Populating CSV Files from Unstructured Text with LLMs for KG Generation with RML

Jan Maushagen¹ Sara Sepehri² Audrey Sanctorum¹ Tamara Vanhaecke² Olga De Troyer¹ Christophe Debruyne³

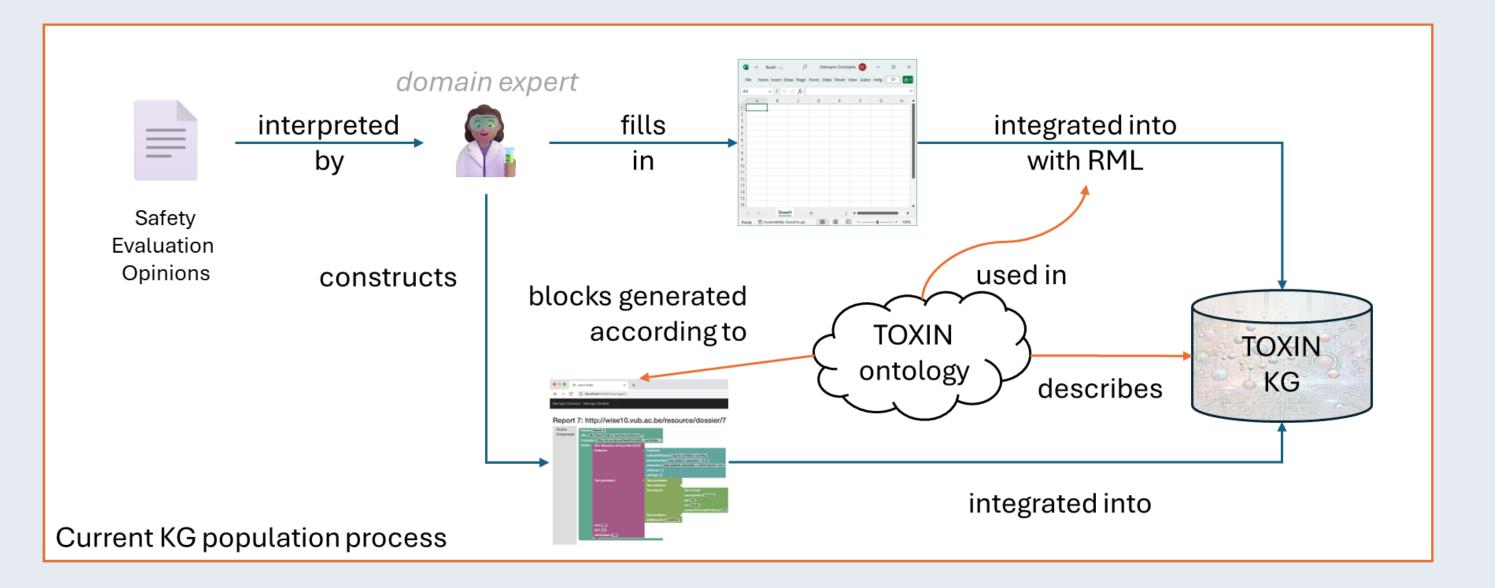
¹ Web & Information Systems Engineering (WISE) Lab, Vrije Universiteit Brussel, Belgium ² Research Group of In Vitro Toxicology and Dermato-Cosmetology (IVTD), Vrije Universiteit Brussel, Belgium ³ Montefiore Institute of Electrical Engineering and Computer Science, University of Liège, Belgium

2024-09-18 at SEMANTICS 2024

Introduction and Research Question

• The TOXIN KG gathers existing safety data of annexed cosmetic ingredients, written up in Safety Evaluation Opinions, to contribute to non-animal systemic toxicity assessments.

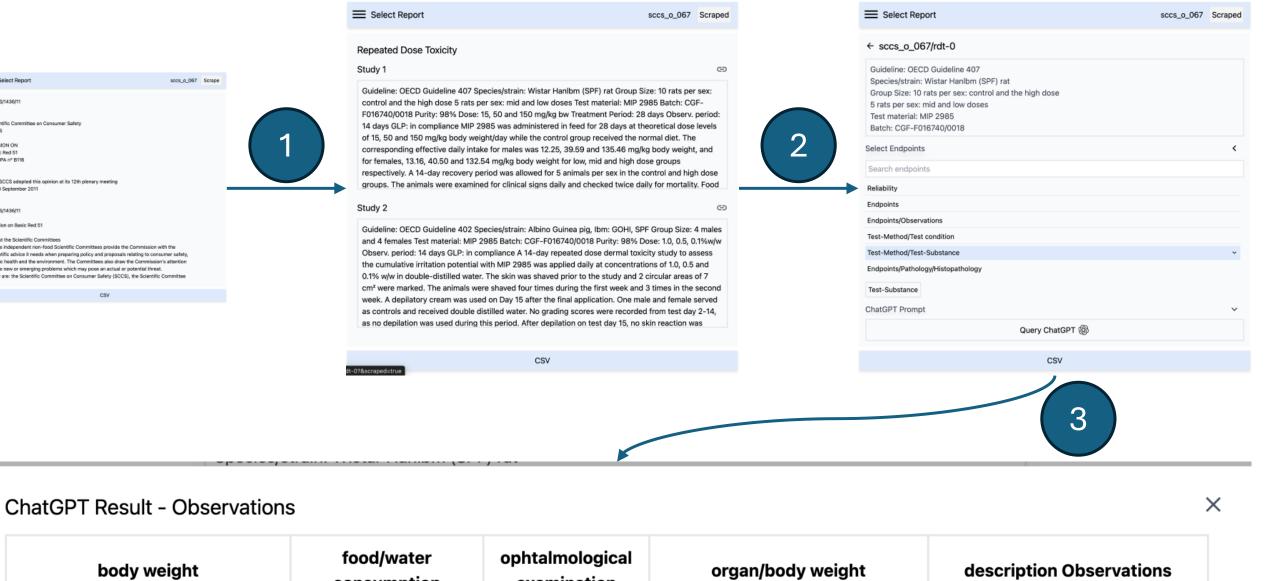
How? Domain experts interpret the expert opinions on chemical compounds (reporting on various studies). It is manual as reports are difficult to interpret, and data needs to be authoritative.



- Experts fill in CSV files or use a jigsaw-inspired editor to populate the KG. [1] The CSV files are less complex and transformed into RDF with [R2]RML.
- Question: Can we help domain experts make the process more efficient by adopting LLMs to fill in those CSV files?

Approach: Generating CSVs with LLMs

- In the current prototype, the text about the experiments (or studies) is extracted using regular expressions (1), and the column headers are used to generate the prompts (2).
- The column headers are grouped under categories. A user can select one or more such categories. Initial testing has quickly shown that the LLM in our experiment, GPT-4, struggled to generate a coherent CSV with many columns.
- We generate the following prompt *for each column*: *"Find the value"* for the following variable "«column name»" based on the category "«category name»" in the following text "«text»". If you can't find the answer in the text, respond with "-". Don't include any commentary



body weight	food/water consumption	ophtalmological examination	organ/body weight	description Observations
Mean body weight and body weight gain were slightly increased in the mid and low dose	The relative food consumption was		Organ weights and organ to body weight ratios were higher in the mid-dose group males, but were	The animals were examined for clinical signs daily and checked twice daily for mortality. Food

text!". The result of which is shown in (3).

The current prototype does not keep track of past interactions; each prompt is executed in a new session.

dose group females when compare C ①	similar in all groups. C ①	G ()	attributed to the higher terminal body weights of the C ①	consumption and body weight were recorded once pre-test C ()
Re-Submit	CSV		Save to CSV	Remove from CSV

Approach: Explain Provenance of Values with LLMs

• A promising feature in the prototype is a button prompting the LLM to point to the part of the text that was used to fill in one of the columns. This feature could assist the project in ensuring the data entered in the CSV is authoritative.

ChatGPT Result - Observations		Source - organ/body weight		
body weight		Guideline: OECD n° 408 Species: Wistar rats, Crl: (WI) BR Group sizes: 15 mal and 15 females (+ 10 males and 10 females for recovery high dose and control group) Material: Fluorgelb II in 0.5% aqueous sodium carboxymethylcellulose		
Mean body weight and body weight gain were slightly		Batch: AZ 212 Purity: > 99% Dose levels: 0, 10, 30 and 90 mg/kg bw in a volur of 10 ml/kg Exposure: 5 days per week for 90-92 days 14 SCCS/1322/10		
increased in the mid and low dose group males, and reduced in high-		Opinion on HC Yellow n° 13 GLP: in compliance Study date: 1991 Fluorgelb II, i 0.5% aqueous sodium carboxymethylcellulose, was administered by oral gava to groups of 15 male and 15 female rats at doses of 10, 30 and 90 mg/kg bw, 5		
dose group females when compare		days per week for 90-92 days. The high dose group and the control group included an additional 10 males and 10 females for observation in a 4-week		
C ()		recovery period. Controls received the vehicle only. There was no effect on either the second		
Re-Submit	Prompt: Provide a text quote from "< <text>>" which is used to answer the</text>	the absolute or relative organ weights in treated rats compared with controls. Islet cell degeneration, accompanied by inflammation or fibrosis of the endocrine pancreas was observed in two of the male rats treated with 90 mg/l bw. These changes were accompanied by a high but not statistically significar		
	following command, namely	blood glucose level and were considered to be material related. No pancreatic changes were found in any intermediate-dose animals. No other treatment		
	<i>"<<previous endpoint="" for="" of="" prompt="" the="" value="" variable="">>"</previous></i> .	related effects were reported. The dose of 30 mg/kg bw/day was considered a NOAEL Ref.: 25 SCCS Comment The NOAEL was adjusted to 21 mg/kg bw/day because of only 5 days treatment per week.		

Conclusions

State-of-the-art has shown some challenges with hallucinations and the validity and well-formedness of the KG.

Limitations and Future Work

Exploring different prompting techniques.

- LLMs have been used to generate KGs, but we wanted to test whether the generation of CSV would render KG generation more efficient and ensure domain-expert involvement.
- An initial exploration of this approach makes us believe it is worthwhile to investigate.

References

A. Sanctorum, J. Riggio, J. Maushagen, S. Sepehri, E. Arnesdotter, M. Delagrange, J. De Kock, T. Vanhaecke, C. Debruyne, and O. De Troyer, "End-user engineering of ontology-based knowledge bases," Behaviour & Information Technology, vol. 41, no. 9, pp. 1811–1829, 2022.

Integrating the prototype into a workflow for domain experts to allow for domain expert validation.

Experiments involving domain experts, i.e., user studies.

Acknolwedgements

The TOXIN project is financially supported by Vrije Universiteit Brussel under Grant IRP19. Some funding came from Cosmetics Europe and the European Chemical Industry Council.





