Engineering of an Artificial Intelligence Safety Data Sheet Document Processing System for Environmental, Health, and Safety Compliance

Kevin Fenton†
Systems Engineering
Colorado State University
Fort Collins, CO USA
kevin.fenton@colostate.edu

Steven Simske
Systems Engineering
Colorado State University
Fort Collins, CO USA
steve.simske@colostate.edu

ABSTRACT

Chemical Safety Data Sheets (SDS) are the primary method by which chemical manufacturers communicate the ingredients and hazards of their products to the public. These SDSs are used for a wide variety of purposes ranging from environmental calculations to occupational health assessments to emergency response measures. Although a few companies have provided direct digital data transfer platforms using xml or equivalent schemata, the vast majority of chemical ingredient and hazard communication to product users still occurs through the use of millions of PDF documents that are largely loaded through manual data entry into downstream user databases. This research focuses on the reverse engineering of SDS document types to adapt to various layouts and the harnessing of meta-algorithmic and neural network approaches to provide a means of moving industrial institutions towards a digital universal SDS processing methodology. The complexities of SDS documents including the lack of format standardization, text and image combinations, and multi-lingual translation needs, combined, limit the accuracy and precision of optical character recognition tools.

The approach in this document is to translate entire SDSs from thousands of chemical vendors, each with distinct formatting, to machine-encoded text with a high degree of accuracy and precision. Then the system will "read" and assess these documents as a human would; that is, ensuring that the documents are compliant, determining whether chemical formulations have changed, ensuring reported values are within expected thresholds, and comparing them to similar products for more environmentally friendly alternatives.

 ACM ISBN 978-1-4503-8596-1/21/08…$15.00 https://doi.org/10.1145/3469096.3474933

CCS CONCEPTS

I.4.1 [**Image Processing and Computer Vision**]: Digitization and Image Capture. I.5.4 [**Pattern Recognition**]: Applications – Text processing. I.7.5 [**Document and Text Processing**]: Document Capture – Optical character recognition (OCR).

KEYWORDS

Meta-algorithmics, Neural Networks, EHS compliance, Safety Data Sheets, Validation, Optical Character Recognition.

ACM Reference format:

Kevin Fenton and Steven Simske. 2021. Engineering of an Artificial Intelligence Safety Data Sheet Document Processing System for Environmental, Health, and Safety Compliance. In *Proceedings of the 21st ACM Symposium on Document Engineering (DocEng'21). Limerick, Ireland, 4 pages.* https://doi.org/10.1145/3469096.3474933

1 Introduction

As of 2015, global chemical business transactions accounted for approximately $1.7 trillion worldwide and more than $450 billion in the United States [1]. SDSs are the primary safety/hazard communication mechanism for these transactions and provide the necessary data for industrial organizations to conduct occupational health and safety assessments and to perform hazardous material, hazardous waste, and air emissions calculations for the regulatory compliance for numerous industrial organizations worldwide.

This research was conducted for the U.S. Department of Defense (DoD) and as part of a system that is expected to save over *$3 million* each year in the reduction of workload performed by over 100 full time personnel loading over 100k SDSs each year, with the expectation that a new direct machine-encoded data transfer system will eventually become mandated. For the DoD, chemical ingredient data is used for usage calculations in millions of containers tracked worldwide each year. While a direct data transfer of machine-encoded text via xml or an equivalent would be ideal, the predominant form of SDS data still exists in millions of SDSs in circulation worldwide. Contractual agreements with vendors will take years to update with new requirements; thus, a transitional

system was necessary to assist industrial institutions with converting existing PDFs until a more useful machine-encoded text easily processed by databases. With thousands of variations of SDSs available, trained neural networks provide a base to allow the system to "learn" and appropriately classify values and calculate the statistical probabilities for output accuracy and precision [2].

Algorithms, functions of these algorithms, and combinations thereof are employed to refine classification until the appropriate level of confidence has been reached or no higher confidence can be reached within the capabilities of the approaches taken. Document structure was analyzed on SDSs from various manufacturers to essentially reverse-engineer text extract methods and ensure algorithm compatibility among the vast differences in layout and document features. Once classified, parsed text segments are scrubbed to remove outlier characters and are validated against a GHS expected value schema to ensure numerical values are within expected thresholds, mandatory components have been identified, and desired calculations are achievable given the provided data.

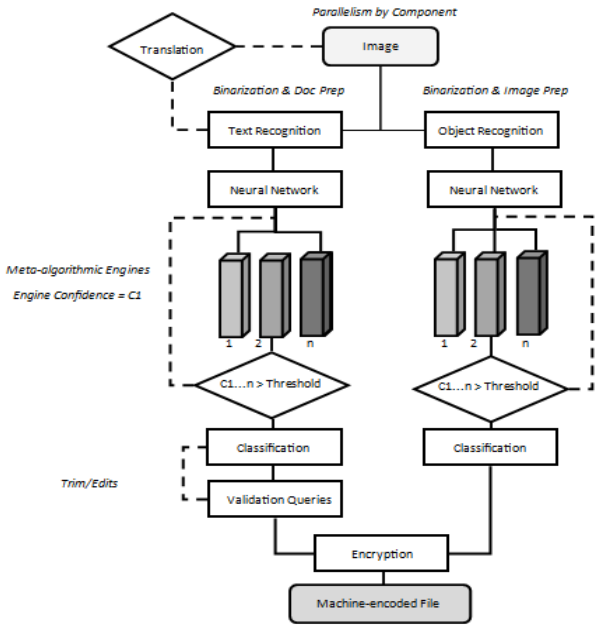

Figure 1: SDS Processing Flow.

1.1 Globally Harmonized System

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 created difficulties for chemical manufacturers to abide by the multitude of chemical hazard communication standards and in the ability for end users to assess environmental and human health exposure threats due to these differences. GHS provided an international approach and framework for the classification and labeling of chemicals to improve environmental and human health and safety protective measures. Two of the important measures associated with GHS implementation are the formatting and information requirements for the SDS. The SDS provides comprehensive information associated with chemical use in the workplace. Each GHS-compliant SDS now contains 16 sections which include information such as product identification, hazard identification, composition, emergency response measures, and storage. Additionally, hazard classification is now required along with labels including signal words, pictograms, and hazard statements. These GHS additions can be used by the system for validation of other physical and chemical data points (e.g. pH values <2 or >11.5 should have a corrosive classification and vice versa). While the sections must be numbered and identified according to the GHS requirements, the general layouts and formatting often differ largely from SDS to SDS and as such require algorithmic applications to apply structure and assign commonality to the derived values.

1.2 Optical Character Recognition

1.2.1 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. Global or whole-image thresholding is the simplest form of binarization and uses a luminosity histogram and assumes a single large peak corresponding to the background of the image and less cohesive sets of luminosities separated by a trough in the histogram. 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. Pre-processing of SDSs is typically minimal in most instances. SDSs are predominantly bimodal in nature, with the exception of GHS and other label pictograms displayed in few colors or tabular data with colored cells. GHS pictograms are all red, white, and black, and can easily be binarized and classified given there are only nine variations. The vast majority of SDS documents offer ideal conditions for OCR.

1.2.2 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 pre-packaged 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 and the potential

for data discrepancies from commonly used words with varying uses.

2 Parallel Processing via Meta-Algorithmics

2.1 Meta-Algorithmics

Meta-algorithmics provides system designers 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.

2.2 Normalized Cross-Correlation

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 [3]. 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.

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

$$\text{Where} \quad T'(x'y,') = T(x',y') - \frac{1}{w*h} * \sum \text{x,yT(x,y)}$$

$$I'(x+x',y+y')$$

$$= I(x+x',y+y') - (\frac{1}{w*h})\sum \text{x"y'I(x+x, y + y")}$$

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 [4]. 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 isolated and classified the pictograms.

2.3 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 [5]. 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.

2.4 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 [6]. 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 *{x1, x2,…xk}* was used to determine the probability of the hidden key sequence *{y1, y2,…yk}*:

$$P(y1, y2, \dots yk | x1, x2, \dots, xk) \quad [7]$$

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.

2.5 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 [8]. 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 but the other did, the former would be accepted and vice versa. 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

2.6 Data Validation and Clean-up

Value validation and clean-up can occur either pre or post value extraction. One common result of the 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 [9].

3 Conclusion

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.

REFERENCES

[1] GHS Requirements. Occupational Safety and Health Administration, 2015, https://www.osha.gov/dsg/hazcom/ghs.html

[2] Marinai, M. Gori and G. Soda, "Artificial neural networks for document analysis and recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 1, pp. 23-35, Jan. 2005, doi: 10.1109/TPAMI.2005.4.

[3] Munsayac, Francisco & Alonzo, Lea & Lindo, Delfin & Baldovino, Renann & Bugtai, Nilo. (2017). Implementation of a normalized cross-correlation coefficient-based template matching algorithm in number system conversion. 1-4. 10.1109/HNICEM.2017.8269520.

[4] OpenCV, Template Matching, https://docs.opencv.org/3.4/de/da9/tutorial_template_match-ing.html, 2020

[5] Evelina Maria De Almeida Neves, A. G. A Multi-Font Character Recognition Based on its Fundamental Features by Artificial Neural Networks. IEEE. 1997.

[6] Yu Bei, Pan, David Z., Matsunawa, T., Zeng, Xuan. "Machine Learning and Pattern Matching in Physical Design". 20th Asia and South Pacific Design Automation Conference. 2015.

[7] Chakraborty, S., Lakshminarayanan, S. Nyarko, Y., Extraction of (Key,Value) Pairs from Unstructured Ads. 2014 AAAI Fall Symposium. 2014.

[8] Simske, Steven J. 2013. *Meta-Algorithmics: Patterns for Robust, Low Cost, High Quality Systems*. Wiley-IEEE Press ISBN: 978-1-118-62669-6

[9] Hazard Communication Standard: Safety Data Sheets. Occupational Safety and Health Administration, 2012, https://www.osha.gov/sites/default/files/publications/OSHA3514.pdf