# Creating Knowledge graphs from the literature: the case of health resilience in Green Building Neighbourhoods

Luc Jonveaux, Filip Kučera, Jorge Velasco Manrique, Marta Ingelmo Gomez, Kathrine Nykjær Brejnrod, Sissa Bekombo Priso

The Integrator-centric approach for realising innovative energy efficient buildings in connected sustainable green neighbourhoods project

## Context

The Horizon 2020 PROBONO project (Grant agreement 101037075) aims at demonstrating "*strong examples of how Green Building Neighbourhoods (GBNs) technological and social innovations can be applied, with a vision focused on building infrastructure and a renewed focus on people and sustainability, taking full advantage of digitization and smart technologies for the benefit of society*". The Task 3.5 of this projects aims at reviewing "Interventions to mitigate diseases outbreaks".
In this document, we summarize how we used new technologies, including Large Language Models (LLMs), to consolidate a Knowledge Graph (KG) of this topic, based on the literature, to demonstrate the feasibility of building a body of knowledge pertaining to a certain domain, with a specific angle. This opens the door to more opportunities the growing space existing between LLMs and KGs.

## Disclaimer

Please note that this document is a research-based exploration compiled by knowledge management researchers, not medical professionals. Our findings are presented with the intention to inform and contribute to the dialogue on public health strategies. They are indicative and should serve as a preliminary guide. We encourage all readers to consult with qualified health professionals for expert advice and to confirm and enrich these insights.

## Results

We started with defining a basic ontology based on classes of interest for mitigation measures, namely 'Risks', 'Mitigations', 'People', and 'Technologies'. These constitute an initial body of knowledge, which is then used to build 'Blueprints', possible interventions to mitigate diseases outbreaks.
To date, the Knowledge Graph (v0.4) contains information on 377 articles, from which were programmaticaly derived 21145 risks, 22950 mitigation measures, 16125 stakeholders and 23140 technologies. The team used these to build 24 blueprints manually, and automated the production of 50 others. We hope that releasing the knowledge graph under an open-source license (CC BY-NC-SA) will drive use of this knowledge graph and that health professionals can use this to derive useful, professionally-approved mitigation measures.

## Acknowledgements

Thank you to the PROBONO team.

## Contact Information

- contact@probonoh2020.eu
- www.probonoh2020.eu

## Objective

The main objective of the task is to review the scientific literature to identify key risks, stakeholders, technologies and mitigations measures both at building and neighbourhood scales.

## Technical details

The present knowledge graph has been created using new tools, helping to streamline, faster and more consistently, information from the literature:

- Parsing of the literature was done with *GROBID*. This provided structured text (XML) from the article PDFs;
- The data was processed and structured in an RDF that was produced with the *Owlready2* python library;
- Vector embedding based on *ChromaDB* because of the early possibility to integrate with *LangChain*, and because of its ease of use.

Once the data was prepared, we explored structuring the information using different solutions:

- *NLTK* was used to extract topics and themes of the articles and *Spacy* with and *CoreferenceResolver* to tackle disambiguation ;
- Text was processed using a combination of OpenAI API (both using GPT3.5 and GPT4 endpoints) as well as running local models (NOUS/LLaMa), using the python requests or *LangChain* libraries.
- We used the content from the articles, stored in *ChromaDB*, used through *LangChain* and deployed through a *FastAPI API*.

## Next steps

This activity yielded expected outcomes, consisting in mapping out risks, technologies and stakeholders, as well as suggesting mitigation measures. This however is an asset that has a highly reusable potential, and this list, albeit listing possible actions from a project perspective, could be undertaken by other parties:

- We plan on integrating more robust graph management solutions, possibly Neo4j or similar, to continually review and enrich the KG;
- We would want to enrich the semantic content of the graph to make it more usable and possible an input to KG-backed LLMs;
- Reusing existing semantic assets (eg Wikidata) might help structure
- Connect this knowledge graph to other KGs or ontologies.

## Solution repository

GitHub repository

## Note(s)

This PDF can be renamed as a ZIP file to extract source code for this document and for the solution.

This project has received funding from the European Union's Horizon 2020 Europe Research and Innovation programme under Grant Agreement No 101037075.
This output reflects only the author's view, and the European Union cannot be held responsible for any use that may be made of the information contained therein.