KNIME Innovation Notes

Automate hazard information extraction for strong risk management

Remove the need for manual work by automatically gathering and harmonizing text-based information

KNIME software is adept at pulling chemical hazard information from Safety Data Sheets (SDS). SDS are standardized documents by which chemical manufacturers communicate a chemical's hazard information to chemical handlers. They typically contain chemical properties, health and environmental hazards, protective measures, as well as safety precautions for storing, handling, and transporting chemicals. Chemical handlers often review the documents for relevant information manually.

But this is not an effective process on a company-wide basis. To put a strong risk management plan together, the Health, Safety, and Environment (HSE) manager has to compile the hazard information about every chemical a company uses. KNIME software can automatically pull hazard information from thousands of SDS without manual effort.

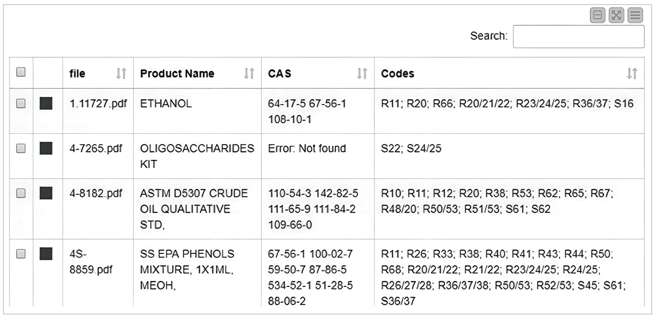
The European Union requires that every hazardous chemical clearly denotes its risks and handling precautions on its SDS. When it comes to using multiple hazardous chemicals, a company needs to gather the risk and precaution information for each one, and make that information available to anyone working with a chemical substance. KNIME software combines text mining and string manipulation to extract the risk information from a collection of SDS. Then, it can categorize that information by how dangerous the substance is, which a user can download.

Here's how.

The user compiles the SDS from various sources, customers, and providers. Then, the user uploads those to a KNIME workflow, either as a single PDF, a library of PDFs, or as a folder containing PDFs. The user also uploads an Excel file with a list of requested phrases to be updated. These materials may also be deployed to KNIME Business Hub if the volume requires more computational power. The Tika Parser garners results, which the software applies text mining nodes to; via string or regex manipulation, the software analyzes each sentence by searching through the Chemical Abstracts Service (CAS) database.

The results report the file name, product name, all the CAS numbers retrieved in each document, and all the retrieved phrases, which are matched with the defined user list. In accordance with regulation, the software saves SDS reporting codes concerning mutagen or cancerogenic dangers in a second Excel file.

Extracting SDS phrases



Search:

☐		file	Product Name	CAS	Codes
☐	■	1.11727.pdf	ETHANOL	64-17-5 67-56-1 108-10-1	R11; R20; R66; R20/21/22; R23/24/25; R36/37; S16
☐	■	4-7265.pdf	OLIGOSACCHARIDES KIT	Error: Not found	S22; S24/25
☐	■	4-8182.pdf	ASTM D5307 CRUDE OIL QUALITATIVE STD,	110-54-3 142-82-5 111-65-9 111-84-2 109-66-0	R10; R11; R12; R20; R38; R53; R62; R65; R67; R48/20; R50/53; R51/53; S61; S62
☐	■	4S-8859.pdf	SS EPA PHENOLS MIXTURE, 1X1ML, MEOH,	67-56-1 100-02-7 59-50-7 87-86-5 534-52-1 51-28-5 88-06-2	R11; R26; R33; R38; R40; R41; R43; R44; R50; R68; R20/21/22; R21/22; R23/24/25; R24/25; R26/27/28; R36/37/38; R50/53; R52/53; S45; S61; S36/37

Fig.1: Graphical output of the workflow. The table shows the file name of the SDS, the product name, all the retrieved CAS numbers, and all the phrases contained in the document.

1000+ PDF files parsed in under 60 minutes:

KNIME Analytics Platform fully accomplished the recovery of the complete range of risk phrases. With a few thousand PDF files, all SDS present in a medium-size company were parsed in less than an hour.

Notable results:

- Significant time saved in repetitive operations (from about two minutes to a few seconds for each SDS)
- Useful for both single and batch processing of SDS files
- Avoidance of deprecated terms due to updating the risk phrases list using an Excel file

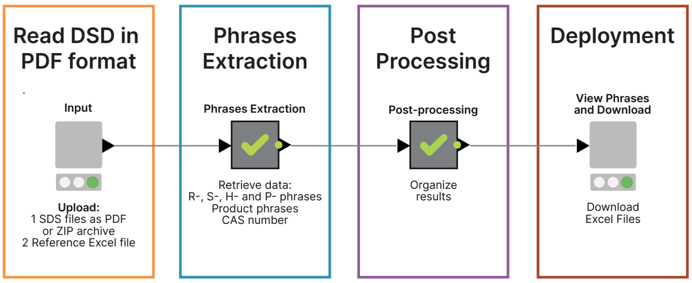


Fig. 2: High-level KNIME workflow

An efficient process for strong risk management plans

KNIME makes this task not only faster, but also reduces the risk of human error. The Tika Parser node enables the retrieval of meta information from each file. The try/catch errors construct effectively avoids workflow mistakes, and regex code in a Java snippet isolates CAS numbers from the PDFs.

Open for Innovation
KNIME

KNIME AG · Talacker 50 · 8001 Zurich, Switzerland · info@knime.com · www.knime.com
©2022 KNIME AG. All rights reserved. The KNIME® trademark and logo and OPEN FOR INNOVATION® trademark are used by KNIME AG under license from KNIME GmbH, and are registered in the United States and the European Union.

Try it out for yourself!

This workflow is available on the KNIME Hub:
https://tinyurl.com/SDS-Risk-Phrase-Extraction