

Propagation de propriétés entre entités contextuellement identiques

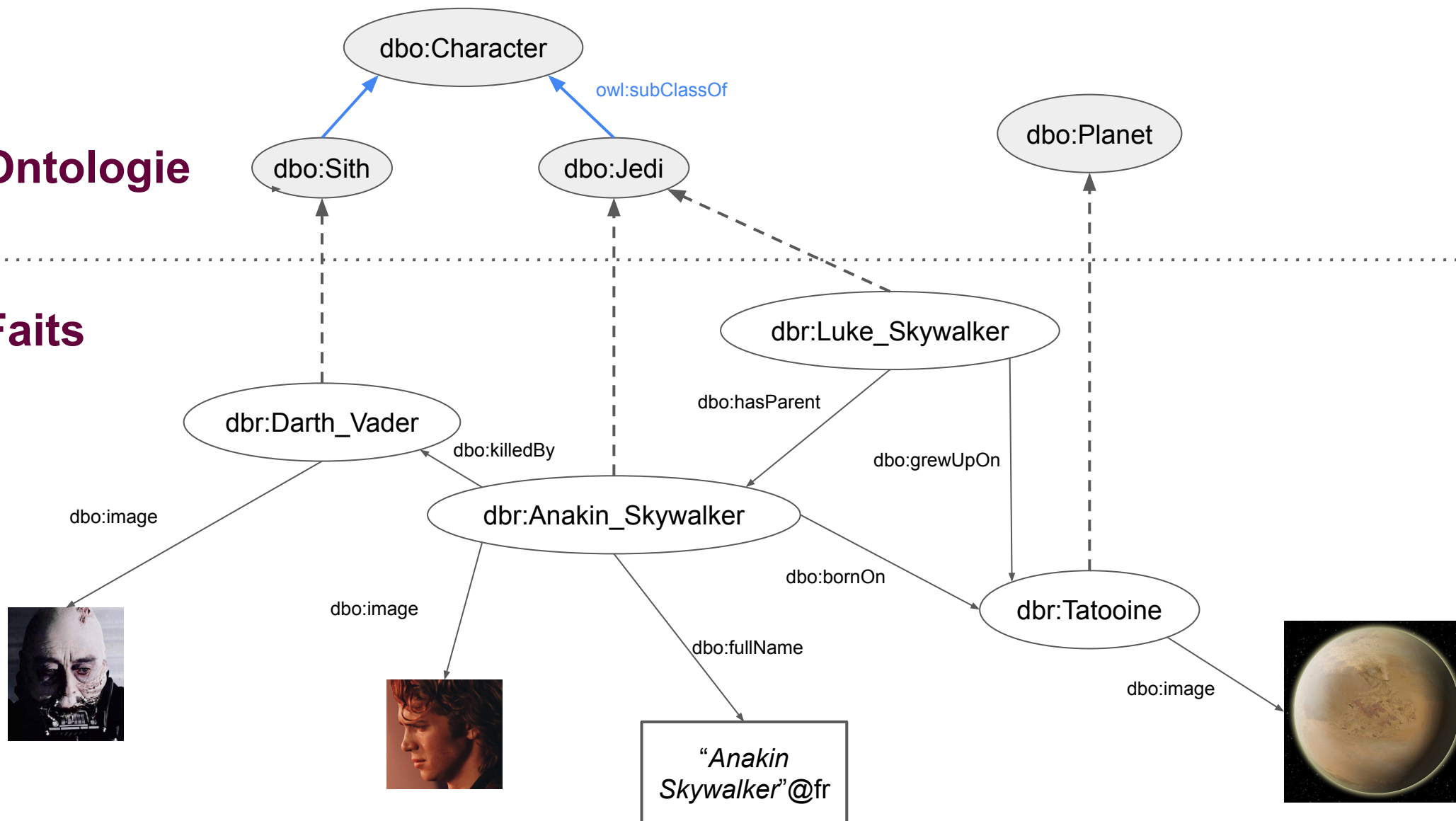
1. Les graphes de connaissances 101
2. Un peu d'histoire et de contexte
3. Etat des lieux
4. Une identité moins rigide ?
5. Comment propager des informations ?
6. Cas d'usage
7. Travaux futurs et conclusion

Les graphes de connaissances 101

Les graphes de connaissances

Ontologie

Faits

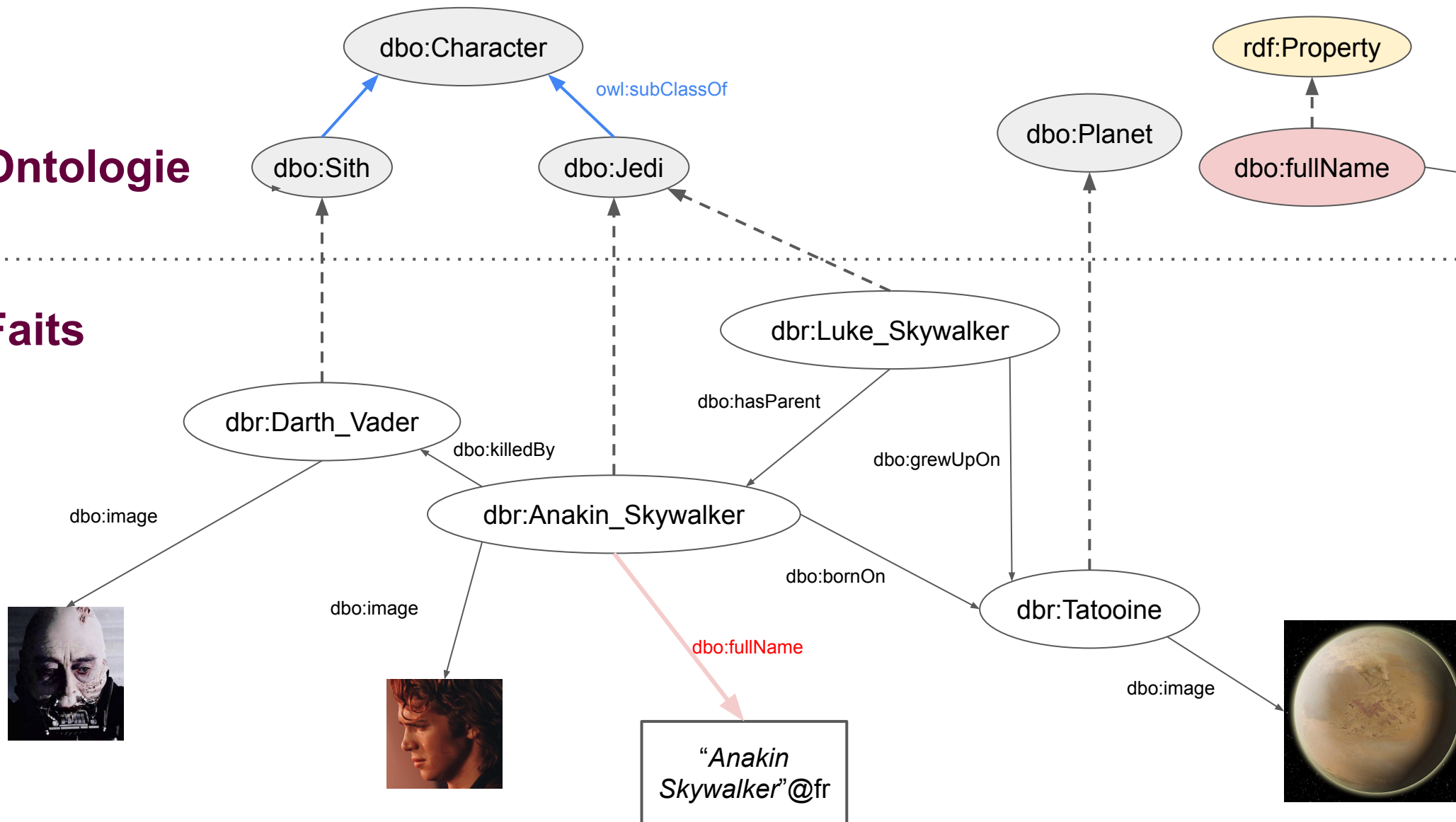


Les graphes de connaissances

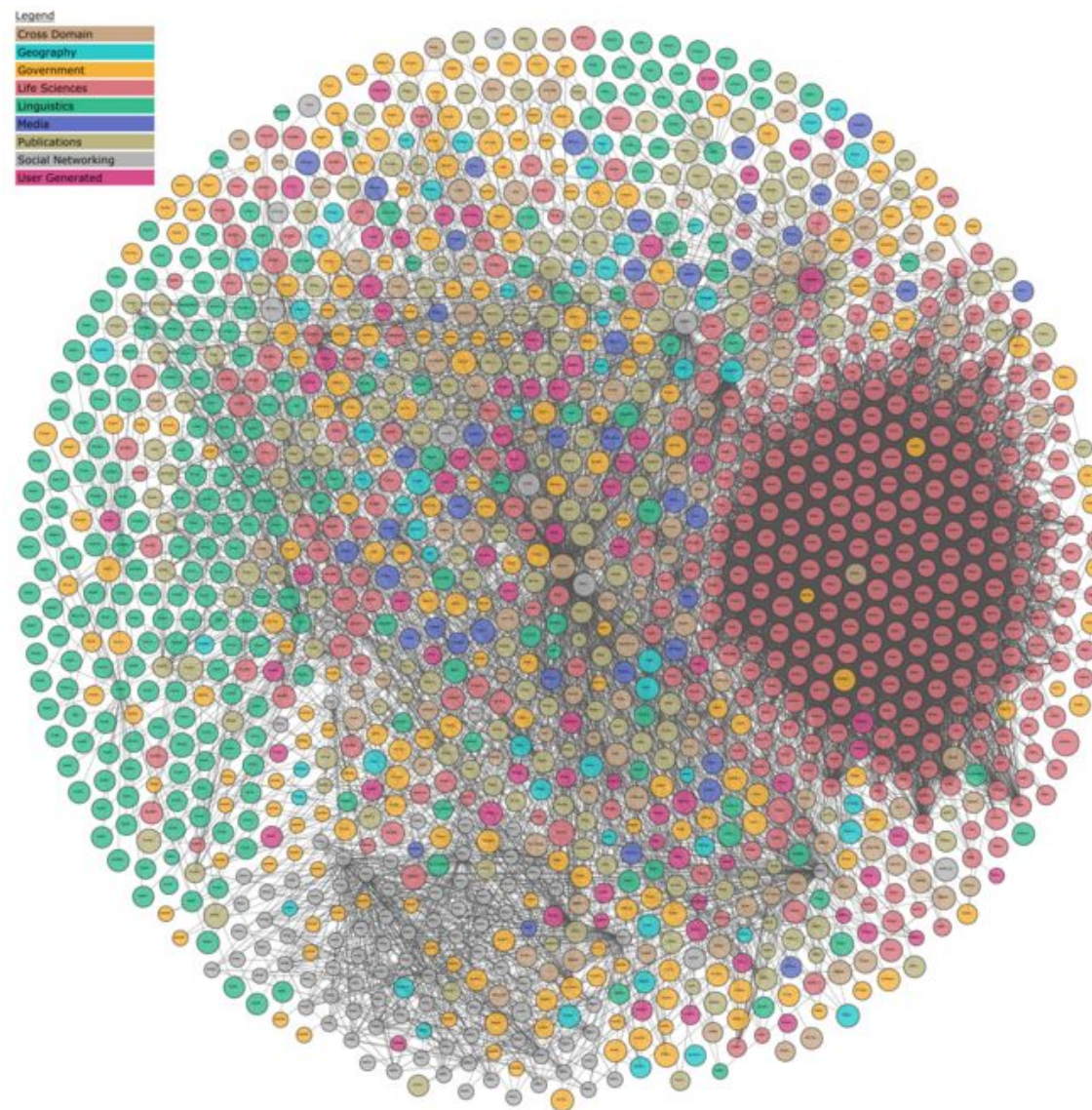
Ontologie



Faits



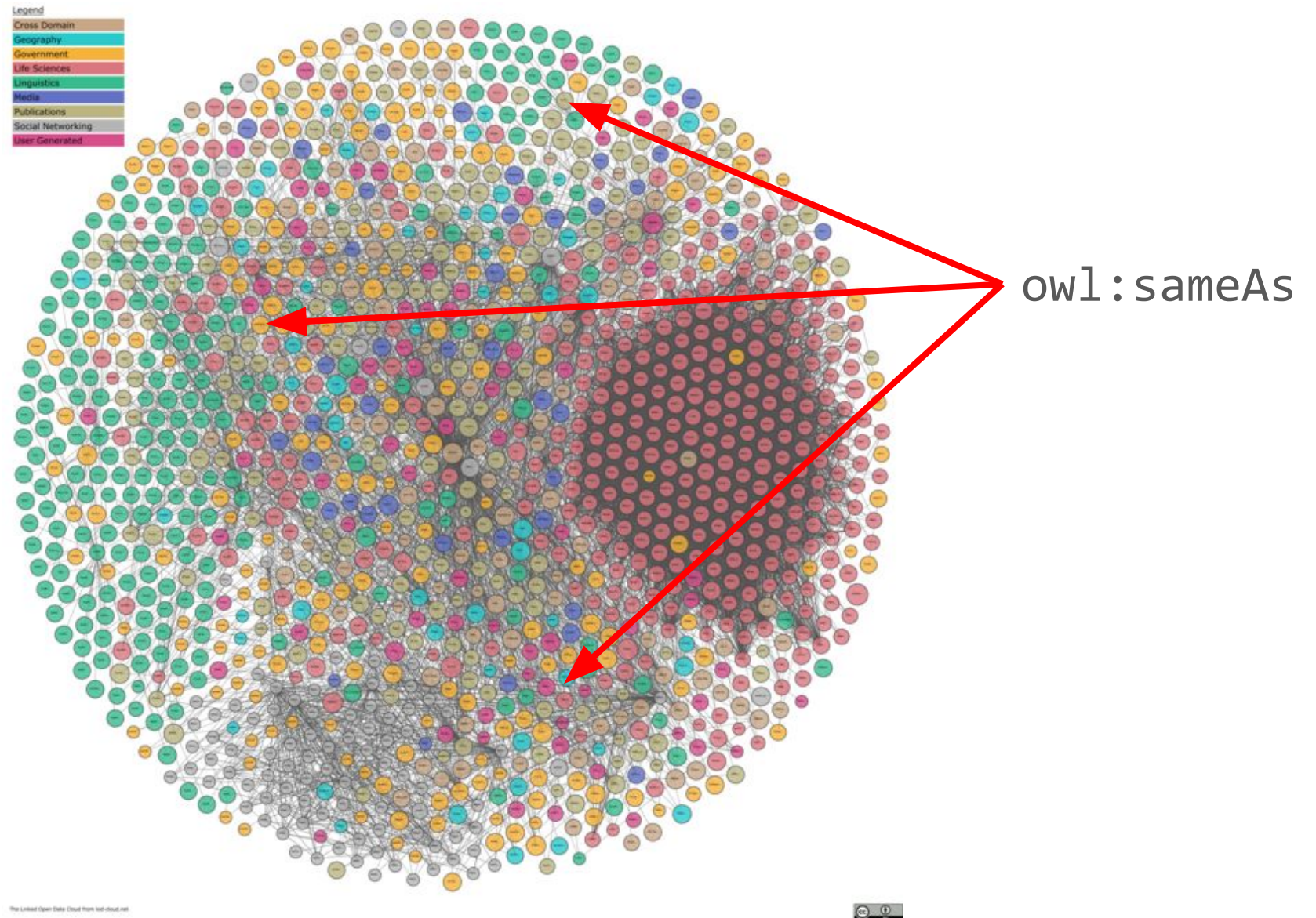
Les données liées



The Linked Open Data Cloud from lastcloud.net



Les données liées



Un peu d'histoire et de contexte



“Le président est arrivé !”



“Le président est arrivé !”

De quel président s’agit-il ?



“Le président est arrivé !”

De quel président s’agit-il ?



Emmanuel Macron

“Le président est arrivé !”

De quel président s’agit-il ?



Emmanuel Macron



Donald Trump

“Le président est arrivé !”

De quel président s’agit-il ?



Emmanuel Macron



Donald Trump



Philippe Augé

Contexte : Et la logique dans tout ça ?

- Logiques classiques $\rightarrow$ domaine, interprétation et modèle

$D = \{ \text{img1}, \text{img2}, \text{img3} \} : I_1(\text{President}) = \{ \text{img1}, \text{img3} \} \text{ ou } I_2(\text{President}) = \{ \text{img2} \}$

Contexte : Et la logique dans tout ça ?

- Logiques classiques $\rightarrow$ domaine, interprétation et modèle

$D = \{ \text{[Image of Emmanuel Macron]}, \text{[Image of Nicolas Sarkozy]}, \text{[Image of Donald Trump]} \} : I_1(\text{President}) = \{ \text{[Image of Emmanuel Macron]}, \text{[Image of Donald Trump]} \}$ ou $I_2(\text{President}) = \{ \text{[Image of Nicolas Sarkozy]} \}$

- Logiques contextuelles $\rightarrow$ John McCarthy

$$\forall c, \text{Country}(c, \text{France}) \wedge \text{Event}(c, \text{UM}, 2025) \rightarrow \text{ist}(c, \forall h, \text{Human}(h) \wedge \text{President}(h) \rightarrow \text{Name}(h, \text{XX}))$$

Contexte : Et la logique dans tout ça ?

- Logiques classiques → domaine, interprétation et modèle

$D = \{ \text{[Image of Emmanuel Macron]}, \text{[Image of Nicolas Sarkozy]}, \text{[Image of Donald Trump]} \} : I_1(\text{President}) = \{ \text{[Image of Emmanuel Macron]}, \text{[Image of Donald Trump]} \} \text{ ou } I_2(\text{President}) = \{ \text{[Image of Nicolas Sarkozy]} \}$

- Logiques contextuelles → John McCarthy

$$\forall c, \text{Country}(c, \text{France}) \wedge \text{Event}(c, \text{UM}, 2025) \rightarrow \text{ist}(c, \forall h, \text{Human}(h) \wedge \text{President}(h) \rightarrow \text{Name}(h, \text{XX}))$$

- Logiques modales

$$\Box((\neg \text{french} \vee \text{convicted}) \rightarrow \bigcirc \neg \text{being_elected_president})$$

L'identité : Un problème très ancien

Le paradoxe
de Thésée



L'identité : Un problème très ancien

Le paradoxe
de Thésée



Principe
d'identité



Le principe d'identité de Leibniz

Le principe d'identité des indiscernables

$$\forall x, \forall y \left(\forall p, \forall o, (\langle x, p, o \rangle \wedge \langle y, p, o \rangle) \Rightarrow x = y \right)$$

Le principe d'indiscernabilité des identiques

$$\forall x, \forall y \left(x = y \Rightarrow \forall p, \forall o, (\langle x, p, o \rangle \Rightarrow \langle y, p, o \rangle) \right)$$

La sémantique de *owl:sameAs* est fondée sur ces principes

Le principe d'identité des indiscernables

$$\forall x, \forall y \left(\forall p, \forall o, (\langle x, p, o \rangle \wedge \langle y, p, o \rangle) \Rightarrow x = y \right)$$

Le principe d'indiscernabilité des identiques

$$\forall x, \forall y \left(x = y \Rightarrow \forall p, \forall o, (\langle x, p, o \rangle \Rightarrow \langle y, p, o \rangle) \right)$$

La sémantique de *owl:sameAs* est fondée sur ces principes

⇒ Les couples propriété-valeur peuvent être **propagés** d'une entité à une autre entité identique

Le principe d'identité des indiscernables

$$\forall x, \forall y \left(\forall p, \forall o, (\langle x, p, o \rangle \wedge \langle y, p, o \rangle) \Rightarrow x = y \right)$$

Le principe d'indiscernabilité des identiques

$$\forall x, \forall y \left(x = y \Rightarrow \forall p, \forall o, (\langle x, p, o \rangle \Rightarrow \langle y, p, o \rangle) \right)$$

La sémantique de *owl:sameAs* est fondée sur ces principes

⇒ Les couples propriété-valeur peuvent être **propagés** d'une entité à une autre entité identique

+ **Accroît** l'exhaustivité des entités et la découverte d'informations

Le principe d'identité des indiscernables

$$\forall x, \forall y \left(\forall p, \forall o, (\langle x, p, o \rangle \wedge \langle y, p, o \rangle) \Rightarrow x = y \right)$$

Le principe d'indiscernabilité des identiques

$$\forall x, \forall y \left(x = y \Rightarrow \forall p, \forall o, (\langle x, p, o \rangle \Rightarrow \langle y, p, o \rangle) \right)$$

La sémantique de *owl:sameAs* est fondée sur ces principes

⇒ Les couples propriété-valeur peuvent être **propagés** d'une entité à une autre entité identique

- ➕ **Accroît** l'exhaustivité des entités et la découverte d'informations
- ➖ Propagation de **mauvaises informations** lorsque les liens sont faux

Le principe d'identité de Leibniz

Le principe d'identité

Le principe d'indisc



$p, \forall o, \langle x, p, o \rangle \wedge \langle y, p, o \rangle$
 $= y \Rightarrow \forall p, \forall o, (\langle x, p, o \rangle \wedge \langle y, p, o \rangle)$



La sémantique de *owl:sameAs* est fondée sur ces principes

⇒ Les couples propriété-valeur peuvent être **propagés** d'une entité à une autre entité identique

- + **Accroît** l'exhaustivité des entités et la découverte d'informations
- Propagation de **mauvaises informations** lorsque les liens sont faux

Le principe d'identité de Leibniz

Le principe d'identité

Le principe d'indiscernabilité



$x = y \Rightarrow \forall p, \forall o, (\langle x, p, o \rangle \wedge \langle y, p, o \rangle)$



La sémantique de *owl* : sémantique est fondée sur ces principes

⇒ Les couples propriété-valeur peuvent être **propagés** d'une entité à une autre entité identique



Accès



Propriété



entités le découvert
es informations lorsqu



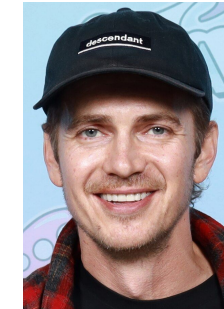
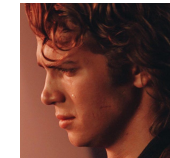
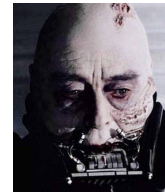
faux

Etat des lieux

La crise d'identité des données liées

L'identité est souvent sous-déterminée:

- connaissance partielle
- dépendance du contexte
- conventions sociales

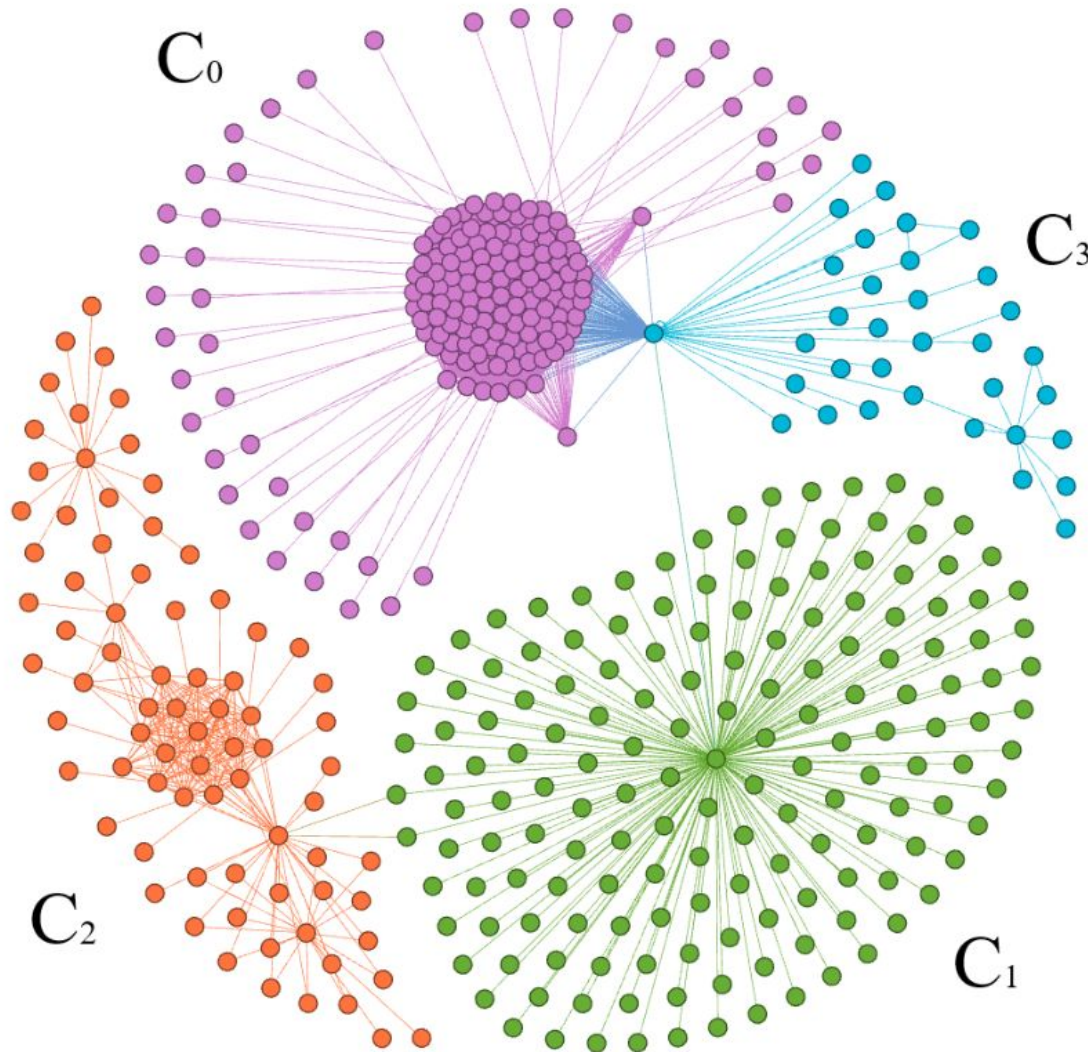


Souvent utilisé pour indiquer :

- l'identité contextuelle ou sociale (par exemple, des noms variant selon le contexte)
- la similitude (par exemple, des éléments apparentés mais non identiques)
- la revendication d'une identité par un agent, qui ne fait pas l'objet d'un consensus universel

Ces usages conduisent à des **déductions incorrectes**, en particulier lorsque les systèmes de raisonnement supposent une **substitutivité totale**.

La crise d'identité des données liées



Les communautés détectées à partir de l'ensemble d'égalité contenant le terme `dbp:Barack_Obama` en utilisant l'algorithme de Louvain.
Les 4 communautés détectées sont distinguées par la couleur de leurs noeuds.

Aspects pratiques :

- **Manque de owl:differentFrom** (seulement ~3.6k vs 558M owl:sameAs en 2015).
- Les générateurs de liens heuristiques ont une **précision limitée** (67-86% dans les évaluations de l'OAEI).
- Le **désaccord humain** sur ce qui compte comme « identique » montre la subjectivité impliquée.

Alternatives à owl:sameAs :

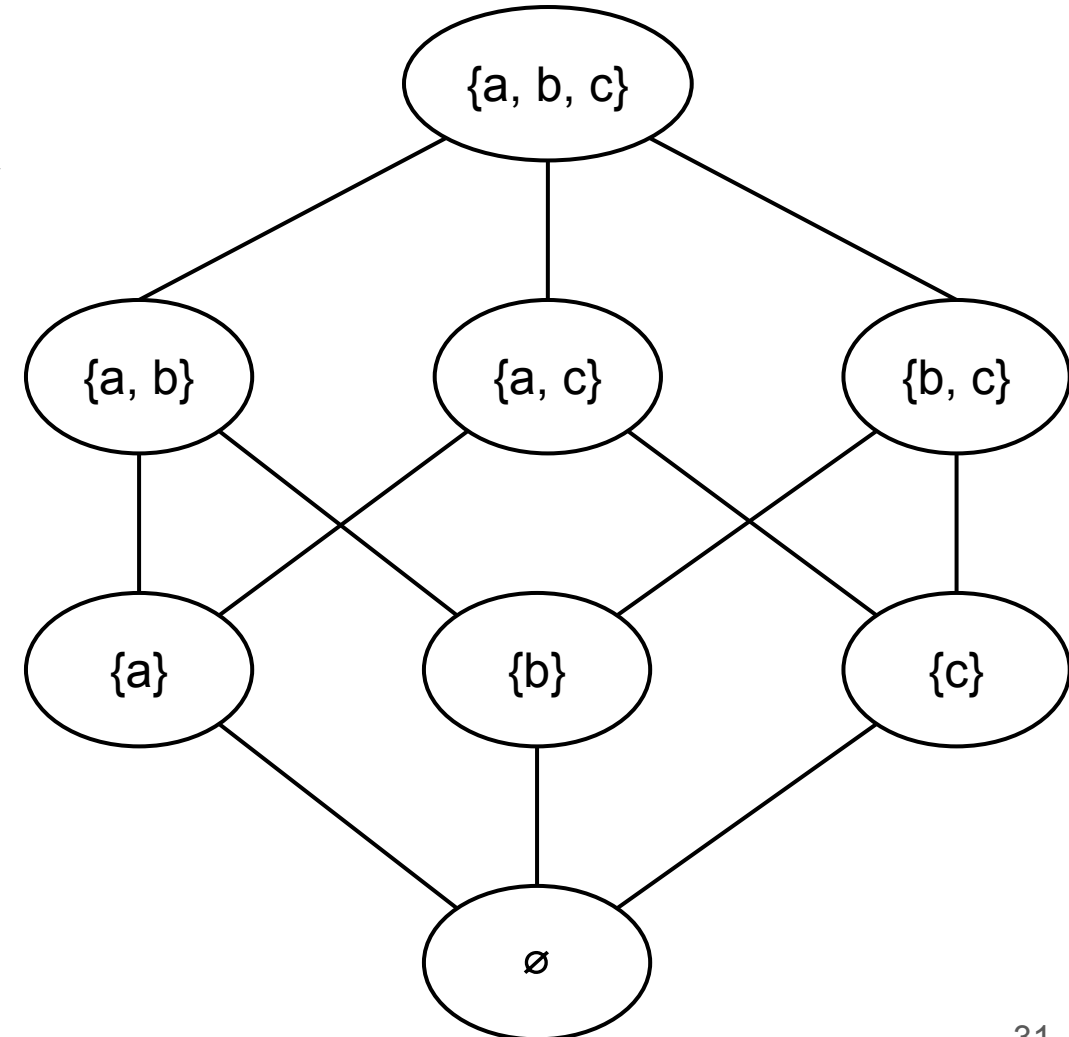
- Les prédicats SKOS : skos:exactMatch, skos:closeMatch - des formes plus douces d'identité.
- wdt:P2888 de Wikidata : utilisé comme skos:exactMatch.
- isLike d'UMBEL : similarité non stricte.
- Similarity Ontology (SO) : définit une hiérarchie de prédicats d'identité/similarité avec des propriétés formelles (réflexivité, symétrie, transitivité).

L'utilisation reste **faible** - la grande majorité des liens utilisent encore owl:sameAs.

Une identité moins rigide ?

- **Contexte d'identité** = ensemble de propriétés (ensemble d'indiscernabilité)
 - Les entités doivent **partager la même valeur pour chaque propriété**
- Les contextes peuvent être représentés par un **treillis**.

Treillis d'identité

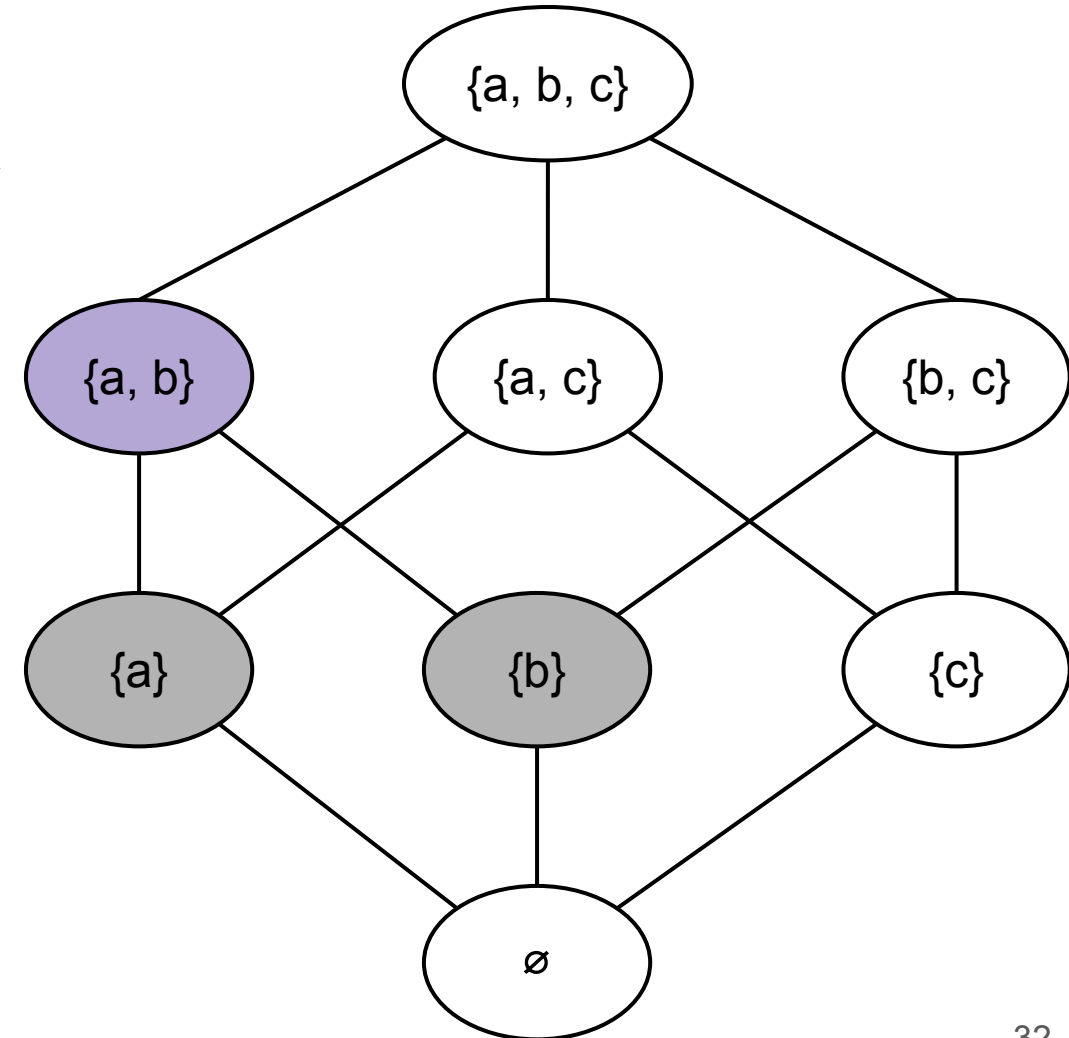


- **Contexte d'identité** = ensemble de propriétés (ensemble d'indiscernabilité)
 - Les entités doivent partager la même valeur pour chaque propriété
- Les contextes peuvent être représentés par un **treillis**.

$\langle x, a, o1 \rangle \wedge \langle \text{graine}, a, o1 \rangle$

$\langle x, b, o2 \rangle \wedge \langle \text{graine}, b, o2 \rangle$

Treillis d'identité



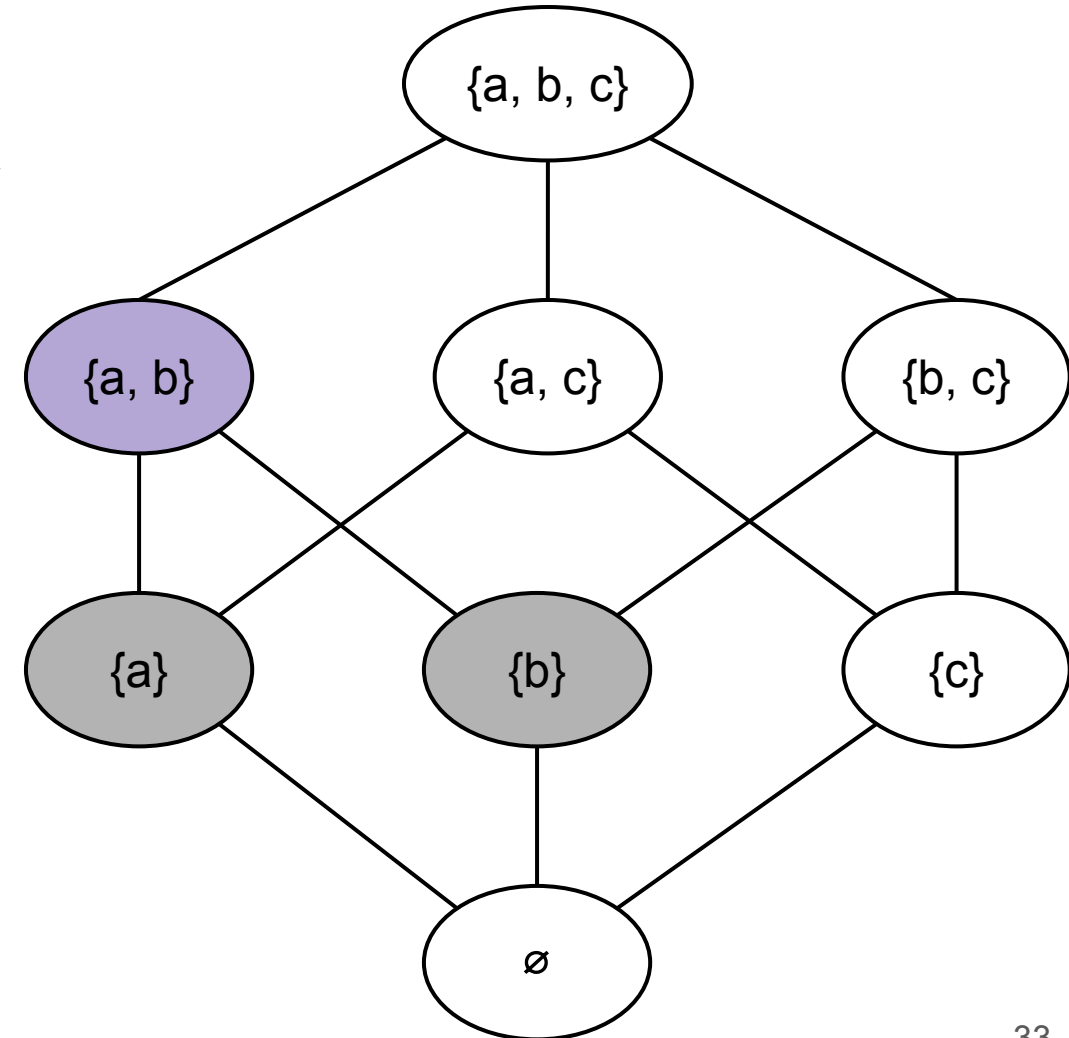
- **Contexte d'identité** = ensemble de propriétés (ensemble d'indiscernabilité)
 - Les entités doivent partager la même valeur pour chaque propriété
- Les contextes peuvent être représentés par un **treillis**.

$\langle x, a, o1 \rangle \wedge \langle \text{graine}, a, o1 \rangle$

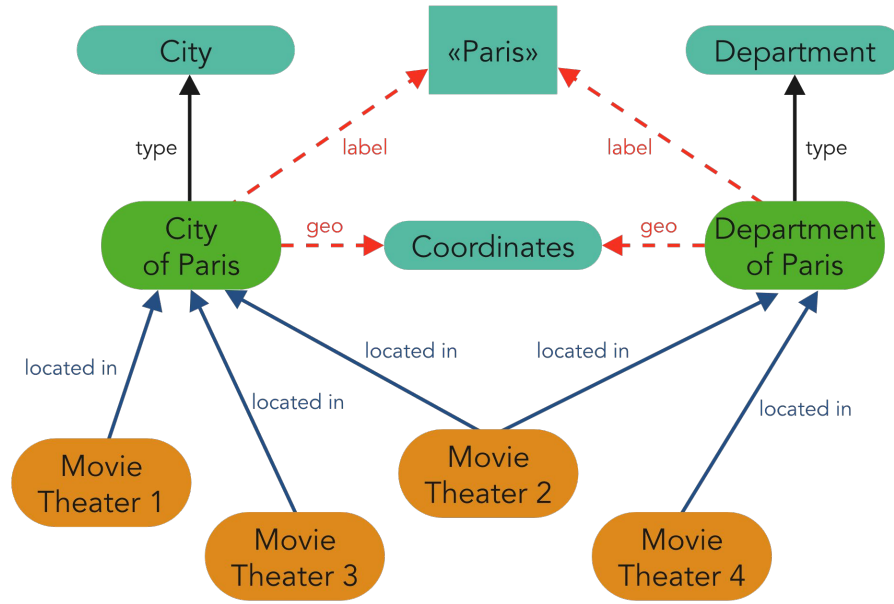
$\langle x, b, o2 \rangle \wedge \langle \text{graine}, b, o2 \rangle$

Si $\langle x, e, o3 \rangle$, a-t-on $\langle \text{graine}, e, o3 \rangle$?

Treillis d'identité

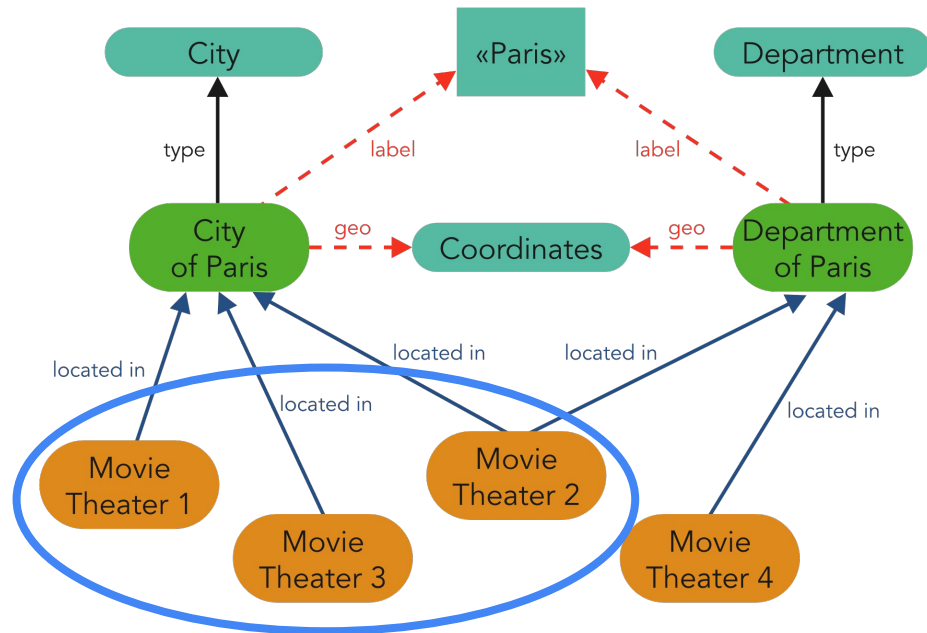


Comment propager des informations ?

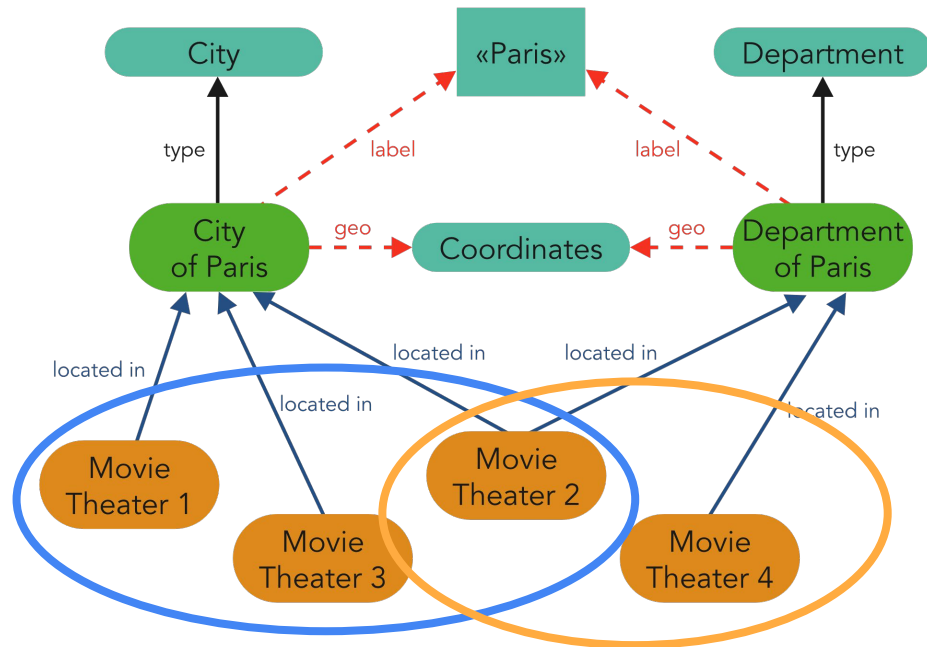


- Calculer semi-automatiquement les propriétés propageables
- Fondée sur les plongements de phrases des descriptions longues des propriétés

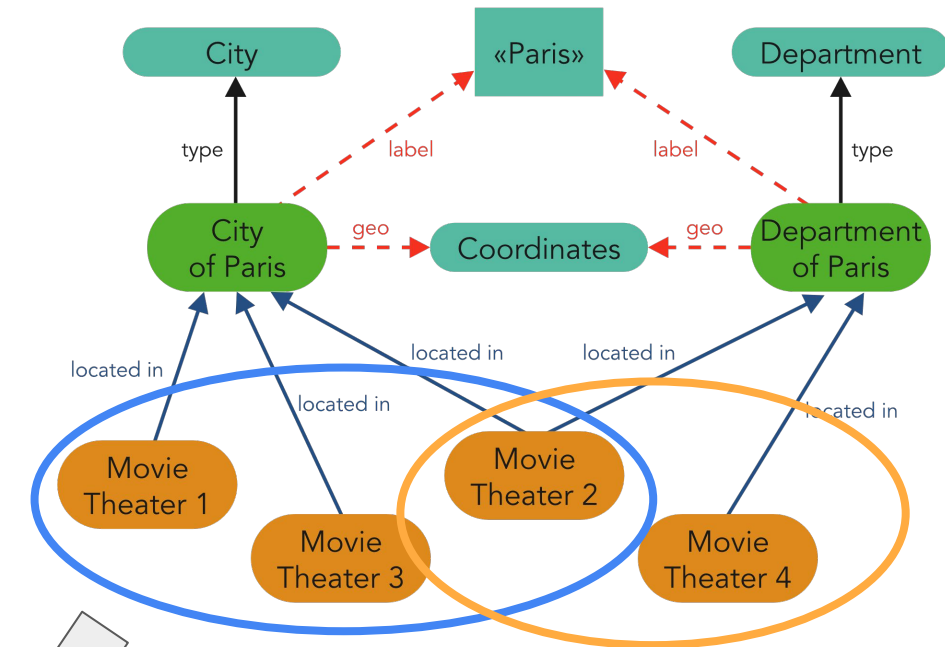
Principe de l'approche



Principe de l'approche



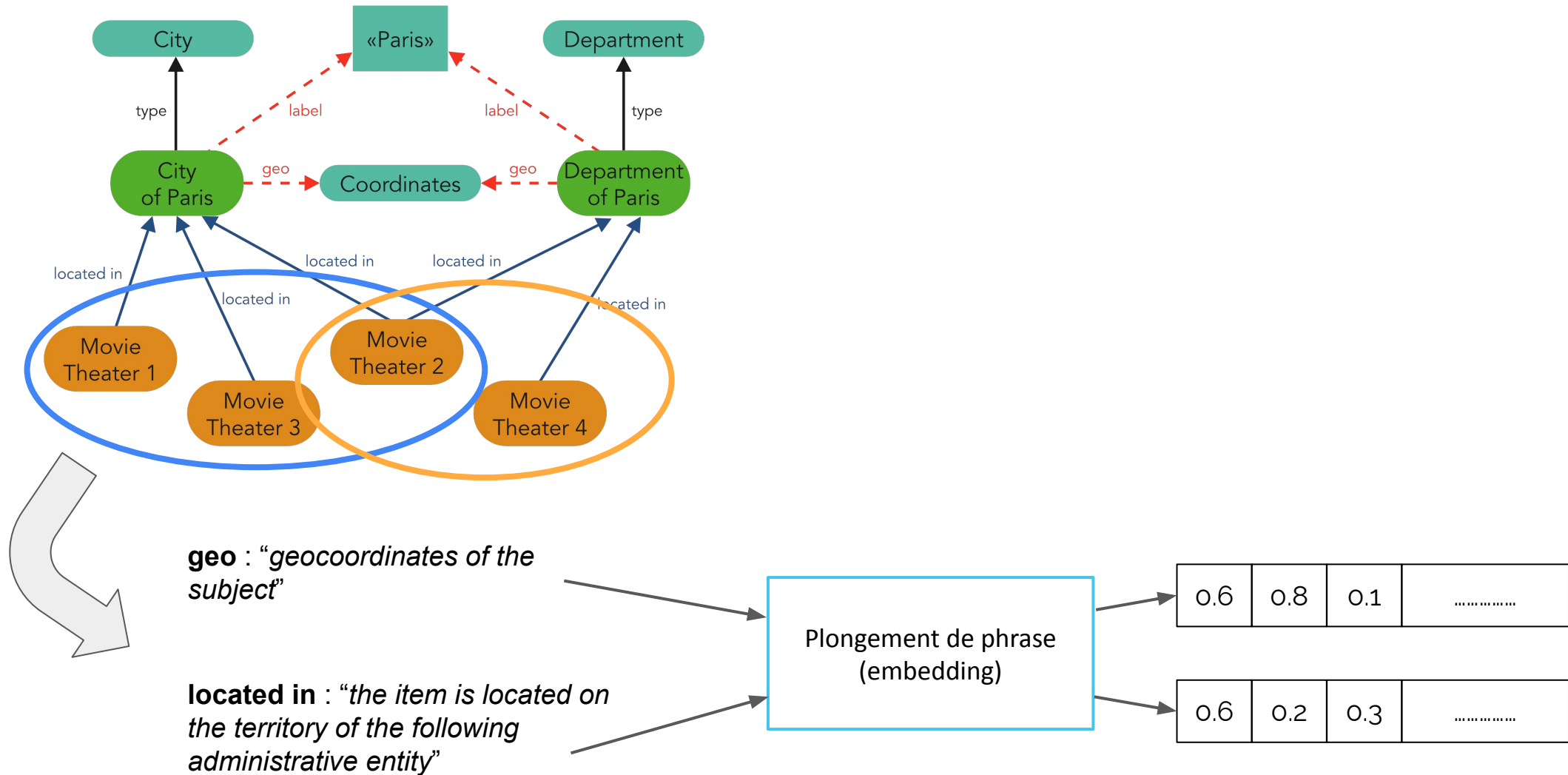
Principe de l'approche



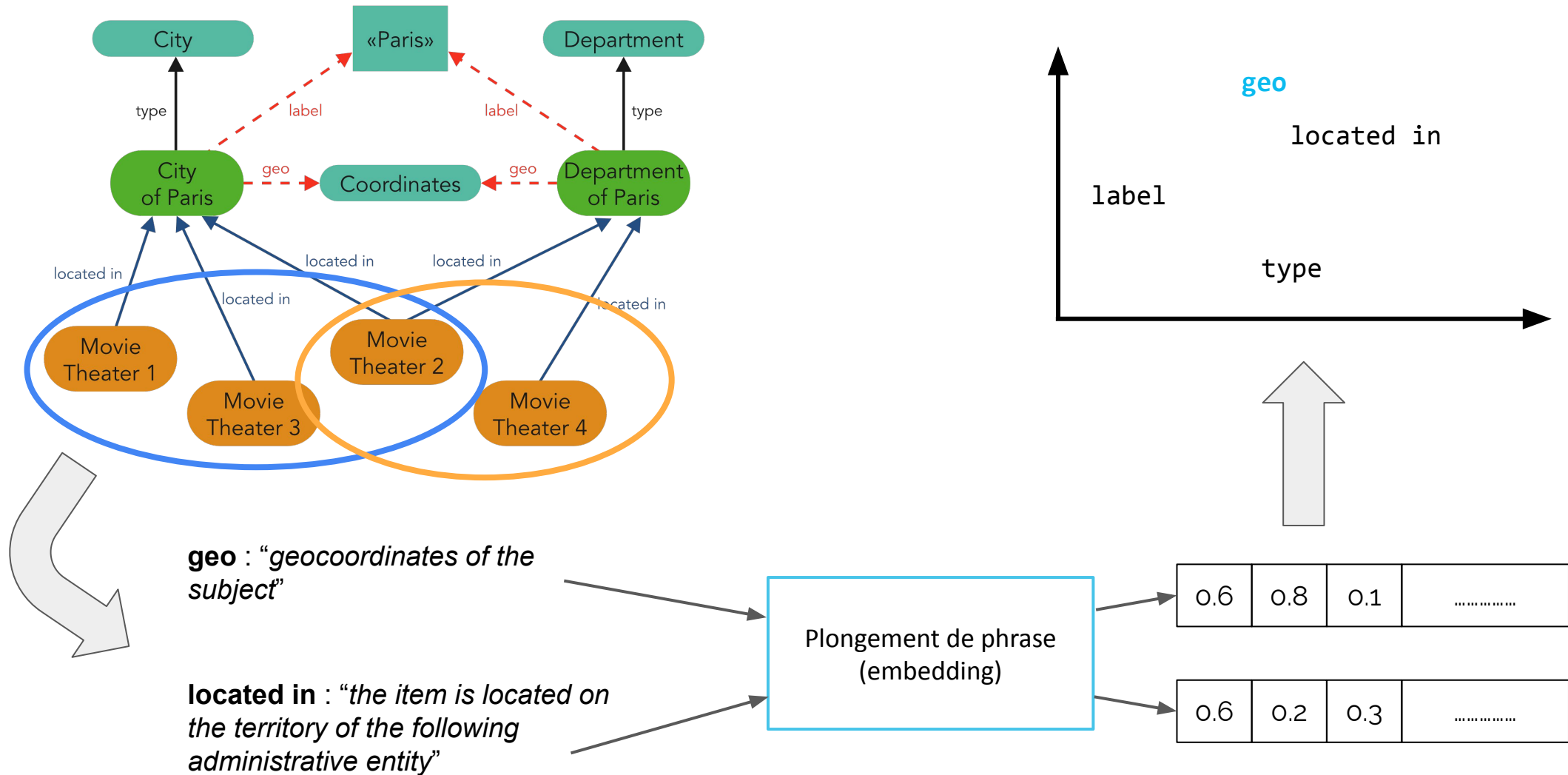
geo : *“geocoordinates of the subject”*

located in : *“the item is located on the territory of the following administrative entity”*

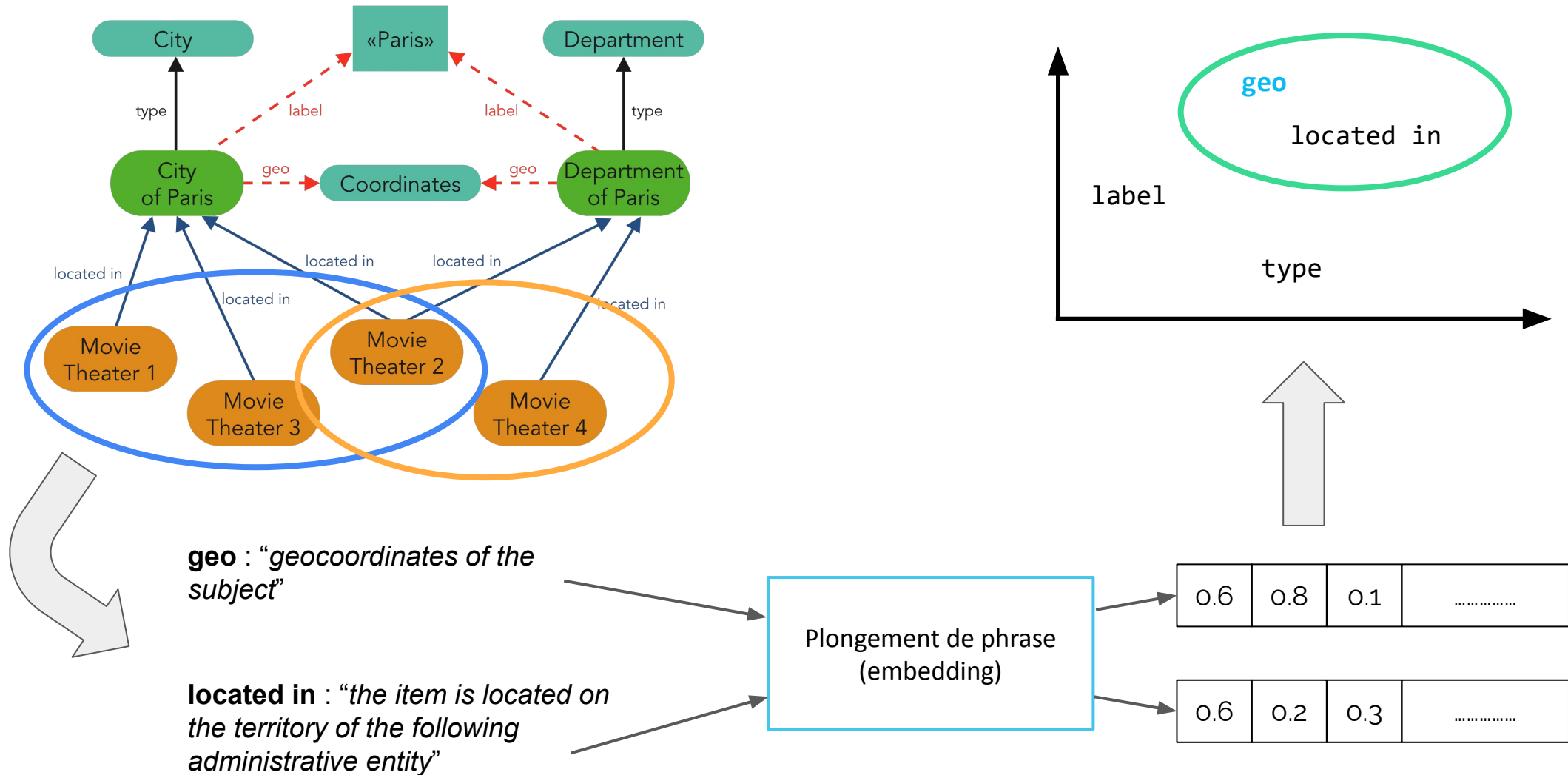
Principe de l'approche



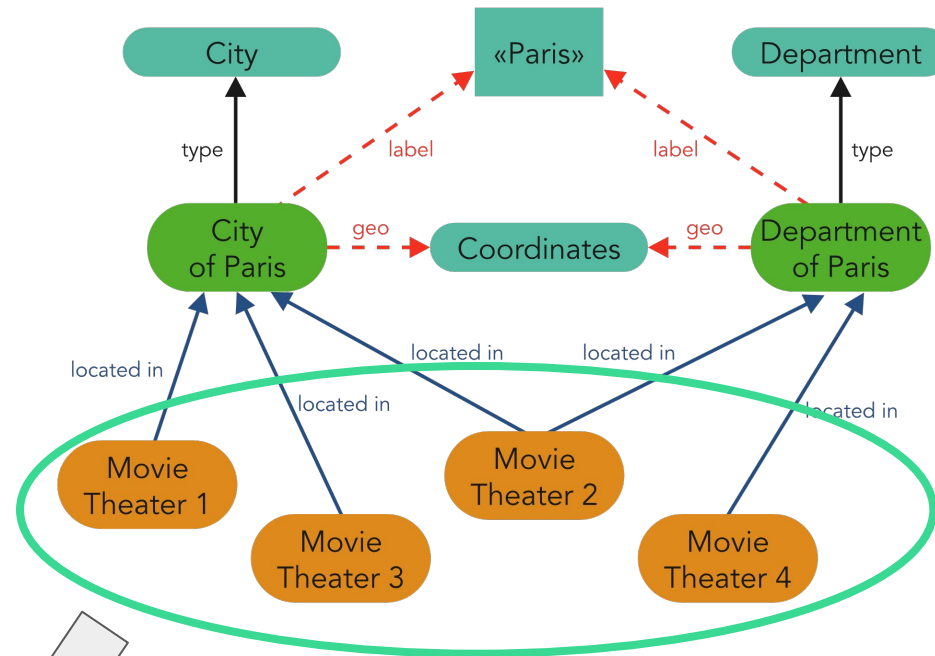
Principe de l'approche



Principe de l'approche



Principe de l'approche



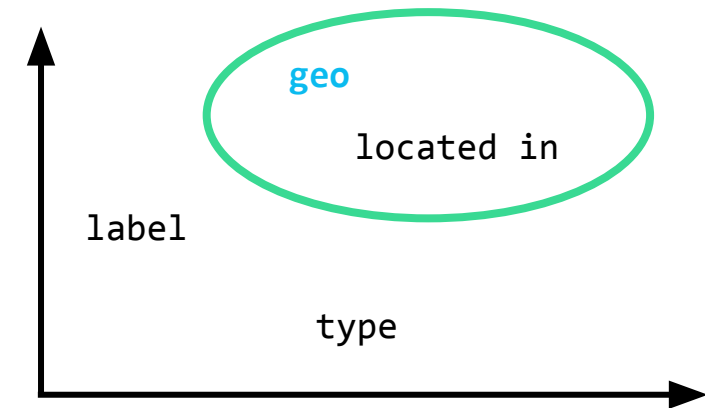
geo : “geocoordinates of the subject”

located in : “the item is located on the territory of the following administrative entity”

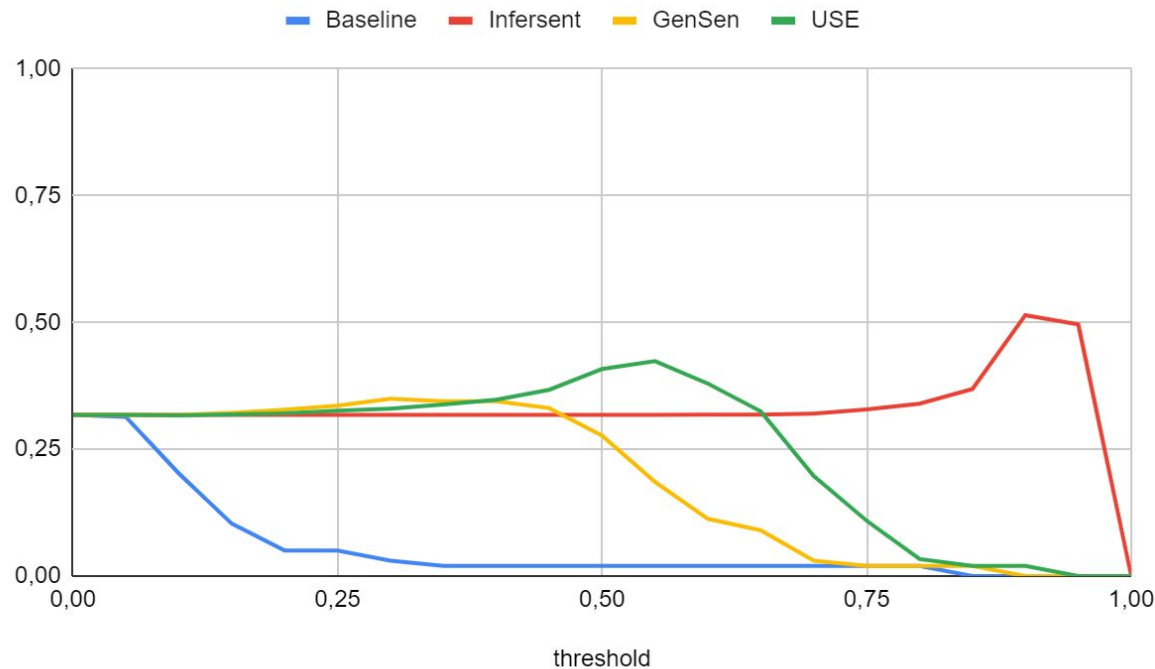
Plongement de phrase
(embedding)

0.6	0.8	0.1	
-----	-----	-----	-------

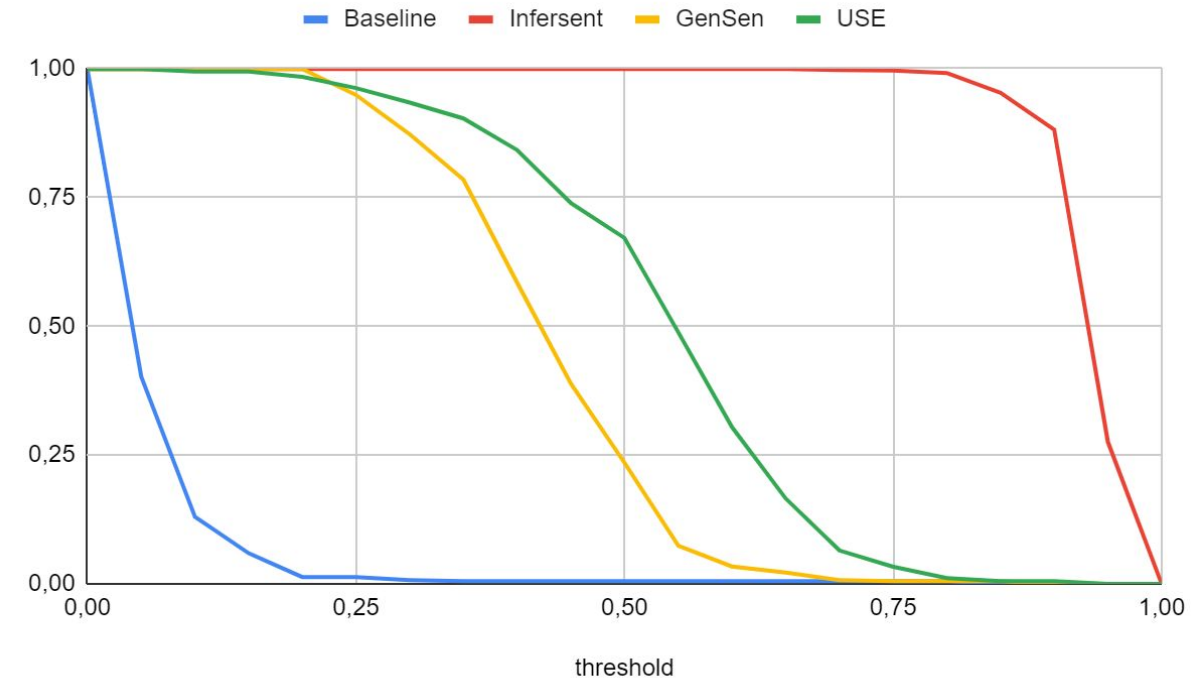
0.6	0.2	0.3	
-----	-----	-----	-------



Précision



Rappel



Gold standard = 100 entités (5 classes)

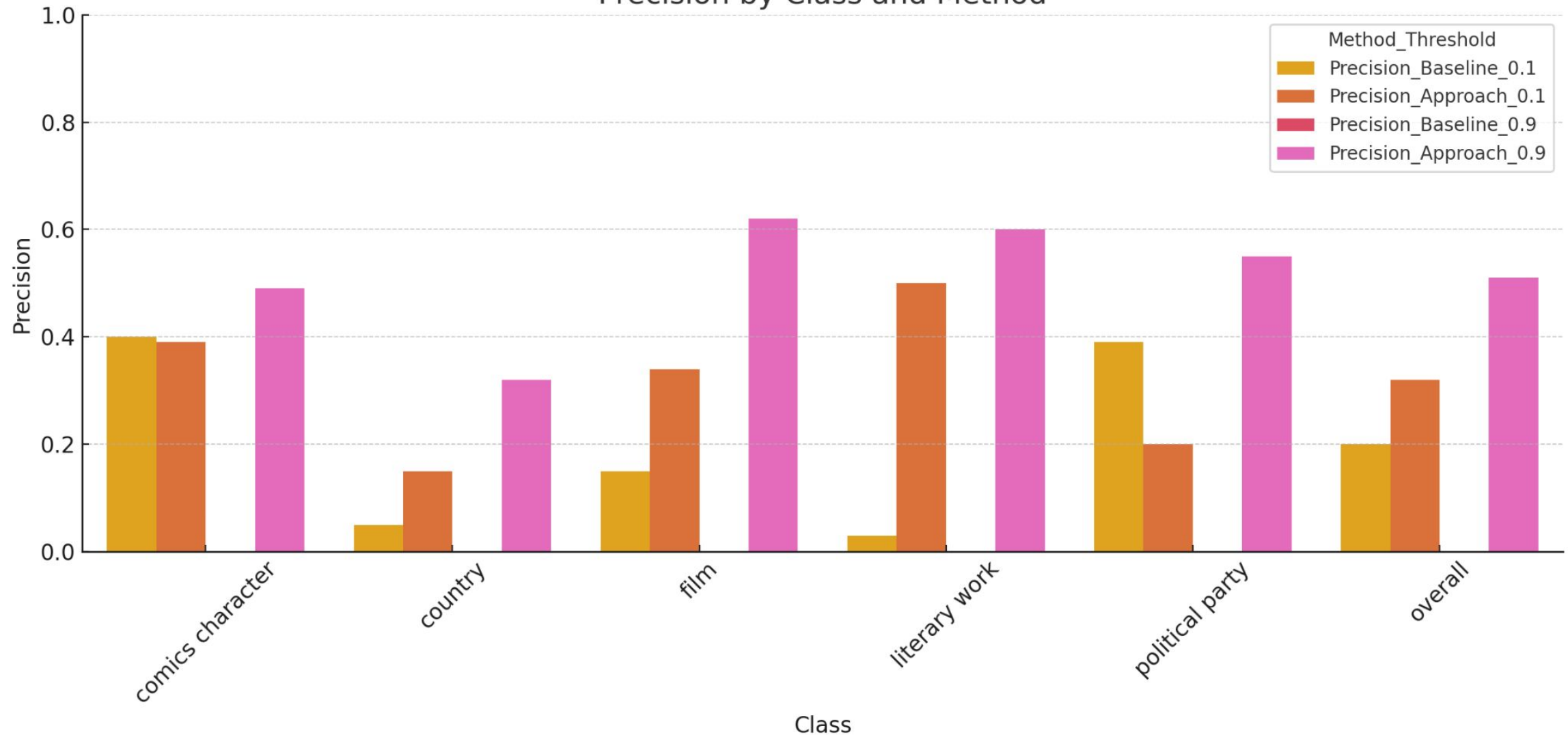
Baseline vs Infersent vs GenSen vs USE
⇒ Le gagnant est Infersent

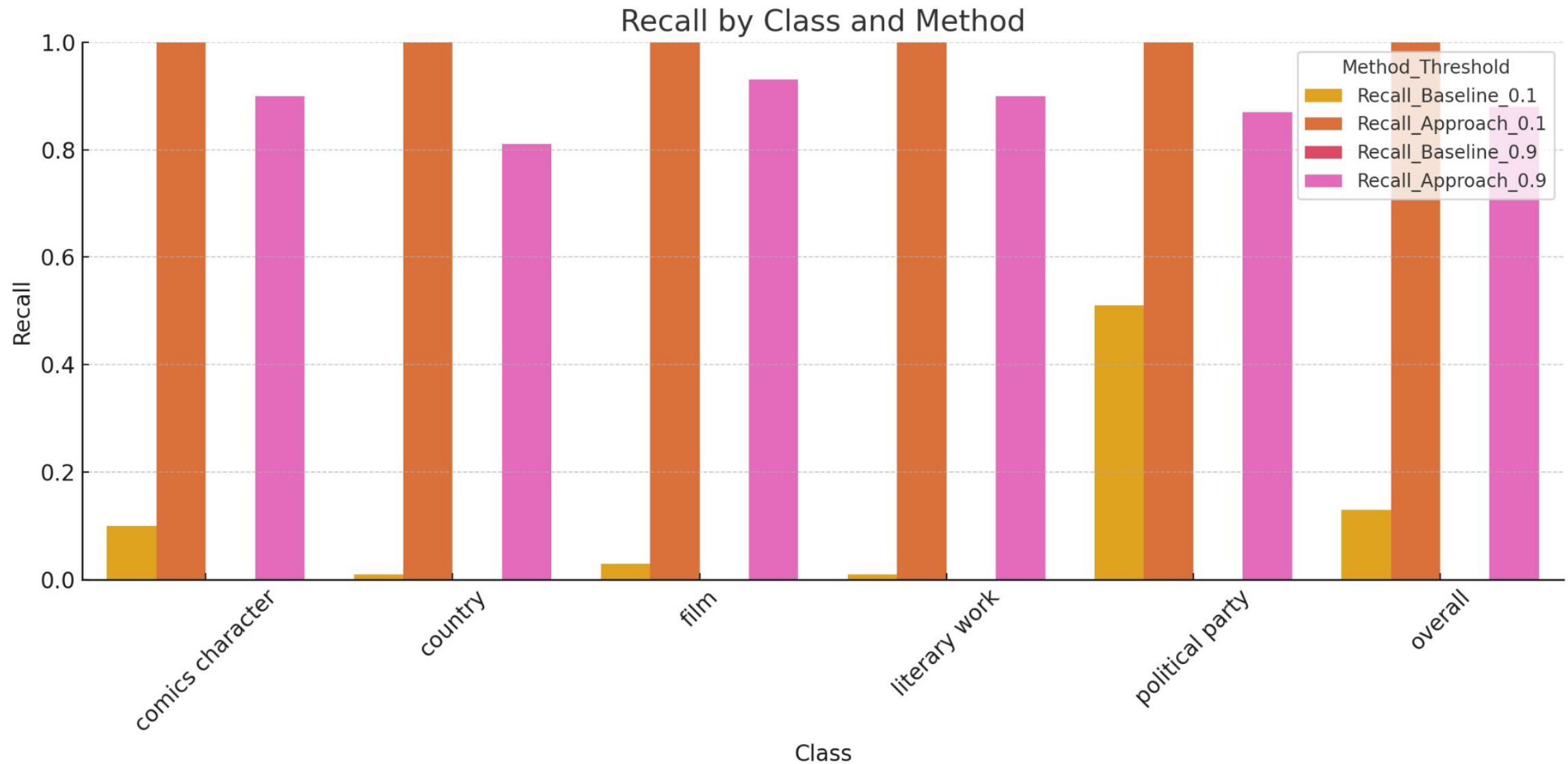


Les descriptions textuelles sont utiles pour découvrir les propriétés qui sont propageables.

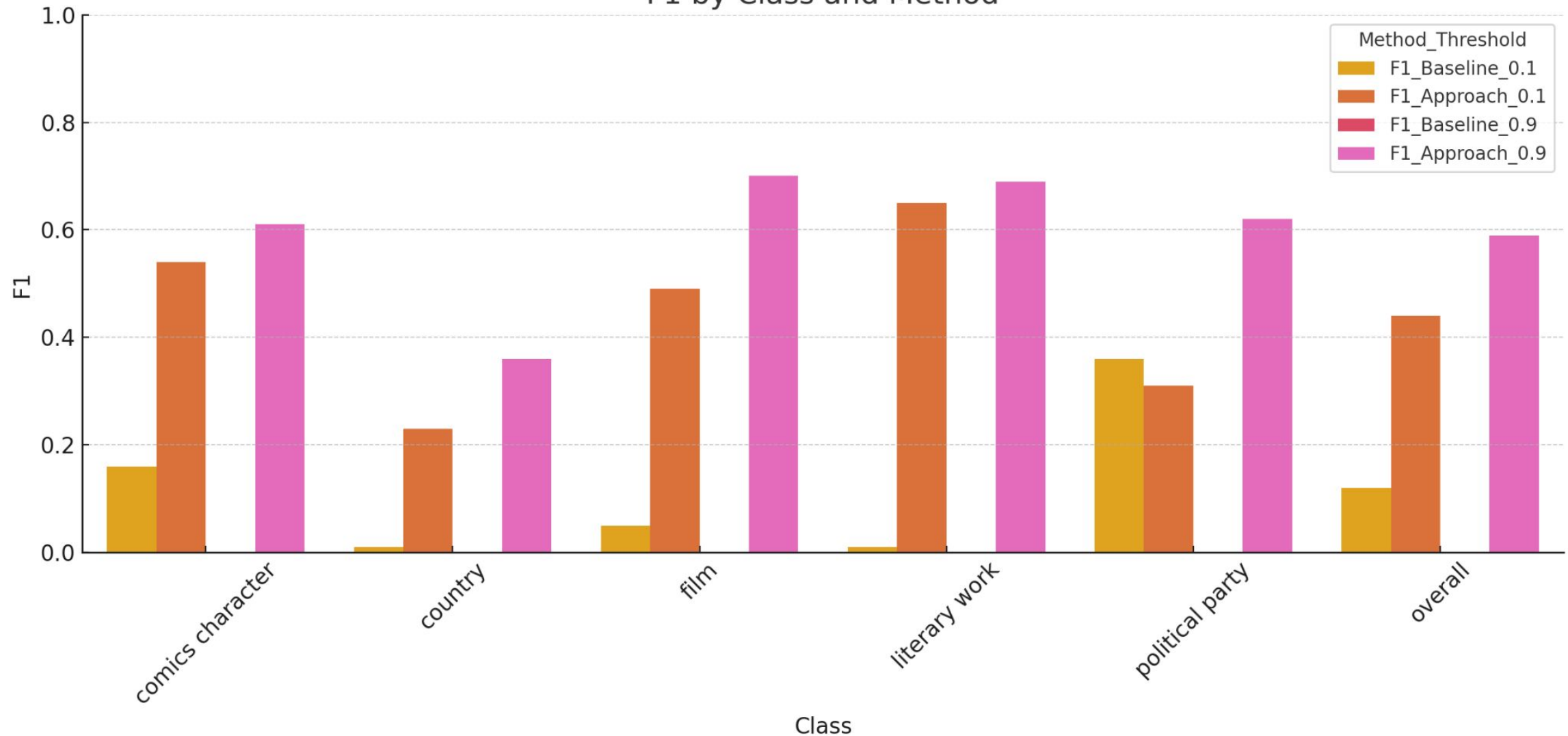
Très dépendant de l'encodeur.

Precision by Class and Method





F1 by Class and Method



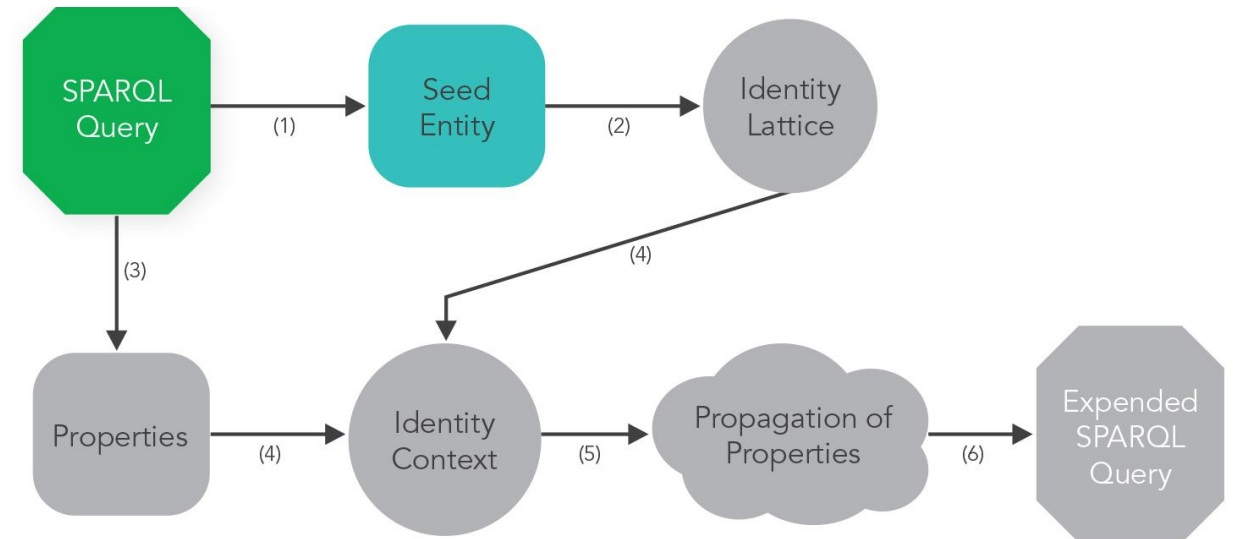
Cas d'usage



certaines images peuvent choquer

Exemple

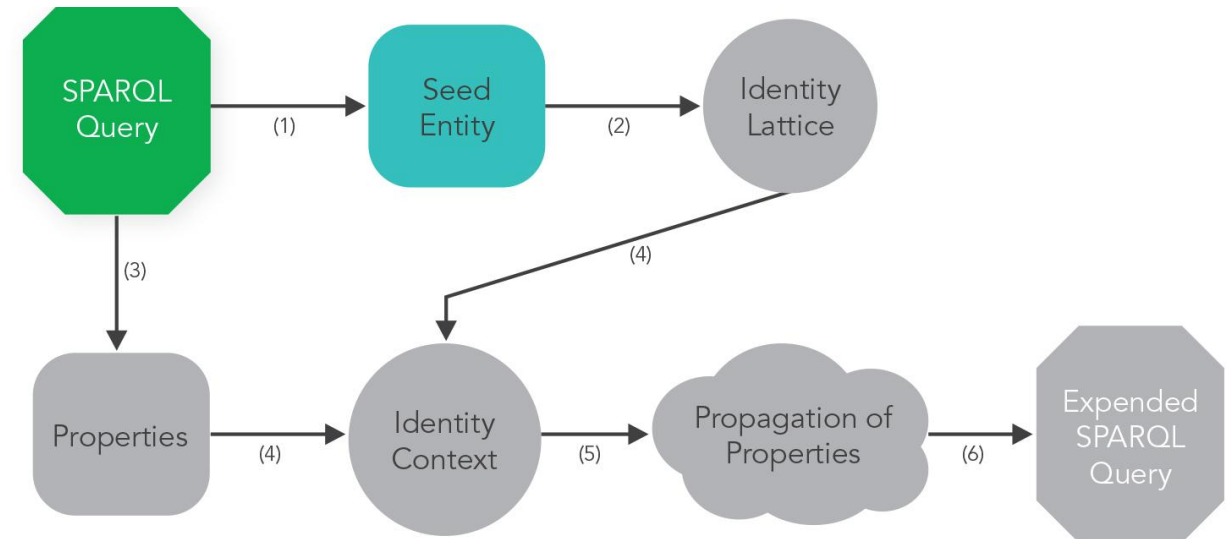
Nombre de républicains condamnés ?



Exemple

Nombre de républicains condamnés ?

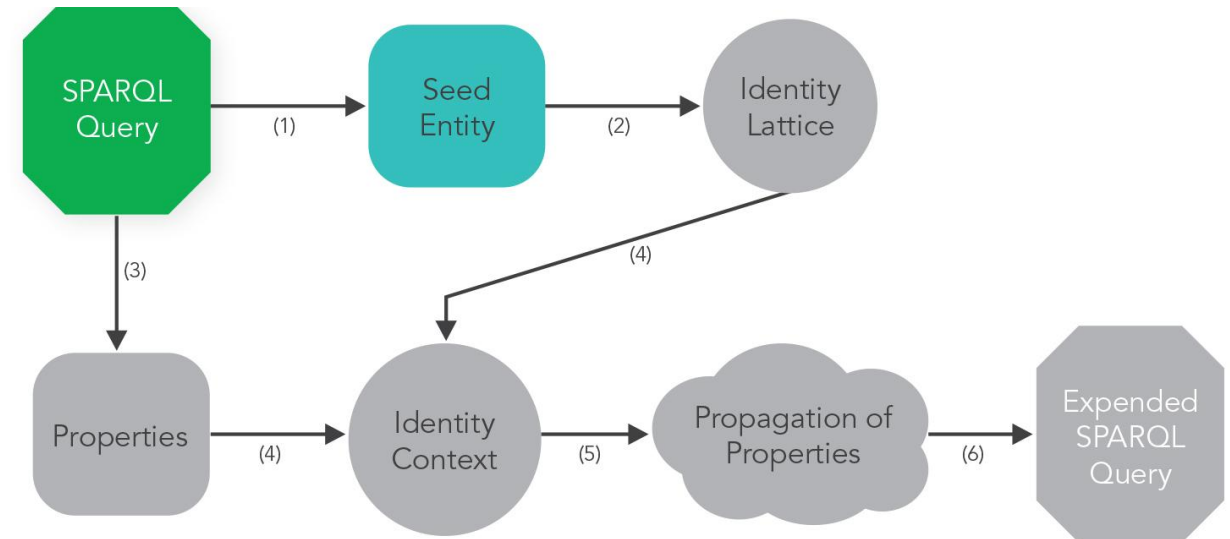
```
SELECT DISTINCT ?politician ?crime
WHERE {
    ?politician :memberOf :TheRepublicans ;
               :convictedOf ?crime .
}
```



Exemple

Nombre de républicains condamnés ?

```
SELECT DISTINCT ?politician ?crime
WHERE {
    ?politician :memberOf :TheRepublicans ;
               :convictedOf ?crime .
}
```



# de résultats sans contexte	2
---------------------------------	---

Exemple

Nombre de rép

```
SELECT DISTINCT
```

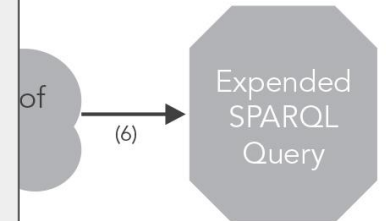
```
WHERE {
```

```
  ?politician
```

```
    :convic
```

```
}
```

de résultat
contexte



Exemple

Nombre de rép

```
SELECT DISTINCT
```

```
WHERE {
```

```
  ?politician
```

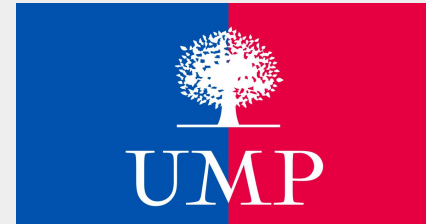
```
    :convic
```

```
}
```

de résultat
contexte



:memberOf (Wikidata)



of

(6)

Expended
SPARQL
Query

Exemple

Nombre de rép

```
SELECT DISTINCT
```

```
WHERE {
```

```
  ?politician
```

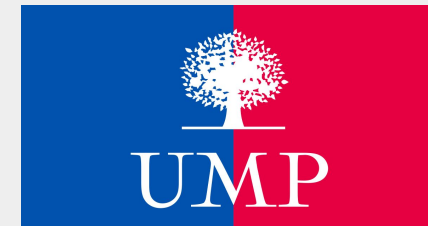
```
    :convic
```

```
}
```

de résultat
contexte



:memberOf (Wikidata)



:memberOf (en réalité)



of

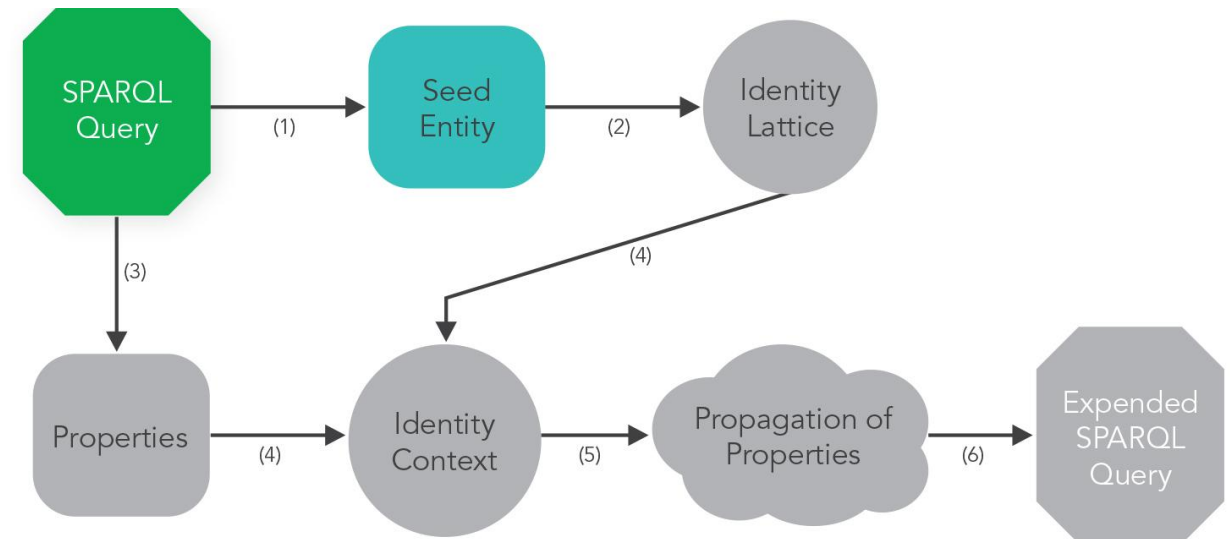
(6)

Expended
SPARQL
Query

Exemple

Nombre de républicains condamnés ?

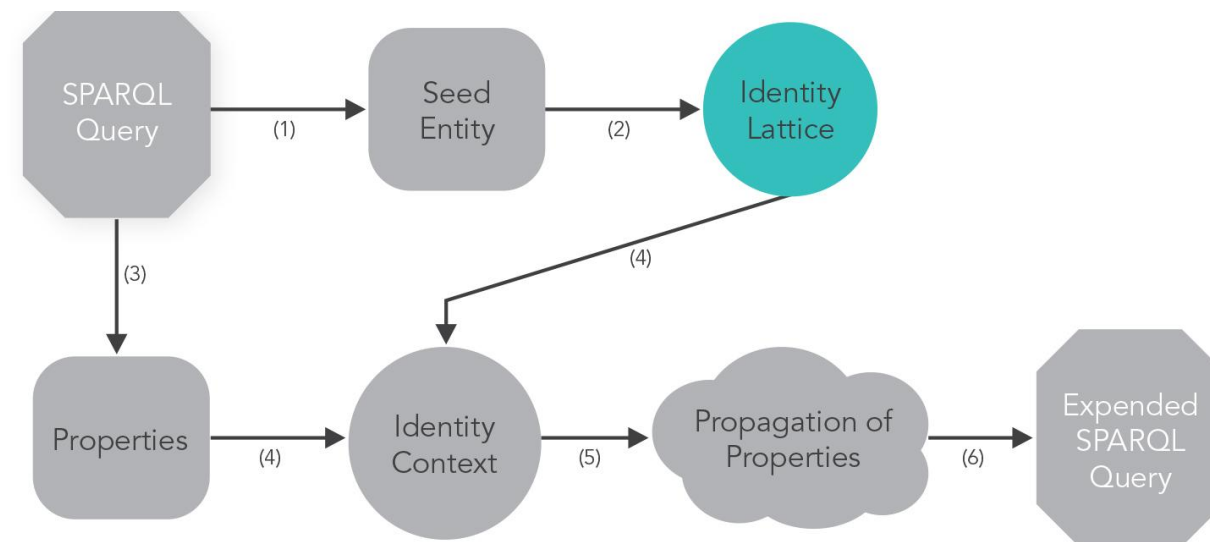
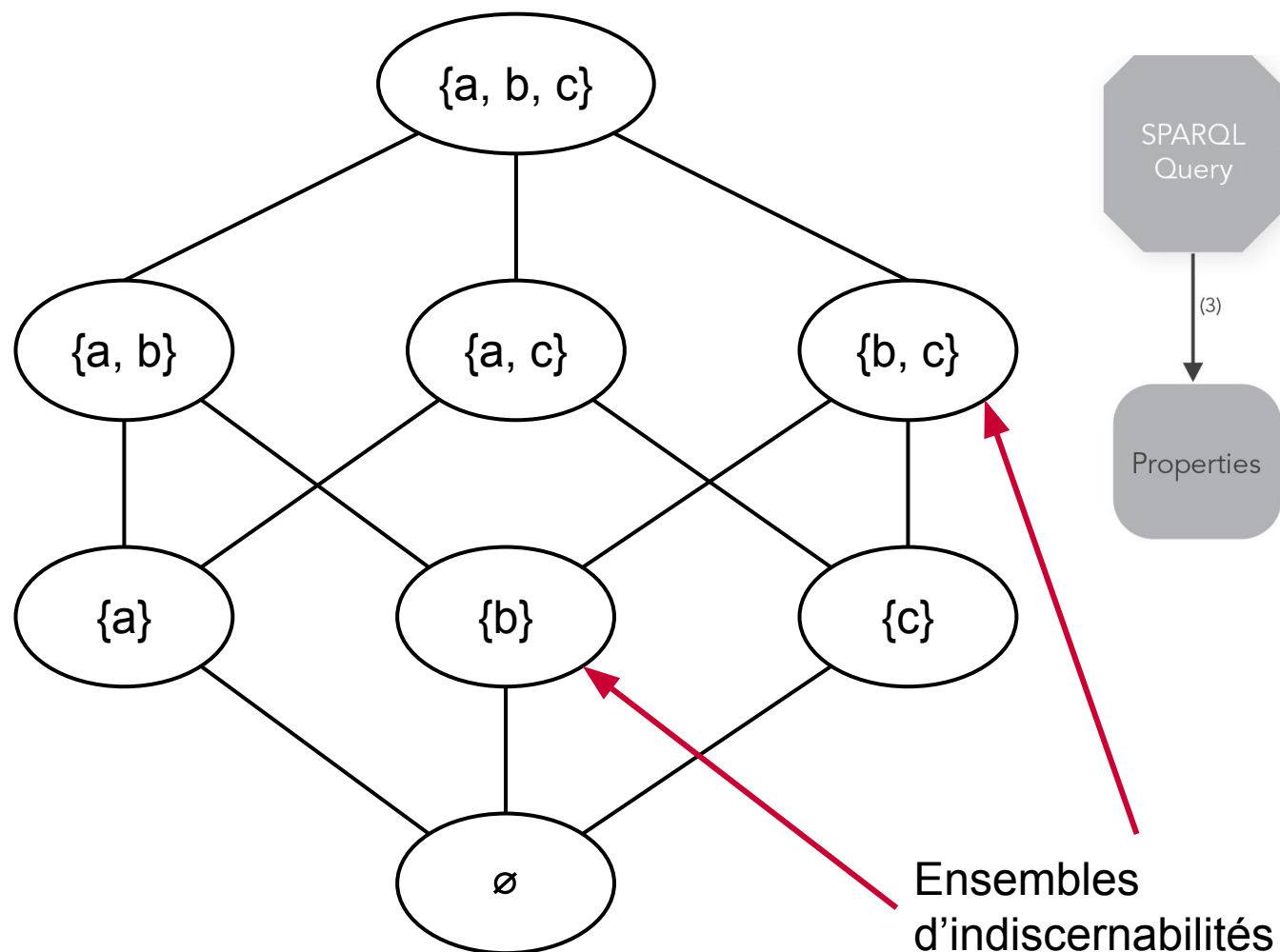
```
SELECT DISTINCT ?politician ?crime
WHERE {
    ?politician :memberOf :TheRepublicans ;
               :convictedOf ?crime .
}
```



# de résultats sans contexte	2
---------------------------------	---

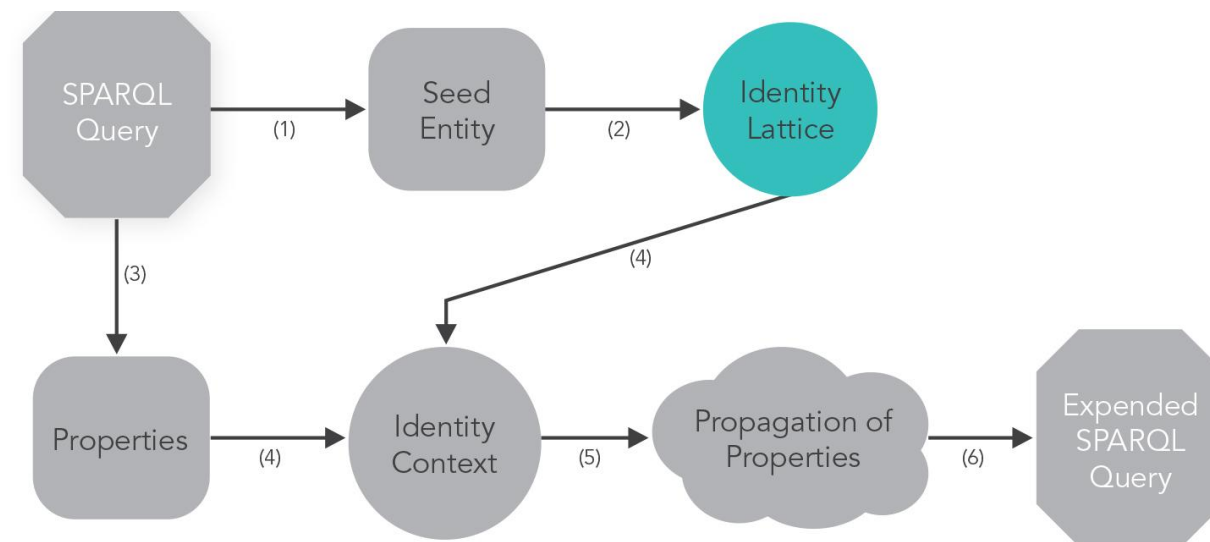
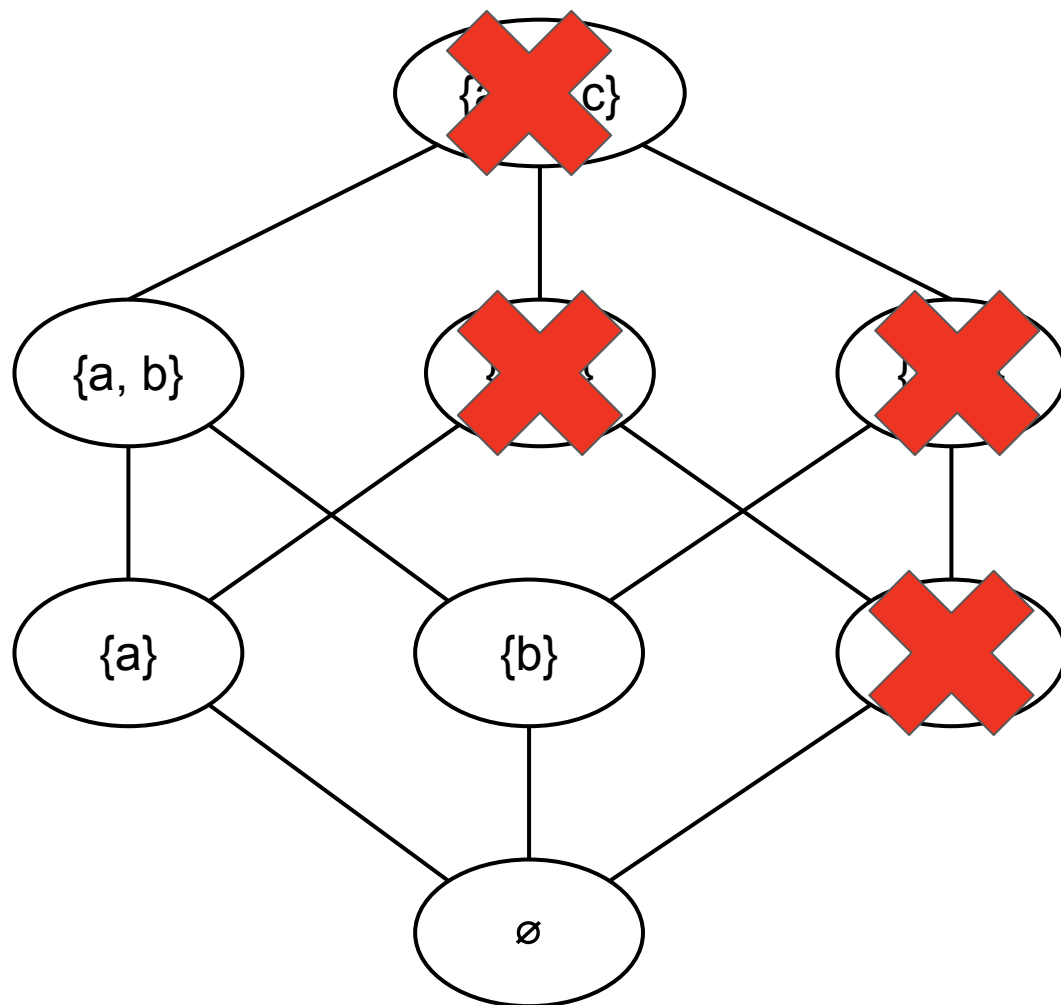
Exemple

Construction du treillis d'identité à partir de "Les Républicains"



Exemple

Elagage du treillis d'identité si pas d'entité contextuellement identique à “Les Républicains”



Exemple

Nombre de républicains condamnés ?

```
SELECT DISTINCT ?politician ?crime
```

```
WHERE {
```

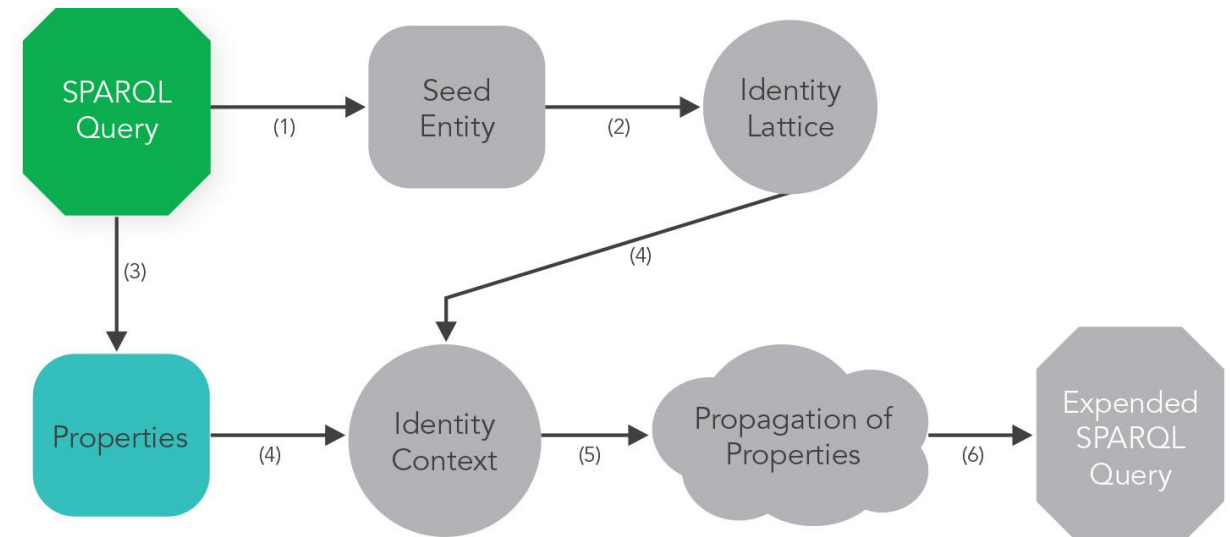
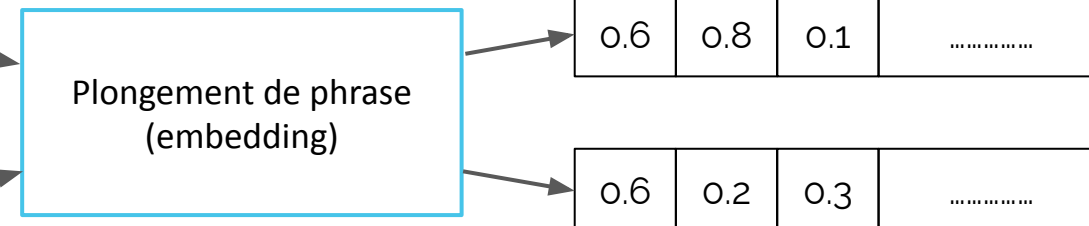
```
  ?politician :memberOf :TheRepublicans ;
```

```
  :convictedOf ?crime .
```

```
}
```

memberOf : “the political party of which a person is or has been a member or otherwise affiliated”

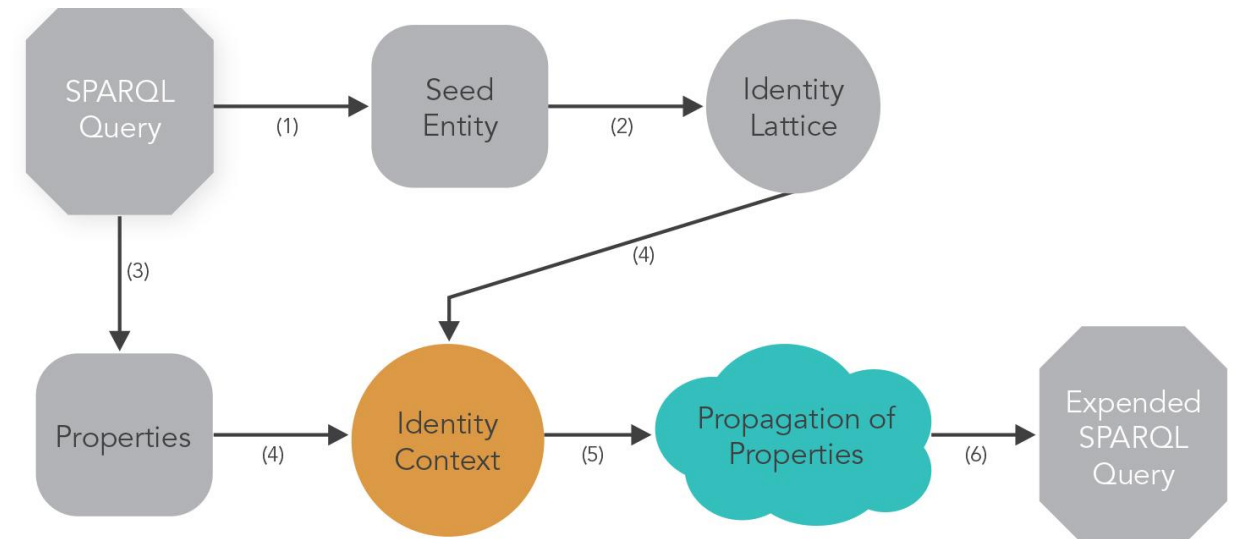
ideology : “political ideology of an organization or person or of a work (such as a newspaper)”



Exemple

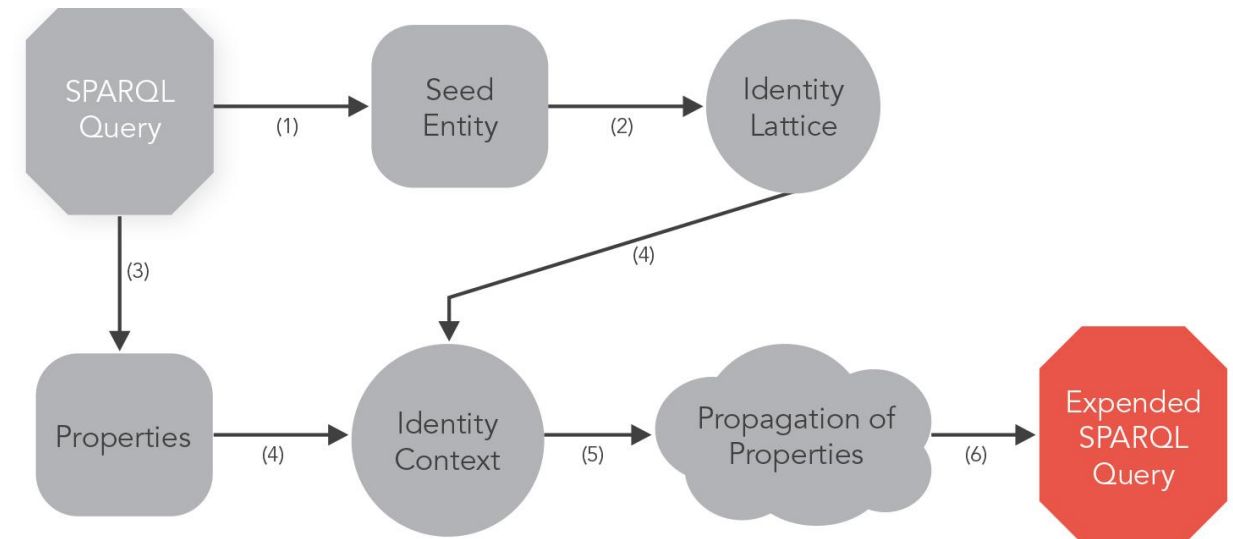
Après **calcul** des ensembles de propagation, on **présente** ceux contenant la propriété à propager

Graine	The Republicans
Ensemble de propagation	member of , political party
Ensemble indiscernable	country, political, ideology
Entités contextuellement identiques	UMP, RPR, UDR, UNR



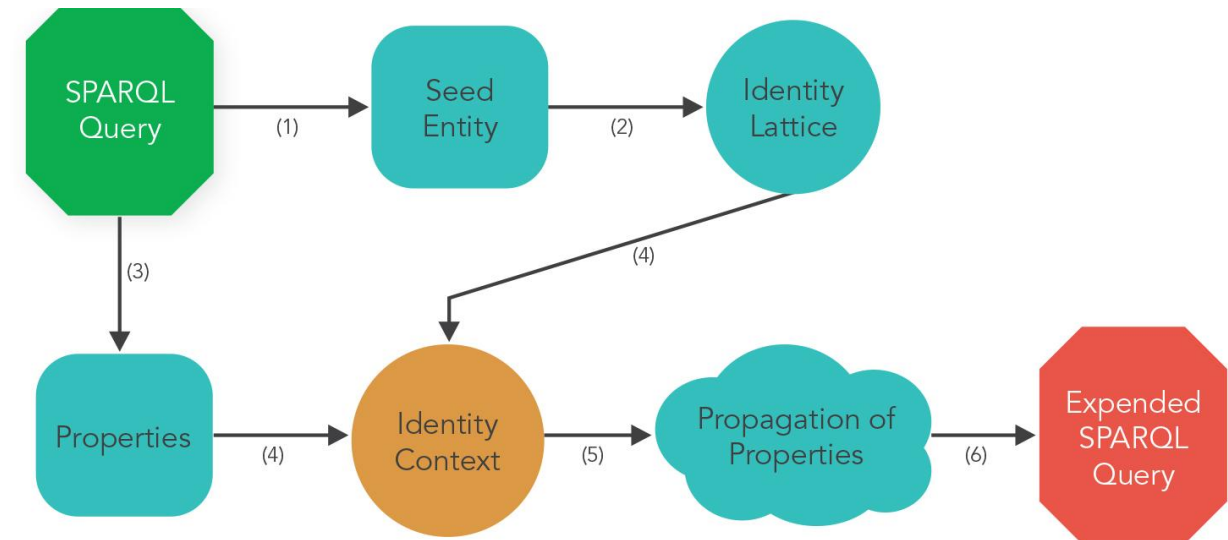
Réécriture de la requête

```
SELECT DISTINCT ?politician ?crime
WHERE {
  VALUES (?party) {
    (:TheRepublicans) (:UMP) (:RPR)
    (:UDR) (:UNR)
  }
  ?politician :memberOf ?party ;
    :convictedOf ?crime .
}
```

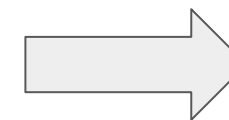


Exemple

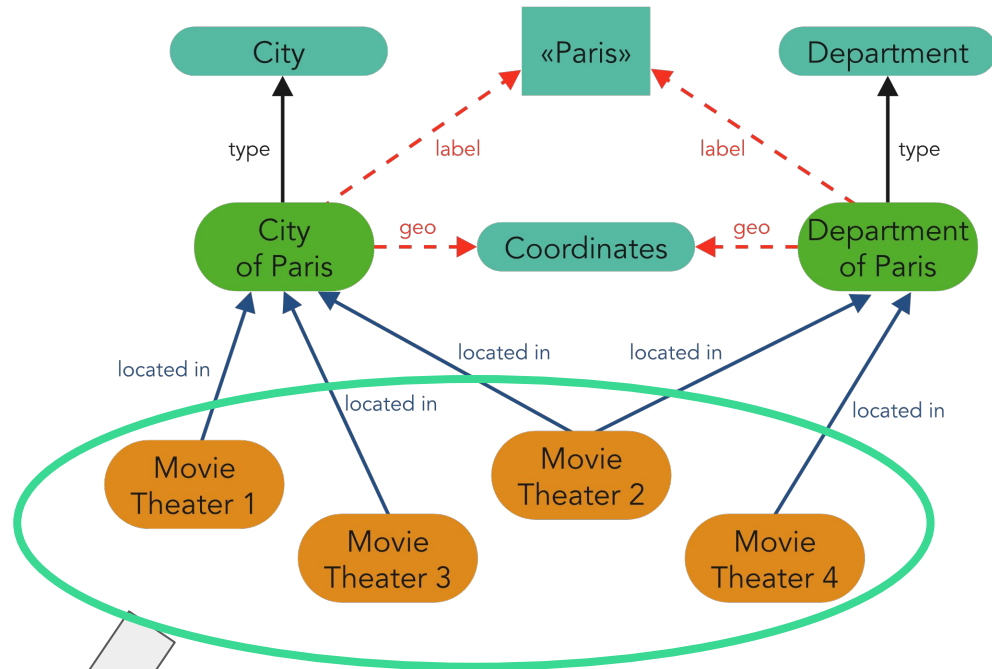
Graine	The Republicans
Ensemble de propagation	member of, political party
Ensemble indiscernable	country, political, ideology
Entités contextuellement identiques	UMP, RPR, UDR, UNR
# de résultats sans contexte	2
# de résultats avec contexte	25



Pour tester la
requête sur
Wikidata



Travaux futurs et conclusion



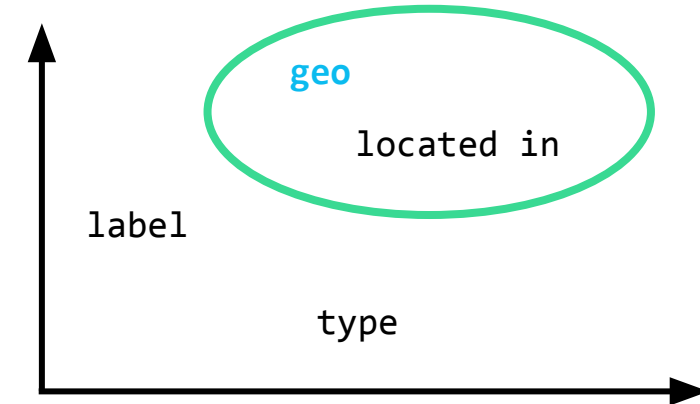
geo : "geocoordinates of the subject"

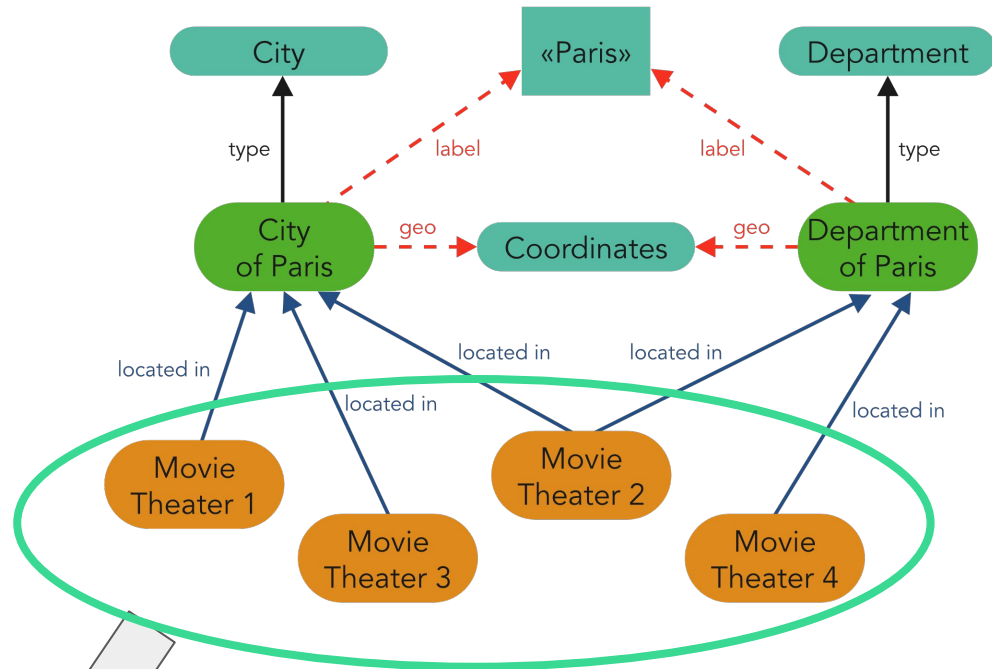
located in : "the item is located on the territory of the following administrative entity"

Plongement de phrase
(embedding)

0.6	0.8	0.1	...
-----	-----	-----	-----

0.6	0.2	0.3	...
-----	-----	-----	-----





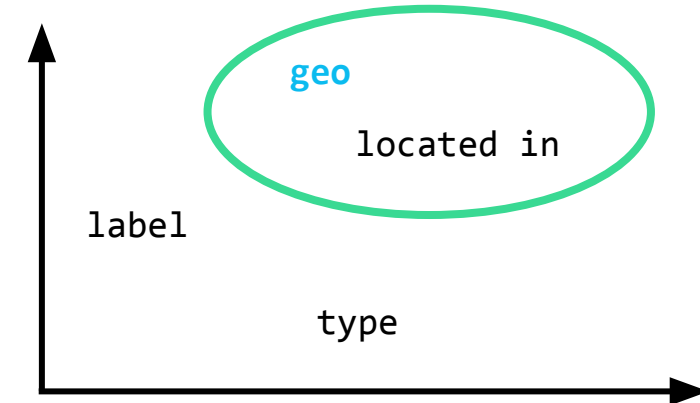
geo : "geocoordinates of the subject"

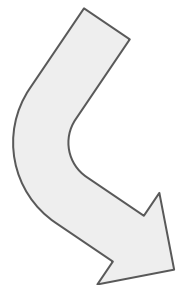
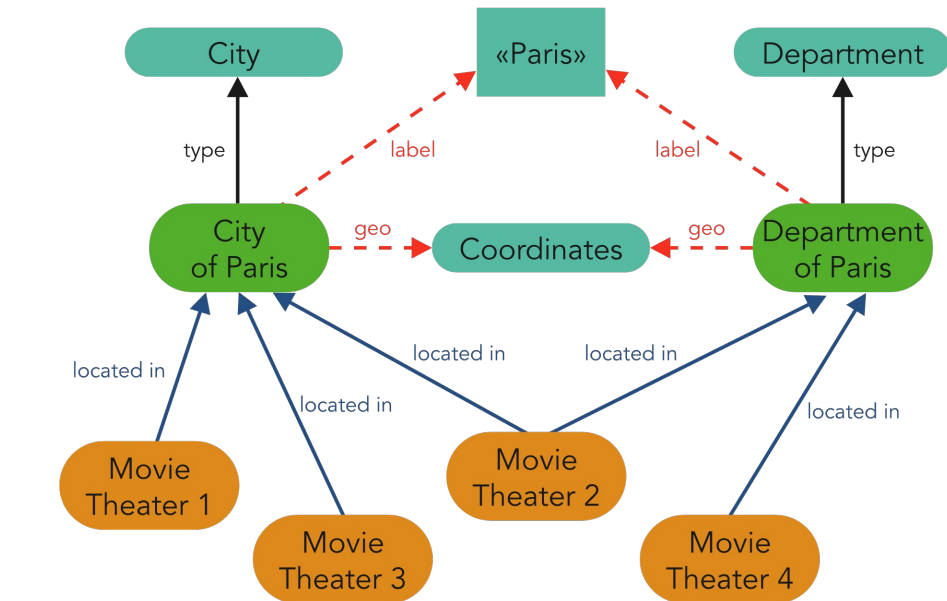
located in : "the item is located on the territory of the following administrative entity"

Plongement de phrase
(embedding)

0.6	0.8	0.1	...
-----	-----	-----	-----

0.6	0.2	0.3	...
-----	-----	-----	-----





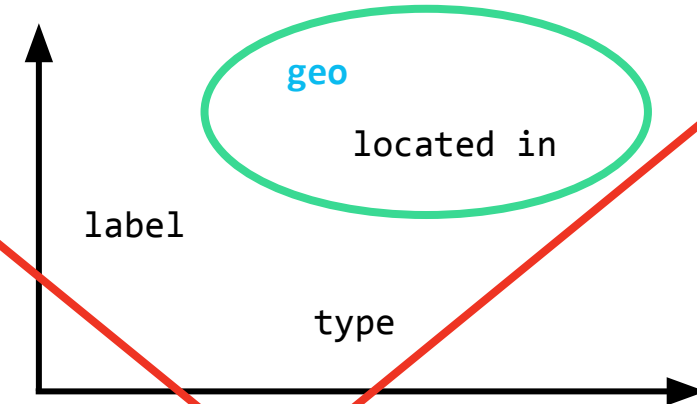
geo : ~~"geocoordinates of the subject"~~

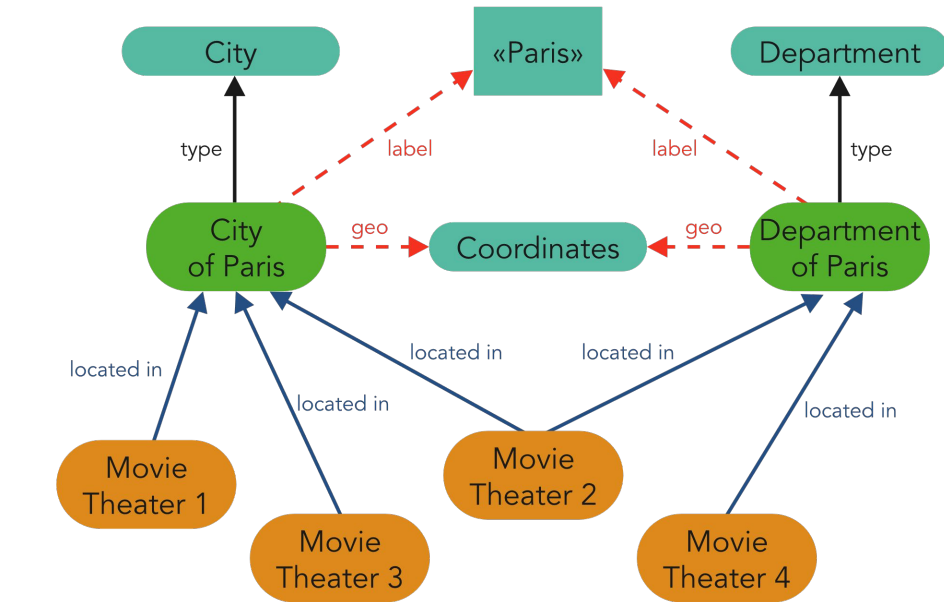
located in : ~~"the item is located on the territory of the following administrative entity"~~

Plongement de phrase
(embedding)

0.6	0.8	0.1	...
-----	-----	-----	-----

0.6	0.2	0.3	...
-----	-----	-----	-----

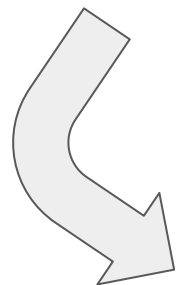
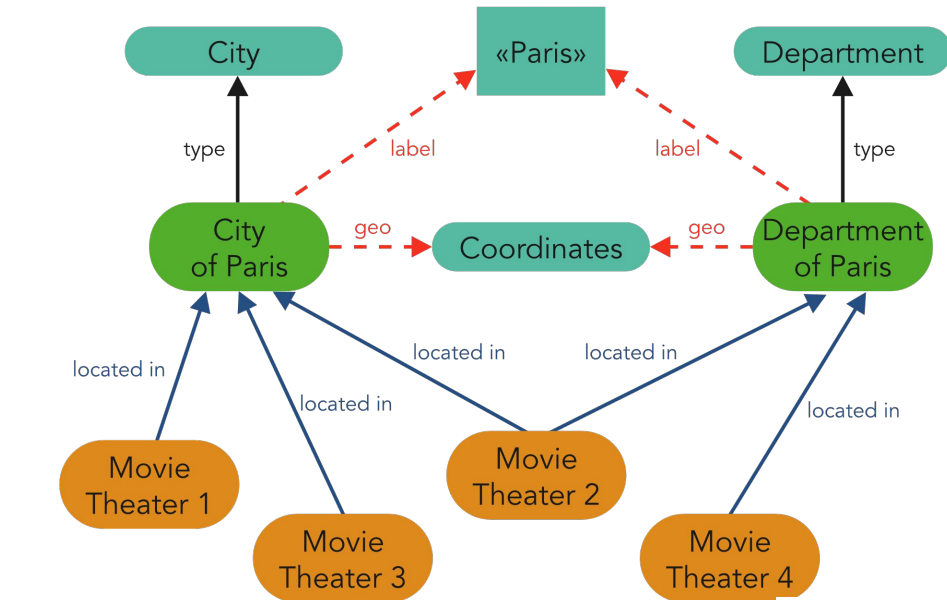




Comment se passer de plongement de phrases ?

geo : ~~"geocoordinates of the subject"~~

located in : ~~"the item is located on the territory of the following administrative entity"~~



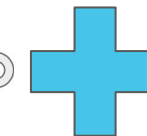
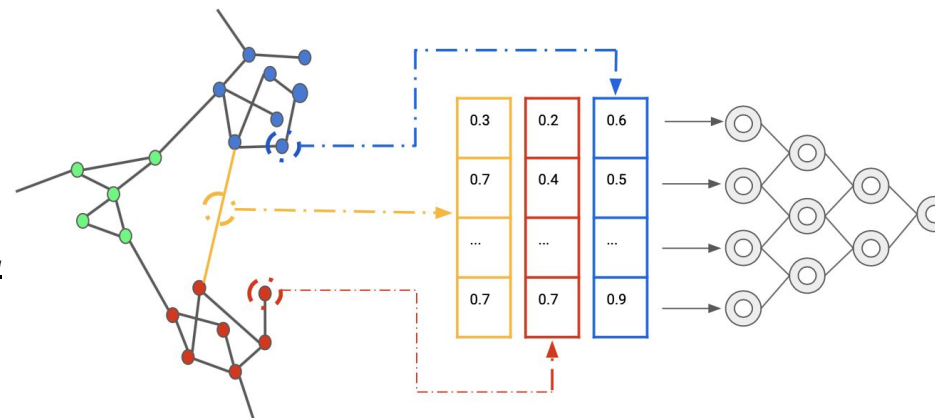
geo : ~~"geocoordinates of the subject"~~

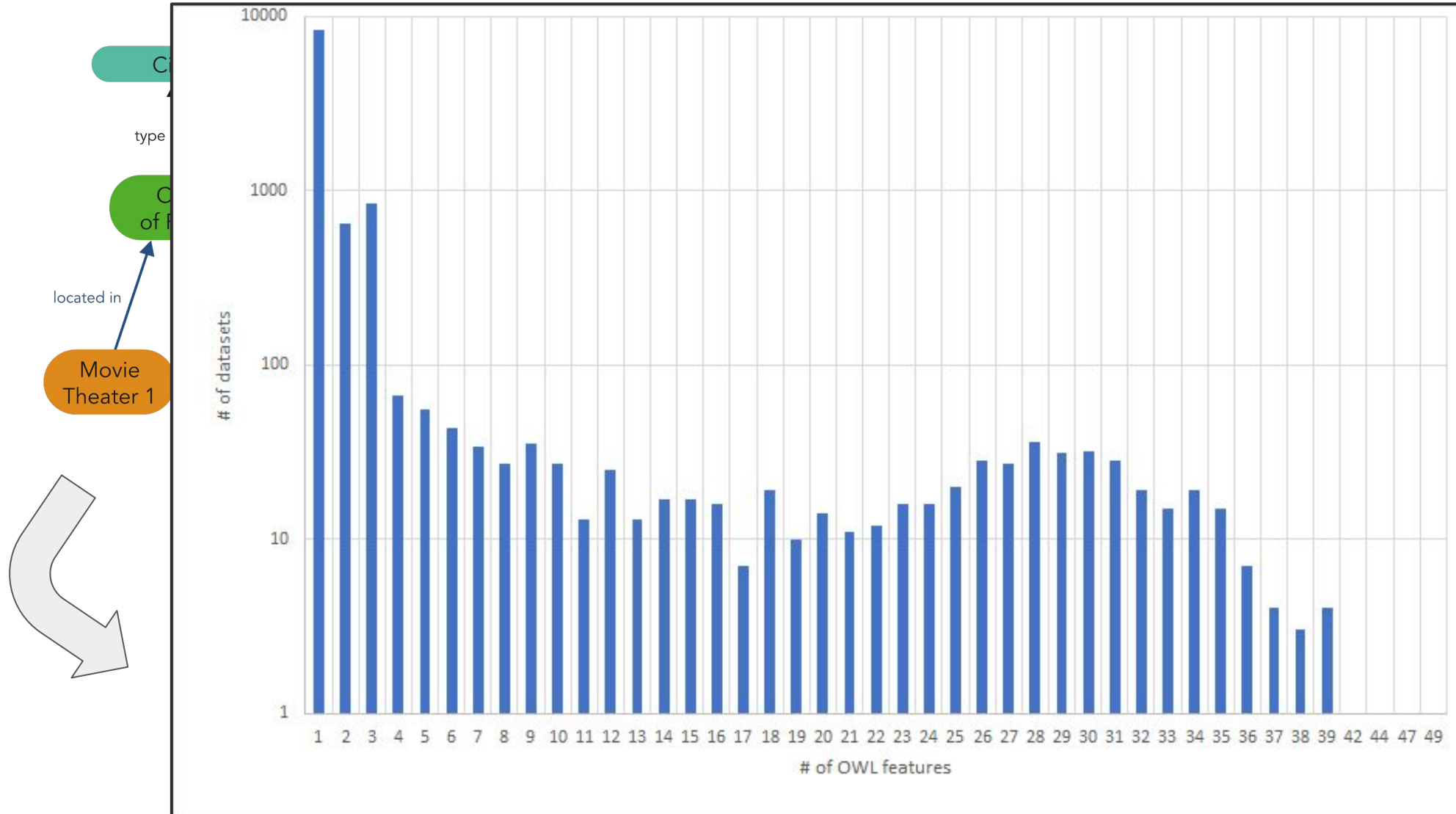
located in : ~~"the item is located the territory of the following administrative entity"~~

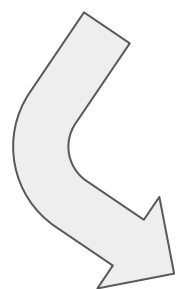
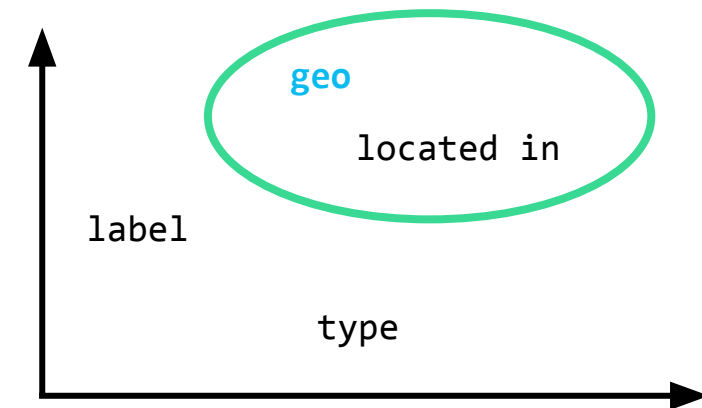
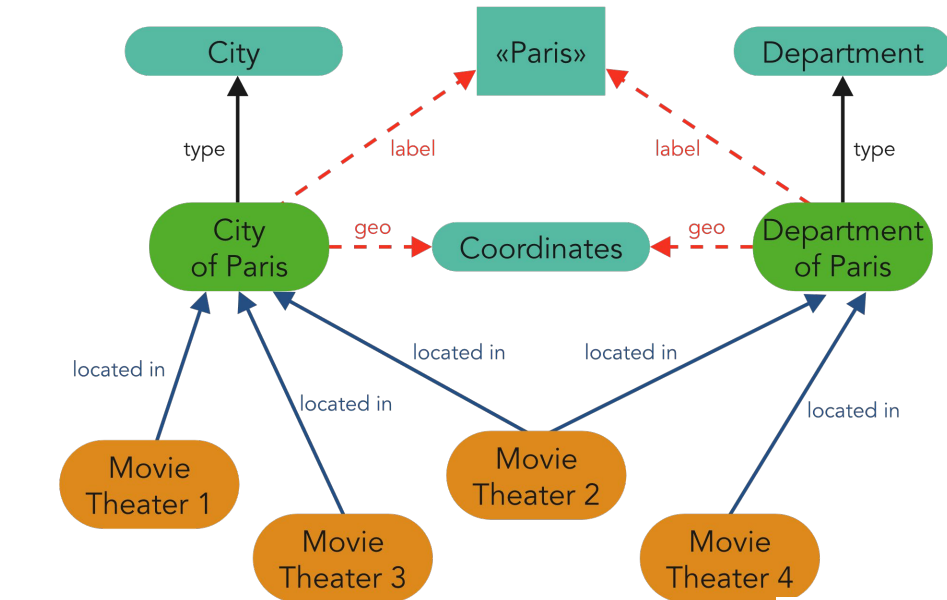
Knowledge Graph

Embedded Representation

Machine Learning Task







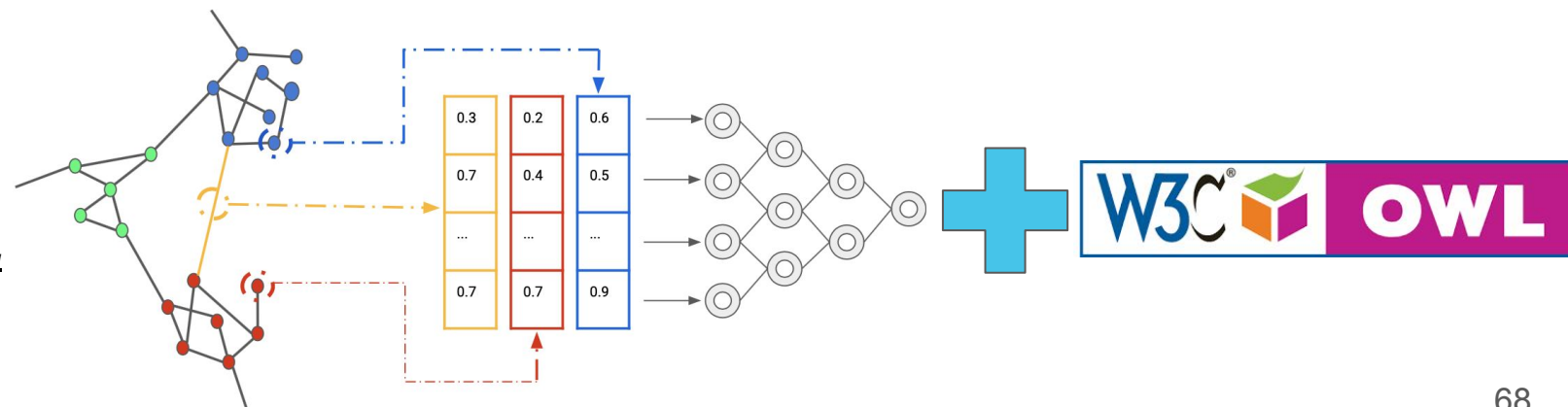
geo : ~~"geocoordinates of the subject"~~

located in : ~~"the item is located the territory of the following administrative entity"~~

Knowledge Graph

Embedded Representation

Machine Learning Task



Projet ANR pour aider les utilisateurs non techniques - en particulier les journalistes - à explorer et à analyser de grands lacs de données hétérogènes :

- CSV, JSON,
- des bases de données relationnelles,
- des PDF et des documents textuels.

Utilisation de graphes de connaissances.



→ Utilisation de l'identité contextuelle et de la propagation



Approche utilisant les plongements de phrases pour trouver des propriétés propageables pour un ensemble d'indiscernabilité donné.

Conclusion



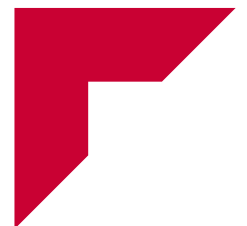
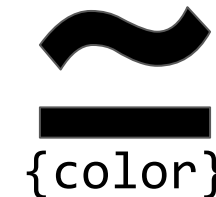
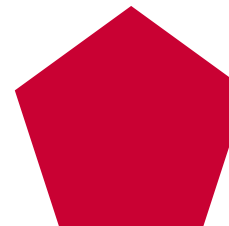
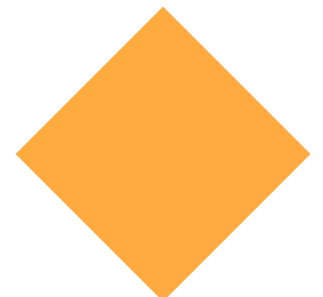
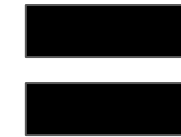
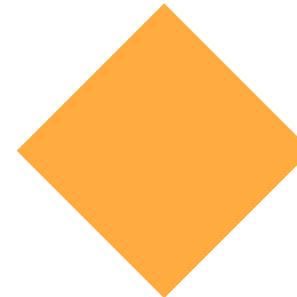
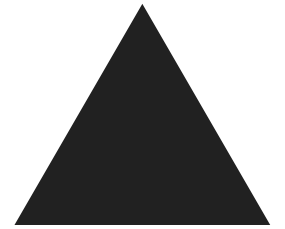
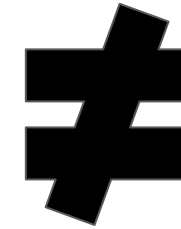
Approche utilisant les plongements de phrases pour trouver des propriétés propageables pour un ensemble d'indiscernabilité donné.



Permet d'utiliser concrètement une version **moins rigide** et donc moins sujette à l'erreur de l'identité.



Le plongement de phrase permet de **découvrir** les propriétés **propageables**, mais les **limites** restent encore floues.



*“While entity disambiguation and resolution is an active research area in the semantic Web, and now in knowledge graphs for several years, it is almost surprising that it continues to be **one of the top challenges** in the industry almost across the board.”*

*“How can **identity be described** in a way that different teams can **agree on it** and know what the other teams are describing?”*

- Halpin et al. *When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data*. 2010
- Beek et al. *A Contextualised Semantics for owl: sameAs*. 2016
- Raad et al. *Detecting Erroneous Identity Links on the Web using Network Metrics*. 2018
- Raad et al. *The sameAs Problem: A Survey on Identity Management in the Web of Data*. 2019
- Noy et al. *Industry-scale knowledge graphs: Lessons and challenges*. 2019
- Paris et al. *A study about the use of OWL 2 semantics in RDF-based knowledge graphs*. 2020
- Paris et al. *Contextual Propagation of Properties for Knowledge Graphs: A Sentence Embedding Based Approach*. 2020
- Coumes et al. *Qiana: A First-Order Formalism to Quantify over Contexts and Formulas*. 2024