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

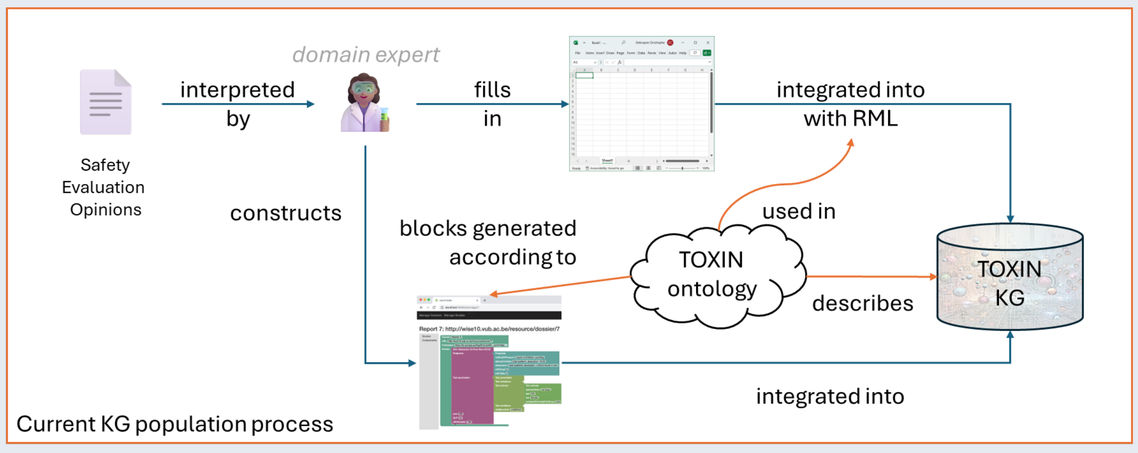
Jan Maushagen[1] Sara Sepehri[2] Audrey Sanctorum[1] Tamara Vanhaecke[2] Olga De Troyer[1] Christophe Debruyne[3]

[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

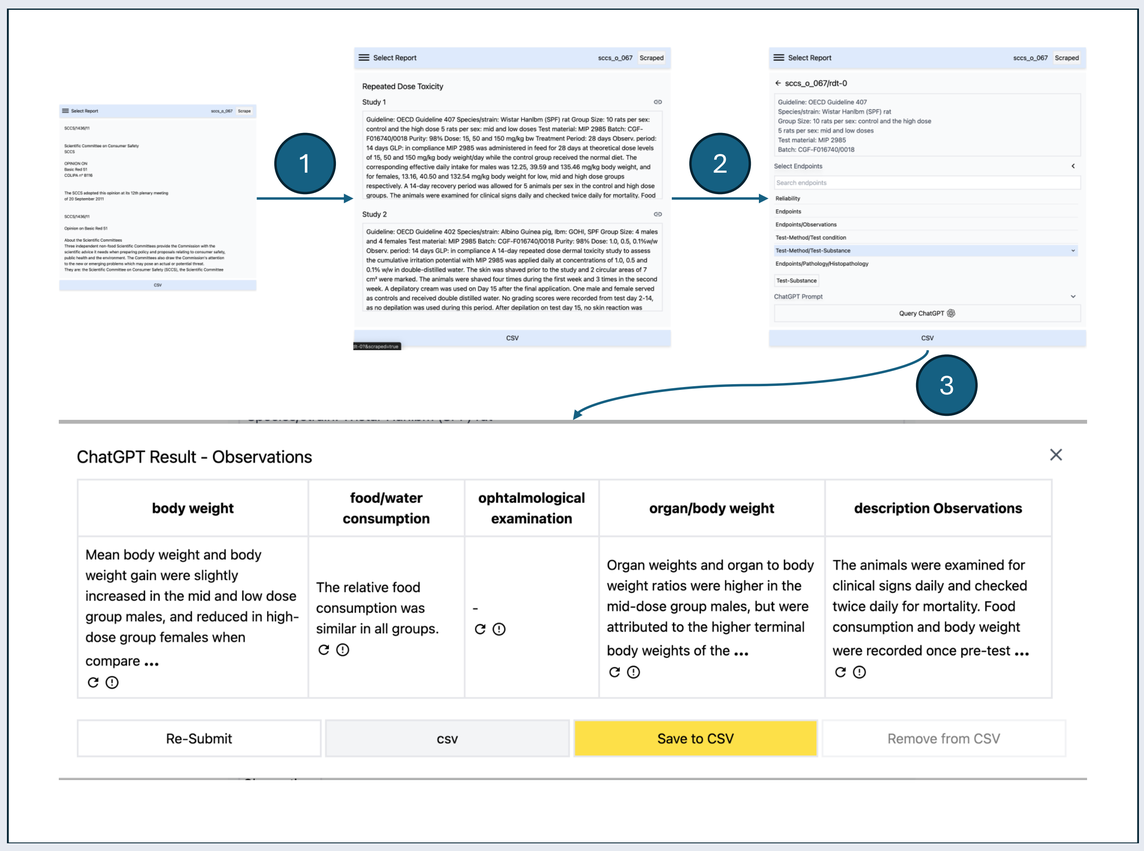
- The TOXIN KG gathers existing safety data of annexed cosmetic ingredients, written up in *Safety Evaluation Opinions*, to contribute to non-animal systemic toxicity assessments.
- How? Domain experts interpret the expert opinions on chemical compounds (reporting on various studies). It is manual as 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?



Current KG population process

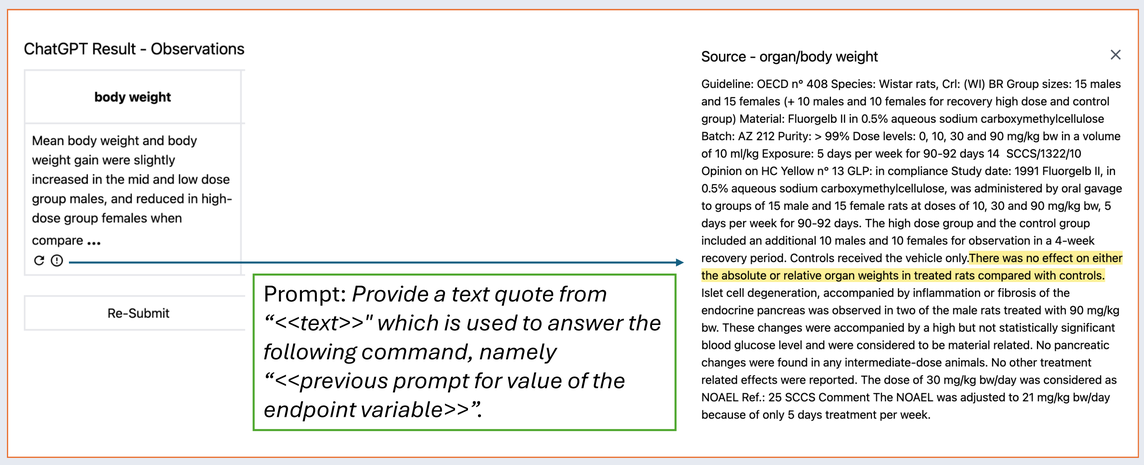
Approach: Generating CSVs with LLMs

- 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 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.



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.



Conclusions

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

Limitations and Future Work

- Exploring different prompting techniques.
- *Integrating* the prototype into a workflow for domain experts to allow for domain expert validation.
- Experiments involving domain experts, i.e., user studies.

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.

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.

VUB IVTD IN VITRO TOXICOLOGY & DERMATO-COSMETOLOGY RESEARCH GROUP
WISE WEB & INFORMATION SYSTEMS ENGINEERING
LIÈGE université Sciences Appliquées