction of Chemicals
SDS – Safety Data Sheet
USAF – United States Air Force
USAFSAM – United States Air Force School of Aerospace Medicine
XML – Extensible Markup Language

1 CHAPTER 1 - INTRODUCTION AND RESEARCH MOTIVATIONS

1.1 INTRODUCTION

For the past 17 years, I've worked in a variety of capacities providing environmental technological and engineering solutions for the United States Air Force. One of the primary programs I have supported over the years has been the Enterprise, Environmental, Safety, and Occupational Health Management Information System (EESOH-MIS), an enterprise system composed of various modules used to track hazardous materials, hazardous waste, pest management activities, environmental restoration, and other compliance and reporting purposes.

For hazardous materials tracking, the use of SDSs that accompany hazardous material products are vital for determining the chemical composition of products and the environmental, safety, and occupational health concerns these products may pose to Air Force personnel. Prior to use of hazardous materials in the workplace, personnel must receive authorization to use hazardous materials. Air Force installation Environmental, Safety, and Occupational Health personnel review these authorization requests to ensure threats are mitigated and reduced to the greatest extent. One of the key documents required to accurately assess chemical impact to human health or the environment is the Safety Data Sheet. Present operations are complicated in that these documents are currently gathered in many ways, from many sources, and often loaded in duplication across different systems using this information. To add further complexity, these documents differ vastly in formatting and data presentation.

It was due to the redundancy of duplicative SDS loading between the authoritative DoD system and EESOH-MIS that led me to meet Mr. Jonathan Luu, the HMIRS program manager at the United States Air Force School of Aerospace Medicine (USAFSAM) to work together on a

solution to interface our systems. It became apparent that a system solution was needed to formulate one centralized and standard processing method and ensure accuracy throughout the entire chemical lifecycle.

During initial contract award involving the procurement of hazardous materials, vendors are contractually required to provide the SDSs for the products they are selling to the government. This is immediately problematic since materials can go to different organizations and are not centralized through a singular contract procurement system. Some products get shipped directly to the installations using them while the majority, however, get processed through the Defense Logistics Agency (DLA). Once the material is distributed by the DLA out to the installations, all transactions are managed in a supply system which does not store SDS information (immediately severing the direct association of SDS to product). As a result of this, end users must often research the SDSs for the products they are using and send these SDSs off to be loaded in various compliance systems and HMIRS. End users of the chemicals are often not EHS personnel so often times the version of the SDS or the entire product association could be incorrect. All data entry into both HMIRS and EESOH-MIS are then performed manually by data stewards which could be streamlined by upload capability of and XML or equivalent electronic data transfer methods. Process improvement opportunities can be gained with the development of a systematic solution to centralize SDS submittals and establish processes to maintain necessary product-to-SDS associations throughout the product lifecycle.

Lastly, the final opportunity for improvement was finding a solution for the digitized recognition and parsing of data on SDSs provided in PDF format for expedited data load in cases where no XML or equivalent could be provided. Even with an immediate adoption of a direct data transfer or XML solution, hundreds of thousands of legacy SDSs still exist in PDF format and the

necessary policy and contractual modifications required to specific XML or equivalent standards would take years to effectively implement. A digitized data processing solution would be required to break SDSs down into basic components, apply machine learning algorithms to isolate and classify SDS components accordingly.

This singular data processing solution to address each of these operational improvement initiatives would need to be a multifaceted system capable of integration of data from various sources and in various formats and ultimately interface with HMIRS. Research for this dissertation included the various technological solutions for effective data parsing and validation as well as guiding environmental and occupational health compliance drivers, and existing chemical management procedures throughout the DoD.

From the beginning of my dissertation, I wanted a project that would not just result in published research but work that could be used for real-world improvements. In addition to improving accuracy and saving money and time in data processing, this work has the potential to impact the entire industry given the leverage the DoD has as one of (if not the) largest procurer of hazardous materials in the world.

1.2 HAZARD COMMUNICATION BREAKDOWN

Over *2 billion* tons of hazardous and toxic chemicals are manufactured in the United States each year (2019 Toxic Release Inventory National Analysis, 2019). According to the 2019 U.S. Environmental Protection Agency (EPA) Toxic Release Inventory (TRI), U.S. Federal Facilities alone managed a total of 30.7 billion tons of TRI-listed chemicals and production-related waste during 2019 (United Nations General Assembly, 2019). Of the 30.7 billion tons managed, 11%, or 3.38 billion tons, were released into the environment (proportions released to air, water, and land). In addition to environmental concerns, each year in the U.S., thousands of workers become sick

from workplace chemical exposures with as many as 50,000 people dying each year from the adverse effects of long-term chemical exposure. (National Institute of Environmental Health Sciences, 2022) One of the primary factors in both the documentation of toxic releases and in proper assessment of workplace exposure hazards is the vendor-communicated *Safety Data Sheet* information which provides the chemical ingredients, composition information, physical and chemical attributes, and hazard information needed to derive such calculations.

Total Disposal or Other Releases: Manufacturing Sectors

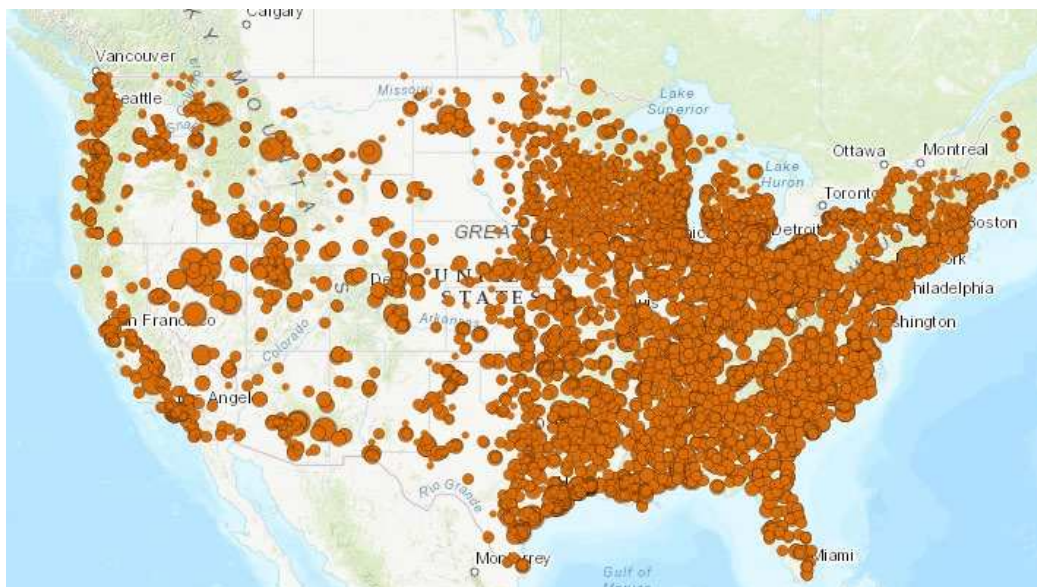
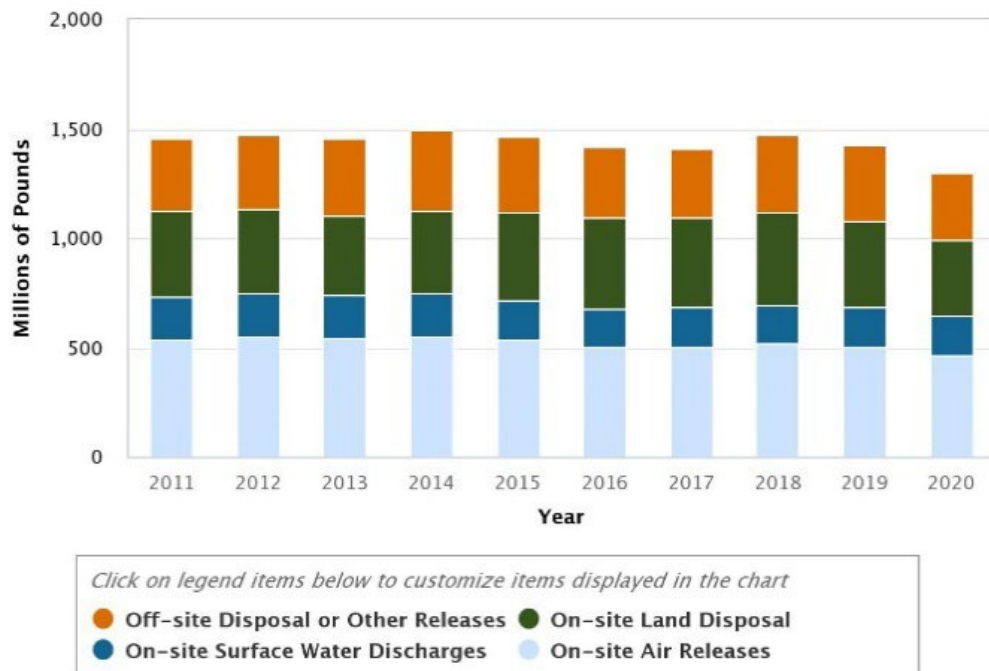


FIGURE 1 2019 TOXIC RELEASE INVENTORY DATA AND LOCATIONS

The current hazard communication system allows for opportunities of missing and/or incomplete data and various errors through the processing and handling of data through the chemical lifecycle. These issues, in turn, lend themselves to other problems including our ability for precise SDS

selection, fast retrieval of SDSs during emergencies, and accurate cradle-to-grave tracking of these hazardous containers. These problems include:

- Missing SDS data from vendors
- Non-GHS compliant SDSs
- SDS data errors
- Manual data entry errors in downstream user databases
- Costs associated with manual entry of data in downstream systems

1.3 LACK OF SDS AVAILABILITY, PRECISION, AND DATA QUALITY

Correct SDSs are often difficult to find due to numerous variations of similar products, variations among manufacturers/vendors/distributors, and hosting of SDSs spread across thousands of chemical manufacturer domains. Due to these inefficiencies, workplace users often receive violations regarding incorrect SDSs, outdated SDSs, and in some cases no SDSs for the chemicals they are using. Perhaps of greater concern is the liability these issues pose in the form of incorrect chemical calculations for environmental compliance and incorrect employee chemical exposure assessments (Glass, 2006) resulting in inappropriate personal protective equipment. For example, chemical usage for a product with a unit of measure of a gallon (Gal) is calculated by:

$$\text{Chemical Lbs. Used} = \sum \text{Gal Used} * \text{Spec Grav} \left(\frac{\text{lb}}{\text{gal}} \right) * \% \text{ Chemical}$$

FIGURE 2 CHEMICAL USAGE CALCULATION
(State of Pennsylvania, 2022)

An incorrect SDS has the potential to alter the ingredient list, chemical composition, and specific gravity (or density); yielding incorrect usage calculations for the total number of product containers associated with that SDS. Liabilities not only exist for management of the products

themselves but also for treatment, cleanup, and disposal. SDS content can be used in support of tort claims in civil litigation suits and in determining where fault lies depending on how the hazardous content was reported on the SDS and how the employer used and managed that information. (Yu, 2020) In the case of *Tolley and Tolley v ACF Industries*, the SDS was used as evidence to support claims that the employer was aware of an isocyanate hazard. (*Tolley and Tolley v. ACF Industries, Inc. et al*, 2002) The term “Safety Data Sheet” resulted in 814 citations in the legal case repository CaseText.com, reflecting the importance of hazard communication in regard to liabilities for manufacturers, employers, and chemical users. SDSs are also used for user-knowledge hazardous waste characterization which could result in incorrect waste treatment or disposal with large-scale contamination concerns of landfills and other processing facilities.

In order to retrieve SDSs for chemical products used in the workplace, chemical users are largely responsible for finding the appropriate SDS that corresponds with each product used. SDSs are commonly housed on manufacturer websites and differ from site to site (including seasonal formulations) with further confusion often introduced as distributors can create their own SDSs, making product matches difficult. Lastly, not all chemicals in product formulations are required by the manufacturer on the SDS. Instead, only those that are deemed as hazardous chemicals by the manufacturer (United States Occupational Safety and Health Administration, 2019). As chemical research progresses and new chemicals are added by the EPA as emerging contaminants (United States Environmental Protection Agency, 2022), liability exists in the now-hazardous chemicals that were previously omitted from older SDSs because they were not deemed hazardous at the time of SDS creation. As a study in 2002 showed, limitations to Material Safety Data Sheets existed since OSHA permitted chemical exclusions when the manufacturer deemed it as non-hazardous or protected as a trade secret (Bernstein, 2002). These limitations hold true today. Per-

and Polyfluoroalkyl Substances (PFAS), for instance, were added under the EPA Toxic Substances Control Act (TSCA) (United States Environmental Protection Agency, 2022) and Toxic Release Inventory (TRI) (United States Environmental Protection Agency, 2022) in 2020. Prior SDS for products containing these chemicals largely omitted these making it now difficult to quantify PFAS inventory and usage in the workforce.

PDF remains the predominant format for the millions of SDSs in circulation today. Downstream users rely on the data from these SDSs for environmental, safety, and occupational health compliance and must typically hand enter this information into their respective compliance systems. Each time manual data entry is performed, potential for data quality errors increases, subsequently increasing the potential for compliance liability. These PDFs often also require validation by SDS data managers to ensure all regulatory fields have been provided and additional follow up with chemical manufacturers for clarification on data fields or requesting additional information for necessary compliance calculations. Direct vendor communication of SDSs using XML or equivalent transfer methods is needed to eliminate the need for separate manual data entry and current outdated communication methods. While some organizations have made significant strides in this area (e.g., SDS-XML, EDASx, SDScomXML), a single universal standard has not been adopted by the Globally Harmonized System (GHS) or the Registration, Evaluation, Authorisation and restriction of Chemicals (REACH) in the European Union.

1.4 LACK OF IMMEDIATE SDS RETRIEVAL ABILITIES

U.S. Occupational Health and Safety Administration (OSHA) listed failures in hazard communication as the second highest most frequently cited standard. (United States Occupational Safety and Health Administration, 2022) 3,624 Hazard Communication (HAZCOM) enforcement citations occurred in fiscal year 2019 totaling approximately \$4,682,380 in proposed penalties

(United States Occupational Health and Safety Administration, 2012). Unfortunately, many current procedures still include maintaining binders of printed SDSs or available copies downloaded on neighboring workstations. These methods are consistently found during inspections to include numerous records that are the incorrect SDS for specific products or are out-of-date, incomplete, and illegible. For EHS systems that require loading of SDS for exposure and environmental reporting calculations, non-EHS personnel are frequently used to find the correct SDS for the products they are using and often not provided sufficient training to determine whether SDSs are GHS compliant and the proper version for the product. OSHA recommends employers “designate a person(s) responsible for maintain SDSs”; it does not reference training for the nuances of precise SDS selection and management. (DeMasi, 2022)

1.5 LOSS IN INVENTORY ACCOUNTABILITY

Lastly, inventory accountability and management are difficult to maintain, as shown in a recent study of chemical storage at research laboratories. (Kuzmina) For hazardous material users, maintaining accurate inventory counts, SDS-inventory associations, proper material segregation, and usage logs can prove an arduous task provided that for many users these are side compliance tasks to their primary duties; often what the materials are being used for. While many of the proposed applications in this paper are not inherently novel, the applications of these in industry, individually and as part of a more robust, integrated system, remain uncommon and their use could help improve or resolve many of the more common EHS violations.

2 CHAPTER 2 - APPLICABLE TECHNOLOGIES REVIEW

2.1 OPTICAL CHARACTER RECOGNITION (OCR)

Optical Character Recognition (OCR) is the process of converting handwritten or printed text into machine-encoded text which can be used to automate the processing of physical documents. While the original concepts of OCR date back over a hundred years ago, OCR technologies had become commonplace in the 1990s and early 2000s as businesses had a need to transform their physical documents into digitized versions. The incorporation of scanners allowed businesses to convert paper documents into PDF or image formats which subsequently led to a desire for auto-recognition of words and sentences to automate the breakdown of text into machine-encoded text which could be stored and made searchable and editable in databases.

Initial OCR technologies largely focused on template-based uses where text coordinates would be entered to identify the specific text location that they wish to recognize. While initial models worked well for well-structured and standardized text, solutions were soon needed to accommodate the vast varieties and formats that documents could come in such as text-image combinations, different/multi languages, unstructured text, handwritten text, tables, etc.

BINARIZATION AND PRE-PROCESSING

Prior to the use of algorithms for classification, images must first be pre-processed to maximize recognition capabilities and provide uniformity in data input. Binarization of the image converts images to black and white (black pixel value = 0, white pixel value = 255) providing a higher contrast to more easily identify characters from their respective backgrounds. A threshold can be used to segment the text from background; typically using a threshold of 127 – half of the allotted

255 pixel range. Common methods of binarization include global thresholding and regional (or local) thresholding.


```

1 import cv2
2 import numpy as np
3
4 import pytesseract
5
6 pytesseract.pytesseract.tesseract_cmd = r"C:\Program Files\Tesseract-OCR\tesseract.exe"
7
8 img = cv2.imread("SDS.JPG")
9 gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
10 adaptive_threshold = cv2.adaptiveThreshold(gray, 255, cv2.ADAPTIVE_THRESH_GAUSSIAN_C, cv2.THRESH_BINARY, 105, 11)
11 text = pytesseract.image_to_string(adaptive_threshold)
12
13 print(text)
14 cv2.imshow("adaptive thresh", adaptive_threshold)
15 cv2.waitKey(0)

```

FIGURE 3 PYTHON BINARIZATION USING GAUSSIAN ADAPTIVE THRESHOLD

SAFETY DATA SHEET



Date of issue/Date of revision : 31 July 2020
Version : 3.01

Section 1. Identification

Product name : CA7255B ACTIVATOR COMPONENT
Product code : CA7255B ACTIVATOR COMPONENT
Other means of identification : Not available.
Product type : Liquid.

Relevant identified uses of the substance or mixture and uses advised against

Product use : Industrial applications.
Use of the substance/ mixture : Hardener.
Uses advised against : Not applicable.

Manufacturer : PPG Aerospace PRC-DeSoto
12780 San Fernando Road
Sylmar, CA 91342
Phone: 818-362-6711
(412) 434-4515 (U.S.)
(514) 645-1320 (Canada)
01-800-50-21-400 (Mexico)




Emergency telephone number

Section 2. Hazards identification


OSHA/HCS status : This material is considered hazardous by the OSHA Hazard Communication Standard (29 CFR 1910.1200).
Classification of the substance or mixture : FLAMMABLE LIQUIDS - Category 2
SKIN IRRITATION - Category 2A
EYE IRRITATION - Category 2A
SKIN SENSITIZATION - Category 1
CARCINOGENICITY - Category 2
SPECIFIC TARGET ORGAN TOXICITY (SINGLE EXPOSURE) (Respiratory tract irritation) - Category 3
Percentage of the mixture consisting of ingredient(s) of unknown acute toxicity: 32.1% (inhalation)

GHS label elements

Hazard pictograms :

SAFETY DATA SHEET



Date of issue/Date of revision : 31 July 2020
Version : 3.01

Section 1. Identification

Product name : CA7255B ACTIVATOR COMPONENT
Product code : CA7255B ACTIVATOR COMPONENT
Other means of identification : Not available.
Product type : Liquid.

Relevant identified uses of the substance or mixture and uses advised against

Product use : Industrial applications.
Use of the substance/ mixture : Hardener.
Uses advised against : Not applicable.

Manufacturer : PPG Aerospace PRC-DeSoto
12780 San Fernando Road
Sylmar, CA 91342
Phone: 818-362-6711
(412) 434-4515 (U.S.)
(514) 645-1320 (Canada)
01-800-50-21-400 (Mexico)




Emergency telephone number

Section 2. Hazards identification

OSHA/HCS status : This material is considered hazardous by the OSHA Hazard Communication Standard (29 CFR 1910.1200).
Classification of the substance or mixture : FLAMMABLE LIQUIDS - Category 2
SKIN IRRITATION - Category 2
EYE IRRITATION - Category 2A
SKIN SENSITIZATION - Category 1
CARCINOGENICITY - Category 2
SPECIFIC TARGET ORGAN TOXICITY (SINGLE EXPOSURE) (Respiratory tract irritation) - Category 3
Percentage of the mixture consisting of ingredient(s) of unknown acute toxicity: 32.1% (inhalation)

GHS label elements

Hazard pictograms :

United States Page: 1/16

FIGURE 4 GLOBAL THRESHOLD RESULTS FROM PYTHON BINARIZATION

Global thresholding is the simplest form of binarization. Global or whole-image thresholding uses a luminosity histogram and assumes a single large peak (e.g., Gaussian distribution) corresponding to the background of the image (typically or white) and less cohesive sets of luminosities separated

by a trough in the histogram (Jyotsna, 2016). This fitting applies a grayscale intensity threshold and sets each pixel to either black or white depending on whether the pixel is closer to the peak or trough in the Gaussian distributed histogram. For global thresholding to be used effectively, the image must possess a bimodal distribution. One of the most common global thresholding methods is Otsu's method. Otsu's binarization method separates pixels into two classes, foreground and background, and chooses the threshold that maximizes between-class variance and minimizes within-class variance (Otsu, 1979). Otsu's binarization method finds threshold value (t) which minimizes the weighted within-class variance.

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t)$$

where

$$\begin{aligned} q_1(t) &= \sum_{i=1}^t P(i) \quad \& \quad q_2(t) = \sum_{i=t+1}^I P(i) \\ \mu_1(t) &= \sum_{i=1}^t \frac{iP(i)}{q_1(t)} \quad \& \quad \mu_2(t) = \sum_{i=t+1}^I \frac{iP(i)}{q_2(t)} \\ \sigma_1^2(t) &= \sum_{i=1}^t [i - \mu_1(t)]^2 \frac{P(i)}{q_1(t)} \quad \& \quad \sigma_2^2(t) = \sum_{i=t+1}^I [i - \mu_2(t)]^2 \frac{P(i)}{q_2(t)} \end{aligned}$$

FIGURE 5 OTSU BINARIZATION METHOD

Another common form of binarization is the adaptive thresholding. Adaptive or local thresholding selects thresholds for each pixel depending on the properties of a localized region within an entire image. For example, on a document that contains many sub-images within the image as a whole, each sub-image could contain a variety of colors making a Gaussian distribution on the entire image impossible. The local maxima and minima method (C(i,j)) applies thresholds to a local part of the image allowing for different threshold values for different parts of the image. One benefit of localized binarization is that it handles uneven lighting and text color variations more

effectively. In some cases, however, the localized thresholding can sometimes accentuate background noise in an image (Milyaey, 2020).

$$C(i, j) = \frac{I_{max} - I_{min}}{I_{max} - I_{min} + \epsilon}$$

I_{max} = maximum pixel value in image; *i_{min}* = minimum pixel value in image; *E*= Constant

FIGURE 6 MAXIMA MINIMA BINARIZATION

Pre-processing of SDSs will be likely minimal in most instances. SDSs are predominantly bimodal in nature with the exception of GHS and other label pictograms displayed in few colors. GHS pictograms are all red, white, and black, and can easily be made binary and classified given there are only nine variations. The vast majority of SDS documents are primarily black text on a white background. A global threshold should prove sufficient in binarization of the document for recognition purposes. There may be some cases of poor-quality scanned documents, however, these cases have grown increasingly rare over the years.

SKEW CORRECTION

Skew analysis is performed using a parallel-architecture approach in which two sets of points are collected and analyzed for their primary set of skew angles. Angles between the sets of points are accumulated in a histogram where the histogram peaks can be smoothed with a moving average filter (He, 2005). Skew angles can be derived from text line angles, image characteristics, or algorithms that perform more dynamic assessments. Upon calculation of skew angles, images are rotated to correct the skew angle.

For SDSs, the vast majority of documents processed will have minimal (if any) skew corrections necessary. SDSs are compiled largely using common operating system document software packages and will come aligned. The exception will be any poorly scanned copies, which again, are becoming increasingly uncommon.

SEGMENTATION AND FEATURE EXTRACTION

Input can be segmented by line separation in which each line is isolated in the input image. Once the lines are isolated, then each character is isolated within each line. Additionally, characters can also be segmented by analysis of pixels along the character boundaries where pixels = '0'. Once segmented, each character can be normalized (e.g., focus, size) and compared to character templates used for ANN training. Character recognition success is directly correlated with the quality of the image. Higher quality images will yield more precise segmentation and character recognition which will subsequently have a direct impact on the success of ANN categorization.

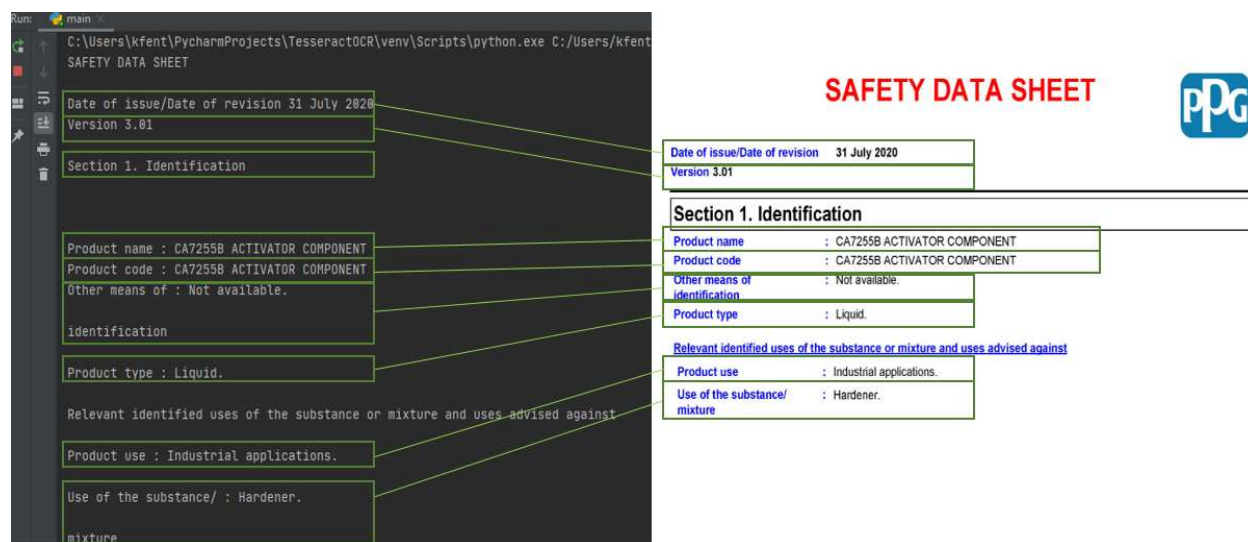


FIGURE 7 PYTHON OPEN CV OCR TEST

LANGUAGE TRANSLATION

Chemical users can purchase and procure chemicals from vendors manufacturing and distributing around the globe and providing SDSs in many languages. For workplaces with multi-lingual personnel, common safety and occupational health requirements often specify the need to maintain

SDSs in the languages spoken at each respective site. Translation services can be costly and often introduce new potential data discrepancies during translation. OCR translation functionality has become increasingly more accurate and reliable and often come prepackaged with many OCR tools. The OCR engine will need to be configured to the possible languages it could be receiving (or through a translation API) bearing in mind that the more languages selected, the slower the processing.

2.2 ARTIFICIAL INTELLIGENCE

ARTIFICIAL NEURAL NETWORKS

The emergence of artificial neural networks and machine learning algorithms allowed developers to incorporate programming that allowed decisions to be made on specific OCR complexities similar to how the human brain would process the information in a sensible fashion. Not only would machine learning allow for comprehension and validation of information but could also be used in meta-algorithmic approaches for mistake-checking to increase the accuracy and precision of the OCR engines. Artificial Neural Networks (ANN) use topological features such as shape and symmetry, boundaries, number of pixels, and pixel assignment to identify and recognize characters and images. One of the primary benefits of an ANN is that it can be trained based upon samples and use these for classification. Neural networks could be used to classify information, using probability calculations and regression models to ascertain decisions, predictions, and degrees of certainty of how well the data classifies. If decisions had not yet been made, feedback loops could be incorporated to allow for additional processing until the degree of certainty has reached a sufficient level.

A common approach to ANN is the use of Feature Vectors to segregate and assign numerical values to various features or properties. Vectors are the sets of numerical values for character identification ultimately used to train a system based upon unique properties. Features are chosen to define characters similar to how the human brain would process the image based upon properties that distinguish these characters. Vector generation is dependent upon calculations and features sufficiently diverse to increase precision. Features are any property of the image that can be used to identify the character, such as Curves, Closed Areas, Horizontal and Vertical lines, Symmetry, Contours, and Projections. (Evelina Maria De Almeida Neves, 1997). Larger amounts of feature sets yield increased recognition precision.

Neural networks and machine learning algorithms will be valuable for the recognition and processing of SDSs given the flexibility that chemical manufacturers and distributors have with how the information can be displayed on an SDS. One large benefit from the GHS standardization efforts is that there are now sixteen distinct sections that are required on an SDS so we can isolate specific information that is required and we can expect on each SDS.

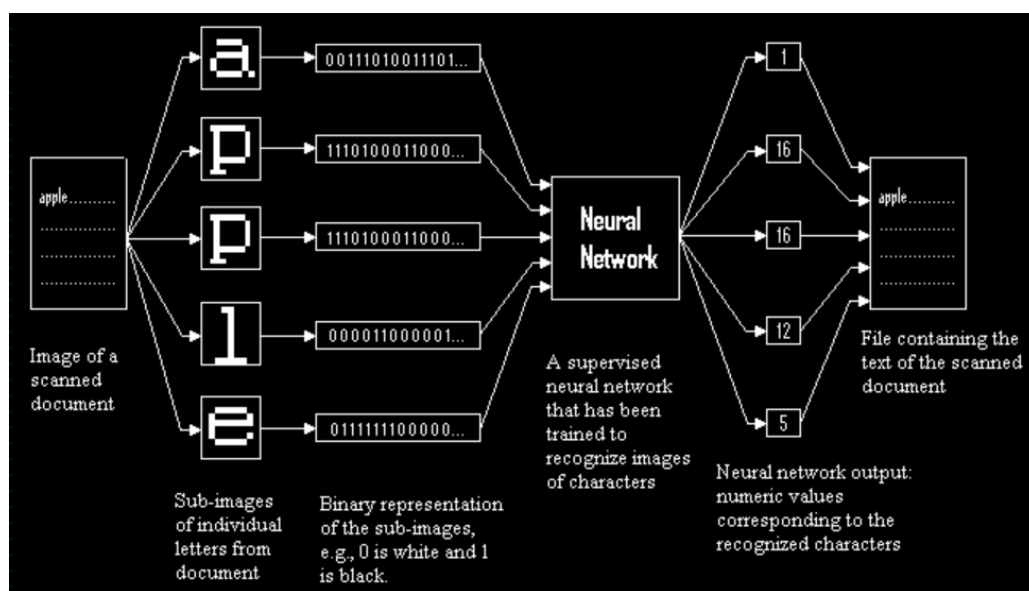


FIGURE 8 BINARIZATION & NEURAL NETWORK
(Neuron Synthesizer ANN, 2019)

SYSTEM TRAINING

Prior to classification, ANNs require training in order to develop machine learning capabilities and map various inputs effectively to their respective outputs. Data is processed by the system through the input layer. Predictions and decisions are then made based upon the methodology chosen for classification. In the example above of the binarization of an SDS flammable pictogram, a sum has been calculated of the positive binary values. Although simplistic, this could act as a feature that can be used for categorization.

Prior to training, however, a critical step is to determine the number of hidden layers containing a number of neurons that should be used in the neural network. This is generally determined largely on the complexity of the task with higher scientific complexity requiring greater number of hidden layers and neurons although care must be taken not to over-train the system to where the benefits become marginalized given the decrease in processing efficiency. For classification tasks like the classification of an SDS pictogram to the correct pictogram class, if the predicted label matches the observed label (or within an appropriate level of confidence) then the prediction is proven correct. Measures of correctness can then be obtained depending on how well the classification has occurred. Weights can then be applied to the level of confidence in a specific feature in correctly classifying the object. With each iteration, the machine can then begin to learn as weights are modified based on the results of previous classification attempts.

MACHINE LEARNING

Machine learning is defined as the process of computers changing the way they carry out tasks by learning from new data, without a human needing to give instructions in the form of a program (Dictionary, 2023). As humans, our brain acts as our model and has been trained over the course of our lives through our life experiences, education, reading, and sensory information we've

absorbed. With computers this is accomplished through large data sets that the computer can use to find patterns or make decisions on other input data. In image recognition, for instance, image recognizers are fed vast data sets of classified images that can be used to determine similarities among new input data and infer a level of confidence that the subjects within the images are the same. Three of the most common forms of machine learning are supervised learning, unsupervised learning, and reinforcement learning.

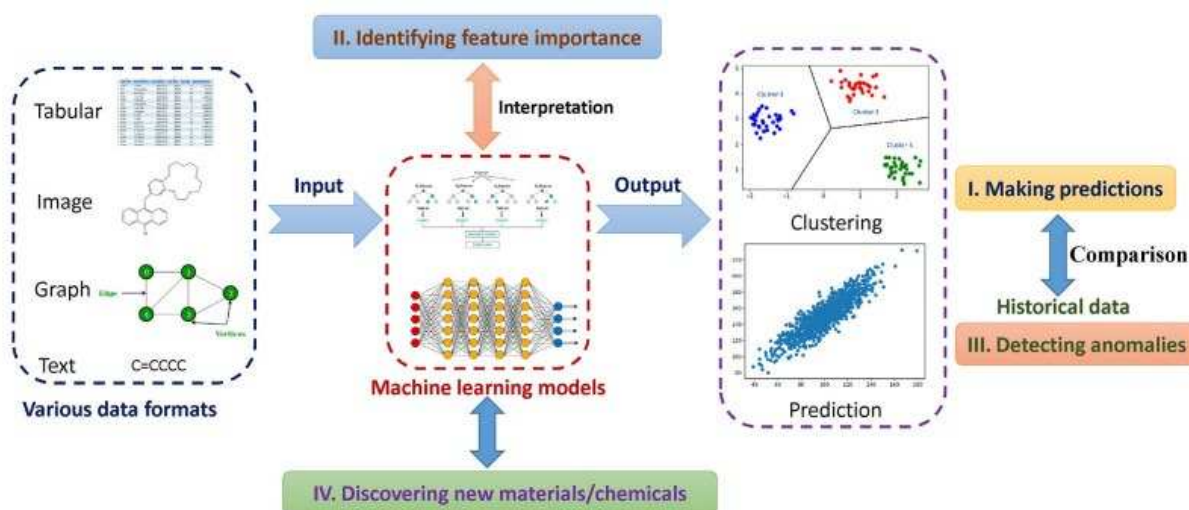


FIGURE 9 EXAMPLE OF ML APPLICATION FOR CHEMICAL/PRODUCT SUBSTITUTION
(Zhong, 2021)

SUPERVISED LEARNING

Supervised learning is defined as a learning method where the algorithm generates a function that maps inputs to desired outputs and is the most common form of training neural networks and decision trees (Ayodele, 2010). For SDSs, one example of a supervised learning method would be the classification of a GHS pictogram. The model used essentially maps the input (SDS pictogram) to the respective output (pre-classified GHS pictograms). Algorithms for supervised

learning include linear regression, logistic regression, support vector machine, k-nearest neighbor, decision tree, random forest, and naïve bayes.

UNSUPERVISED LEARNING

Unsupervised learning differs in that the training data is unlabeled and there is no fixed output variable. The model instead learns patterns and identifies specific features that can be used to determine the likelihood of a probable output. One example in SDSs would be the syntax mapping of chemical names or formulas. While not necessarily explicitly labeled, the syntax from chemical formulations and nomenclature differs from other verbiage on an SDS and could grouped with a clustering algorithm. Common algorithms used include clustering algorithms such as k-means for deducing association through clusters.

REINFORCEMENT LEARNING

Reinforcement learning approaches learning with no predefined target and uses actions and rewards to approach potentially numerous paths towards an end state. With defined rules, the algorithm explores different options and possibilities, monitoring and evaluating each result to determine the optimal approach (Wakefield, 2023).

2.3 “NATURAL LANGUAGE PROCESSING” (NLP)

Natural Language Processing (NLP) is a branch of artificial intelligence that combines computational linguistics with statistical, deep learning, and machine learning models. NLP allows computers to process text and voice recognition using a rule-based model of human language and understand the meaning, intent, and sediment of the language expressed. Regarding SDSs, this allows us to transform SDSs into a “living” document capable of receiving and interpreting feeds

on changes to regulatory chemical listings, product recalls, product newsfeeds, and alternative products. Incorporation of NLP into a centralized SDS repository would offer significant enhancement and auxiliary services dramatically expanding on HAZCOM to chemical end users. The consistent monitoring and tracking of chemical updates in suitable research publications, EPA publications, and manufacturer recalls are largely infeasible given the tens if not hundreds of thousands of chemicals large organizations can potentially manage. NLP allows for precision chemical literature curation and can provide recommendation reviews and prioritized classification of these feeds for human review. SDS data points, such as Chemical Abstract Service (CAS) numbers, chemical names, product trade names, and part numbers, would be classified and then run through a named entity recognizer before external document analysis and recommendations would be researched. Given the numerous synonyms for chemicals, an additional semantic dictionary mapper would be needed to replace synonym references to the primary chemical name or CAS. References, in turn, would be run through document analysis and recognition process to identify potential items of interest but also perform filtration on irrelevant terminology and reduce inflectional forms of these key terms (Sharma, et al., 2020). Vector space models would then be used for identification of SDS term frequency and inverse document frequency for term importance analysis and pairwise cosine similarity between each document (Salton & McGill, 1986) (Salton, Wong, & Yang, A vector space model for automatic indexing, 1975). If pairwise frequency thresholds are sufficient, the document can become a candidate for recommendation. Lastly, SDS data stewards would have a final review and relevant feedback could then be propagated back into the system for machine learning and relevance weighting and relevant NLP feedback tied to the chemical of interest for knowledge distribution.

2.4 COMPUTER VISION

Computer vision is a field of artificial intelligence that enables computers and systems to derive meaningful information from digital images, videos, and other visual inputs – and take actions or make recommendations based on that information (IBM, 2023). Algorithms used in this process largely rely on pattern recognition based upon large data sets used for training (Babich, 2020). In terms of SDSs, computer vision provides image recognition capabilities allowing any embedded images within an SDS to be recognized and classified. Computer vision can be used for GHS pictogram classification, NFPA pictogram classification, etc. This functionality also serves for alternative uses as any images embedded in documents can also be recognized and classified accordingly.

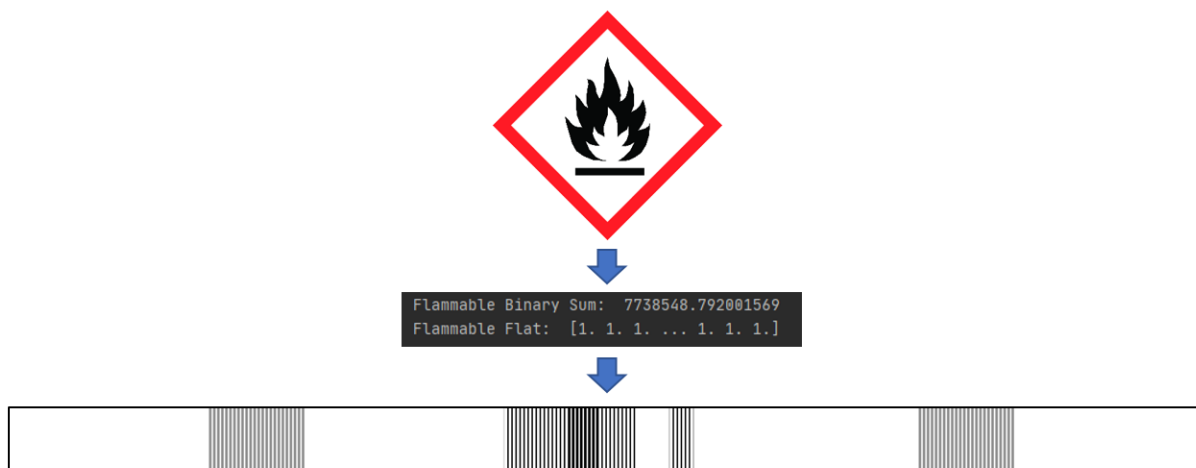


FIGURE 10 GHS PICTOGRAM BINARIZATION

While the DoD data processing system is being developed specific to SDS documents and requirements, ultimately the meta-algorithmic system itself is essentially an advanced document decomposition and data processing tool that can just as easily be trained using different models for

different document types altogether. Computer vision allows the system to find images, interpret patterns among the pixels and feature characteristics, and classify objects within those images.

3 CHAPTER 3 - APPLICABLE REGULATORY DRIVERS REVIEW

Essential for system conceptual development, it's critical to know not only what information an SDS provides but also how the data is ultimately used and the value of this data to downstream users. SDS data maintained in HMIRS is used to meet compliance and reporting needs for the regulatory requirements listed below. Hazardous material product chemical constituent data provides the chemical ingredients and composition information needed for usage and inventory calculations. For hazardous waste compliance, waste characterization based on user knowledge uses SDSs to appropriately classify waste to determine appropriate treatment or disposal measures. For spills, unintentional releases, and emergency response, SDSs provide an immediate source of chemical, hazard, and response information for mitigation and response measures. For occupational health assessments, SDSs provide exposure information needed to assist occupational health personnel in providing adequate personal protective equipment (PPE) and minimizing personnel exposure and hazards in the workplace.

3.1 EMERGENCY PLANNING AND COMMUNITY RIGHT-TO-KNOW ACT (EPCRA)

The primary uses of HMIRS data are for environmental, safety, and occupational health regulatory compliance. In the United States, the Emergency Planning and Community Right-to-Know Act (EPCRA) exists to provide communities information for emergency response and to help government organizations protect these communities from potential risks. EPCRA was implemented in the United States after an accidental release of methylisocyanate killed over 2,000

people in Bhopal, India. To reduce the chances of similar accidents occurring, EPCRA was designed to increase the public's knowledge of chemicals used at facilities and releases into the environment.

Reportable sections of EPCRA applicable to DoD installations include sections 301 to 303 in which information needs to be reported of what chemicals and quantities are stored for emergency response procedures. Sections 311 and 312 provide chemical inventory balances and quantities of chemicals stored at facilities along with SDSs associated with these products to state/local government officials and fire departments. Section 313 is used to report chemicals that have exceeded regulated thresholds annually. For chemical usage and inventory reporting requirements, SDSs provide product chemical ingredients, % of ingredients, densities, specific gravities, Volatile Organic Compounds (VOC) %, Hazardous Air Pollutants (HAPs), and other fields vital for calculations. For chemical usage calculations, liquid and gaseous volumes must first be converted to weight-based units (lb or kg) using the density or specific gravity (and whether the SG is relative to air or water) from the SDS. Weight-based product totals are multiplied by the ingredient percentages from the SDS to determine totals by chemical. Chemical Abstract Service (CAS) numbers and chemical names from the SDS are used to classify chemicals, determine approval and authorization for use, identify conditions of use, and restrictions.

3.2 COMPREHENSIVE ENVIRONMENTAL RESPONSE, COMPENSATION, AND LIABILITY ACT (CERCLA)

The Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA) was created in the United States to ensure proper management of hazardous materials releases. Informally known as Superfund, CERCLA was in response to toxic waste dumps such as in Love Canal, NY, which received national attention from contamination of a school and local community

due to mismanaged and unreported hazardous material releases. CERCLA holds facilities responsible for cleanup of these toxic releases. It requires the immediate reporting of accidental hazardous substance releases into the environment in quantities greater than reportable quantity thresholds set by the EPA. If a product is spilled, SDSs are often used for immediate spill response information (section 6 of the SDS).

3.3 RESOURCE CONSERVATION AND RECOVERY ACT (RCRA)

The Resource Conservation and Recovery Act (RCRA) was created to provide framework and guidance on proper management of hazardous and non-hazardous solid waste. Also created over historic mismanagement of hazardous material disposal, RCRA provides regulatory guidance on the management of hazardous and non-hazardous solid waste at facilities and ensures proper treatment or disposal. The EPA provides the national regulatory framework for solid waste control and the states can provide further restrictions of management rules if needed. In hazardous waste management, SDSs are used to characterize waste using “User Knowledge”, relying on SDSs to determine composition of waste streams. Hazard identification is listed in section 2 and Physical/Chemical properties are listed in section 9 of the SDS. Transportation information used for hazardous waste shipping manifests documentation can be found in section 14 of the SDS.

3.4 CLEAN AIR ACT (CAA)

The Clean Air Act was created to regulate air emissions from stationary and mobile sources, regulate emissions from hazardous air pollutants, and established national standards to protect public health and welfare. Section 112 of the Clean Air Act establishes hazardous air pollutant emission standards and requires reduction in emissions of these pollutants. Clean Air Act reporting relies on chemical ingredients, ingredient %s for usage calculations, and other SDS air-specific

data fields (e.g., HAP, VOC, Vapor Pressure, Vapor Density, etc.) primarily from section 9 of the SDS.

3.5 OCCUPATIONAL SAFETY AND HEALTH ADMINISTRATION HAZARD COMMUNICATION (OSHA HAZCOM)

The Occupational Safety and Health Administration created the Hazard Communication Standard to protect workers through the dissemination of information regarding chemicals and toxic substance hazards and the personal protective equipment and controls that should be used to reduce exposure to these hazards to the greatest extent possible. Information dissemination is largely done through the distribution of SDSs and hazardous labels which convey hazards to downstream users in a standardized fashion. General industry toxic and hazardous substance regulations are listed in Code of Federal Regulations 29 CFR 1910 subpart Z.

3.6 GLOBALLY HARMONIZED SYSTEM (GHS)

As of 2005, global chemical business transactions accounted for approximately \$1.7 trillion worldwide and more than \$450 billion in the United States¹; the largest procurer of which is the Department of Defense. The Globally Harmonized System (GHS) was created as a result of a United Nations international mandate adopted in 1992. Vast differences in how different countries labeled chemicals and documented hazards was resulting in difficulties for chemical manufacturers to abide by the multitude of chemical hazard communication standards and also the ability for end users to assess environmental and human health exposure threats due to these differences. GHS provides an international approach and framework to the classification and labeling of chemicals to improve environmental and human health and safety protective measures.

Implementation in the United States effectively began on December 1, 2013, as the initial implementation phase required employers to train employees of new GHS label elements and SDS formatting changes. On June 1, 2015, all chemical manufacturers, importers, distributors, and employers were required to comply with all modified provisions of the GHS final rule with an exception granted to distributors with legacy inventory that had been labeled under the previous system. Final legacy inventory was required to become compliant by December 1, 2015. Lastly, on June 1, 2016, employers were required to update alternative workplace labeling and hazard communication and provide additional training to employees on newly identified physical or health hazards.

While not a regulation or standard, GHS establishes a template for hazard classification system and communication provisions with directions on its application into each country's own regulatory process. Its primary objective is the identification of intrinsic chemical substance hazards and mixtures and to convey the hazard information in an organized fashion. Countries that have chosen to adopt GHS have agreed to its criteria and provisions and incorporate into their own host country requirements. The uniformity that GHS provides helps address issues with inconsistent documentation of hazards and protective measures for individuals using chemicals due to varying chemical classification methodologies, labeling differences, and numerous differences on how the chemical information is relayed to downstream users via the Safety Data Sheets.

GHS SAFETY DATA SHEETS

One of the important measures associated with GHS implementation is the formatting and information requirements on the SDS. The SDS (formerly known as MSDS) provides comprehensive information associated with chemical use in the workplace. The SDS provides information vital for appropriate environmental and human exposure assessment including the ingredients, fire and emergency response, transportation, ecological concerns, and more to provide a clear description of the hazards to handlers of the product. Each GHS-compliant SDS contains 16 sections which are similar to ISO, EU, and ANSI requirements (only that sections 2 and 3 are reversed). While the sections must be numbered and identified according to the GHS requirements, the general layouts and formatting may differ from SDS to SDS.

Section	Title	Compliance	HMIRS Requirements
1	Identification		
2	Hazard(s) Identification	Section 3 in EU	
3	Composition/Information on Ingredients	Section 2 in EU	
4	First-Aid Measures		Not required in HMIRS
5	Fire-Fighting Measures		Not required in HMIRS
6	Accidental Release Measures		Not required in HMIRS
7	Handling and Storage		Not required in HMIRS
8	Exposure Controls/Personal Protection		Not required in HMIRS
9	Physical and Chemical Properties		
10	Stability and Reactivity		Not required in HMIRS
11	Toxicological Information		Not required in HMIRS
12	Ecological Information	Non-mandatory by OSHA	Not required in HMIRS
13	Disposal Considerations	Non-mandatory by OSHA	Not required in HMIRS
14	Transport Information	Non-mandatory by OSHA	
15	Regulatory Information	Non-mandatory	
16	Other Information		Not required in HMIRS

FIGURE 11 GHS SDS SECTIONS

3.7 EUROPEAN UNION REGISTRATION, EVALUATION, AUTHORISATION, AND RESTRICTION OF CHEMICALS (REACH)

The Registration, Evaluation, Authorisation, and restriction of Chemicals (REACH) was legislation introduced in 2001, adopted in 2003, and that came into effect in 2008 as companies began the material registration process. The purpose of REACH was to provide standardization of hazard communication for hazardous products used throughout Europe. A new agency, the European Chemicals Agency (ECHA) was created to provide oversight and daily management of the REACH legislation. Upon implementation, REACH required all hazardous material manufacturers to register specific identified chemical substances over the course of 11 years.

Upon registration, the ECHA evaluates the products and their chemical constituents to determine the potential harm to human health and the environment. The new legislation puts more of the workload on the manufacturers to prove that they have performed due diligence to ensure the safety of their products and have documented environmental and health concerns. Each manufacturer must perform an analysis of potential alternatives or substitution where possible prior to authorization by the ECHA.

Prior to REACH, extremely hazardous substances had been manufactured in large quantities and, in many cases, had not been accompanied with sufficient information to conduct adequate health and hazard assessments. The legislation filled many of these gaps and has encouraged manufacturers to consider safer alternatives. Any substance manufactured or imported in quantities of 1 metric ton or more per year must be registered with the ECHA and have registration dossiers completed identifying risks associated to the substances and how these risks will be managed. “Without registration, substances cannot be manufactured or imported into the EU”.

Substances of very high concern are initially identified on a “Candidate List” and, if no authorization is given, placed on the “Annex XIV” list which states that these substances cannot

be placed on the market or used after a given “sunset date”. The DoD continuously monitors the usage of all chemicals on both the “Candidate” and “Annex XIV” lists to ensure these chemicals are phased out of use appropriately. Substances deemed as highly hazardous by the ECHA can also receive restrictions if they pose an unacceptable risk to human health or the environment. Additionally, substances can be banned if the risks are too great and cannot be effectively mitigated.

3.8 SDS FORMATTING AND POTENTIAL FOR OPTICAL CHARACTER RECOGNITION AND TEXT PARSING

Electronic SDSs provided to the HMIRS office presently come in PDF format. A benefit from the adoption of GHS guidelines is that the sections of the SDSs have now been standardized. Each GHS-compliant SDS is made up of sixteen required sections; four of which are not mandatory in the United States but are required in some of our host nations for overseas installations. The majority of vendor SDSs are optimal for OCR providing largely bimodal text and images, commonly white background with black text.

How the information is displayed (e.g., paragraphs, tables, bullets, etc.) varies from SDS to SDS. Coordinates cannot be assigned to text fields because no standardization exists for margins and text formatting. DoD Outside Continental United States (OCONUS) regulations require workplaces with foreign personnel to maintain SDSs in their host language so SDSs that are both multi-language and in different languages will also need to be processed. SDSs can contain combinations of text and images such as hazardous classification and National Fire Protection Association (NFPA) labels. SDS physical and chemical properties can also be listed in both text and numerical values making derived calculations more difficult.

3.9 PRECISION OF CALCULATED VALUES

Downstream users of HMIRS use the SDS calculated values such as ingredient percentages, densities, specific gravities, pH values, vapor pressures, flash points, etc. for both hazard assessment and chemical usage calculations. Miscalculated values used for regulatory compliance and reporting purposes has the potential to open the DoD to inaccurate reporting, liability concerns and strict penalties. Something as simple as specific gravity decimal misplacement could significantly alter usage calculations exponentially. Additionally, extrapolation as to whether the specific gravity is relevant to air or water is vital for differentiating calculations based upon liquid or gaseous values.

Validation metrics will be required to ensure values are within expected thresholds, e.g., pH values should be between 0 and 14, percentages should not exceed 100%, etc. Additionally, excessive outliers should be documented and flagged for additional validation. OCR functionality will need the ability to extract numerical values and units for numerous physical and chemical properties presented in a variety of formats including tables and free text and validate to ensure the values are within expected ranges. The image below depicts some of the complications associated with physical and chemical values from SDSs. This same data may exist in tables, bullets, paragraph form, etc. Physical and chemical value cells contain numerical values, ranges, text, combinations of values and additional text, units, and special characters.

Section 9. Physical and chemical properties	
Boiling point	: >37.78°C (>100°F)
Flash point	: Closed cup: -4.44°C (24°F)
Auto-ignition temperature	: Not available.
Decomposition temperature	: Not available.
Flammability (solid, gas)	: Not available.
Lower and upper explosive (flammable) limits	: Not available.
Evaporation rate	: Not available.
Vapor pressure	: Not available.
Vapor density	: Not available.
Relative density	: 1.2
Density (lbs / gal)	: 10.01
Solubility	: Insoluble in the following materials: cold water.
Partition coefficient: n-octanol/water	: Not available.
Viscosity	: Kinematic (40°C (104°F)): >0.21 cm ² /s (>21 cSt)
VOC	: 258 g/l

SECTION 9 – PHYSICAL AND CHEMICAL PROPERTIES		
Properties Listed Below are for Electrolyte:		
Boiling Point:	210 - 245° F	Specific Gravity (H2O = 1): 1.215 to 1.320
Melting Point:	N/A	Vapor Pressure (mm Hg): 10
Solubility in Water:	100%	Vapor Density (AIR = 1): Greater than 1
Evaporation Rate: (Butyl Acetate = 1)	Less than 1	% Volatile by Weight: N/A
pH:	~1 to 2	Flash Point: Below room temperature (as hydrogen gas)
LEL (Lower Explosive Limit)	4.1% (Hydrogen)	UEL (Upper Explosive Limit) 74.2% (Hydrogen)
Appearance and Odor: Manufactured article; no apparent odor. Electrolyte is a clear liquid with a sharp, penetrating, pungent odor.		

FIGURE 12 SDS PHYSICAL & CHEMICAL PROPERTY SECTIONS FORMATTING

3.10 HAZARDOUS MATERIALS INFORMATION RESOURE SYSTEM (HMIRS)

HMIRS was developed by DLA to comply with the Department of Labor's OSHA requirements for SDS listed in 29 CFR 1910.1200. Department of Defense Instruction (DODI) 6050.05 specifies HMIRS as the official DoD repository for SDSs, providing DoD personnel with readily available information regarding hazardous material transportation data, hazard warning labels, occupational health, safety, and environmental data. Hazard Characteristic Codes (HCC) are provided in HMIRS for the identification and classification of items that DoD services wish to track as hazardous materials and provides useful information regarding the storage segregation and compatibility requirements of these materials. Typically, SDSs are loaded into HMIRS upon receipt at a designated Defense Logistics Agency receipt and distribution point. HCCs are assigned by HMIRS environmental, safety, and occupational health professionals as they review the SDSs for completeness, technical accuracy, consistency, and load the SDSs. GHS labels are also loaded into HMIRS in the event that the chemical warning labels are lost or rendered illegible.

3.11 HMIRS XML STANDARD

An XML standard was finalized in 2021 with the intent of inevitably requiring all DoD vendors to provide SDSs according to the new standard. Future hazardous material contracts would need to be appended to include verbiage specifying the preferred vendor SDS submittal method complying with the XML standard. While it will likely take many years for the XML standard to be incorporated by the majority of DoD chemical vendors, the proposed optical character recognition process will be the primary intermediary approach until finalization and provide a means of transitioning vendors towards a more centralized approach.

The XML standard will provide the framework for data import into HMIRS. Regardless of SDS submittal method into the centralized repository – XML, OCR-derived XML, or manual, standardized validation algorithms will be used to ensure that all are consistent and are compliant with regulatory and XML standards.

4 CHAPTER 4 - SYSTEMS ENGINEERING METHODOLOGIES

4.1 SDS PROCESSING AS A “SYSTEM”

One common definition of a system is “a set of interrelated components working together toward some common objective.” (Kosaaikoff, 2003). Systems engineering methodologies will be used for conceptual development and design phases to result in a proof-of-concept prototype. Systems engineering focuses on complex systems, converging interdisciplinary engineering components and governing the interfaces of these components through design and development to ensure desired objectives are met and system as a whole succeeds. This research focuses on the integration of these interrelated components and application of systems engineering procedures to not only stakeholder needs and requirements but achieve these objectives through the design and development of an efficient data processing system with the potential to generate significant efficiencies over current processes and significant savings in streamlining and optimization.

4.2 STAKEHOLDER INFORMATION

The following lists the primary stakeholders that are directly involved with the loading of SDS chemical data or use the data in their daily operations.

- Aerospace Industries Association
- Chemical Vendors/Distributors
- Defense Health Agency

- Defense Logistics Agency Aviation
- Defense Logistics Agency J4/J6
- Department of Veterans Affairs
- Hazardous Material Information Resource System Program Office
- Headquarters Air Force
- United States Air Force Material Command
- United States Air Force School of Aerospace Medicine
- United States Army Public Health Command
- United States Coast Guard
- United States Marine Corps
- United States Navy Public Health
- United States Navy Supply Systems Command

4.3 NEEDS ANALYSIS

Optical Character Recognition (OCR) and Computer Vision

- Automated optical character recognition capability to recognize and parse text from SDSs to increase the speed, accuracy, and efficiency associated with SDS data entry.
- Ability to classify sections of SDSs provided with varying layouts and formats into standardized data points.
- Ability to identify and classify images and map top corresponding HMIRS label pictograms.
- Ability to translate SDSs provide in multi-languages or various languages used at DoD overseas installations.

Data Mapping

- Provide recommended hazard characteristic code based upon physical and chemical properties.
- Format extracted text to align with the HMIRS XML standard for common HMIRS interface upload ability.

Metrics/Analytics

- Performance metrics to identify level of confidence in accuracy of parsed and classified text and discrepancies from data points with low confidence thresholds.
- Validation analytics to ensure chemical and physical numerical data values are within expected ranges.

User Interface

- Vendor User interface to allow vendors three submittal methods – populated standardized XML files, SDSs to be processed with OCR, and a form to allow vendors to build new SDSs and submit.
- Results of all three submittal methods will conform with the same XML standardized format.
- Data Steward User interface for data stewards to accept inbound XML files; both OCR and standardized XML submittals
- Data Steward User interface to validate data population and discrepancies prior to accepting into HMIRS.

4.4 CONCEPT OF OPERATIONS (CONOPS)

CURRENT STATE

Currently, vendors are required to provide SDSs for all hazardous materials they are contracted to provide upon contract award. SDSs are provided in PDF format in bundles to the Defense Logistics Agency (DLA) upon shipment. DLA HMIRS technicians then proceed to check to see if the SDS already exists in HMIRS and, if not, manually enter values from the SDS into HMIRS and upload the associated SDS PDF. Additional hazardous material procurement exists from local purchase as well as other procurement methods (e.g., contracted services). Downstream users can purchase hazardous material online or on the local economy using Government Purchase Cards. The downstream users are then expected to check their environmental systems to see if the SDS has already been loaded (e.g., EESOH-MIS for US Air Force) or provide the SDS to the environmental system's data stewards for loading (which then gets forwarded to the HMIRS office for loading).

SDS DATA ENTRY

The HMIRS data fields are broken up in a series of tabs representing the 16 sections on an SDS along with additional tabs for DoD-specific information as well as a tab to upload supporting documentation (e.g., additional languages, supporting technical data sheets, manufacturer specifications, etc.). HMIRS personnel are required to enter the 41 mandatory fields listed in Appendix B along with various other fields that are contingent upon the type of material and the physical and chemical properties. While the entry of all SDS data is desirable, unfortunately limited manpower resources have reduced data entry to sections 1, 2, 3, 9, 14 and 15 of the SDSs. Once the required fields and any conditional fields applicable to the material being loaded have been entered, the data steward will also associate the label type to the material, associate a hazard characteristic code (HCC) and upload the PDF of the SDS along with any supporting documentation.

PROCESS INEFFICIENCIES

The proposed system has the potential to address multiple current process inefficiencies ranging from initial SDS receipts to downstream user SDS requests and breaks in the process flow requiring these users to research the SDSs that match their products.

- Vendors provide SDSs in predominantly hard copy format requiring teams of personnel to perform manual data entry.
- No automated standard currently exists to provide these in a machine-encoded text.
- Not all SDS information can be loaded due to resource limitations for manual data entry.
- Contract association with the product are lost as materials are distributed requiring re-association of product to SDS downstream.
- Lack of automated standard currently limits direct interface or data transfer with downstream user systems.
- Small vendors may not have resources to comply with an XML standard.
- Hundreds of thousands of SDSs still exist in PDF or hard copy format .

CONCEPTUAL STATE

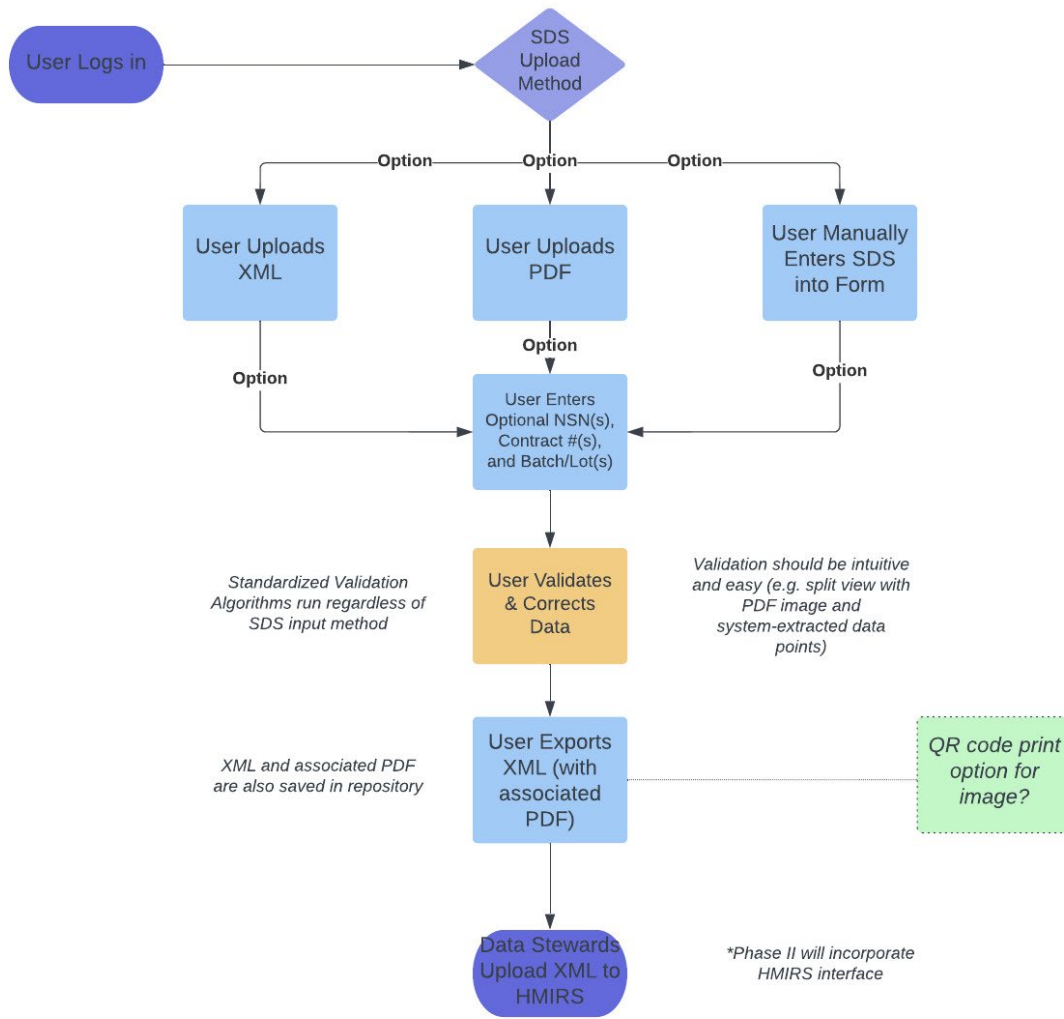


FIGURE 13 SDS UPLOAD METHODS

Primary SDS Submittal option - Vendor Contract Procurement (Preferred Method)

With an addition to FAR 23.3 and DFARS, Subpart 223.3 section G, language would state that preferred SDS submittal method by vendors would be in XML format. Submissions of XML files (standardized by DLA) would allow for the most seamless transfer of data from vendor to HMIRS. This process would be the preferred submittal option providing minimal opportunities for human error factors. Embedded analytics in HMIRS could be used to validate that numerical values are

within expected thresholds, and at minimum, all required fields had come across correctly. Data steward workload would be reduced significantly from manually loading all SDSs to an automated process where they can quickly review data transferred in bulk. In addition to SDS data, valid contractual data would be provided via XML as well allowing product-SDS associations to be used by downstream user systems. USAFSAM has a contract with DLA for the development of an XML standard and upload ability within HMIRS. Preliminary standard and development completion are expected in early 2021.

Second Alternative SDS Submittal Option - OCR

Small vendors may not have the resources necessary to modify in-house databases to allow for an XML SDS export. Additionally, for large vendors, it may take years for contracts to be updated through attrition and have new language incorporated specifying the preferred submittal method. Lastly, hundreds of thousands of legacy SDSs only created in PDF format could still potentially come through for future processing. The OCR option would be the secondary submittal option. A front-end public-facing user interface would be used by vendors to submit all contractually required SDSs. The OCR system would read the SDS, provide character and image recognition, classify SDS sections as appropriate, map label images to GHS label types, and export the file in a machine-encoded XML format. The XML format will be based upon the DLA XML standard that will be used for all SDS processing alternatives. While still greatly reducing data steward workload, OCR XML file transfers would still require a data steward to review for potential corrections needed (e.g. text trimming). The user interface should provide a hazard characteristic code based upon the physical and chemical values provided. Analytics would be used for validation of physical and chemical properties to ensure values are within expected ranges and thresholds. Once reviewed, the data steward could approve the record and transfer results to

HMIRS. SDSs processed in foreign or multi-languages would additionally be run through a translating engine upon text recognition and parsing.

Third Alternative SDS Submittal Option – Vendor Manual Creation

The third option for SDS submittals would allow for small business vendors to enter and create SDSs directly on a front-end form producing both a GHS-compliant PDF of the SDS as well as an XML export aligned with the XML standardization used in the previous alternatives. The vendor would be able to populate all required 16 sections of the GHS safety data sheets, inputting all required fields and contingent fields dependent on the material properties. GHS label and HCC selection would also be available. Values entered that exceed expected norms or thresholds would be prompted to the vendor for correction/clarification.

Fourth Alternative SDS Submittal Option – Downstream User Interface

The fourth option would allow for downstream user systems the option to interface with HMIRS complying with the HMIRS XML standard (upon development and finalization). Downstream users also receive SDSs primarily for products purchased outside of standard DoD supply chains. For instance, using a government purchase card, a hazardous material user could purchase materials on the local economy. SDSs for these materials must still be provided to be received in environmental tracking systems. The downstream user interface would allow these users to provide these additional SDSs to be included in the HMIRS repository. Additionally, SDSs and other information needed to associate materials to SDSs (e.g., contract numbers, material stock numbers, etc.) would also be able to flow directly to the downstream user systems.

XML Standardized Data Output and Storage

All SDS of the aforementioned SDS submittal methods will produce data conformed to the XML standard developed by DLA to allow consistent data transfer to HMIRS regardless of origin. Each submittal method will also go through the same standardized validation algorithms for consistency.

4.5 PROTOTYPE RESOURCES AND FUNDING

When first researching dissertation topics, one thing that was very important to me from the beginning was to striving to impact real change in chemical industrial business practices. In order to do so, we would need to go beyond conceptual for development of a real-world prototype to show the value of this research to Air Force leadership and make this system part of the existing process. To build a prototype proof-of-concept system capable of processing thousands of varied documents with an acceptable level of accuracy would require funding. Numerous funding requests were put in for various Air Force research initiatives. On March 10th, one of the requests put forth to the United States Air Force 711 Human Performance Wing was approved for FY22 Studies and Analysis funding. The funding mechanism used was a cooperative agreement managed by the United States Army Corps of Engineers (USACE). On September 22, 2002, we received an award of \$280,000 at Colorado State University for a “Prototype of Novel Method for Collection of Natural Resource and Human Protection Information from Chemical Safety Data Sheets”. In December of 2021, the Colorado State University programs leads (Dr. Simske and I) procured the services of DashTech, Inc. for the coding and software development.

4.6 INTELLECTUAL PROPERTY

In order to secure intellectual property rights for the novel approach taken for document analysis and meta-algorithmic data parsing, a patent application was submitted on behalf of the inventors:

Kevin Fenton, Steve Simske, and Jonathan Luu. The application was submitted by the Air Force Research Laboratory and was approved by Air Force legal patent officials before proceeding to the United States Patent Office. Patent application 17/832800 was filed with the U.S. patent office on June 6, 2022.

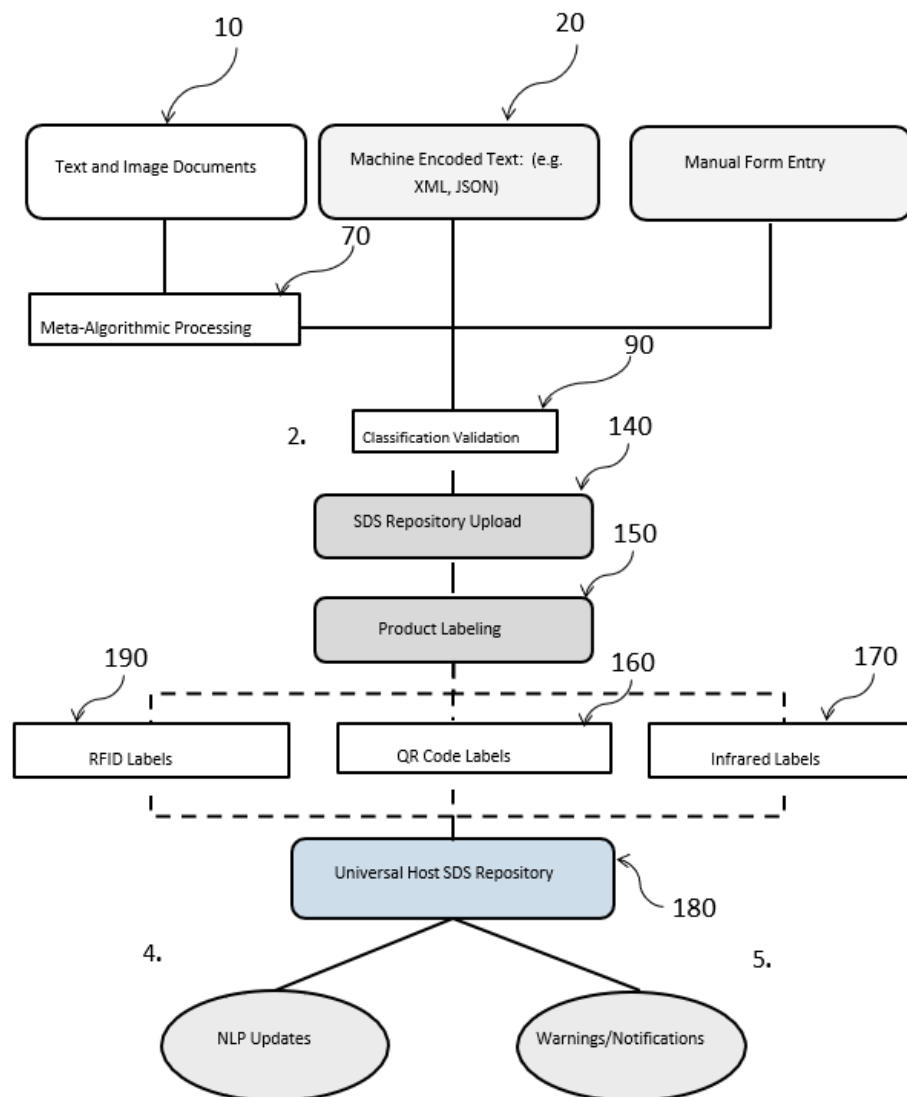


FIGURE 14 PATENT GRAPHIC 1

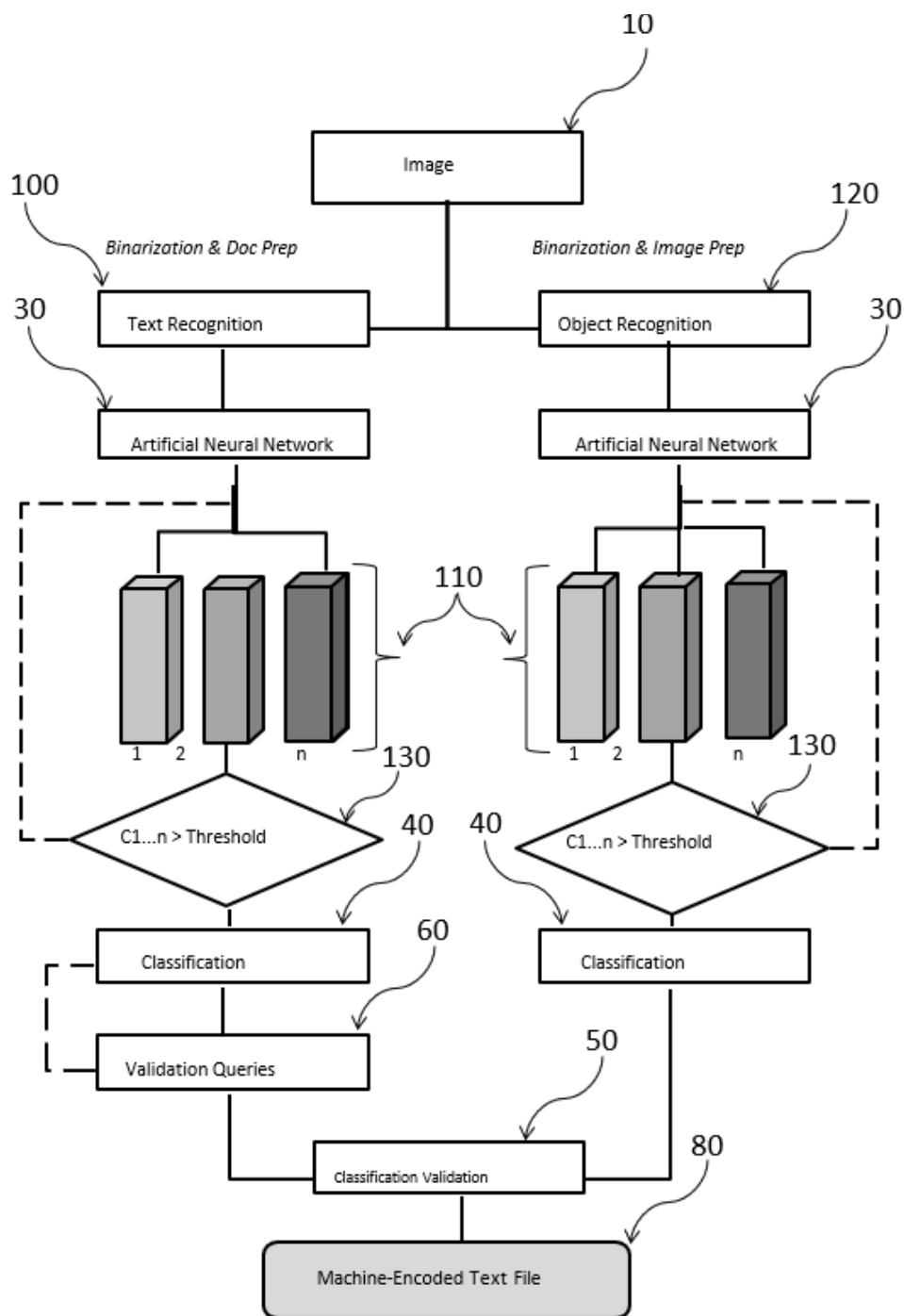


FIGURE 15 PATENT GRAPHIC 2

The full patent application can be found in Appendix A.

5 CHAPTER 5 - CONCEPTUAL DESIGN

5.1 PARALLEL PROCESSING VIA META-ALGORITHMICS

Meta-algorithmics provides system designers with a methodology for formulating results from multiple algorithms into a data analysis approach that can more effectively encapsulate the complexity of artificial intelligence tasks. The aggregation and analysis of the output of multiple algorithms, particularly when performed with neural network classification using diverse settings, can often yield higher accuracy and precision than any single algorithm. Given the diverse formats and structures of SDSs, a meta-algorithmic approach is well suited to analyze classification from various perspectives and then to maximize the specific output of the results, improving upon the results of the neural network. For the meta-algorithmic assessment in this experiment, a combination of normalized cross-correlation was used for sub-images and convolutional neural networks, machine learning key-value pattern arrays, and tessellation and recombination of the combined previous algorithmic methods for text identification and classification.

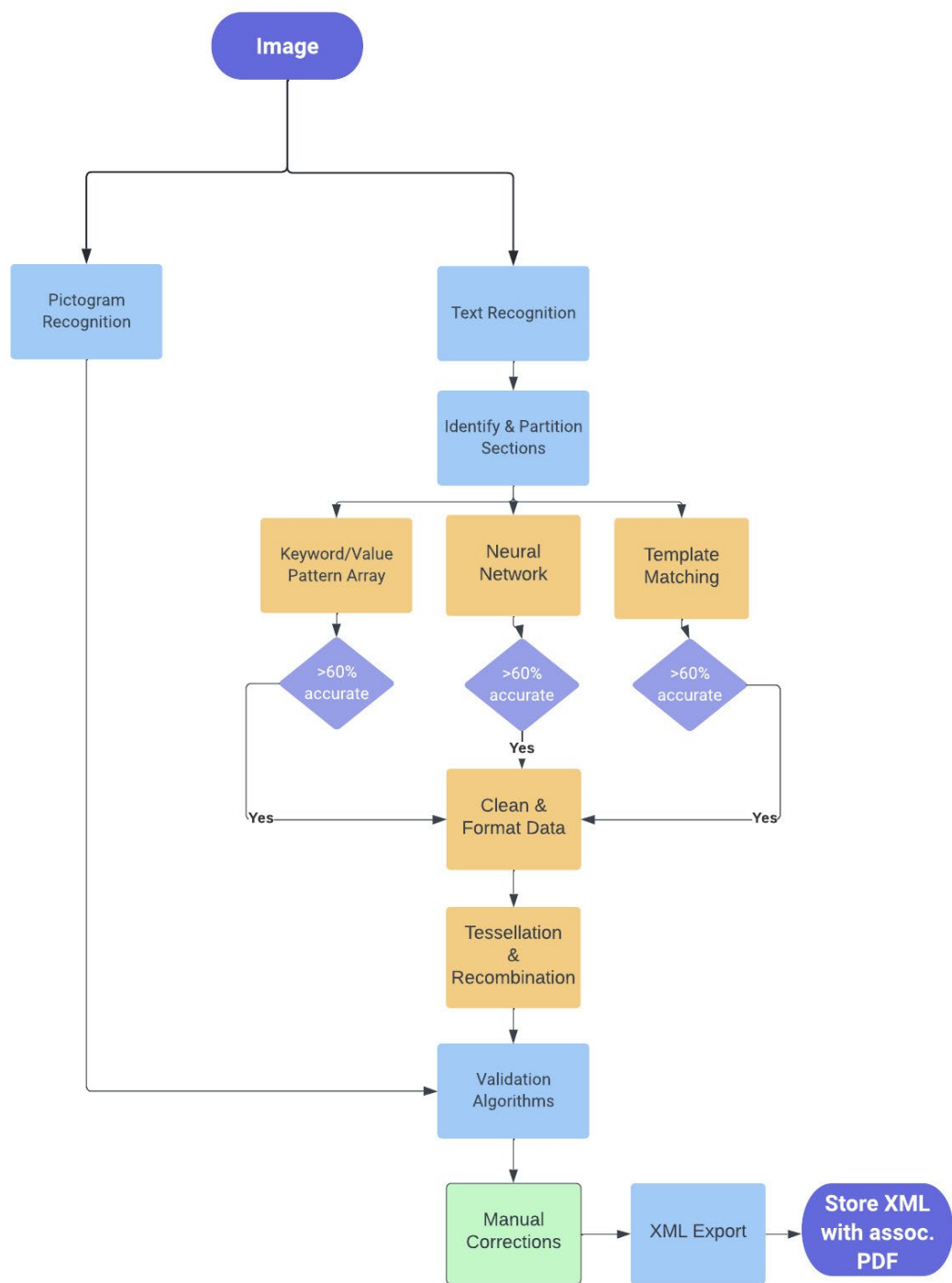


FIGURE 16 SDS OCR AND AI PROCESSING SYSTEM

5.2 NORMALIZED CROSS-CORRELLATION

Normalized Cross-correlation (NCC) is a signal processing method used to derive the validity of similarity between a sub-image embedded within a parent image (Munsayac, 2017). NCC is often used for pattern recognition tasks to determine the likelihood that an image exists within another image. NCC lends itself nicely to SDS image validation for determining the likelihood that the various GHS pictograms exist within a given SDS. The algorithm uses a distortion function that measures the degree of similarity between the sub-image and the parent image. Given the minimum distortion or maximum correlation, the location of the sub-image within the parent image is determined and the degree of likelihood of a match is calculated. In order to identify a match, the template image (i.e. GHS pictogram) is slid across the parent image (i.e. SDS image) in order to detect the area with the highest match.

$$R(x, y) = \sum x', y' (T'(x', y') * I'(x + x', y + y'))$$

Where

$$T'(x', y') = T(x', y') - \frac{1}{w * h} * \sum x, y T(x, y)$$
$$I'(x + x', y + y') = I(x + x', y + y') - \left(\frac{1}{w * h}\right) \sum x'', y'' I(x + x'', y + y'')$$

FIGURE 17 NORMALIZED CROSS CORRELATION EQUATION

Pixels are moved (or slide) left to right, up to down, with each template location T being matched over the parent image location I. Results are stored in matrix R with (x,y) in R containing the match metric (OpenCV, 2020). Python and OpenCV were used to match pictograms and Python partitioning used to maximize algorithmic efficiency by partitioning section 2 from the SDS and efficiently isolate and classify the pictograms.

5.3 CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network (CNN) document layout analysis was used to provide further pattern recognition on the inconsistent structure of SDSs. Using the analysis of spatially related values, the CNN provided a likelihood of accuracy and precision of given pattern recognitions. Using a model with a 70 percent training to 30 percent testing ratio, a sample size of $n = 500$ (50 SDSs * 10 distinct data points) SDSs were used to assess the accuracy and precision of CNN in SDS value extraction. Bounding boxes (bbox) were used to identify specific selected SDS values for extraction and provide feature vectors for image properties used for character identification such as curves, closed areas, symmetry, contours, and projections (Evelina Maria De Almeida Neves, 1997). The neural network was then trained with a variety of SDSs from various manufacturers and layouts. Input features such as boundaries, locations, and edges were used to create predictions and the predictions subsequently used to back-propagate and adjust weights as needed. With increasing n , patterns began to emerge with each training session strengthening the model. The CNN performed better in cases where the data was less structured, and terminology differed.

5.4 MACHINE LEARNING KEY-VALUE PATTERN ARRAYS

Configurable machine learning key-value pattern arrays can be used to search and partition OCR-derived text to isolate and extract values associated with each key and provide prediction and matching capabilities (Yu Bei, 2015). Although language and format may differ for key fields from SDS to SDS, commonality exists between many of the fields even in widely different SDS items. SDS fields may be indexed with extraction and validation rules that provide the system instructions on how to isolate and extract specific data fields. The initial index can be loaded based on user

knowledge and known repeating key words. As more and more SDSs fields are validated, machine learning algorithms can assign a weight based upon the frequency of the value used on multiple SDSs. The larger the index database, the greater the probability of targeted value acquisition. Once indexed, validated values can be used for either direct pattern matching or “fuzzy” pattern matching which determines a best fit option. Key-value pairing provides structure to an unstructured data set, making machine tasks much simpler to process. The keys in this case are the attributes of the SDS (e.g., product name, manufacturer, pH, flash point, specific gravity, etc.) and the values represent their corresponding specific values. The precision of the extrapolation of these attributes can be further improved by the division of the required sixteen distinct GHS sections. An SDS schema can be used to create a library of these keys for value extraction. To create the schema, a SQL query was run against the DoD SDS repository to identify the chemical vendors with the most SDSs created in the system within the past year. For the initial index, key value labels were extracted from SDSs from the most prevalently used vendors. The keys used reflected the most commonly used terminology on the predominant vendor-SDSs. 31 keys were used for the first set of values extracted and 16 keys for the second. Each set of SDS attributes $\{x_1, x_2, \dots, x_k\}$ was used to determine the probability of the hidden key sequence $\{y_1, y_2, \dots, y_k\}$:

$$P(y_1, y_2, \dots, y_k | x_1, x_2, \dots, x_k) \text{ (Chakraborty, 2014)}$$

The key-value pattern array worked very well for the first set of keys used; with the values reflecting a 9% increase in accuracy over the CNN. A significant drop occurred on the second data set analyzed as the values and terminology used varied significantly and the formats became more unstructured. Increased accuracy would be expected as the key-value dictionary is expanded but the pattern arrays run into difficulty with less structured fields that often pull in additional unexpected data or omit critical data due to varying formatted cut off points and layout differences.

5.5 TESSELLATION AND RECOMBINATION

Tessellation and recombination is a process by which components are broken down into their lowest level, assessed and compared, and then recombined or reintegrated in a more efficient or useful manner (Simske, 2013). In our example, values retrieved from the CNN and the Key-value pattern arrays are broken down into words and characters and assessed for optimal recombination based upon their extracted text commonalities. If one of the processes neglected to retrieve a value then the other is accepted. When both algorithmic applications retrieved values, the tessellated values are recombined, and the overlapping words and characters are used to increase the likelihood that the correct value has been ascertained.

TABLE 1 META-ALGORITHMIC ACCURACY ASSESSMENT

Sample	CNN, Fields 1-5	CNN, Fields 6-10	KV Fields 1-5	KV Fields 6-10	T & R, Fields 1-5	T & R, Fields 6-10
1	0.50	0.94	0.63	0.56	0.88	0.94
2	0.88	0.89	0.69	0.89	0.94	1.00
3	0.69	0.94	0.94	0.25	0.94	0.94
4	0.63	0.90	0.81	1.00	0.88	1.00
5	0.50	0.83	0.58	0.75	0.67	1.00
μ	0.64	0.90	0.73	0.69	0.86	0.98
σ	0.16	0.04	0.15	0.30	0.11	0.03

5.6 DATA VALIDATION AND CLEAN UP

Value validation and clean-up can occur either pre or post value extraction. One common result of neural network data extraction is the inclusion of special characters. The coding language can be modified to remove any special or unexpected characters at the beginning or end of each extracted value. Likewise, similar clean-up scripts can be incorporated post-extraction. Coding can also be used to standardize and validate date formats. Validation queries can be used, specifically with numeric values, to ensure that extracted values are within expected ranges (e.g., pH between 0 and

14, flashpoint between 0 and 200 degrees Celsius, ingredient values between 0 and 100%, etc.). Additionally, validation queries can be used to ensure that the SDS itself meets GHS minimum standards (e.g., 16 identified sections and required minimum information for an SDS (United States Occupational Health and Safety Administration, 2012)).

5.7 RESULTS

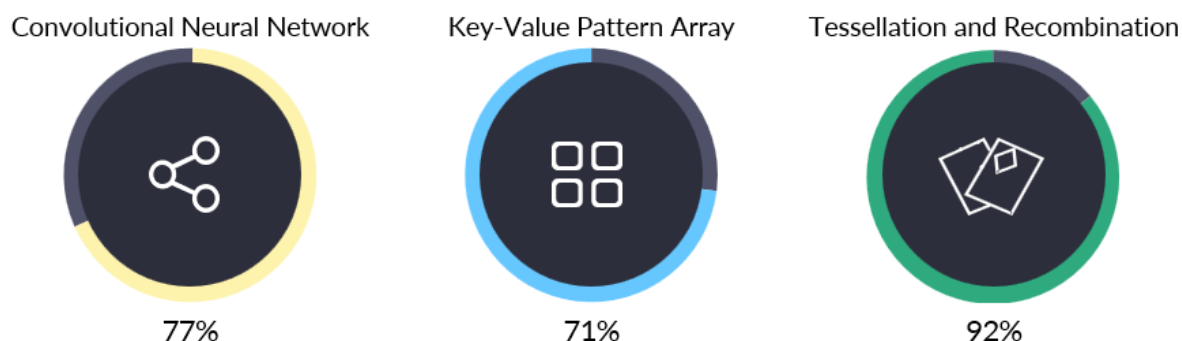


FIGURE 18 AI SDS META-ALGORITHMIC PROCESSING RESULTS

The varying algorithmic analysis applications complemented each other by assessing the SDS data from different perspectives. The tessellation and recombination application, however, yielded a significant higher accuracy rate with a 15% increase over the CNN results and a 21% increase over the key-value pattern array matches allowing for error rate reductions of up to 100%. By aggregating the values and assessing commonalities between the previous algorithms, the tessellation and recombination algorithm produced much more inclusive and accurate results. As the sample quantities for neural network training and validation increase and the key-value library continues to expand, the proposed system has the potential to improve upon the error rate incurred during manual data entry while adding significant time and monetary savings. While additional complexities may emerge as other SDS data values are added to the meta-algorithmic assessment, the research has indicated that the incorporation of this approach may yield a highly efficient

method of SDS conversion as the industry moves toward a standardized electronic transmission approach and provide a conversion method for the vast amount of current and historical SDSs still in the predominant PDF formats.

5.8 DATA ANALYTICS, HAZARD CLASSIFICATION CALCULATION, AND META-ALGORITHMIC VALIDATION

Benefits: Automated hazard calculation for storage compatibility, validation of GHS compliance and required fields, and improved SDS data extraction accuracy. Analytics provide data quality control and actionable data for business practice improvement opportunities.

Regardless of input type (i.e., XML, PDF, manual input), analytical and validation layers can be applied to the SDS processing to ensure data standardization, document completeness, and GHS validation while also calculating hazard characteristic codes (HCC) and performing meta-algorithmic validation (advanced hybridization of two or more algorithms) between computer vision-derived pictogram classification and HCC classification. The HCC is a code used by the United States Department of Defense to classify materials by their primary hazard for proper segregation and storage of hazardous materials. These codes are calculated based upon values extracted or derived from SDSs. The use of HCC assures uniformity in the identification of and management of hazardous materials and will assist in the proper recognition and safe storage by compatibility (Defense, 1999). Once calculated using SDS values (whether XML or OCR-based machine learning), these codes can be used for both validation purposes and to enhance the advanced labeling methods described below (e.g., use of RFID for storage proximity warnings for incompatible hazardous materials). The HCC can be calculated using the OCR-derived fields (e.g., flashpoint, boiling point, pH, etc.) and can then be used to validate the computer vision-derived hazardous classification pictogram on the SDS (e.g., C1 – Acid, Corrosive, Inorganic HCC

matches the Corrosive GHS pictogram). Once verified, the HCC has additional benefits such as being used in conjunction with RFID tags to ensure storage compatibility.

With a centralized SDS repository, product comparison queries can be run for potential chemical replacement initiatives. For product comparison, naturally, a common denominator must exist by which we can assess similar materials. One such way in the federal logistics system is the national stock number (NSN). Despite varying manufacturers, product names, and chemical characteristics, similarly used products with closely associated size ranges are often assigned the same NSN. This allows for queries of a chemical of concern on one SDS that can be used to trace back to all other product chemical detail associations for various products falling under the same NSN to determine if another product would effectively serve as an eco-friendlier substitute with a more benign chemical formulation. Additionally, resources such as the General Services Administration Green Procurement Compilation provides listings of potential alternatives for sustainable acquisition covered by mandatory and non-mandatory federal environmental programs (e.g., Bio-Preferred, Safer Choice, Energy Star) (Administratoiu, 2022).

Transactional analytics can also be employed for chemical usage monitoring and accuracy assessments. While manual data entry is still largely used in industry for chemical reporting systems, minor data entry errors can yield significant errors and liabilities in reporting. An incorrect specific gravity (SG) or documentation of whether the SG is relative to air or water, for instance, can result in exponential errors in chemical usage. Simple threshold analysis can ensure these values are within expected thresholds likewise for SDS numerical values. The SDS preparation dates can also serve as a useful metric to analyze chemical usage accuracy. Since the GHS adoption by chemical manufacturers in 2015, all manufacturers have been required to abide by GHS standardization requirements ensuring that the oldest preparation dates for products

created should be no older than 2015. Many manufacturers also commonly revise or release new SDSs for their products on a nearly annual basis. While some exceptions certainly exist, most deteriorative hazardous products commonly have an associated SDS within a year or two of manufacture. Trend analysis of older SDS associations to recently used materials can often be used to indicate potential EHS data reporting discrepancies.

5.9 ENTITY RELATIONSHIP DIAGRAM

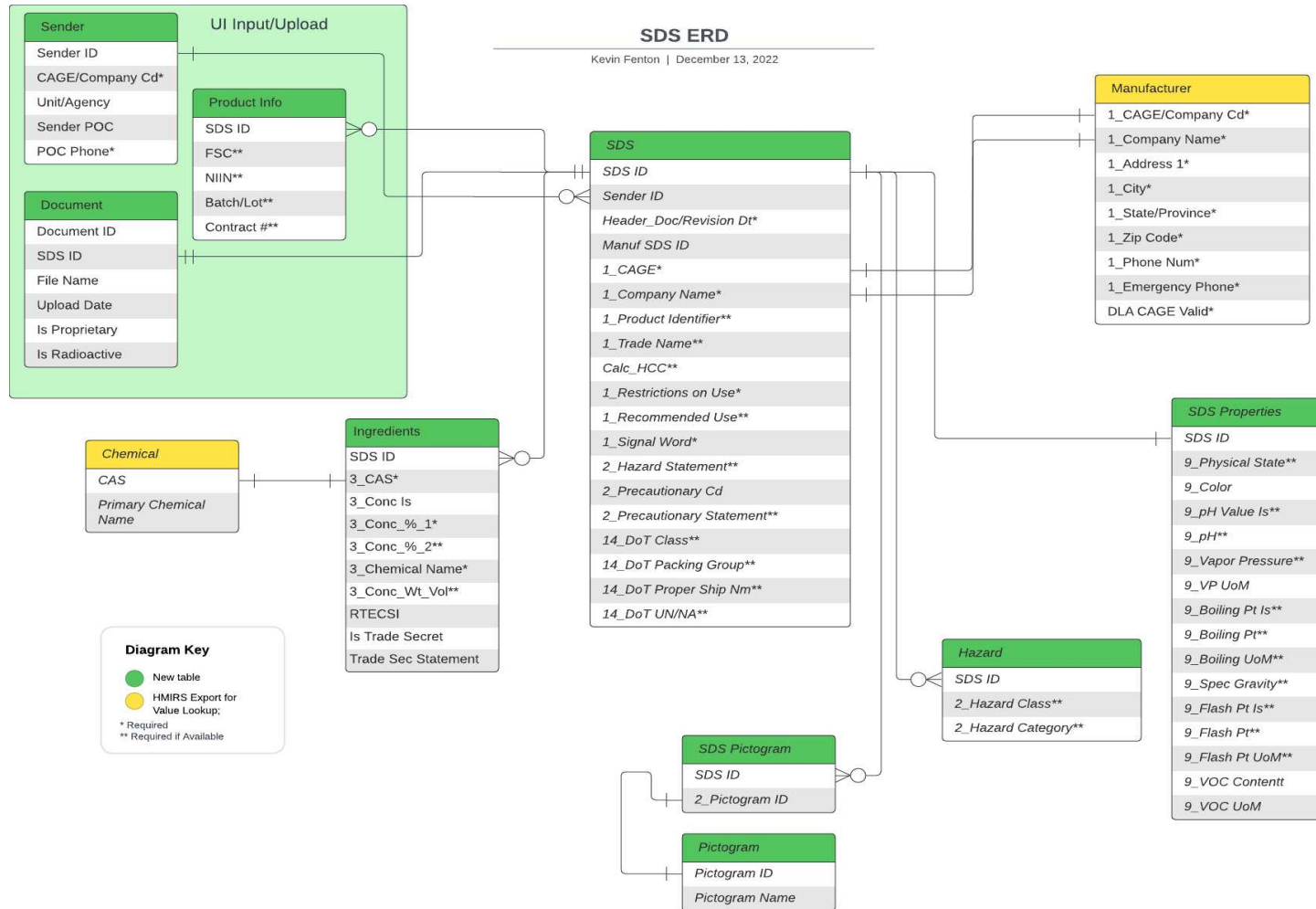


FIGURE 19 CONCEPTUAL ERD

5.10 GRAPHICAL USER INTERFACE DESIGN

Chemical & Hazard Extraction Manager (C.H.E.M.)
A.I. SDS Processing

C.H.E.M. Login

Username:

Password:

[Create Account](#) [Forgot Username/Password](#)

AFRL THE AIR FORCE RESEARCH LABORATORY

Colorado State University

DASH TECHNOLOGIES

FIGURE 20 GUI DESIGN 1

Chemical & Hazard Extraction Manager (C.H.E.M.)
A.I. SDS Processing

SDS Upload

Contract #: *Not required

AFRL THE AIR FORCE RESEARCH LABORATORY

Colorado State University

DASH TECHNOLOGIES

FIGURE 21 FIGURE 20 GUI DESIGN 2



SDS Upload

File 1: SDS001.pdf	National Stock Number(s): 8010011239874	+	Batch/Lot: ASDF/1234
File 2: SDS002.pdf	National Stock Number(s):	+	Batch/Lot:
File 3: SDS003.pdf	National Stock Number(s):	+	Batch/Lot:

Process

AFRL
THE AIR FORCE RESEARCH LABORATORY

Colorado State University

DASH
TECHNOLOGIES

FIGURE 22 GUI DESIGN 3

C.H.E.M.
A.I. SDS Processing

77%

Required Fields

94%

Req. Field Accuracy

91%

Total Accuracy

→ Save & Next
📄 Export XML

SAFETY DATA SHEET

Revision Date: 24-Feb-2020
Revision Number: 2

1. Identification

Product Name: Hexavalent Chromium, standard solution, Specpure®, Cr(+6) 1000µg/ml

Cat No.: 42234

Synonyms: No information available

Recommended Use: Laboratory chemicals.

Uses advised against: Food; drug; pesticide or biocidal product use.

Details of the supplier of the safety data sheet:

Company:
Alfa Aesar
Thermo Fisher Scientific Chemicals, Inc.
30 Bond Street
Ward Hill, MA 01835-8099
Tel: 800-343-0660
Fax: 800-322-4757
Email: tech@alfa.com
www.alfa.com

Emergency Telephone Number
During normal business hours (Monday-Friday, 8am-7pm EST), call (800) 343-0660.
After normal business hours, call Carechem 24 at (866) 928-0789.

2. Hazard(s) Identification

Classification:
This chemical is considered hazardous by the 2012 OSHA Hazard Communication Standard (29 CFR 1910.1200)

Germ Cell Mutagenicity	Category 1B
Carcinogenicity	Category 1A
Reproductive Toxicity	Category 1B

SDS Date: 2/24/2020 99%

Product Name: Hexavalent Chromium, standard solution, Specpure, Cr(+6) 1000mg/ml 92%

Company: Thermo Fisher Scientific Chemicals, Inc. 97%

Address: 30 Bond Street 92%

City: Ward Hill 92%

State: MA 92%

Zip Code: 01835-8099 92%

Recommend Use: Laboratory Chemicals 90%

Category Code(s): Category 1B 88%

FIGURE 23 GUI DESIGN 4

5.11 REQUIREMENTS

General system requirement tasks include:

- Software prototype capable of processing Safety Data Sheets received from USAF operations to DoD HMIRS system via an optical character recognition tool which will use image recognition, machine learning, and a meta-algorithmic validation process to read, parse, and export required text fields.
- Processing the records in a timely manner not to exceed 2 min per SDS.
- Quality Assurance Analysis (QA) of the data elements specified for entry by the USAF HMIRS PMO. The software should allow the uploader the ability to review and correct extracted data prior to XML export.
- All required missing data points will be notified to the user and require completion.
- All data points extracted will conform to the HMIRS SDS XML standard.
- Written/oral presentations or reports for AF HMIRS PMO or other AF leadership.
- Implementation of Quality Assurance analysis of the data quality and data elements with the goal to achieve greater than 90% accuracy from parsed data values.
- Documenting accuracy, precision, and recall metrics.
- Providing status updates of project objectives and completion timeline.
- Providing final software system source code to the AF HMIRS PMO.

Additionally, the cooperator shall prove or disprove the hypothesis that the following system requirements can be met for a functional system within the acceptable quality levels:

Requirement	Acceptable Quality Level
System shall be a standalone DoD-owned system with no subscription costs	Web-based application accessible to DoD personnel
System shall be able to process up to 15k Safety Data Sheets (SDSs) in PDF format monthly	95% of SDSs should process without readability errors
System shall clean extracted data by removing special characters and spaces at the beginning and end of strings.	Extracted data should be cleaned appropriately for use.
System shall standardize and conform extracted date fields.	Extracted date values should match date formats from the HMIRS data dictionary and XML standard.
System performs image recognition on GHS pictograms	95% accuracy required for pictogram recognition
System shall perform Optical Character Recognition to identify SDS text	95% of SDSs should process without readability errors
System shall use a neural network to appropriately identify and classify parsed text	Minimum training set of 1k SDSs. Training should be based upon most frequently used SDSs.
System shall use a key-value array to parse out text values according to loaded key library	100% of identified key values should be used for value identification.
System shall add to list of keys in key library as text classifiers repeat	Key value listing should be editable.
System shall break down neural network values and key-value values using tessellation and recombination. If one does not produce a value, the other is automatically taken.	System shall return the values with the highest level of confidence and only display values with a 70% or higher confidence level unless otherwise directed.
System shall display the tessellated and recombined values on a user-friendly interface and allow the user to quickly scan through and make corrections as necessary	System shall display values in a user-friendly interface that can quickly and easily be corrected by the user.
System shall validate SDS signal warnings against GHS pictograms and notify user of potential mismatch	Image recognition will be performed on pictograms. Pictogram type will be compared to signal warning fields for an incompatibility check.
System shall notify user of all validation criteria mismatches	HMIRS data dictionary and XML standard will be used to ensure all fields meet data type, character, and character length parameters.

System shall notify user of missing required fields	Required fields will be provided from the HMIRS data dictionary and XML standard. User will be notified when required fields are missing and when GHS compliance has not been met or within acceptable thresholds.
System shall allow user to export parsed SDS fields into an XML file using a provided XML standard	XML export will be formatted in accordance with the HMIRS XML standard that will be provided.
System shall reside on a predetermined domain	Domain must be accessible by DoD, contractors, and civilians.
System shall allow users to create accounts and restrict access based upon approval by DoD project leads	Username and password entry along with dual authentication will be required for system access on the specified domain.
System shall allow users with system access to either drag SDSs from a folder or browse and begin processing immediately with a user-friendly interface	System should be user-friendly and offer an efficient and intuitive user interface.
If multiple SDSs are provided at once, the system shall process each document and export XML in turn and allow for all processed XML exports to be retrieved	Individual SDSs will need to be parsed individually and exported in individual XML files.
System shall be able to document multiple chemical ingredients, values, units, and range symbols per SDS.	System will need to accurately read and parse multiple chemical ingredients including Chemical Abstract Service (CAS) numbers, values, units, and range symbols from various SDS formats.
System shall at a minimum process the minimum required SDS fields identified by the DoD (~45 fields)	Required fields will be provided by HMIRS USAF POC.
System shall process each SDS in < 2 min time per document	System should process fields in a reasonable timeframe and gracefully prompt user and handle errors as necessary.
System shall be trained on SDSs provided by the DoD	System will need to be trained on minimum of one thousand SDSs.
System shall determine hazard characteristic code (HCC) based on SDS properties.	Characterized in accordance with DALI 4145.11.
System shall have an overall classification accuracy rate of 90% on required fields	Overall accuracy rates will be based on 45 required fields extracted from system testing SDSs.

Additional required system features:

- Ability to document one or more National Stock Numbers, Batch/Lot, and Contract number for each SDS.
- Ability to enter the same required SDSs fields into a form and have the same validation criteria run against the form as the OCR-parsed text; also including NSN/batch/lot/contract assoc.
- Ability to upload an XML in the same format and same standard that we will be exporting our OCR-parsed fields to and have the same validation criteria run against these as well; also including NSN/batch/lot/contract assoc.
- Compare parsed Chemical Abstract Service numbers or Chemical Names to a table containing chemical information.
- Compare parsed Manufacturer names or CAGE codes to a table containing vendor information.
- Obtain what fields beyond the minimum 45 as possible; given time, resource, and processing ability.

5.12 VENDOR CONTRIBUTION QUERY RESULTS

Although HMIRS has received hundreds of thousands of Safety Data Sheets from thousands of chemical manufacturers, some manufacturers are obviously more prevalent than others. A query was run to list a count (by manufacturer) of SDSs loaded in the last two months. The top 25

companies accounted for nearly 50% of all SDSs loaded. This data allowed us to train the model more heavily on these companies to increase overall system accuracy.

TABLE 2 MOST PREVALENT MANUFACTURERS

Manufacturer	Count
PRC - DESOTO INTERNATIONAL, INC. PPG AEROSPACE	3837
SHERWIN-WILLIAMS COMPANY	3092
HENTZEN COATINGS INC	1898
SIGMA ALDRICH	1486
AIRGAS, INC	1435
3M COMPANY	757
PRC-DESOTO INTERNATIONAL INC (DBA PPG AEROSPACE DEFT)	517
LINDE INC. (PREVIOUSLY PRAXAIR, INC.)	511
PRC-DESOTO INTERNATIONAL, PPG INDUSTRIES COMPANY	509
3M COMPANY 3M GOVERNMENT MARKETS	499
HENKEL CORPORATION (ONE HENKEL WAY)	495
MILSPRAY LLC	486
AKZO NOBEL AEROSPACE COATINGS INC.	470
CAAP CO INC	411
PRC DESOTO INTERNATIONAL INC/PPG AEROSPACE	355
THE FLAMEMASTER CORPORATION	346
FISHER SCIENTIFIC CO CHEMICAL MFG DIV	300
RUST-OLEUM CORP	283
ALFA AESAR/ THERMO FISHER SCIENTIFIC CHEMICALS INC	245
PPG INDUSTRIES INC (ONE PPG PLACE)	244
EXXON MOBIL CORPORATION (SPRING-TX)	219
ROYAL ADHESIVES & SEALANTS LLC WILMINGTON	210
HACH COMPANY (P.O. BOX 389)	180
DUNN-EDWARDS CORP	168
DAP INC	167

Percentage of SDS contributed by top 25, 50, 100, and 1000 vendors:

TABLE 3 TOP MANUFACTURER SDS CONTRIBUTORS

Top 25	49.53%
Top 50	56.20%
Top 100	63.48%
Top 1000	89.21%

5.13 DOCUMENT PROCESSING FLOW

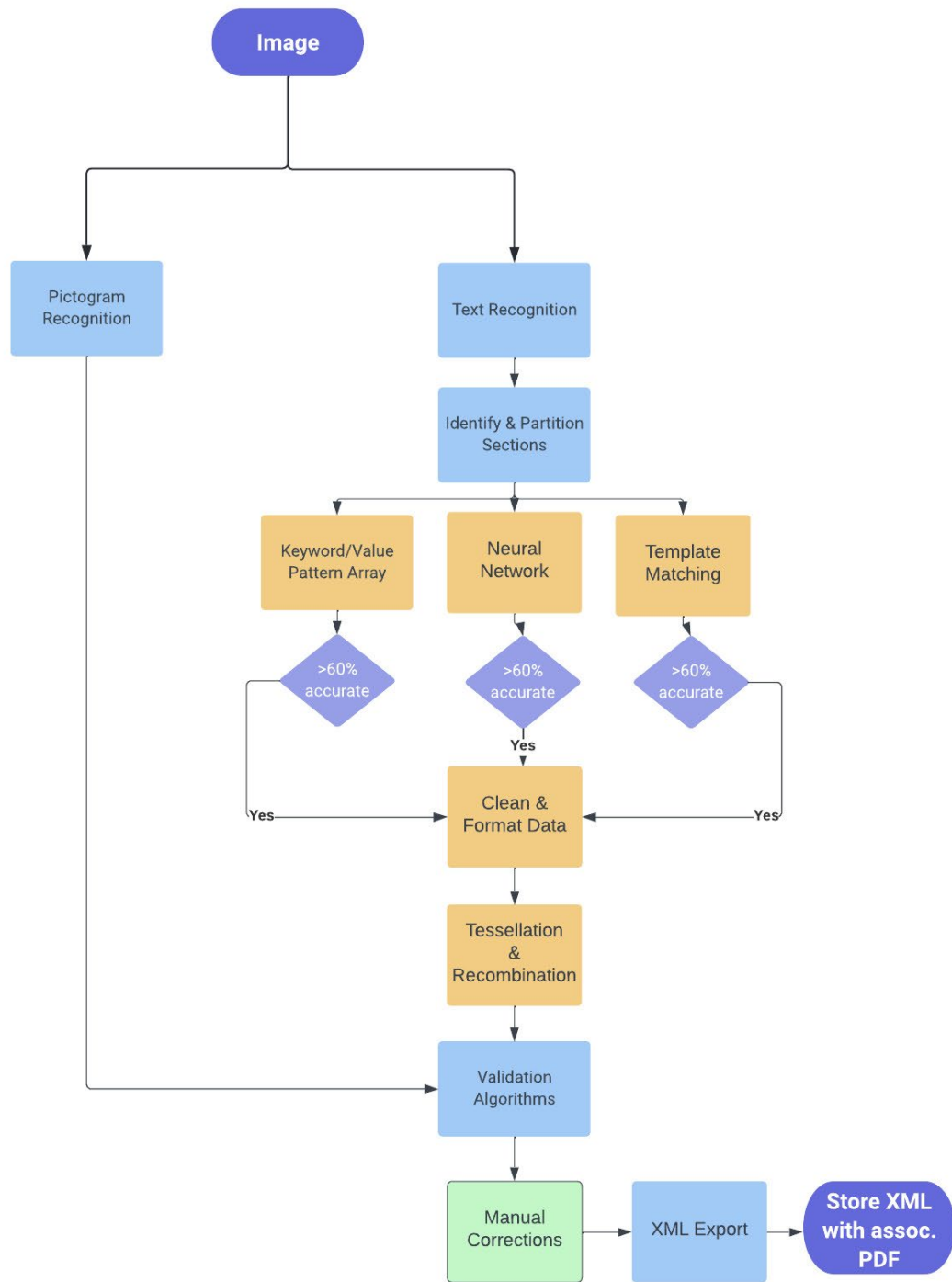


FIGURE 24 ADVANCED DEVELOPMENT PHASE

The proposed system will be prototyped to offer the key stakeholders a visualization of the system's operational and functional capabilities as well as implement the proposed architectural design of system components in the conceptual phase into an operational system that can perform the desired functions and meet system requirements. The use of a UML support tool can assist in converting the design information into the software architecture. The prototype should include a test environment mirroring real world processes to ensure the system can meet its objectives.

System Integration

Although the prototype will likely not interface directly to HMIRS, validation will occur to ensure the output, interface protocols, and languages are all compatible with HMIRS capabilities. Output files can be produced from the prototype to manually feed into HMIRS and ensure data transfers occur successfully.

Test and Evaluation

Documentation of Technical Performance Measures (TPMs) will allow us to adequately assess how well the system meets the desired objectives. Accuracy in the data sheet parsing, translation, and level of record completeness will likely need to be observed. Additional TPMs could potentially include the amount of time it takes to load a set number of records over a given time period. These TPMs can be used as a baseline to ensure the system not only meets customer expectations but also justifies the system need through cost benefit analyses. The system prototype can be used for functional evaluation and data exports can be validated to ensure interface processing compatibility.

5.14 SYSTEM DELIVERABLES

- Stakeholder Analysis
- Needs Analysis
- Concept of Operations
- Feasibility Analysis
- Conceptual Design / Architecture
 - Process Flow Diagram
 - UML/ERD
- Operational and Functional Requirements Document
- Functional Analysis Document
- Test and Evaluation Results
- Prototype/Model

6 CHAPTER 6 - SYSTEM OPTIMIZATION

6.1 UNIVERSAL SDS DATABASE AND REPOSITORY

6.2 Universal SDS Database and Repository

Benefits: Single accessible source of chemical data for all chemical users, elimination of separate duplicative systems across industry, product comparisons for safer product selection, personnel and time savings from researching SDSs across thousands of vendor websites, data quality control, point source for systems to electronically access SDS data, singular source for precise SDS access methods

One change that would provide immediate benefit on many of these issues, is also the most difficult to implement and appears to require a paradigm shift in the regulatory compliance measures imposed on chemical vendors. This shift is to a single universal SDS repository for all chemical users. Although some repositories such as SDS.com exist, existing SDS repositories are not all inclusive, can require membership payments, and have varying levels of quality assurance. All chemical users, industrial or otherwise, should be entitled a single centralized universal repository for all SDS with full chemical disclosure with easy and intuitive access of hazards for any hazardous material. A single national or universal SDS repository will allow (and require) a standardized SDS formatting, significant reduction in varying or missing information, reduction in back-and-forth communication with downstream users and manufacturers, complete Globally Harmonized System requirement compliance, significant time and monetary savings in duplicative loading of SDS across numerous existing repositories, and a single system that could be harnessed for product to SDS association as described with the following EHS technological advances.

The entire systematic approach for chemical user hazard communication can be improved upon with a singular SDS database and the incorporation of 21st century technologies. While SDSs largely remain on chemical vendor websites or various locations, the complexity for systematic improvement remains challenging. Pollutant reduction assessments, including greenhouse gases, could be performed much more efficiently by product comparison and allow for greater visibility of the usage of these products nationwide. A universal SDS system could easily be created using an existing SDS repository as a model to be expanded upon. For example, the Department of Defense (DoD) Hazardous Material Information Resource System (HMIRS) already serves as an effective repository for hundreds of thousands of SDSs while ensuring quality control and includes

an xml standard for streamlined digital transfer. A system such as this could easily be expanded upon to serve as the singular hazard communication platform for the public.



FIGURE 25 CENTRALIZED SDS REPOSITORY

For an effective so-named Universal Safety Data Sheet system, the system will need to address the many possible ways SDSs can be received and how these varying SDS submittals methods can be consolidated into a single management approach. Ideally, SDSs would be transmitted via a user interface directly to the centralized repository in a standardized XML or equivalent data transfer form. However, efforts are also underway for the OCR and AI SDS “reading” methods described below for seamless conversion of SDSs in PDF form to machine-encoded text.

Additionally, for new SDSs creations, a manual creation option would allow vendors to efficiently create new GHS-compliant SDSs with immediate loading into the universal SDS repository. Finally, the last SDS feature will allow other systems using SDSs to communicate the

information directly to the centralized universal system via interfaces following the same XML or equivalent data standard. Regardless of the SDS submittal method, standardized validation algorithms can be run to ensure SDSs are complete, GHS compliant, and contain accurate data values. Chemical management responsibility has quickly surged to the forefront of global needs and a universal SDS system would be a large step in our ability to reduce our environmental global footprint while increasing safety and occupational health standards. From climate change to chemical exposure reduction, data quality and chemical analysis precision are paramount in our ability to gather actionable data and use it for effective pollution and exposure identification, reduction, and potential remediation.

6.3 LABEL ADVANCEMENTS

Benefits: Precise SDS-to-product association, increased tracking ability, product authenticity, and anti-counterfeit measures. Improvements to precise SDS selection and information availability to the user can be made by moving from traditional legacy product barcodes to the following modern applications that can be used to quickly direct users to the product's SDS.

6.4 BASIC SDS QUICK RESPONSE (QR) INTEGRATION

In the advent of the COVID-19 global pandemic, the use of QR codes has seen a dramatic rise with many restaurants replacing hard copy menus with digitally accessible versions online using common cell phone camera applications. QR codes are available in two forms, static and dynamic, dependent on whether a user wants data essentially hardcoded into a barcode or using a URL that can be modified after printing (Moore, Davis, Spear, & Bombaci, 2021). Numerous QR code generators exist for chemical vendors or EHS personnel to create these codes at little to no additional cost. These codes could be printed on custom organizational barcodes or, ideally, by the

vendors on the products themselves. The ability to immediately retrieve the correct SDS for any hazardous material by any personnel with a cell phone helps alleviate the reliance on often unmaintained SDS binders and greatly expedites the time for personnel to access these SDSs in the event of an emergency (Langerman, 1995). Additionally, direct association of the SDS has the added benefit of remaining tied to the product outside of the workplace, should the product move from its original intended use location. Lastly, association of a product to the chemical SDS is mandatory for many environmental, health, and safety compliance systems that use the ingredients and hazards on the SDS to calculate usage for regulatory reporting, determine exposure for personnel, and employ appropriate safety measures. Direct QR code association from a vendor would greatly increase the speed and efficiency in loading SDS information into downstream user systems by eliminating the need for manual research on vendor sites. QR code storage capacity depends on the size and version used. Modern QR codes (currently version 40) contain 31,329 squares encoding up to 3 KB of data translating to 7089 numeric characters or 4269 alphanumeric (additionally Kanji/Kana, Arabic, and other languages can be stored with varying capacities) (Abas, Yusof, Din, Azali, & Osman, 2020). The average website URL contains approximately 40–50 alphanumeric characters so the potential clearly exists for schema development that could provide numerous relevant hazardous material data points to a user.

6.5 QR CODE APPLICATION EXPANSION

In addition to retrieving SDSs from a cloud-based server, the additional benefits of the expansion of product-related data that can be retrieved can greatly enhance industrial operations. One example is the direct transmission of an SDS to a downstream user system. If an XML or equivalent file is hosted on the server, the machine-encoded file could be transmitted directly to subsequent systems (assuming an XML standard is met). If the file is in PDF, a combination of

OCR and AI meta-algorithmics (described above) can be used to reverse engineer and parse and validate SDS fields into a usable format by the downstream system. This eliminates not only the need for the chemical user to scour the internet looking for the correct SDS but also immediately submits it to a database in the machine-encoded text, which would greatly expedite if not eliminate the need for SDS manual data entry for hazardous material tracking needs. The SDS is automatically routed for environmental, safety, and occupational health review (which could also be performed via mobile application). Second, the QR code can be used to retrieve other product related data to enhance inventory operations. Information such as unique SDS identifiers, product batch/lot information, container numbers, NSNs, container/unit/package information (e.g., 16 oz bottle), manufacturer, trade name, noun, manufacture date, and expiration date could all be retrieved by the same API connection. For the DoD, expiration dates could include not only the original product expiration but also return updated expiration and service life dates based upon lab result testing or user extension, whereas traditional barcodes are simply printed with the original dates and require database searches for these updates.

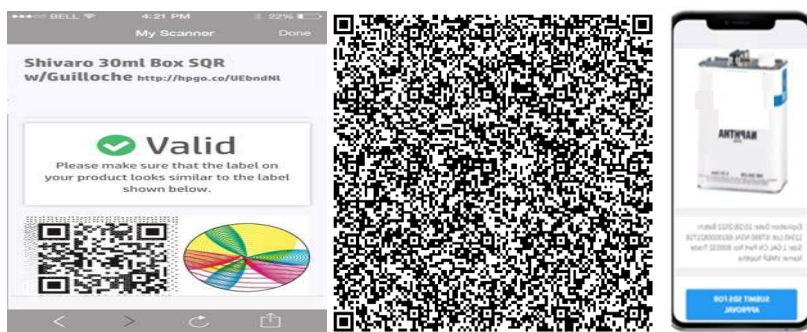
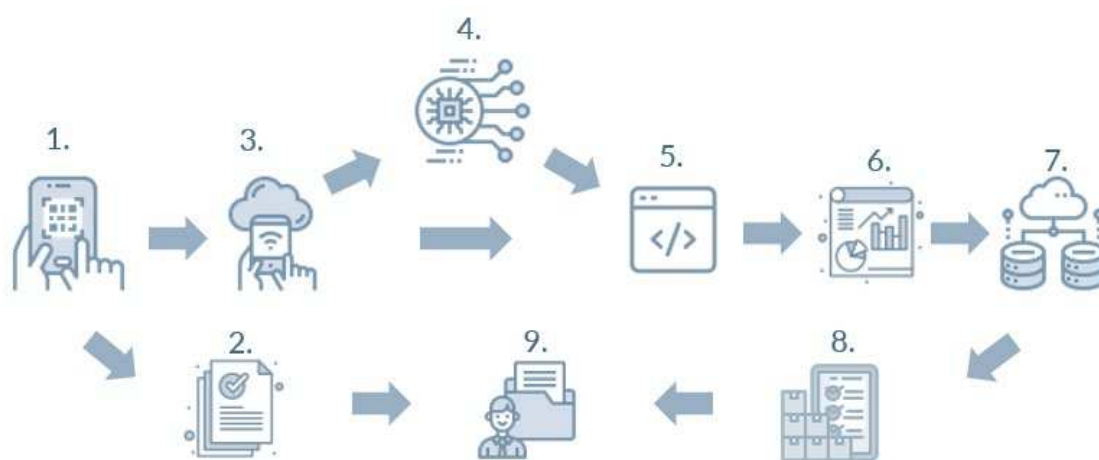


FIGURE 26 EXAMPLES OF VARIOUS EXPANDED QR CODES CAPABLE OF PRODUCT AUTHENTICATION AND SDS RETRIEVAL

Object storage services are used to provide supplemental material such as images (SDSs, technical data sheets, and specifications) and any other associated documents to the downstream user. QR

codes have also advanced for uses in supply control measures and counterfeit detection (Gaubatz & Simske, 2010) (Picard, Landry, & Bolay, 2021). Several anti-counterfeit designs have been created including QR codes containing secure graphics that can verify the authenticity of the product and provide detailed information on the location and devices used to scan the product. While this technology was originally developed for security needs, it can also be used to meet a host of EHS needs and requirements and can be used to forward SDSs to an AI processing system, which would read and load the SDS into the host repository while maintaining supply chain authentication.



1. Chemical Product QR Scan
2. Immediate return of OSHA compliant SDS
3. Application sends SDS in machine-encoded text or pdf for validation
4. If pdf, system performs OCR and machine learning algorithms and parses SDS fields into machine-encoded text
5. Machine-encoded text exported to standardized import file
6. Analytics performed on incoming machine-encoded text for field validation, ensures GHS-compliance, and complete SDS
7. Data loaded into databases for EHS review and approval
8. Shelf-life extensions and other updates sent to user
9. Immediately accessible SDS inventory for all users

FIGURE 27 QR CODE SDS RETRIEVAL AND AI PROCESSING METHODOLOGY

The proposed methodology can not only provide numerous benefits to chemical users but to the manufacturers as well. Negative implications are primarily cost-driven, so the cost-benefit analysis would depend on the potential workload savings and the monetary losses associated with counterfeit products. The many benefits to chemical vendors include the following:

- reduction of back-and-forth correspondence to chemical users on SDS locations and SDS data points
- marketing of new added benefits of streamlined SDSs and product information to users
- benefitting from brand protection while greatly reducing counterfeiting and offering validation of product authenticity
- increased customer engagement in their products
- significant reduction in counterfeit products; greater product security, flexibility in supply chain logistics, and anti-fraud requirements.

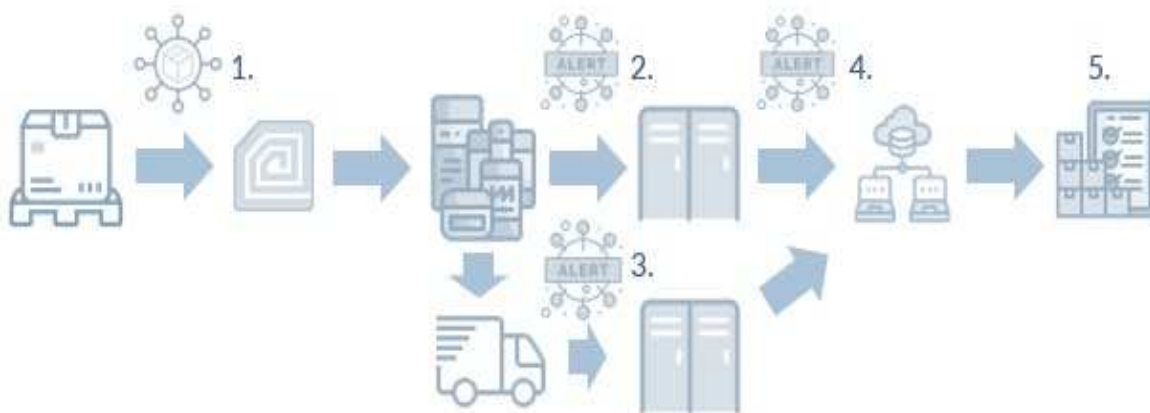
CONTAINER TRACKING OPTIMIZATION

Benefits: Automated tracking ability, cost/time savings from manual tracking reduction, compatibility validation, and warning opportunities for inventory concerns.

6.6 RADIO FREQUENCY IDENTIFICATION (RFID) INTEGRATION

Another possible labeling improvement that could either work in conjunction with QR code systems or as an alternative to traditional barcodes is the use of RFID tags on hazardous materials and hazardous waste, either by the vendor or by downstream user labeling upon receipt. Hazardous material and hazardous waste management typically require a database for large operations to accurately and consistently monitor where containers are and closely observe the amount of time containers exist at these sites and approach regulatory thresholds. Movement of containers from

site to site within the database thus requires manual transactions that must occur for transfer documentation. The use of RFID technology can be used to eliminate the need to track hazardous material containers from supply points to end users or waste containers as they move from satellite accumulation points to centrally managed hazardous waste storage areas or treatment, storage, and disposal facilities. Rather than relying on follow-up manual transfers to occur, sensors at each location would automatically document the movement and record the appropriate transaction in the host database. RFID tags also eliminate the need for a line of sight between tags and readers allowing expedited inventory management and accountability. Many inventories in the DoD are still performed either by manual count or scanning using legacy barcode scanners. RFID tags enable constant accurate accounting of inventory. Similarly, emergency response personnel can be equipped with technologies that not only provide what chemicals are stored in a response location based on receipt transactions but also what products and chemicals RFID tags are communicating that exist within the RFID boundaries (Figure 6).



1. RFID loaded with EHS data – shelf life, hazard codes, SDS association, NSN, size, etc.
2. Inventory tagged with RFID
3. Inventory movement from site-to-site triggers database location transfer
4. Alerts triggered for expired material, incompatible storage, shortages, etc.
5. Accurate current inventory available to site managers and first responders with updates of hazards or inventory discrepancies

FIGURE 28 RFID HAZARDOUS MATERIAL/WASTE TRACKING

Microchips in RFID tags can be either read-only or read-write. The latter allows users to add data to the RFID tag, such as the waste profile number for the waste, the accumulation start date, and the site-specific regulatory start date. Likewise, for hazardous materials, SDS IDs, expiration dates, and other inventory data points can be loaded against the RFID tag. The benefits of incorporating this technology could yield significant time savings of personnel, monetary savings of reduced data entry and tracking, and increased EHS accountability of these containers. Additionally, dangers associated with improper hazard segregation could be mitigated by cross-analysis of incompatible storage items through RFID proximity. For instance, acids should not be stored with bases or flammable products should not be stored with oxidizers. RFID technologies could be used to trigger a warning system of possible incompatible storage. Similar warnings can be presented to inventory managers on a host of other potential inventory concerns such as expiration, shortages,

max allotment exceedances, and temperature threshold concerns. Active RFID tags can store between 16 bytes and 128 KB, and passive UHF tags are capable of 32 KB at a frequency between 865 and 956 MHz allowing for longer ranges, (Pais & Symonds, 2011) both offering ample capacity for the limited EHS inventory fields desired.

6.7 ENHANCED COMPUTER VISION USING AUTOMATED NEURAL NETWORK IMAGE PRE-PROCESSING

Upon researching computer vision methods and writing computer vision python scripts, I discovered that image recognition applications do not take an automated approach of applying varying image transformations to improve upon returned features and levels of confidence (Bishop, 2006). Manual optimization is largely impractical due to the exponential number of image modifications and combinations possible. Even minor image alterations, however, can lead to significant changes in how computer vision interprets a desired image. An automated solution allows for an assessment of each image from various perspectives allowing the neural network to benefit from the most advantageous transformation combination (Wang, 2019).

APPROACH

The ideal system will include algorithms that automatically run each potential combination of n filters selected for input alterations (see Figure 1). A cloud vision API was used for classification and assessing which pre-processed method was more effective in maximizing overall accuracy (Krizhevsky, 2017). The neural network was trained with images of the desired features to be identified and classified. For our example, we used the Google Cloud Vision API, a large-scale neural network which offered pre-trained learning models and classifications into millions of

predefined categories. Google Cloud Vision was used to determine the classification and percentage likelihood that the image has been correctly classified.

The inherent algorithms determined feature characteristics such as edges, boundaries, locations, and number pixels that were used to measure this degree of confidence of classification (Liu, 2019). Each image transformation brought about variations in these characteristics. Each differently pre-processed image contained its own unique characteristics that were measured; and through this input layer, predictions and comparisons were made to the trained images (Das, 2016). Comparisons of the feature characteristics were made to the millions of available images accessible through the Google Cloud Vision API.

For this experiment, sample photos were altered with various combinations of image transformations using the ImageMagick software library. NodeJS was used to automate the process and collect the results. Given the time allocation needs for establishing a functional auto-optimization algorithm and tool, the approach for this experiment will be to simply simulate the benefits on a smaller scale by showing how variations of image alterations can impact image classification confidences. 10 images containing one or more subjects were used with multiple combinations of alterations for a total of 140 variations.

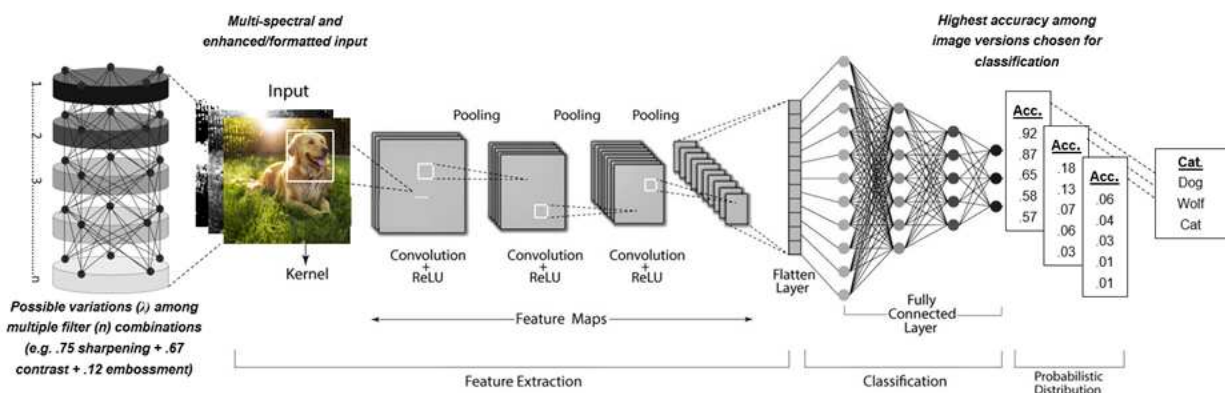


FIGURE 29 ENHANCED NEURAL NETWORK PROCESS

Each of the original 10 images was analyzed using the Google Cloud Vision application to establish the baseline on what each classification confidence percentage is based upon the well-proven application consisting of millions of pre-trained images.

First, by using each individual filter method and then by performing combinations of each of the filters. Each altered variation was then fed into Google Vision to determine the new percentage of confidence and the impact that each filter had on Google Vision’s ability to appropriately classify each subject to the highest degree. The images chosen were meant to provide a variety of perspectives – closeup and distant subjects, single and multiple subjects, and clear and hidden subjects.

$$\max_{\sigma} \text{NNacc}(\sigma_{(1:n)} : S(\sigma_n))$$

where

$\sigma_{(1:n)}$ = image pre-processing filters

$S(\sigma_n)$ = possible permutations of filters

FIGURE 30 OPTIMIZED PRE-PROCESSING EQUATION

RESULTS

The various imaging operations in combination with CNN approaches added features not generally associated with the combination of original images + CNN approaches. This yielded computer vision resilience in areas not currently addressable, such as remote sensing areas with limited input data sets. Essentially, the images and their transforms created a larger data set, with the transformed images operating as “simulated images” in some cases to increase the effective level of training.

Of the 140 various images used, the original only recorded the highest level of confidence on 3 occasions. Nearly each of the combinations had significant impacts on confidence for the different images used. For instance, the sharpen filter moved pixels away from their clustered values to add definition and enhance edges reflecting an increase in entropy in the image. Despeckle did the opposite as it removed noise and decreased entropy. This was beneficial for images where the decrease in classification confidence was less about defined edges and more about excess background noise that was impacting the ability to clearly classify the primary subjects.

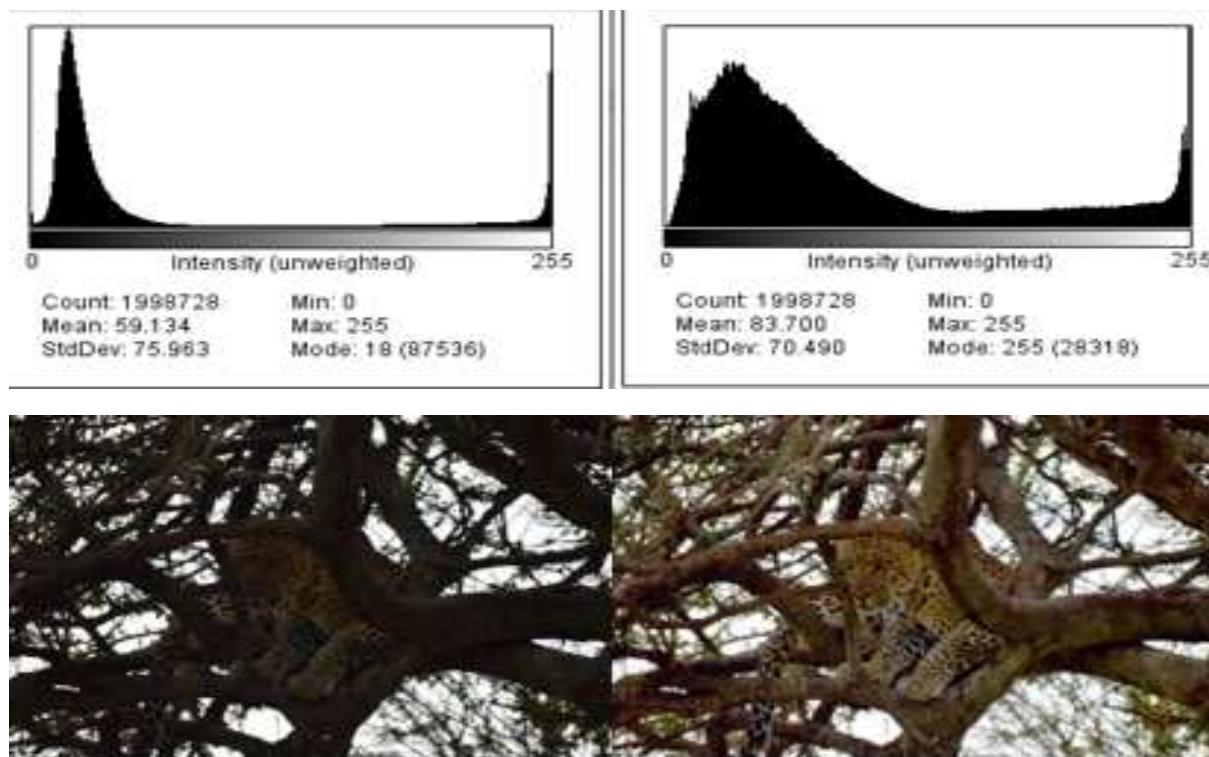


FIGURE 31 OPTIMIZED IMAGE

While the edge detection and embossment filters performed relatively poorly overall in terms of the level of confidence, both did add to the number of subjects appropriately classified for the

image that had numerous subjects in the background. The results proved that there was no single optimal filter that can be used for various image classification and the need for an optimization algorithm. Even with a limited data set, 4 different filters or combinations each produced optimal results for at least one image. Given Google Vision's immense training set, a 5% average increase in confidence was significant and worthy of further research. An automated process would expand on this further and allow for even greater accuracy with the addition of even more filters and the numerous potential combinations that can be made.

CONCLUSIONS

The implications of this research could yield significant improvements to computer vision operations by increasing system levels of confidence, increasing the number of useful image features, and providing simulated images to improve training breadth, resilience, and robustness. At a minimum, the research will provide insight to optimum combinations of various image transformation methods for image pre-processing prior to neural network classification operations. The next step will be algorithm refinement and additional experiments on mass combinatorial image transformation applications and analysis of the original image and all its transformations in parallel to increase the overall accuracy of image recognition, object extraction, steganographic extraction, etc. The final algorithm would return the highest accuracy-producing image variations dependent on the selected image transformation filters.

7 FUTURE DEVELOPMENT AND FURTHER SYSTEM INTEGRATION

7.1 CURRENT DEVELOPMENT

As of the date of this publication, the SDS data processing prototype is in development. The research for this dissertation provided all the necessary preliminary and conceptual design documents to begin immediately upon award. From the onset of the prototype development, steps have been taken to prepare for full system development and integration. System expansion will focus primarily on additional data points, data transfer to HMIRS, and security measures needed to comply with DoD domain cybersecurity requirements. Digital signatures for the template and final xml along with system role restriction and dual-factor authentication were some of the groundworks laid to prepare for further DoD development. Simultaneously, efforts are underway to find a suitable hosting platform for the system for full integration allowing both an easy and efficient way for chemicals to provide this information while also applying necessary security precautions and cybersecurity restrictions to streamline this data to other systems on the highly-protected DoD domain.

7.2 FULL CHEMICAL DISCLOSURE

The timing of the prototype development coincides with a larger DoD effort to obtain full chemical disclosure for the products it procures. With the advent of recent per- and poly- fluoroalkyl substances (PFAS) gaining national attention regarding widespread water contamination problems and human health concerns, the need for full chemical disclosure has become increasingly more important. SDS hazard communication regulations do not require full chemical disclosure, only those chemicals the manufacturer deemed hazardous. These chemicals largely come from regulatory agency documented chemicals of concern. In the case of PFAS, these chemicals were

not listed by the EPA as emerging contaminants until recently. Now there are visibility concerns of the prevalence of these chemicals and where, how, and by whom they are being used. It quickly became apparent that as new research discovers additional hazards, a gap exists where these previously unregulated chemicals were not always communicated on SDSs making impact assessments much more difficult without this visibility.

EMERGING CONTAMINANT GOVERNANCE COUNCIL TECHNICAL INTEGRATION

The research that I had completed for this dissertation allowed me to become a valuable resource to the Emerging Contaminant Governance Council (ECGC) Risk Management Assessment technical integration team. While full disclosure would eventually be required of all vendors on the products the DoD procures, a technical solution would need to be employed for effective data collection and data follow through for the full chemical lifecycle of these products within the DoD. Data standardization of SDS information would be one component necessary to ensure all required data had been obtained efficiently and validated. With an XML solution in place, our OCR/AI system would provide the transition piece necessary to bring the hundreds of thousands of existing legacy SDSs into XML standardization. Full disclosure would be documented as users upload SDS and quickly allow missing chemicals to be added to existing formulations.

The other problem requiring resolution was maintaining the association of SDS to the product throughout the chemical lifecycle. QR code solutions are currently be researched on how to best integrate into DoD processes. With this, the DoD has a unique opportunity to impact the industry if it were to require chemical/SDS QR codes on all the product it procures. With the DoD being the largest procurer of chemicals in the United States, this could potentially have a profound impact on the industry itself. If this were to come to fruition, regardless of procurement method, any

chemical user (DoD or otherwise) could have immediate access to health and safety information for chemical products.

8 SUMMARY

8.1 MILESTONES

1. Document Engineering 2021 Conference, Limerick Ireland

- Presented findings for preliminary modeling and prototype efforts at the Doc Eng '21 conference (presented virtually).
- Engineering of an Artificial Intelligence Processing System for Environmental, Health, and Safety Compliance published in the DocEng21 conference proceedings by the Association for Computing Machinery.

2. Filed patent request with the U.S. Patent office.

- Patent request was submitted to and accepted by the Air Force Research Laboratory legal office.
- Patent filed June 6, 2022, under patent # 17/832800; currently patent-pending.

3. Awarded \$280,000 award for novel A.I. SDS processing system.

- A proposal was submitted to and accepted by the 711th Human Performance Wing Studies and Analysis program.
- Contract coordinated through U.S. Army Corps of Engineers.

4. Society of Environmental Toxicology and Chemistry (SETAC) 43rd annual North America presentation and abstract publication.

- Presented chemical management system optimization strategies at the Society of Environmental Toxicology and Chemistry (SETAC) 43rd annual North America.
- Published abstract in the SETAC conference proceedings.

5. Society of Environmental Toxicology and Chemistry (SETAC) 43rd annual North America poster publication.

- Co-authored a poster for CO₂ prediction at a cement manufacturing plant using manufacturing data at the SETAC 43rd annual North America.
- Poster published with the SETAC conference proceedings.

6. Archiving 2022 conference presentation and publication.

- Published article for Enhanced Computer Vision using Automated Optimized Neural Network Pre-processing.

7. Journal of the American Chemical Society Omega publication.

- Published article for the Mitigation of Chemical Reporting Liabilities through Systematic Modernization of Chemical Hazard and Safety Data Management System.
- Article routed through various levels of DoD leadership.

8. Participation of Emerging Contaminant Governing Council (ECGC) Risk Management Assessment (RMA) technical integration team.

- Incorporated ideas from research into chemical hazard communication improvement efforts for full product chemical disclosure and precise association of SDSs to products.
- Pilot program to be initiated using proposed research strategies.

9. Project co-lead for proposed novel system development using research documentation for conceptual design.

- Kickoff meeting occurred January 2023.
- To date, 21 fields have been accurately extracted from SDSs.
- Projected completion date: June 2023.

8.2 CONCLUSION

The research proved that the novel meta-algorithmic approach was effective in SDS data processing. The research and incorporation of system engineering principles provided the foundation for system development to begin and a patent pending for the novel approach. The research also led to five publications, a contract award for system development, and incorporation of the research findings in a larger effort for DoD hazard communication improvement efforts. Additionally, the system optimization technologies all represented possibilities in employing 21st technologies to increase safety and minimize human and environmental exposures. Strategies such as these can be used to minimize greenhouse gas emissions and our chemical footprints by getting a more accurate account of the products we use and translating this into actionable data. These process additions would effectively allow us instant HAZCOM in emergencies and more efficient, accurate, and robust HAZCOM allowing us access to the latest changes in regulatory tracking and chemical research. Proactivity would also extend to safety measures informing personnel of chemical temperature or incompatibility warnings before reaction and adding efficiency gains for

chemical lifecycle tracking. While each of the discussed methods can be employed either as a combined system or independently in existing industrial management processes, all of which would benefit from a centralized and universal SDS repository providing an easily accessible and standardized data source while minimizing data quality issues garnered from data transfers. Applications like machine learning and the proposed meta-algorithmic approach provide a quicker and more efficient means by which to transition to fully automated HAZCOM. The number of hazardous communication failure findings remains staggering and associated hazard data quality loss and availability is insufficient and unacceptable for emergencies, both immediate and the longer-term climate change emergencies. We have the means to currently implement these suggestions and move the industry forward; it only requires action.

This research will hopefully promote this necessary change by incorporation of these methods in the processes of the largest user of hazardous materials in the United States.

BIBLIOGRAPHY

- Abas, A., Yusof, Y., Din, R., Azali, F., & Osman, B. (2020). Increasing Data Storage of Coloured QR Code Using Compress, Multiplexing, and Multilayered Technique. *Bull. Electr. Eng. Inform.*, 2555-2561.
- Administratoiu, U. S. (2022). *GSA Green Procurement List*. Retrieved from GSA Green Procurement List: <https://sftool.gov/greenprocurement>
- Babich, N. (2020). *What is Computer Vision & How Does it Work? An Introduction*. Adobe. Retrieved from <https://xd.adobe.com/ideas/principles/emerging-technology/what-is-computer-vision-how-does-it-work/>
- Bernstein, J. (2002). Material Safety Data Sheets: Are they reliable in identifying human hazards? *Journal of Allergy and Chemical Immunology*.
- Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Cambridge: Cambridge.
- Chakraborty, S. L. (2014). Extraction of (Key Value) Pairs from Unstructured Ads. *2014 AAAI Fall Symposium*.
- Das, R. T. (2016). Framework for Content-Based Image Identification with Standardized Multiview Features. *ETRI Journal* vol. 38, 174-184. doi:10.4218/etrij.16.0115.0102
- Defense, U. S. (1999). *Storage and Handling of Hazardous Materials, DLAI 4145.11*. United States Department of Defense.
- DeMasi, A. E. (2022). Safety Data Sheets: Challenges for Authors, Expectations for End Users. *ACS Chemical Health and Safety*.

- Dictionary, C. (2023, 03 02). *Cambridge Dictionary*. Retrieved from <https://dictionary.cambridge.org/us/dictionary/english/machine-learning>
- Evelina Maria De Almeida Neves, A. G. (1997). "A Multi-Font Character Recognition Based on its Fundamental Features by Artificial Neural Networks". *IEEE*.
- Gaubatz, M., & Simske, S. (2010). Towards a feature set for robust printing-imaging cycle device identification using structured printed markings. *IEEE International Workshop on Information Forensics and Security*, (pp. 1-6).
- Glass, D. S. (2006). The Challenges of Exposure Assessment in Health Studies of Gulf War Veterans. *Phil. Trans. R. Soc.*
- He, J. D. (2005). A Comparison of Binarization Methods for Historical Archive Documents. *IEEE*.
- IBM. (2023). *What is Computer Vision?* Retrieved from <https://www.ibm.com/topics/computer-vision>
- Jyotsna, S. C. (2016). Binarization Techniques for Degraded Document Images - a Review. *2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)* (pp. 163-166). Noida.
- Kosaaikoff, A. S. (2003). *Systems Engineering Principles and Practice*. John Wiley & Sons.
- Krizhevsky, A. S. (2017). Imagenet Classification with Deep Convolutional Neural Networks. *Neural Networks*, 84-90.
- Kuzmina, O. H. (n.d.). Chemical Management: Storage and Inventory in Research Laboratories. *ACS Chemical Health*.
- Langerman, N. (1995). Material Safety Data Sheets - Who Uses Them? *Chemical Health & Safety*, 26-29.

- Liu, Y. H. (2019). Anti-Noise Image Source Identification. *Concurrency and Computation: Practice and Experience* vol. 31. doi:10.1002/cpe.5104
- Milyaey, S. B. (2020, November 8). *Image Binarization for End-to-End- Text Understanding in Natural Images*. Retrieved from Information Age: www.information-age.com/optical-character-recognition-tools-ocr-ai-123479324/
- Moore, C., Davis, K., Spear, S., & Bombaci, S. (2021). Link and Learn: How to Use QR Codes to Communicate Science. *Colorado State University*.
- Munsayac, F. A. (2017). Implementation of a normalized cross-correlation coefficient-based template matching algorithm in number system conversion. *HNICEM*, 1-4.
- National Institute of Environmental Health Sciences. (2022). *Chemical Exposure*.
- Neuron Synthesizer ANN. (2019). Retrieved from Sylvain Kepler: http://sylvain.kepler.free.fr/studio/html/neuron_synthesizer_ann.htm
- OpenCV. (2020). *Template Matching*. Retrieved from OpenCV.org: https://docs.opencv.org/3.4/de/da9/tutorial_template_matching.html
- Otsu, N. (1979). A Threshold Selection Method from Gray Level Histograms. *IEEE Trans. Systems, Man and Cybernetics* vol. 9, 62-66.
- Pais, S., & Symonds, J. (2011). Data Storage on a RFID Tag for a Distributed System. *Int. J. UbiComp*, 26-38. doi:10.5121/iju.2011.2203
- Picard, J., Landry, P., & Bolay, M. (2021). Proceedings of the 21st ACM Symposium on Document Engineering (DocEng '21). *Association for Computing Machinery*, (pp. 1-4). New York.
- Salton, G., & McGill, M. (1986). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, Inc.

- Salton, G., Wong, A., & Yang, C. (1975). A vector space model for automatic indexing. *Commun. ACM*, 613-620.
- Sharma, B., Willis, V., Huettner, C., Beaty, K., Snowdon, J., Xue, S., . . . Michelini, V. (2020). Predictive article recommendation using natural language processing and machine learning to support evidence updates in domain-specific knowledge graphs. *JAMIA Open*, 332-337.
- Simske, S. (2013). *Meta-Algorithmics: Patterns for Robust, Low Cost, High Quality Systems*. Wiley-IEEE Press. doi:978-1-118-62669-6
- State of Pennsylvania. (2022). *Chemical Calculator (state.pa.us)*. Retrieved from Pennsylvania Department of Environmental Protection: state.pa.us
- Sweet, A. K. (2003). *Systems Engineering Principles and Practice*. 1.
- Tolley and Tolley v. ACF Industries, Inc. et al (2002).
- United Nations General Assembly. (2019). *Principles on Human Rights and the Protection of Works from Exposure to Toxic Substances, Human Rights Council 42nd Session*.
- United States Environmental Protection Agency. (2022). *Emerging Contaminants and Federal Facility Contaminants of Concern*. Retrieved from <https://www.epa.gov/fedfac/emerging-contaminants-and-federal-facility-contaminants-concern>
- United States Environmental Protection Agency. (2022). *TRI-Listed Chemicals*. Retrieved from <https://www.epa.gov/toxics-release-inventory-tri-program/tri-listed-chemicals>
- United States Environmental Protection Agency. (2022). *TSCA Chemical Substance Inventory*.
- United States Occupational Health and Safety Administration. (2019). *2019 Toxic Release Inventory National Analysis*.

- United States Occupational Health and Safety Administration. (2012). *OSHA Brief Hazard Communication Standard*.
- United States Occupational Safety and Health Administration. (2019). *OSHA Top 10 Cited Standards*. Retrieved from <https://www.osha.gov/top10citedstandards>
- United States Occupational Safety and Health Administration. (2022). *Hazardous Communication Standard*.
- Wang, R. L. (2019). Blur Image Identification with Ensemble Convolution Neural Networks. *Signal Processing*, 73-82. doi:10.1016/j.sigpro.2018.09.027
- Yu Bei, P. D. (2015). Machine Learning and Pattern Matching in Physical Design. *20th Asia and South Pacific Design Automation Conference*.
- Yu, K. (2020). Your Client Is In Flames, The Structure is In Flames, Now What? *Plaintiff Magazine*, 6.
- Zhong, S. Z. (2021). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. *Environmental Science and Technology*, 1-14.

APPENDIX A – PATENT APPLICATION

APPARATUS FOR AUTOMATED SAFETY DATA SHEET PROCESSING

STATEMENT OF GOVERNMENT INTEREST

The invention described herein may be manufactured and used by or for the Government for governmental purposes without the payment of any royalty thereon.

BACKGROUND OF THE INVENTION

This invention relates generally to the field of hazardous substance safety. More specifically, this invention relates to the identification and safe handling procedures for those hazardous substances found in and utilized in the workplace. More specifically yet, this invention relates to efficient systems and methods designed to automate the voluminous amount of information related to the identification of hazardous substances, their contents, and their safe handling procedures.

Background. Environmental, Safety, and Occupational Health (ESOH) program managers employ system safety risk management standard practices to identify, assess, and mitigate environmental, safety, and occupational health risks, and authorize hazardous materials use on U.S. Government installations including USAF installations. The process aims to ensure that the Air Force provides safe and healthful workplaces and conducts operations that minimize risk to mission accomplishment. At the same time, the Air Force preserves resources, protects the environment, and safeguards military and civilian personnel and the public.

However, workplace hazard communication violations are still among the top citations during Environmental, Health, and Safety (EHS) inspections by regulatory agencies. Precise Safety Data Sheet (SDS) selection, fast retrieval of SDSs during emergencies, SDS data management, and material tracking remain areas that can be improved.

Hazardous chemical ingredients are prevalent in many of the products we use day-to-day in both household and industrial applications. The potential hazards and impacts of these chemicals on human health and the environment are primarily communicated to the public through Safety Data Sheets (SDSs) from the chemical vendors or distributors. These documents provide a standardized approach for how and what information is provided to product users to assist them with assessment of precautionary measures, hazard mitigation, emergency response or cleanup procedures, and ESOH management; including many regulatory-driven sections to effectively communicate these hazards to the chemical users.

The U.S. Occupational Health and Safety Administration (OSHA) listed failures in hazard communication as the second highest most frequently cited standard. 3,624 HAZCOM enforcement citations occurred in fiscal year 2019 totaling approximately \$4,682,380 in proposed penalties. Each year in the U.S., thousands of workers become sick from workplace chemical exposure with as many as 50,000 people dying each year from adverse effects of long-term chemical exposure. Unfortunately, many current procedures still include maintaining binders of printed SDSs or available copies downloaded on neighboring workstations. These methods are consistently found during inspections to include numerous records that are the incorrect SDS for specific products or are out-of-date, incomplete, illegible, etc.

Implementation of OEH processes and principles is defined to include mandated use of the Defense Occupational and Environmental Health Readiness System – Industrial Hygiene

(DOEHRS-IH). This meets the requirement for longitudinal evaluation and documentation of “quantifiable data on personal occupational, environmental, and deployment-related exposures of all service personnel (active duty, reservist, National Guard) throughout their military careers and after leaving military service”, in accordance with Presidential Review Directive 5, Improving the Health of Our Military, Veterans, and Their Families and DoDI 6490.03, Deployment Health, which requires the creation and maintenance of an exposure assessment record for each Airman’s full career.

Downstream users rely on the data from these SDSs for environmental, safety, and occupational health compliance. PDF remains the predominant format for the millions of SDSs in circulation today. For ESOH systems that require loading of SDS for exposure calculations, environmental reporting calculations, etc., non-ESOH personnel are frequently used to find the correct SDS for the products they are using and often not provided sufficient training to determine whether SDSs are GHS compliant and the proper version for the product. Personnel must typically hand enter information into their respective compliance systems. Each time manual data entry is performed, potential for data quality errors increases, subsequently increasing the potential for compliance liability. These pdfs often also require validation to ensure all regulatory fields have been provided and additional follow up with chemical manufacturers for clarification on data fields or requesting additional information for necessary compliance calculations.

Moreover, prior to the use of any hazardous chemicals in a workplace, common practices typically require some level of authorization and approval for use by the EHS personnel. In order to accurately determine impacts on human health and safety, and the environment, SDSs are typically reviewed to assess hazards and ingredients. These hazard assessments allow the EHS personnel to determine appropriate Personal Protective Equipment, exposure limits, need for engineering

controls, potential environmental impacts or regulatory reporting concerns, etc. SDSs are obtained through various ways. For large contracts, a chemical vendor will often provide all contracted material SDSs per contract requirements with the product prior to shipment. For hazardous material procured through local economic methods, online services, etc., the industrial personnel are sometimes required to provide the SDS for the products they will use to the EHS team for analysis. If no comparison is done between the SDS and product by the EHS team, potential liability has already emerged by SDS selection by minimally or untrained non-EHS personnel. Mistakes are common and easy to make if a person isn't validating the proper trade names and product codes and ensuring that the revision date is the applicable version for that SDS. Formulations often change over time and the selection of an SDS seven years old may have considerably different chemical formulations leading to significant compliance concerns with incorrect EHS assessments and other potential liabilities such as mischaracterization of waste. Direct vendor communication of SDSs using XML or equivalent transfer methods is needed to eliminate the need for separate manual data entry and current outdated communication methods.

OBJECTS AND SUMMARY OF THE INVENTION

Before an essential product is introduced into a process, precise SDS selection for products used in the workplace is of vital importance for EHS assessments. The hazards communicated and the ingredients listed form the basis of occupational health and safety assessments, calculations for environmental reporting, and how emergency response staff addresses and responds to spills and accidents.

For EHS systems that require loading of SDS for exposure calculations, environmental reporting calculations, etc., non-EHS personnel are frequently used to find the correct SDS for the products they are using and often not provided sufficient training to determine whether SDSs are GHS compliant and the proper version for the product.

In order to retrieve SDSs for chemical products used in the workplace, chemical users are largely responsible for finding the appropriate SDS that corresponds with each product used. SDSs are commonly housed on manufacturer websites and differ from site to site and further confusion can be introduced as distributors can sometimes create their own SDSs, making product matches difficult. This often leads to the lack of immediate access of SDS to personnel in a workplace.

Additionally, pdf (Portable Document Format) remains the predominant format for the millions of SDSs in circulation today. These pdfs often also require validation to ensure all regulatory fields have been provided and additional follow up with chemical manufacturers for clarification on data fields or requesting additional information for necessary compliance calculations. Direct vendor communication of SDSs using XML or equivalent transfer methods is needed to eliminate the need for separate manual data entry and current outdated communication methods.

Among U.S. Government agencies the USAF, for example, receives, reviews, and transcribes product hazard data (PHD) from approximately 2,500- 4000 Safety Data Sheets (SDS) monthly through the Hazardous Material Management Process (HMMP) team consisting of representatives from Environmental, Safety, and Occupational Health. The SDS and PHD is assessed for its use throughout the lifecycle of the product from cradle to grave. The product's data is manually reviewed, hand entered and assigned to their respective compliance systems. This manual process introduces unstandardized interpretation of hazards, relying on individual

judgement, which creates inefficiencies to cost, data quality, and timeliness to meet the readiness mission.

Additionally, other downstream users rely on the data from these environmental, safety, and occupational health compliance systems and if not properly integrated, must also hand enter this information into their respective systems. Each time manual data entry is performed, potential for data quality errors increases, subsequently increasing the potential for compliance liability.

The USAF, the Department of Defense (DoD), and industrial chemical institutions, are required to meet the demands of Federal Laws and Regulations, Presidential and Congressional inquiries, as well as Environmental, Safety, and Occupational Health (ESOH) needs.

The current process for collecting, analyzing, and assessing SDS PHD data is ineffective and inefficient. The motivation for the present invention, therefore, is the reduction and/or elimination of the manual efforts associated with reading, analyzing, translating, and validating SDS data by leveraging current technologies to convert SDSs into machine-encoded text for database use. Data stewarding efforts for the loading of SDSs are currently paid for by the Air Force, Defense Logistics Agency, and other services for occupational health, safety, and environmental tracking and reporting needs. For the DoD as a whole, this invention is expected to save over \$3 million annually. While this invention can significantly reduce costs and increase operational efficiency for the DoD, this invention will also have implications in the civilian sector as it could reduce their data entry costs as well and help move the entire industry to an electronic standard.

It is therefore an object of the present invention to provide a

It is a further object of the present invention to provide a

It is still a further object of the present invention to provide a

It is yet still a further object of the present invention to provide

An additional object of the present invention is to provide a

Briefly stated, the present invention achieves these and other objects through

According to an embodiment of the present invention, a

According to another embodiment of the present invention,

The above and other objects, features and advantages of the present invention will become apparent from the following description read in conjunction with the accompanying drawings, in which like reference numerals designate the same elements.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring to FIGURE 1 The invention presented here serves two main functions. First, the invention translates the entire SDSs from numerous chemical vendors, and in various formats (.pdf, .doc, .txt, .jpg, .gif, .png, etc), each with distinct formatting (linebreaks, paragraphing, spacing, logos, headers, etc.) **10**, to machine-encoded text **20** with a high degree of accuracy and precision. The invention then "reads" and assess documents as a human would; ensuring that the documents are compliant, ensuring reported values are within expected thresholds (FIGURE 2, **130**), and that there are no conflicts in hazardous material classification, and comparing to similar products for more environmentally friendly alternatives.

Referring to FIGURE 2, the invention's trained neural networks **30** provide a base to allow the invention to "learn" and appropriately classify **40** values and calculate statistical probabilities for output accuracy and precision. Meta-algorithmic processing (see FIGURE 1, **70**) though algorithms, functions of these algorithms, and combinations thereof are employed to refine classification **40** until the appropriate level of confidence has been reached or no higher confidence

can be reached within the capabilities of the approaches taken. Once classified, parsed text segments are scrubbed to remove outlier characters and are validated **50** (see also FIGURE 1, **90**) through queries **60** against a Globally Harmonized System (GHS) schema to ensure numerical values are within expected thresholds **130**, mandatory components have been identified, and desired calculations are achievable given the provided data. A machine-encoded file **80** (see also FIGURE 1, **20**) is then produced from the data.

Optical Character Recognition and Neural Networks

For the millions of SDSs that still reside in pdf form, one obstacle of many EHS systems is the need to manually enter chemical information from SDSs into their respective systems for reporting calculations and tracking needs. Still referring to FIGURE 2, the present invention's use of Optical Character Recognition **100** allows for these documents to be broken down to machine-encoded text **80** (see also FIGURE 1, **20**) using a variety of prominent OCR tools (e.g. Adobe Pro, Abbey, and Google Tesseract). One difficulty with SDS documents is the vast differences in format from manufacturer to manufacturer. Although GHS and REACH regulations have provided some structure in required sections, how and where the data is relayed to readers varies widely. Artificial Neural Networks **30** allow the present invention to take OCR **100** another step further and use machine learning applications to provide structure to unstructured data sets and essentially learn to "read" an SDS and parse desired fields as required. The benefits of incorporating this technology in the invention can yield significant time and monetary savings for organizations that require data entry teams to manually load thousands of these documents each year. It is within the scope of the present invention to have additional configurations with additional algorithms 1, 2, ..., n **110** added for alternate data extraction and validation uses. As regulatory needs change, the invention could easily be adapted to new text identification and classification, regulatory validation needs, and

calculations. Additionally, if SDS or other relevant hazard or chemical information become available on product barcodes or other labeling, the invention can be expanded to receive incoming barcode/label submissions as well as generate outgoing barcodes/labels **120**. The invention can also be expanded to accept xml or equivalent files and store together with OCR-derived SDSs. Also, the invention can be tailored for additional environmental compliance documents such as chemical product technical or specification documents, hazardous waste manifests, etc. as these documents become available. Finally, the invention can incorporate Natural Language Processing (NLP) to receive updates from applicable web sites (e.g. chemical vendors, regulatory agencies, reputable news sources) of articles/publications involving the specified product/chemical.

OCR Language Translation Services

If the SDS is in a foreign language or parts of the SDS are in a foreign language, the present invention translates the SDSs into the desired host language. Using parallel processing paths depicted in **FIGURE 2**, the invention then performs Optical Character Recognition (OCR) **100** on the documents and image recognition **120** on contained images. Artificial neural networks **30** are then used to identify patterns and levels of confidence of document attribute classifications. Additionally, multiple algorithms **110** including machine learning key-value pattern arrays, syntax learning, and tessellation and recombination compare **130** against neural network identified attributes and improve confidence in classification **40**.

SDS Pictogram Image Recognition, Classification, and Validation

Similar to the use of neural networks for SDS text parsing and classification, neural networks are also used by the present invention for image recognition of GHS and other regulatory pictograms on an SDS. For images **10**, neural networks **30** and normalized cross-correlation is used to classify images. Once text has been classified to the greatest extent of confidence, validation queries **60**

are used to validate numerical fields, calculate Hazard Characteristic Codes (HCC), validate completeness of the documents, and validate classifications **50** against pictogram images **10** contained in the document. For instance, an ANN performs recognition on a GHS pictogram and classifies **40** it as a corrosive pictogram. Validation **50** occurs through the text recognition of the pH and determine whether the pH value accurately supports this classification. Lastly, the present invention encrypts **80** the document file to ensure alterations do not occur in data transfer. This is used to ensure proper labelling and validation of the SDS itself.

Although serving as the enabling technology of the present invention, the utility of the present invention extends beyond Safety Data Sheet (SDS) recognition and interpretation. Referring to FIGURE 1, the output encrypted document file (FIGURE 2, **80**) once classified and validated **90** (see also FIGURE 2, **50**) is uploaded into an SDS repository **140**. The invention thereafter employs several labeling technologies to incorporate SDS information on the products themselves. Specifically, precise SDS selection and information availability to a user can be realized by the present invention by moving from traditional legacy product barcodes to several modern applications than can be used to quickly direct users to the product's SDS. The present invention utilizes several of these modern applications in conjunction with product labeling **150**.

In the advent of the COVID-19 global pandemic, the use of QR codes has seen a dramatic rise with many restaurants replacing hard copy menus with digitally-accessible versions online through the use of common cell phone camera applications. The present invention utilizes QR code labels **160** which are available in two forms – static and dynamic, dependent on whether a user wants data essentially hardcoded into a barcode or through the use of a URL that can be modified after printing. Numerous QR code generators exist for chemical vendors or EHS personnel to create

these codes at little to no additional cost. A simple mobile application could be used to transmit an embedded SDS directly to the present invention's universal host SDS repository **180**.

Chemical product labelling can be difficult, however, regardless of size products still must provide sufficient hazard, use, and ingredient information and other relevant product information. Product real estate can be sparse and not offer manufactures room for other desirable information such as SDS links or direct SDS information. The present invention reads IR-transparent barcodes **170** and can transmit the informational contents therein to the invention's universal host SDS repository **180**. This feature offers an additional barcoding and/or labeling solution by allowing vendors to provide this additional information without compromising on required or other valuable information.

Hazardous material and hazardous waste management typically requires a database for large operations to accurately and consistently monitor where containers are and closely observe the amount of time containers exist at these sites and approach regulatory thresholds. The present invention's use of RFID technology can be used to eliminate the need to track hazardous material containers from supply points to end users or waste containers as they move from satellite accumulation points (SAP) to centrally managed hazardous waste storage areas (HWAS) or treatment, storage, and disposal facilities (TSDFs). The invention's RFID tags or labels **190** eliminate the need for line of sight between tags and readers allowing expedited inventory management and accountability. Microchips in RFID tags can be either read-only or read-write. The latter feature allows users of the present invention to add data to the RFID tag **190** such as the waste profile number for the waste, the accumulation start date and the site-specific regulatory start date. Likewise, for hazardous materials, SDS IDs, expiration dates, etc. could be loaded against the RFID tag **190**. The benefits of this feature of the invention yields significant time

savings of personnel, monetary savings of reduced data entry and tracking, and increased EHS accountability of hazardous material containers. RFID tag information is likewise uploadable to the present invention's universal host SDS repository **180**.

Still referring to FIGURE 1, although some repositories such as SDS.com exist, these are not all inclusive and can require membership payments. Chemical users should be entitled a single centralized universal repository for all SDS with easy and intuitive access. For an effective Universal Safety Data Sheet system, the present invention accommodates the many possible ways SDSs can be received and how these varying SDS submittals methods can be consolidated into a single management approach. In a preferred embodiment, SDSs are transmitted via a user interface directly to the invention's universal host SDS repository **180** in a standardized xml or equivalent data transfer form.

The present invention's user interface to the universal host SDS repository **180** provides an additional benefit for new SDS creation whereby smaller chemical vendors can create GHS-compliant SDSs directly in the system. Likewise, even within the DoD, lab research is performed where new chemicals and products are created requiring the creation of a corresponding SDS. This feature of the invention would accommodate these situations and allow for the SDS data to be stored in the universal host SDS repository **180** via same user interface feature as are the standardized xml and OCR-neural network processed submittals. Additionally, this user interface feature of the present invention allows other systems using SDSs to communicate the information directly to the universal host SDS repository **180** via following the same XML or equivalent data standard. Also, user interface feature can be further optimized with hazardous material manufacturer's support via embedding of SDS links within the product barcodes to allow instant SDS access utilizing QR code readers found on any modern smart phone. In the event of an

emergency, rather than logging on to a computer or looking through binders for the appropriate SDS, the specific SDS associated with a hazardous material accident can be produced instantly with ubiquitous handheld smartphone technology.

What is claimed is: An apparatus for automated processing of hazardous material safety data, comprising:

an optical character recognition device for translating hazardous material safety data sheet information to digital data;

an image recognition device for translating hazardous material safety data sheet pictograms into digital data;

a device capable of using Natural Language Processing (NLP) for context-specific notifications, warnings, and information

a device for extracting as digital data hazardous material safety data sheet information from any one of the group of hazardous material label types consisting of: radio frequency identification (RFID) labels, QR Code labels, infrared labels or a combination thereof;

a device for creating hazardous material label types consisting of either RFID labels, QR codes labels, infrared labels, or a combination thereof

a processor device that, when executing computer-implementable instructions, performs the steps to:

classify said digital data;

calculate statistical probabilities for the accuracy and precision of said digital data;

calculate statistical probabilities for the accuracy and precision by meta-algorithmic comparison and validation

calculate the Hazard Characteristic Code (HCC) that will be based upon SDS data points and ultimately used for labeling and storage incompatibility monitoring

refine said classification of said digital data to ensure a confidence level is attained;

validate said digital data through queries against a Globally Harmonized System and Host SDS repository schema;

create a machine-encoded text file from said validated digital data; and

upload said machine-encoded text file to a hazardous material safety data repository.

The apparatus of claim 1, wherein said computer-implementable instructions further cause said processor device to perform the steps to compute and validate at least one of a plurality of calculations based upon a meta-algorithmic computation from one or more individual algorithmic methods on said digital data so as to build upon any individual algorithmic application and increase system accuracy.

The apparatus of claim 2 wherein any one of said plurality of said meta-algorithms is calculated based upon results from the selected group of algorithms consisting of:

a conventional neural network parsed text field algorithm;

a key-value pattern array parsed text algorithm;

a parsed text derived from natural language syntax and segmentation algorithm;

a tessellation and recombination algorithm; and

a language translation algorithm.

The apparatus of claim 3, wherein said one or more artificial neural networks comprise one or more of said meta-algorithms.

The apparatus of claim 4, wherein said one or more artificial neural network comprises functions of one or more of said meta-algorithms.

The apparatus of claim 5, wherein said step of validating comprises ensuring that:

- numerical values are within expected thresholds;
- mandatory components have been identified; and
- desired calculations are achievable for said digital data.

The apparatus of claim 1, wherein said machine-encoded text file is encrypted prior to said step of uploading to said hazardous material safety data repository.

The apparatus of claim 1, wherein said step of extracting said digital data from QR code labels comprises extracting SDS data directly embedded in said QR code label.

The apparatus of claim 1, wherein said step of extracting said digital data from QR code labels comprises extracting SDS data indirectly from a URL embedded in said QR code label.

The apparatus of claim 1, further comprising a device for adding to said RFID label any one of the descriptors selected from the group consisting of:

waste profile number for hazardous waste;

accumulation start date for hazardous waste;

site-specific regulatory start date for hazardous waste;

SDS identification number for hazardous materials; and

expiration date for hazardous materials;

storage temperatures;

SDS links;

Hazard Characteristic Codes,

And Container numbers.

The apparatus of claim 9, wherein SDS data is accessed from said QR code label via a smartphone QR code reader.

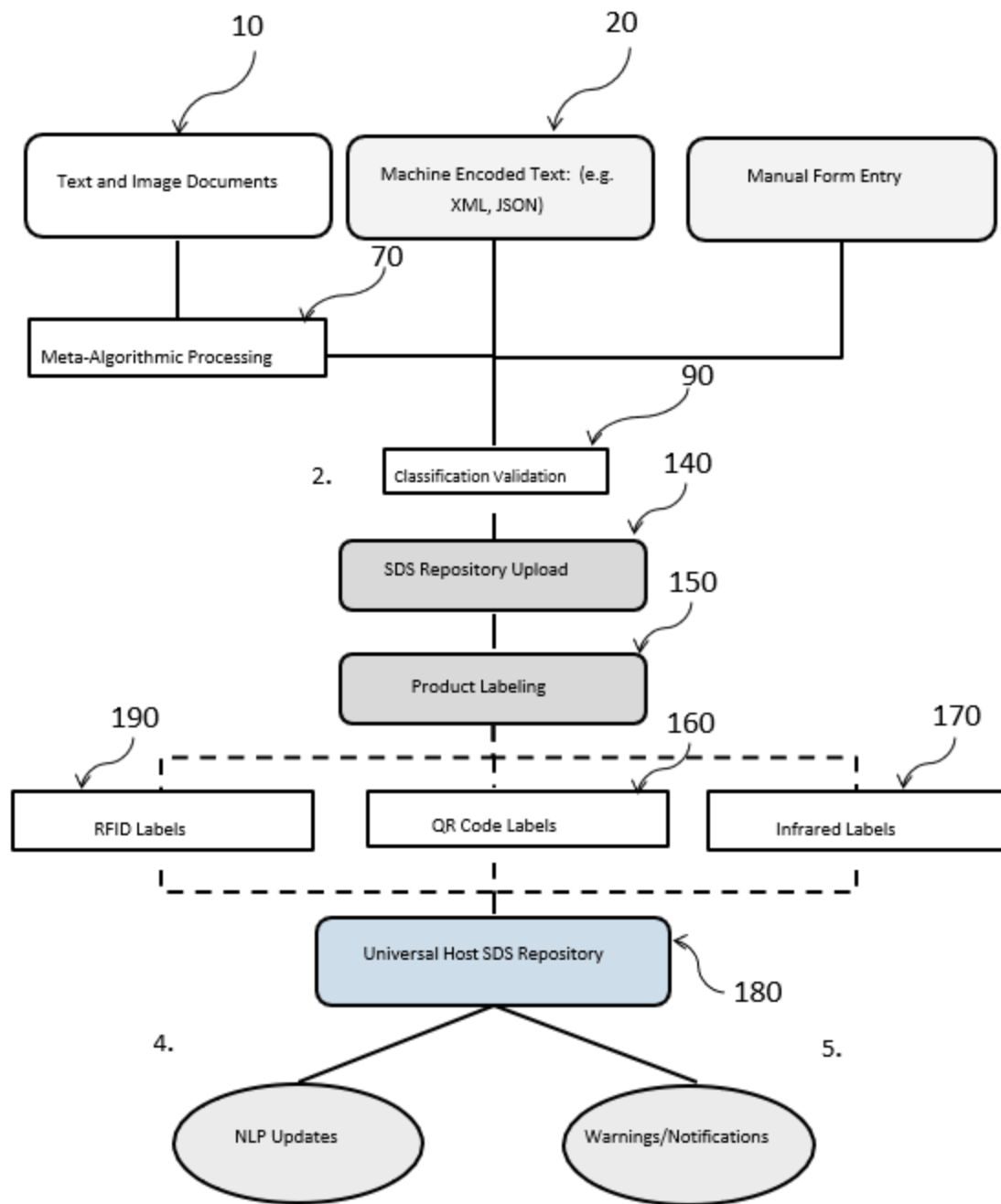


FIGURE 32 PATENT IMAGE 1

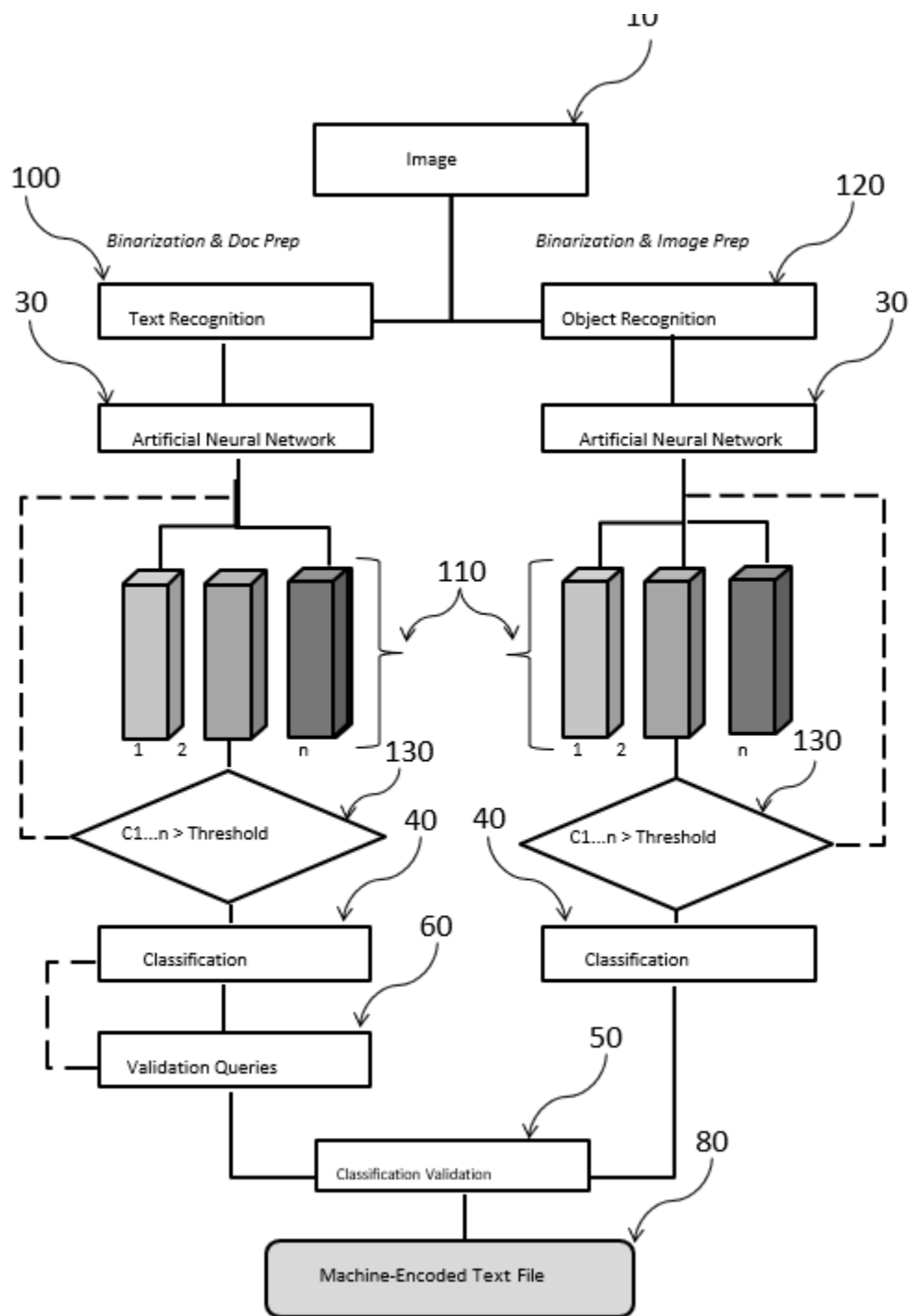


FIGURE 33 PATENT IMAGE 2

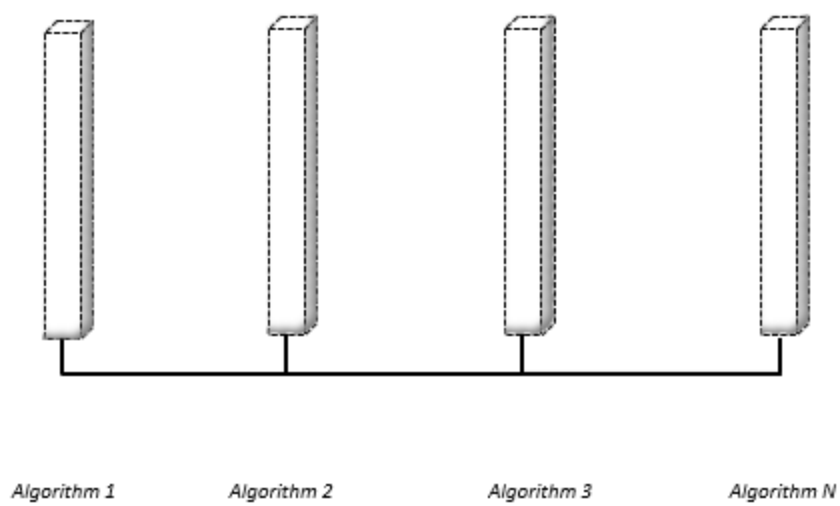




FIGURE 34 PATENT IMAGE 3

APPENDIX B: CSU AWARD



CESWF Grant or Cooperative Agreement Award

1. INSTRUMENT TYPE: Cooperative Agreement		2. AUTHORITY: <input checked="" type="checkbox"/> 16 USC 670c-1 Sikes Act <input type="checkbox"/> 10 USC 2701 DERP <input type="checkbox"/> 10 USC 2684a Preserve Habitat		<input type="checkbox"/> 10 USC 2684 Cultural Resources <input type="checkbox"/> 10 USC 4001 R&D <input type="checkbox"/> 33 USC. 2339 WRDA	
3a. AGREEMENT NUMBER: W9126G-22-2-0042		4a. MODIFICATION NO:		5. ISSUED BY: USACE-SWF	
3b. MASTER AGREEMENT: Rocky Mountain CESU Region		4b. Reserved		7. ASSISTANCE LISTING # (CFDA): <input checked="" type="checkbox"/> 12.005 CESU	
6. DODAAC: W9126G		8. PROGRAM MANAGER: David Leptien			
9. PROGRAM / PROJECT TITLE: Prototype of Novel Method for Collection of Natural Resource and Human Protection Information from Chemical Safety Data Sheets (CSDS).					
10a. ISSUED TO (RECIPIENT'S) ADDRESS: Colorado State University Office of Sponsored Programs 601 Howes St. Ft. Collins CO 80521-2807 POC: Kellie Reifstenzel Phone: (970)-491-6684 Email: kellie.reifstenzel@colostate.edu			10b. RECIPIENT TYPE (CHOOSE ONE APPROPRIATE BOX): <input type="checkbox"/> State Gov't <input type="checkbox"/> Local Gov't <input checked="" type="checkbox"/> Institute of Higher Education <input type="checkbox"/> Hospital <input type="checkbox"/> Other Nonprofit Organization <input type="checkbox"/> For Profit (Lg Business) <input type="checkbox"/> For Profit (Sm Business) <input type="checkbox"/> County Gov't <input type="checkbox"/> Municipal Gov't <input type="checkbox"/> Private Higher ED Institution <input type="checkbox"/> Small Business <input type="checkbox"/> Other (Specify):		
10c. RECIPIENT'S CAGE CODE: 4B575					
10d. RECIPIENT'S UNIQUE ENTITY ID: LT9CXX8L19G1					
10e. RECIPIENT'S TAX ID NUMBER: 46000545					
11. ORIGINAL PERIOD OF PERFORMANCE START DATE: 22 September 2022					
12. CURRENT FUNDED PERIOD OF PERFORMANCE END DATE: 21 June 2023					
13. FUNDING:		This action awards a new Cooperative Agreement where all services shall be performed in accordance with the accepted Project Proposal, Master CESU Agreement and Terms and Conditions			
PREVIOUS	\$				
THIS ACTION	\$280,000.00				
TOTAL FUNDED	\$280,000.00				
15. APPLICABLE ENCLOSURES(S), IF CHECKED: <input type="checkbox"/> STATEMENT OF OBJECTIVES <input checked="" type="checkbox"/> TERMS and CONDITIONS <input checked="" type="checkbox"/> BUDGET <input type="checkbox"/> OTHER <input checked="" type="checkbox"/> PROPOSAL NOTE: All agreement terms and conditions not specifically changed via amendment/modification remain in effect for all requirements in the SOO, including any added via amendment/modification.					
16. IMPORTANT: Recipient <input type="checkbox"/> is not, <input checked="" type="checkbox"/> is required to sign this document and return 1 copy to the issuing office. IN WITNESS WHEREOF, the parties by their authorized representatives execute this Cooperative Agreement/Modification and agree to the terms and conditions contained herein, all assurances and certifications made in the application, and all applicable federal statutes, regulations, and guidelines. The Recipient agrees to administer the funded program in accordance with the approved application and budget(s), supporting documents, and other representations made in support of the approved application.					
17. FOR THE RECIPIENT (IF APPLICABLE) NAME AND TITLE OF AUTHORIZED SIGNER/ SIGNATURE / DATE  Digitally signed by Kellie Reifstenzel Date: 2022.09.15 15:45:19 -06'00'			18. FOR THE UNITED STATES OF AMERICA GRANTS OFFICER / SIGNATURE / DATE  Digitally signed by AUSTIN.ALICE.MILNER.1076269790 Date: 2022.09.15 17:59:30 -05'00'		

24 August 2022 AA