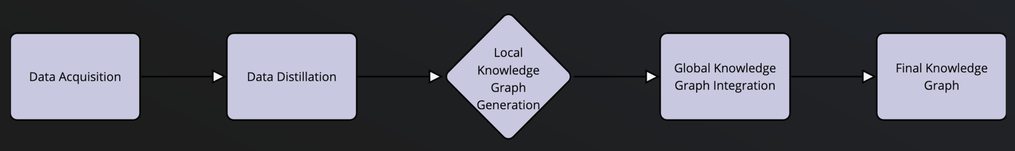
METHODOLOGIES

V0: Building KG (Buehler, 2024)

tl;dr

The baseline methodology for knowledge graph construction closely follows the approach introduced by Buehler (2024). The knowledge graph generation process follows four main phases:

1. **Data Acquisition:** Collecting research papers from various sources
2. **Data Distillation**: Converting research papers into structured text chunks
3. **Local Knowledge Graph Generation**: Extracting triples from each text chunk
4. **Global Knowledge Graph Integration**: Combining local graphs and refining the result



Knowledge Graph Construction

Phase 1: Data Acquisition

The process begins with a curated dataset of scientific papers on longevity and aging. This dataset was seeded using **HALL** (Li et al., 2024), a collection of 136 publications from the past two decades. To expand the dataset, we first extract citations from these papers, retrieving earlier research referenced by HALL papers. Next, we identify papers that have cited works from the HALL dataset, capturing newer research that builds upon these publications. Together, these steps construct a **one-degree citation network**, linking foundational studies with their academic impact over time.

Phase 2: Data Distillation

Step 1: Text Chunking

To ensure efficient processing, markup documents are split into manageable text chunks. We first divide each scientific publication into chunks of 800-1000 tokens, with a 100-token overlap to maintain context.

Step 2: Context Generation

Once text chunks are created, an LLM generates distilled insights for each chunk:

- **Summary**: A concise overview of key information from the chunk.
- **Bulleted Insights**: A detailed breakdown of significant findings.

This distilled content serves as the **"raw context"** for subsequent graph generation.

Phase 3: Local Knowledge Graph Generation

For each chunk's raw context, the system extracts knowledge in the form of triplets:

- **Triplet Extraction**: The system prompts the LLM to identify key concepts and relationships, outputting them as triplets in the form:

```
{
  "node_1": "concept A",
  "node_2": "concept B",
  "edge": "relationship between A and B"
}
```

- **Refinement**: The initial triplets are refined to ensure consistent labeling and terminology via iterative prompting.

Phase 4: Global Knowledge Graph Integration

Multiple local graphs originating from is then merged into a global knowledge graph:

- **Node Embedding**: Sentence transformers (e.g., "all-MiniLM-L6-v2") generate vector embeddings for each node
- **Similarity Detection**: Similar nodes are identified using cosine similarity with a configurable threshold (default 0.85)
- **Node Merging**: Similar nodes are merged, with the highest-degree node preserved and connections redirected. Mergin process preserves all edge relationships.

References

Buehler, M. J. (2024). Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Machine Learning: Science and Technology*, *5*(3), 035083. https://doi.org/10.1088/2632-2153/ad7228

Hao Li, Song Wu, Jiaming Li, Zhuang Xiong, Kuan Yang, Weidong Ye, Jie Ren, Qiaoran Wang, Muzhao Xiong, Zikai Zheng, Shuo Zhang, Zichu Han, Peng Yang, Beier Jiang, Jiale Ping, Yuesheng Zuo, Xiaoyong Lu, Qiaocheng Zhai, Haoteng Yan, Si Wang, Shuai Ma, Bing Zhang, Jinlin Ye, Jing Qu, Yun-Gui Yang, Feng Zhang, Guang-Hui Liu, Yiming Bao, Weiqi Zhang, HALL: a comprehensive database for human aging and longevity studies, *Nucleic Acids Research*, Volume 52, Issue D1, 5 January 2024, Pages D909–D918, https://doi.org/10.1093/nar/gkad880

Previous
V0: Hypothesis Generation (Ghafarollahi & Buehler, 2024)

Last updated 26 days ago