Graphes de connaissances spatio-temporels pour l'interdisciplinaire

Isabelle Mougenot

UM - Espace Dev

2025

Plan général

1. Graphes de Connaissances
2. Structures de graphes de données
   - Systèmes de Gestion de Données
   - Espace et temps
   - TPG
     - Le standard du W3C RDF
   - LPG
     - Exemple de Neo4J
   - Choix de la structure
3. KG et IA génératives
   - RAG

Définition générale

Repris de "Knowledge Graphs, Hogan et al, 2021, arxiv.org/abs/2003.02320

Graphe de connaissances (GC) / Knowledge Graph (KG) =
Un graphe de données destiné à la pérennisation et à la diffusion de connaissances portant sur le monde réel :

- avec des nœuds qui représentent des entités d'intérêt
- et des arêtes qui représentent les relations entre ces entités.

Dans la théorie des graphes, un graphe est un ensemble de nœuds (ou sommets) et un ensemble d'arêtes (ou de liens entre nœuds)

Exemple élémentaire

Hannah Arendt dans un graphe (de données)

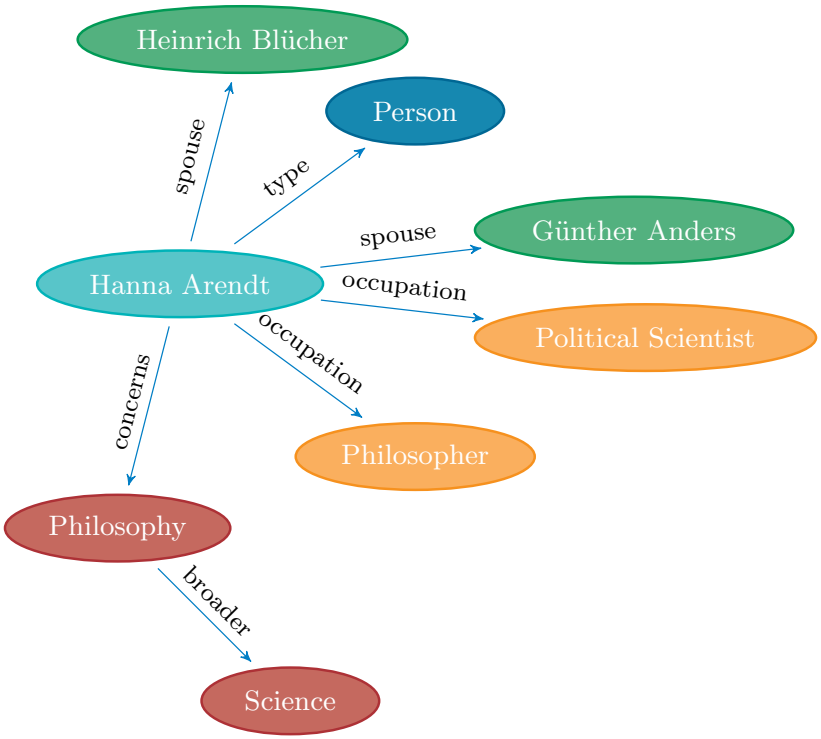


Figure: Graphe étiqueté et orienté

Adéquation avec le système "mental" humain

Associer et catégoriser : des mécanismes cognitifs[1] naturels

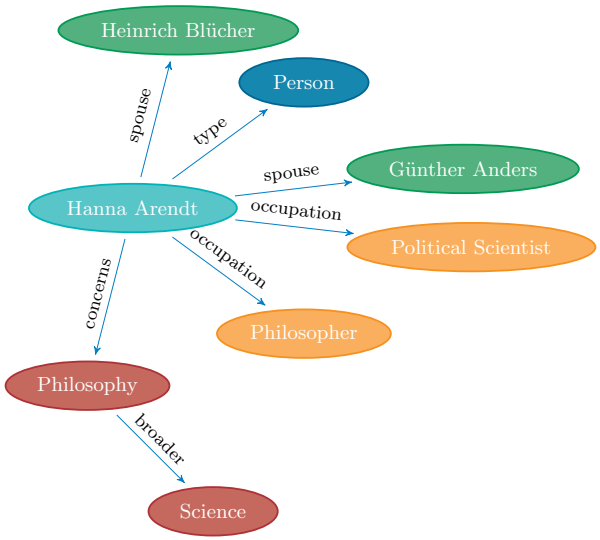


Figure: catégories et associations représentées à l'aide de graphes (que l'on sait traiter efficacement)

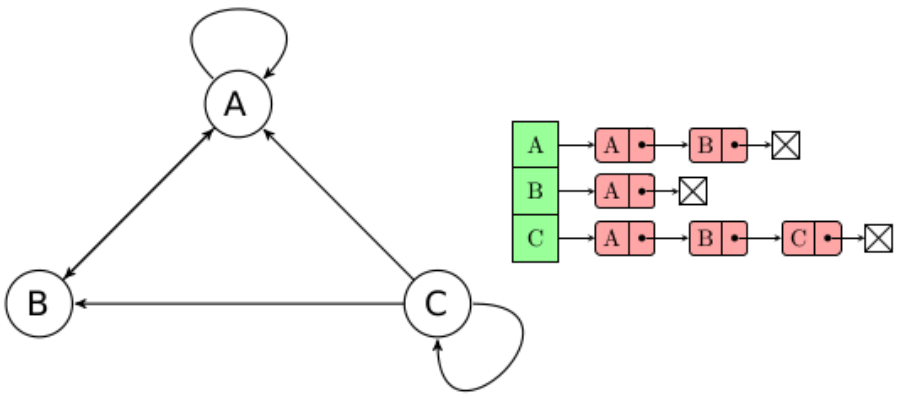
[1] processus psychiques liés à l'esprit

Adossement à la théorie des graphes

- graphe G=<V;E> : où V, ensemble des sommets et E, ensemble des arêtes,
- graphe orienté : les arêtes sont des arcs
- sous-graphe G'=<V';E'> de G=<V;E> est un graphe tel que V' $\in$ V et E' $\in$ E
- chemin C entre 2 nœuds v1 et v2 : séquence de nœuds et d'arêtes permettant de rejoindre v2 à partir de v1
- un graphe est dit connecté si il existe un chemin reliant toute paire de nœuds
- un cycle est un chemin fermé (C(vi;vi))
- un arbre est un graphe connecté et acyclique

Les structures support

ont contribué à rendre les graphes incontournables



	A	B	C
A	1	1	0
B	1	0	0
C	1	1	1

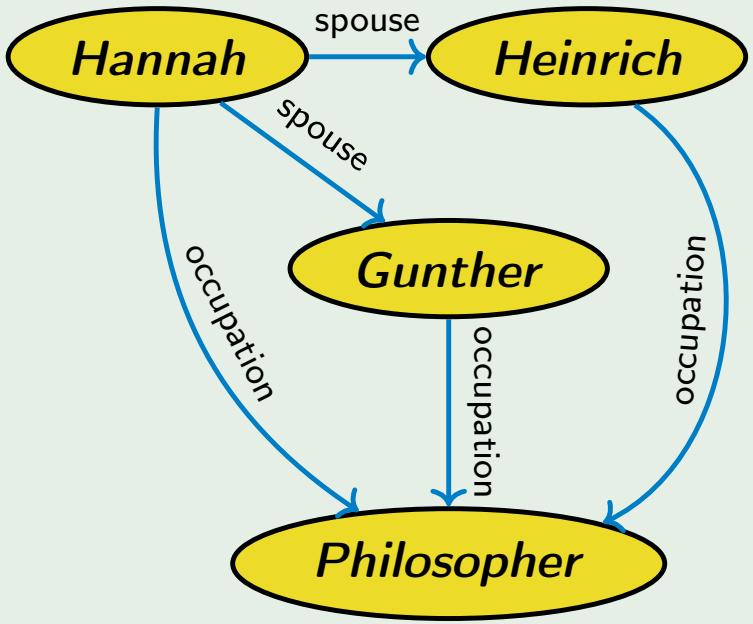
Figure: Liste et matrice d'adjacence

KG et théorie des graphes

- graphe orienté/oriented graph : les arêtes sont des arcs
- multi-graphe : possiblement plusieurs arêtes entre deux nœuds
- multi-graphe orienté : possiblement plusieurs arcs (dans les deux directions) entre deux nœuds
- hypergraphe : arêtes/arcs avec plus de deux extrémités (hyperarcs liant plus de 2 nœuds à la fois)
- graphe étiqueté/labeled graph : un label/étiquette (juste du texte ou pouvant être structuré) donné aux arêtes
- multi-graphe orienté étiqueté : multi-graphe pour lequel un label est donné à chaque arc
- graphe attribué/property graph : les nœuds et les arcs peuvent posséder des attributs/propriétés

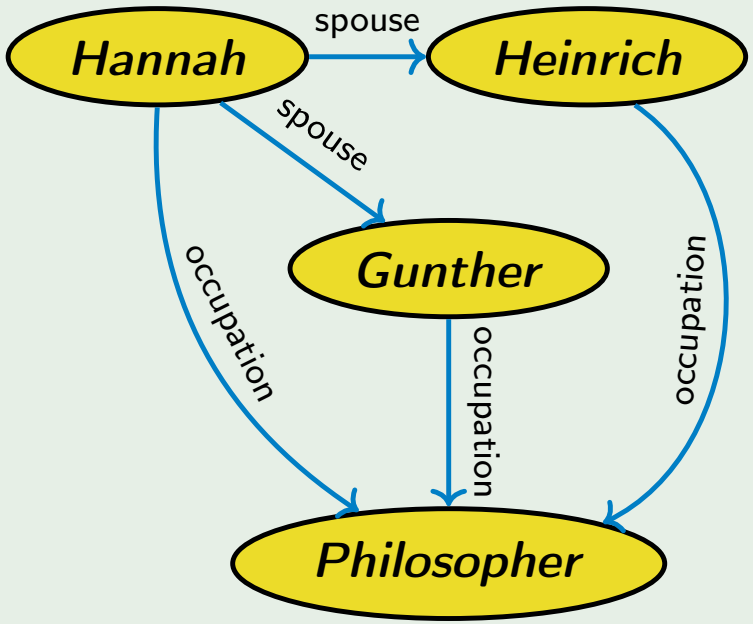
Identifier la structure ?

Example



Identifier la structure ?

Example

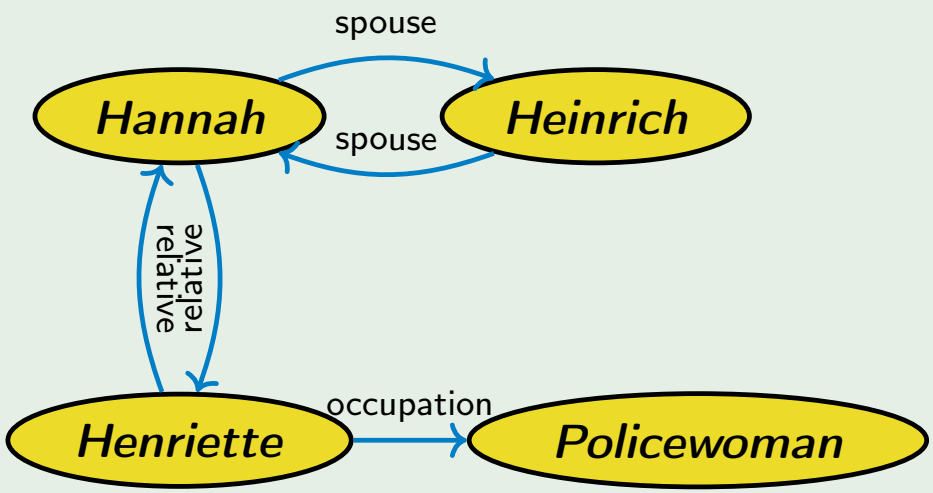


Solution

graphe orienté étiqueté

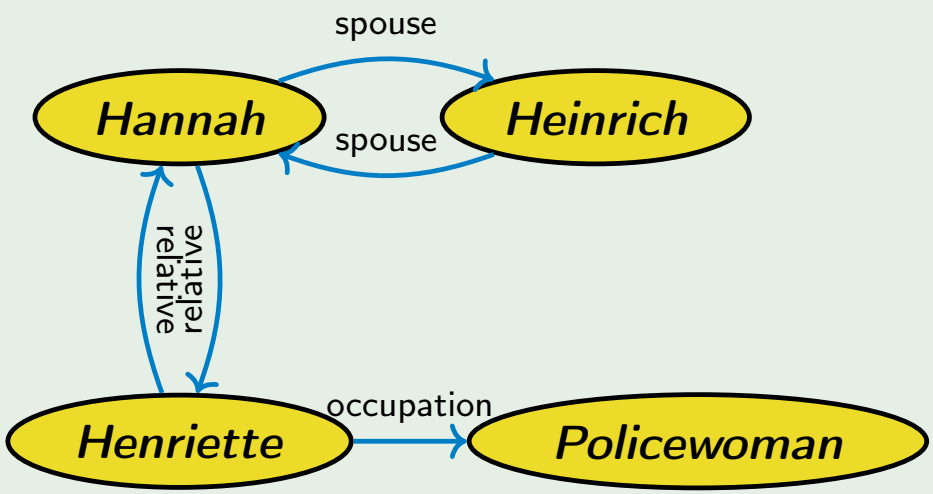
Identifier la structure ?

Example



Identifier la structure ?

Example

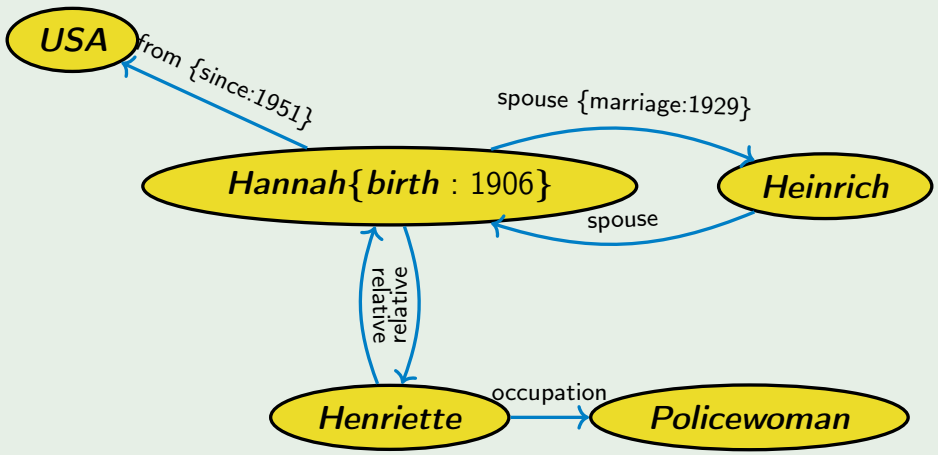


Solution

multi-graphe orienté étiqueté

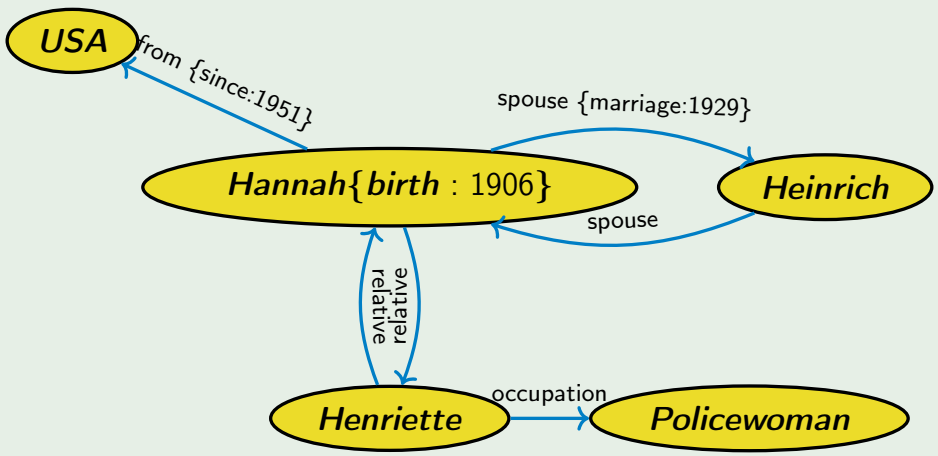
Identifier la structure ?

Example



Identifier la structure ?

Example



Solution

multi-graphe orienté étiqueté attribué

Le choix de la structure pour les systèmes de gestion de données

Deux grandes directions : multi-graphe orienté et typé (∼typed property graph ou TPG), multi-graphe attribué et étiqueté (∼labeled property graph ou LPG)

- Neo4J : multi-graphe attribué et étiqueté
- JanusGraph : multi-graphe attribué et étiqueté
- TigerGraph : multi-graphe attribué et étiqueté
- Triplestores (Stardog, RD4J, Jena TDB, . . .) : multi-graphe orienté et typé

La plus grande différence

Doter les associations/arcs de propriétés valuées permet de contextualiser (notamment au travers de l'espace et du temps)

- Neo4J, JanusGraph, TigerGraph :
  1. plus centrés sur les objets et leurs inter-relations
  2. les propriétés concrètes des entités ne sont pas des arcs du graphe
- Triplestores adossés aux standards du W3C : RDF, RDFS, OWL
  1. plus centrés sur les schémas, l'intégration de schémas et le partage de connaissances
  2. les propriétés sont des ressources à part entière
- GraphDB : solution mixte qui dote les propriétés d'attributs valués en s'adossant à RDF-star et SPARQL-star

Des définitions pour l'espace et le temps

Des dimensions transversales aux disciplines qui confèrent de l'évolutivité aux entités

- Espace :
  1. "L'espace, c'est ce qui arrête le regard, ce sur quoi la vue butte" - G. pérec dans Espèces d'espaces - 1974
  2. "L'univers correspond à tout ce qui existe (galaxie, étoiles, planètes) et l'espace est le territoire où s'étend cet Univers - CNES `https://cnes.fr/dossiers/espace`
- Temps
  1. Continuité indéfinie, milieu où se déroule la succession des évènements et des phénomènes, les changements, mouvements, et leur représentation dans la conscience - Dictionnaire en ligne le Robert

KGs spatio-temporels

Les solutions vont s'avérer plus ou moins génériques pour la prise en charge de ces deux dimensions orthogonales au sein des KGs

1. TPG : Composants sémantiques standards (GeoSPARQL, OWL Time)
2. LPG : des propriétés spatiales comme temporelles ad hoc

Graphes de connaissances de type "TPG"

Parmi lesquels les LOD & LOV : partager les données, voire les concepts, plutôt que les documents sur le Web

- Passer d'un système documentaire à un système de données/connaissances de manière à en faciliter la manipulation et l'interprétation par des agents logiciels.
- Décrire de manière normalisée les ressources du web en s'appuyant sur des standards de données, l'idée est de se rapprocher de systèmes de mutualisation de ressources.

Ontologies

Ontologie en informatique = spécification explicite d'une conceptualisation partagée (R. Studer, 1998)

Une ontologie adossée aux standards RDF, RDFS et OWL du W3C est un KG, qui puise sa raison d'être dans le partage, l'intégration, et le raisonnement

1. GeoSPARQL (R. Battle, D. Kolas, 2011) : composant ontologique pour la dimension spatiale
2. OWL Time (J. Hobbs, F. Pan, 2006) : composant ontologique pour la dimension temporelle

relations topologiques entre entités spatiales / entités temporelles

Architecture du web

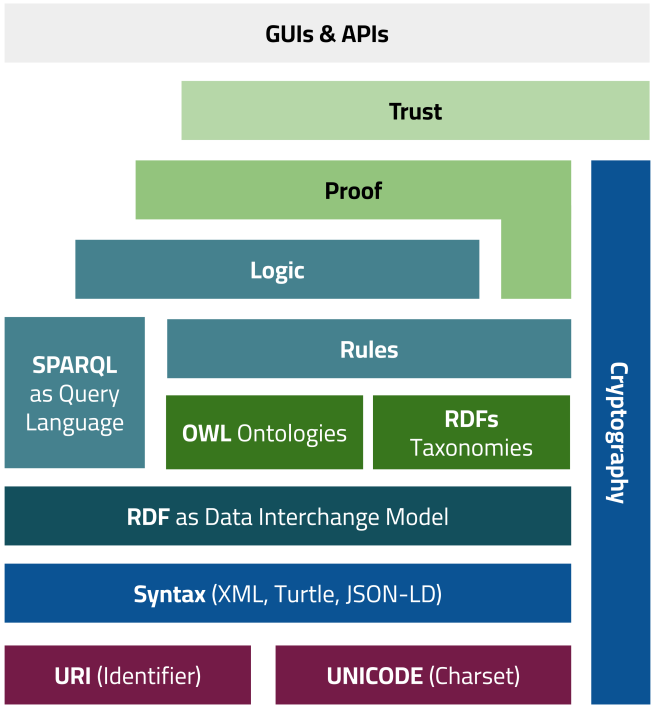


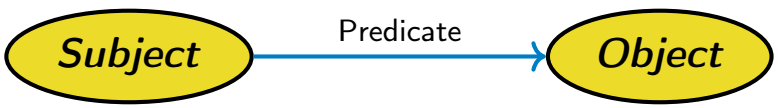
Figure: Empilement de couches (layer cake)

Resource Description framework

Langage RDF (Resource Description Framework) du W3C : initiative pour décrire des *ressources* (notamment Web) au travers de méta-données

Principes :

- décrire une ressource ou des relations entre ressources
- apporter sa perception sur une ressource au travers d'annotations (couples propriété-valeur) sans modifier la ressource
- exploiter de manière décentralisée chaque annotation



GeoSPARQL Core Component

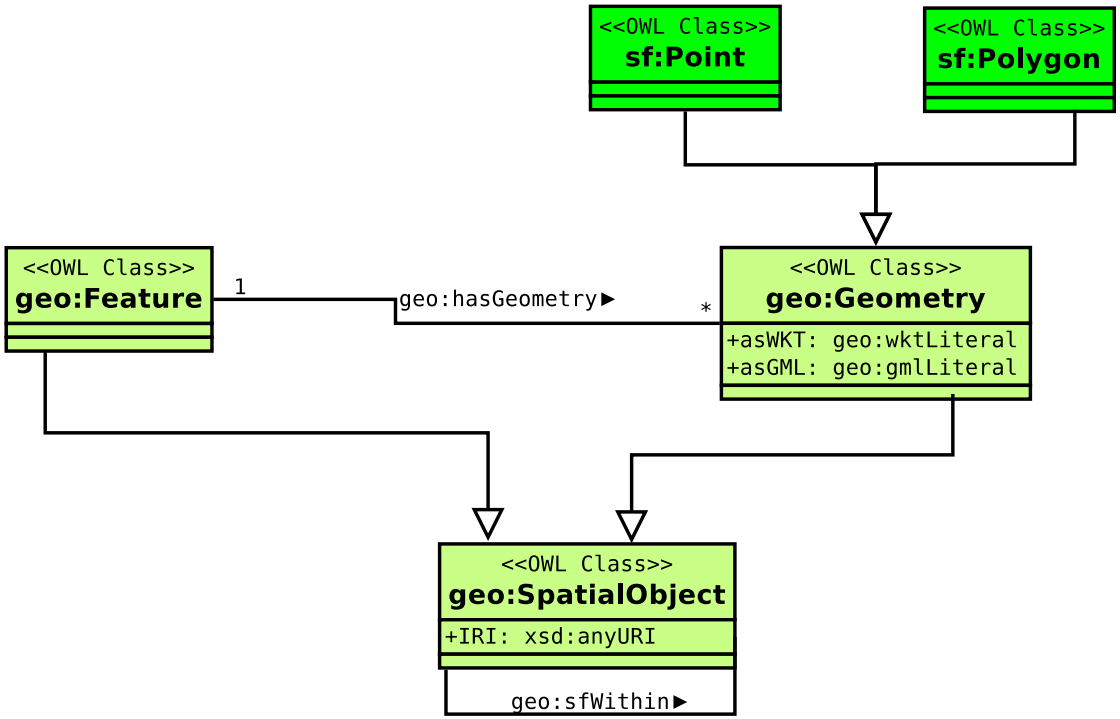


Figure: Diagramme de classes : entité spatiale et ses géométries

OWL Time

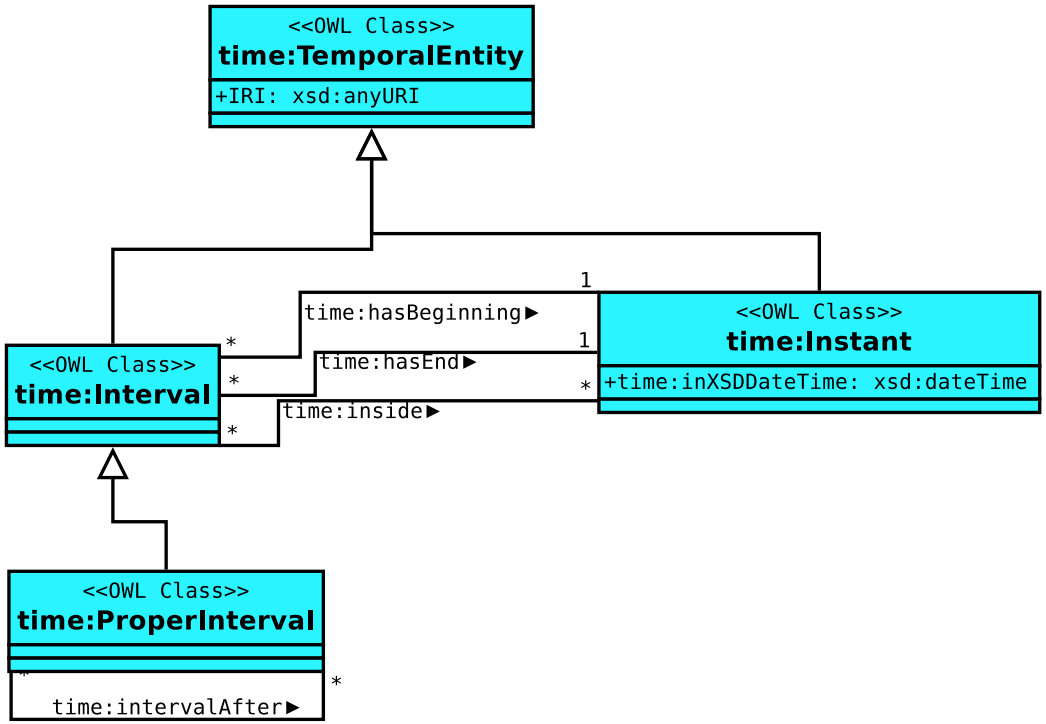


Figure: Diagramme de classes : entités temporelles et leurs liens

"Universalité du savoir" (ici avec Wikidata)

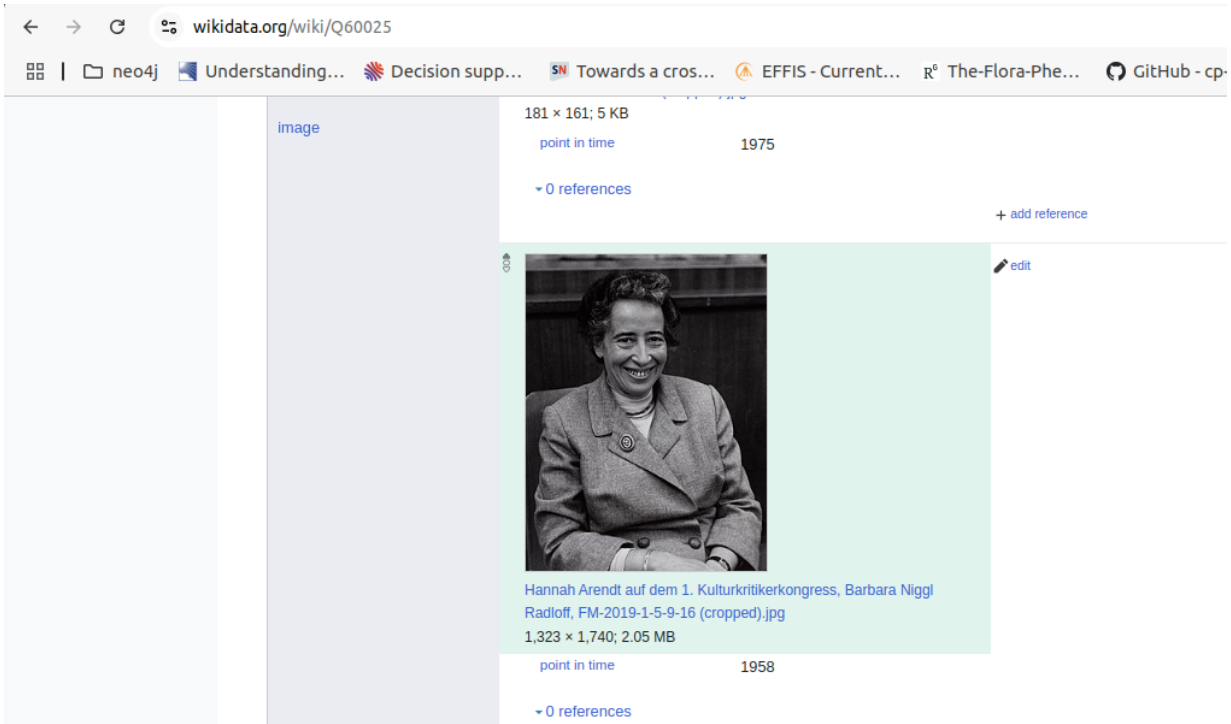


Figure: une page Wikidata pour Hannah Arendt

voir `http://www.wikidata.org/entity/Q60025`

Exemple élémentaire

Hannah Arendt (Q60025) dans Wikidata

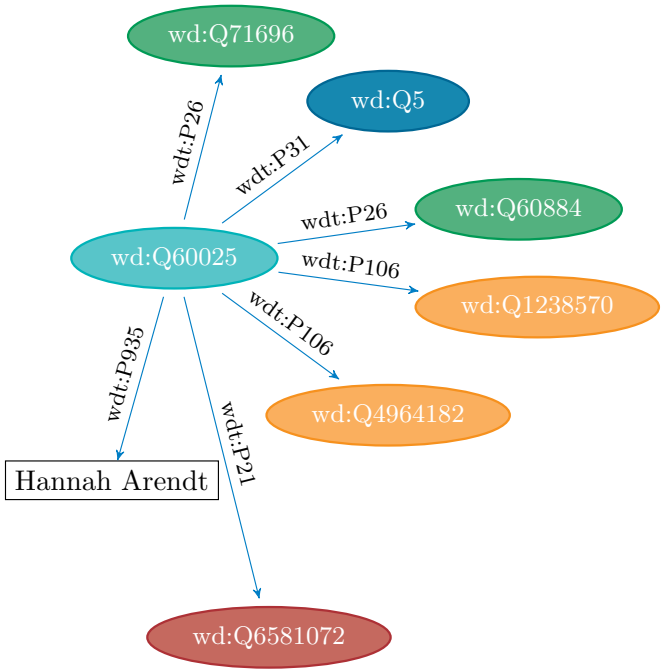


Figure: Graphe partiel

Web de données - les syntaxes concrètes de RDF

Listing 1: Q60025 en langage N3

```
@prefix  wdt:    <http://www.wikidata.org/prop/direct/> .
@prefix  wds:    <http://www.wikidata.org/entity/statement/> .
@prefix  xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix  rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix  wd:     <http://www.wikidata.org/entity/> .
@prefix  wdtn:   <http://www.wikidata.org/prop/direct-normalized/> .

wd:Q60025  rdfs:label  "Hannah Arendt"@de ;
        wdt:P31                 wd:Q5 ;
        wdt:P106                wd:Q4964182 ,wd:Q1397808, wd:Q1238570, wd:Q2306091,
                wd:Q201788, wd:Q15994177, wd:Q11774202, wd:Q36180, wd:Q1622272 ;
        wdt:P26                 wd:Q60884 ,wd:Q71696 ;
        wdt:P21                 wd:Q6581072 ;
        wdt:P569                "1906-10-14T00:00:00Z"^^xsd:dateTime ;
        wdt:P570                "1975-12-04T00:00:00Z"^^xsd:dateTime ;
        wdt:P935                "Hannah Arendt" ;
        wdtn:P245               <http://vocab.getty.edu/ulan/500217300> ;
        wdtn:P268               <http://data.bnf.fr/ark:/12148/cb118890622#about> ;
        wdtn:P269               <http://www.idref.fr/080879128/id> .
```

Le problème posé : absence du contexte temporel et spatial

Figure: Wikidata avec GraphBuilder

Exemple de chemins sur le lien "spouse (wdt:P26)" à partir de l'entité wd:Q60025

`https://angryloki.github.io/wikidata-graph-builder/?item=Q60025&property=P26`

Requête SPARQL sémantiquement proche

```
SELECT DISTINCT ?hannah ?spouse ?spouseLabel WHERE {
  BIND(wd:Q60025 AS ?hannah)
  ?hannah wdt:P31 wd:Q5;
    wdt:P26* ?spouse.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],mul,en". }
}
```

Listing 2: Conjoints d'Hannah ou conjoints de ses conjoints

Les résultats de la requête précédente

Consultation depuis le point d'accès SPARQL de Wikidata

hannah	spouse	spouse label
wd:Q60025	wd:Q1266986	Elisabeth Freundlich
wd:Q60025	wd:Q5086051	Charlotte Lois Zelka
wd:Q60025	wd:Q71696	Heinrich Blücher
wd:Q60025	wd:Q60025	Hannah Arendt
wd:Q60025	wd:Q60884	Günther Anders

`https://query.wikidata.org/`

Requête SPARQL qui fait appel à une chosification de l'information associée au mariage

Listing 3: Des nœuds complémentaires mobilisés

```
SELECT ?spouseLabel ?start ?end ?place WHERE {
  BIND(wd:Q60025 AS ?hannah)
  ?hannah p:P26 ?marriage.
  ?marriage pq:P580 ?start;
    ps:P26 ?spouse.
  OPTIONAL { ?marriage pq:P582 ?end. }
  OPTIONAL { ?marriage pq:P2842 ?place. }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

Les résultats de la requête précédente

Consultation depuis le point d'accès SPARQL de Wikidata

spouseLabel	start	end	place
Günther Anders	1929-01-01T00:00:00Z	1937-01-01T00:00:00Z	wd:Q64
Heinrich Blücher	1940-01-16T00:00:00Z	1970-10-30T00:00:00Z	

`https://query.wikidata.org/`

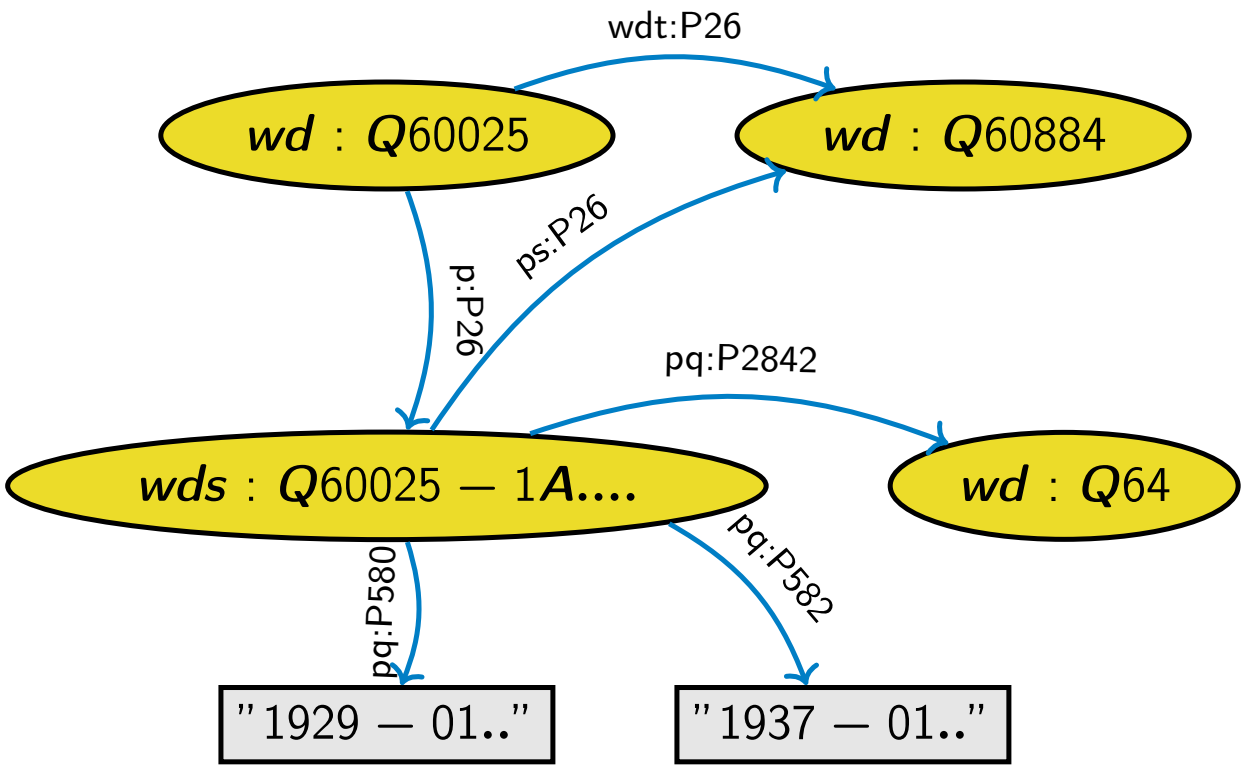
Une information présentée sous deux formes (un arc et un arc+un nœud)

Listing 4: Q60025 en langage N3

```
@prefix  wdt:    <http://www.wikidata.org/prop/direct/> .
@prefix  wds:    <http://www.wikidata.org/entity/statement/> .
@prefix  xsd:    <http://www.w3.org/2001/XMLSchema#> .
@prefix  rdfs:   <http://www.w3.org/2000/01/rdf-schema#> .
@prefix  wd:     <http://www.wikidata.org/entity/> .
@prefix  p:      <http://www.wikidata.org/prop/> .

wd:Q60025  rdfs:label  "Hannah Arendt"@de ;
        wdt:P31                wd:Q5 ;
        wdt:P26                wd:Q60884 ,wd:Q71696 ;
        p:P26                  wds:Q60025-676A67E9-2FA0-4B4A-9D40-A72A6D813446 ,
                       wds:Q60025-1A90A9A4-506B-4E8D-ADDE-E2D34179CB6C .
```

Respecter la structure "Typed Property Graph"



Requête SPARQL qui fait appel à une fonction GeoSPARQL

Listing 5: Philosophes nés à Hanovre et alentours

```
SELECT ?philospher ?birthPlace WHERE {wd:Q1715 wdt:P625 ?hanoverLoc.
 ?philospher  wdt:P106 wd:Q4964182; wdt:P19 ?birthPlace .
 ?birthplace  wdt:P625 ?birthPlaceCoor.
FILTER(geof:distance( ?birthPlaceCoor,  ?hanoverLoc) < 30 ) }
```

Requête syntaxiquement correcte mais qui conduit à un dépassement de mémoire avec le point d'accès Wikidata

Graphes attribués (LPG) : une information plus concise

```
create (hannah:human {name:'Hanna Arendt', gender:'female', birth:'1906-10-14',
      death:'1975-12-04'}) -[sp1:spouse {start:'1929', end:'1937', place:'Berlin'}]->
 (gunther:human {name:'Gunther Anders', gender:'male', birth:'1902-07-12',
       death:'1992-12-17'}), (gunther) -[sp2:spouse {start:'1929', end:'1937',
       place:'Berlin'}]-> (hannah), (hannah) -[o1:hasOccupation]->
       (philosopher:occupation {name:'philosopher'}), (gunther)
       -[o2:hasOccupation]-> (philosopher)
```

Listing 6: Création de sommets et arcs avec Neo4J/Cypher

Les nœuds comme les arcs possèdent des attributs propres et sont annotés avec des labels (étiquettes textuelles). Le graphe est moins volumineux mais le modèle est moins générique

Multi-graphe attribué étiqueté avec Neo4J

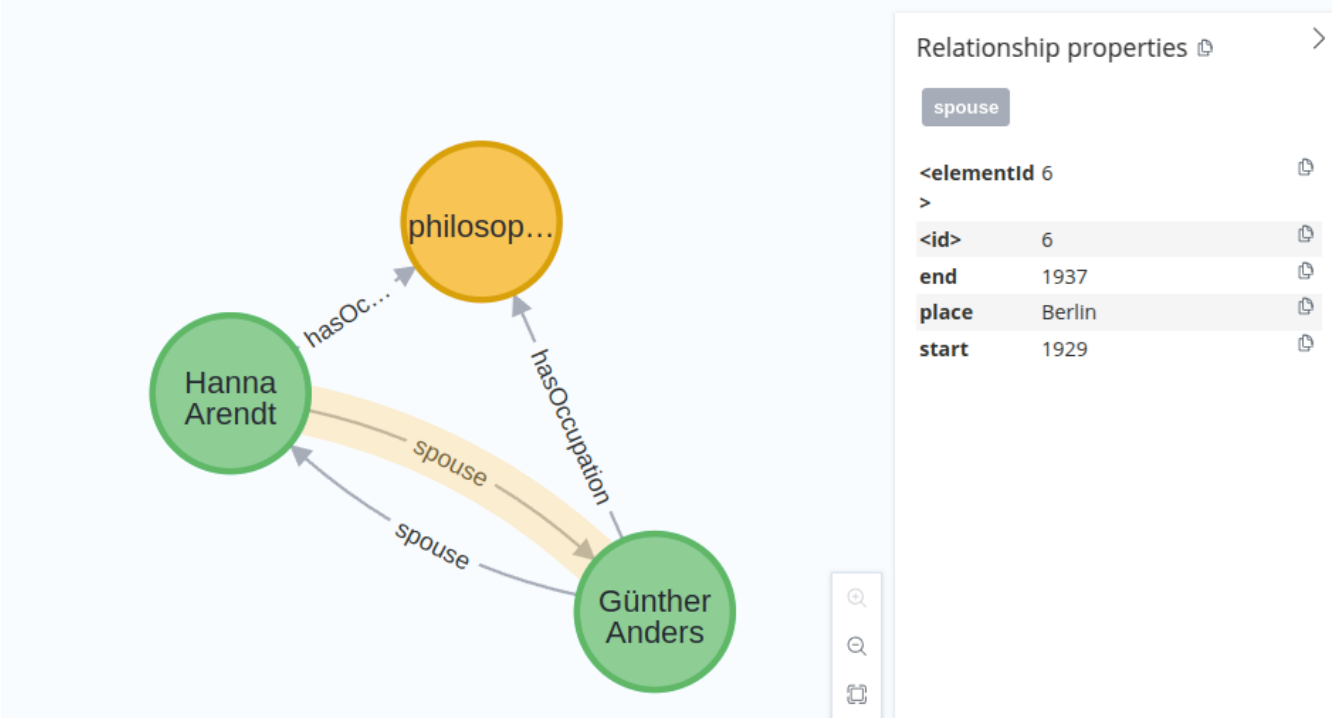


Figure: sommets et arcs dotés d'attributs

Les possibles

En fonction des besoins ciblés

1. partager, intégrer, focus sur les schémas → multi-graphe orienté typé (TPG)
2. persister de gros volumes de sommets, analyser des données factuelles, focus sur les données → multi-graphe attribué étiqueté (LPG)

Les deux structures peuvent être mobilisées dans des approches d'IA génératives

Combiner KGs et GenAI

Différentes orientations (non exhaustif)

1. RAG (génération augmentée de récupération) qui peut mobiliser des KGs (GraphRAG)
2. XAI (IA eXplicable) passant par des techniques de visualisation de graphes
3. Graph Learning : techniques d'apprentissage qui conduisent à des KGs
4. Machine Learning on Graphs : les KGs comme structures support

Focus essentiellement sur RAG

Les LLMs ou grands modèles de langage

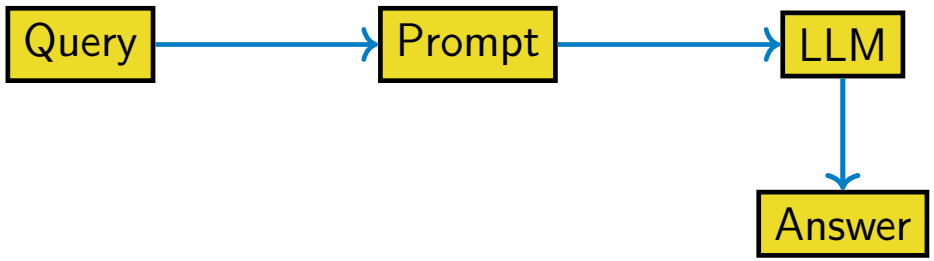
Les grands modèles de langage (LLM) : nouvelle classe de modèles de traitement du langage naturel (NLP) qui permettent :

1. d'obtenir des réponses à des questions ouvertes ou encore des résumés de contenus
2. de converser, de traduire
3. de générer du contenu et du code

Les LLMs sont entraînés sur des jeux de données très volumineux et font appel à des algorithmes d'apprentissage statistique complexes à partir desquels ils extrapolent les motifs et les structures du langage humain

voir `https://www.databricks.com/fr/glossary/large-language-models-llm`

Modalités de fonctionnement LLM

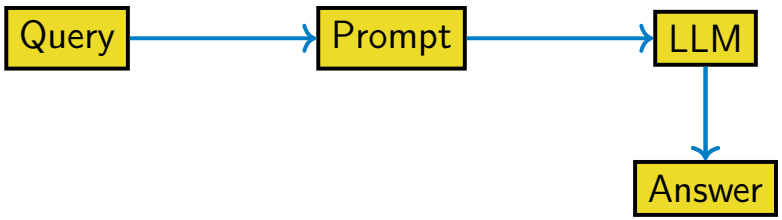


limites :

1. défaut d'actualisation des données,
2. données lacunaires

Problèmes dit d'hallucination

Modalités de fonctionnement LLM



Hannah Arendt et Günther Anders ne se sont pas mariés.
Ils ont eu une relation amoureuse, mais ils ne se sont jamais unis par le mariage.

Question posée à ChatGPT

quels sont la date et le lieu du mariage de Hannah Arendt et de Günther Anders ?

Contextualisation

L'idée est de tirer parti des LLMs mais aussi de tous les apports de la Recherche d'Information (IR)

→ conduit à la "génération augmentée de récupération" – plus connue sous le vocable RAG (pour retrieval augmented generation) – qui connecte des modèles d'IA générative à des sources de données internes ou personnalisées (entreprise, laboratoire, . . .)

L'objectif est une prise en considération des spécificités d'une entreprise et de ses secteurs d'activité, ou bien d'un contexte de R&D

Illustration RAG

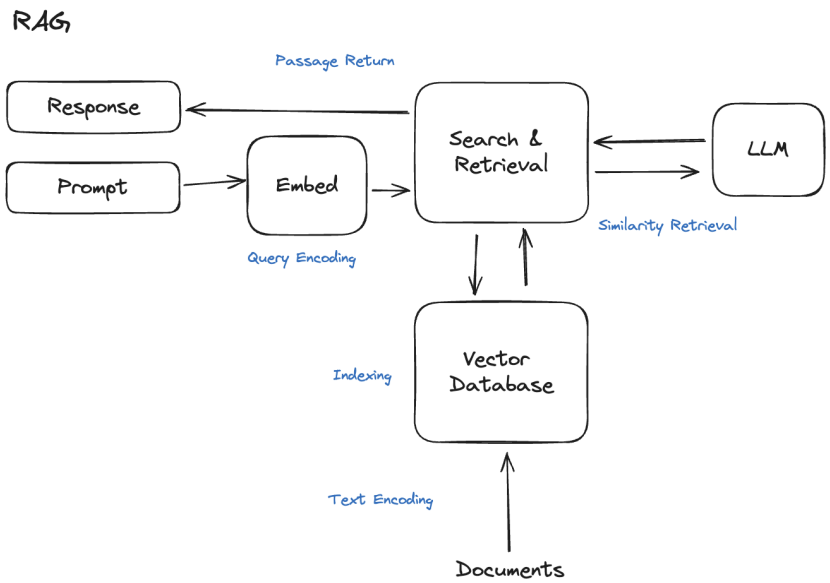


Figure: Combiner LLM et source documentaire

Crédit illustration :

superlinked.com/vectorhub/articles/improving-rag-knowledge-graphs

Les avantages du "GraphRAG"

Les sources de données internes (entreprise, laboratoire, ...) sont des graphes de connaissances (LPG ou TPG)

L'objectif est d'exploiter les liens entre entités et d'augmenter encore la prise en charge du contexte par les moteurs de recherche. Les usagers peuvent bénéficier ainsi de résultats fiables pleinement adaptés à leur recherche et au moindre coût

Illustration RAG

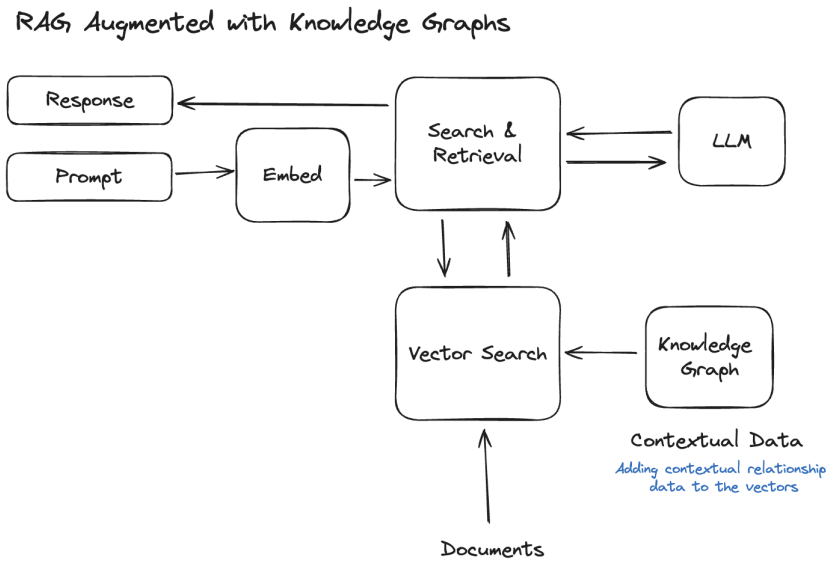


Figure: Combiner LLM et KG

Crédit illustration :

superlinked.com/vectorhub/articles/improving-rag-knowledge-graphs

Conclusion

Les KGs ne semblent pas menacés en terme de légitimité à véhiculer des données structurées

Les KGs spatio-temporels en organisant les entités au travers du temps et de l'espace peuvent s'avérer de vrais atouts dans des approches RAG pour compenser les limites connues des LLMs

Questions ?