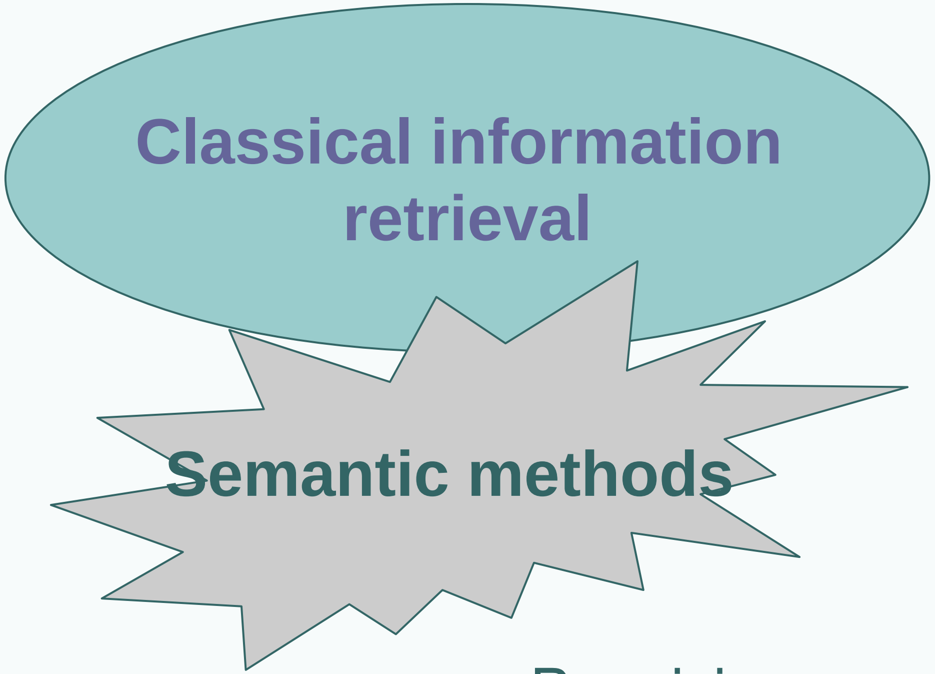
# Semantic information retrieval

## A Description logics based approach

Naouel Karam

karam@isima.fr

# Context



Classical information retrieval

- Statistical methods
- Considerable number of results
- Non appropriate results
- Lot of work to the user

Semantic methods

- Promising approach
- NLP

# The proposed approach

- **Input**
  - Query Q as a natural language description
  - A set of documents {D1,…,Dn}
- **Problem**
  - Sort {D1,…,Dn} according to their semantic distance w.r.t. Q
  - Need to detect the related information between the query and the documents

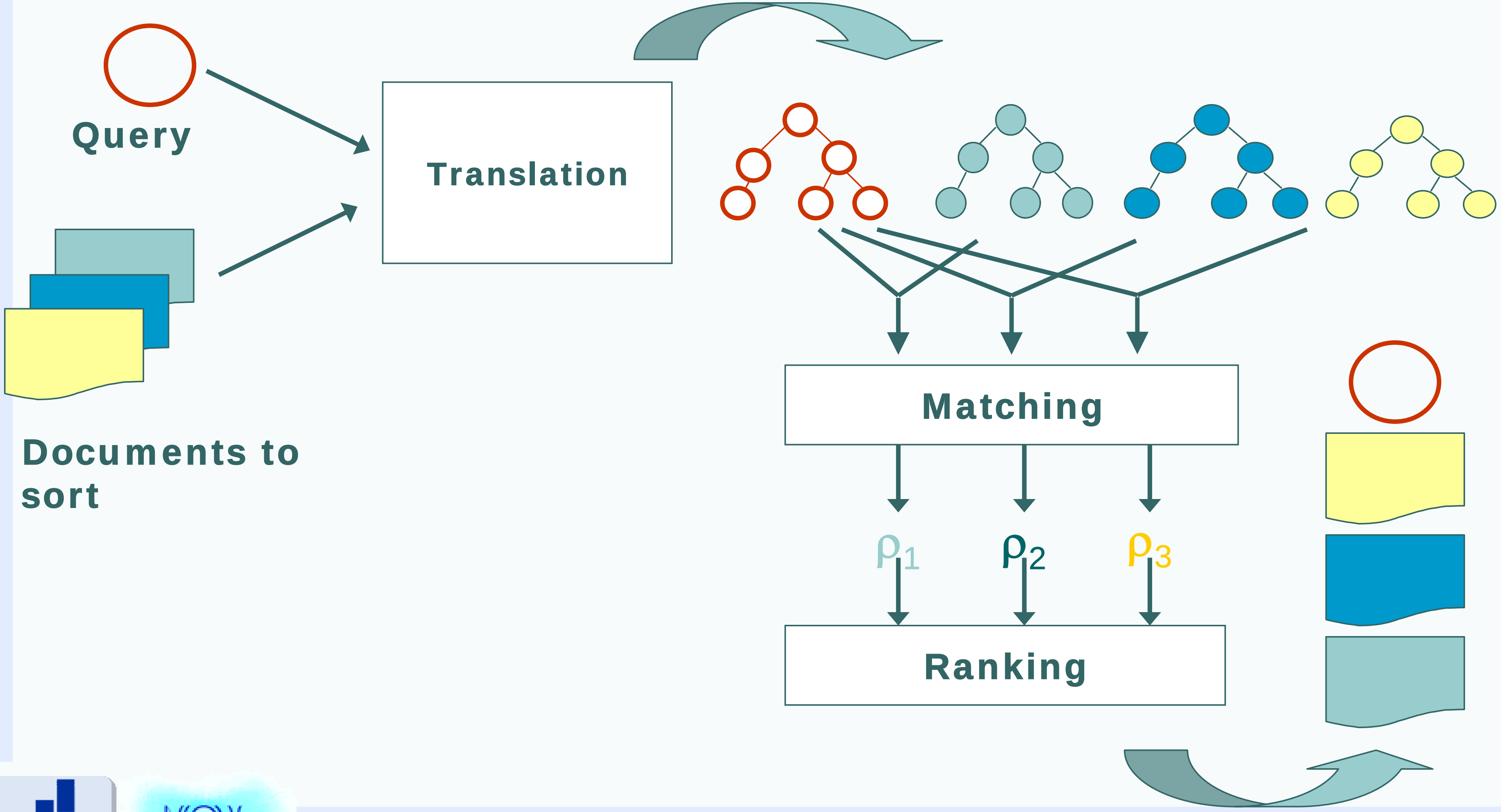
Matching Problem

# The proposed approach - An example

Researchers are embedded in a laboratory. They examine guinea pigs and discover factors that give rise of protein receptors. They study only mice.

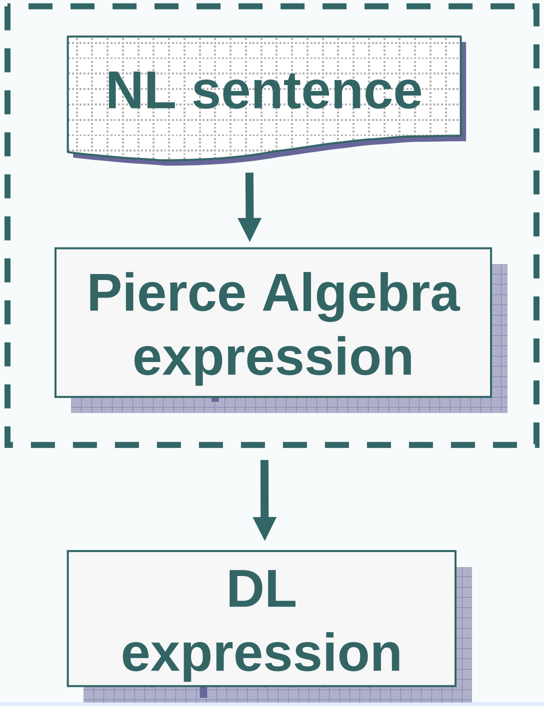
Scientists are attached to a research laboratory. They discover genes which produce specialized protein receptors. These genes are found in cells.

# The global process



# The translation step

- Formal representation
  - Description Logics (DL)
  - 2 reasons:
    - Well defined semantics / correct algorithms
    - Link with Natural language already established
- Based on existent work [Schmidt 92, 96]
  - Correspondence between syntactic constructions and a semantic representation
  - Connection: Pierce algebras



NL sentence

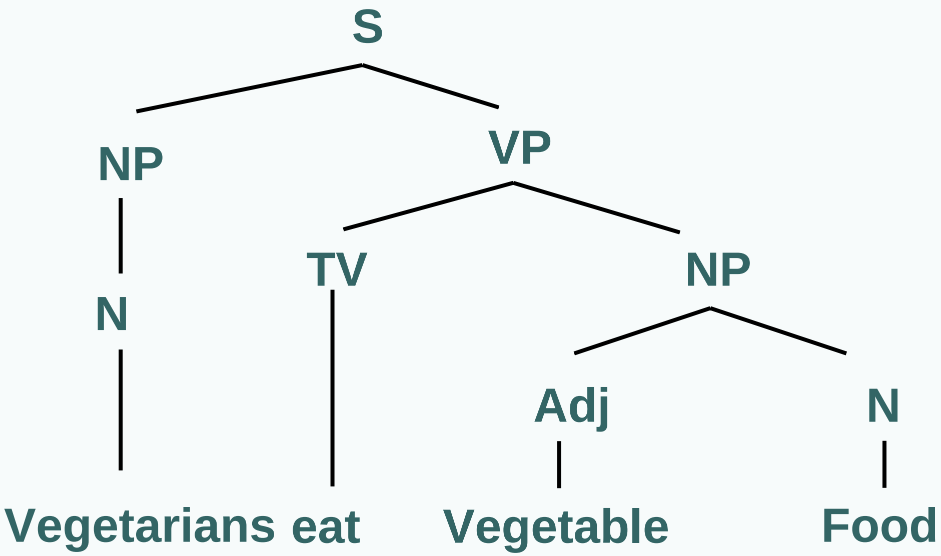
Pierce Algebra expression

DL expression

# The translation step – NL and Pierce algebras

- Syntax

| | Production rules |
|---|---|
| (i) | S → NP + VP |
| (ii) | VP → TV + NP |
| (iii) | NP → Adj + N |
| (iv) | NP → N |



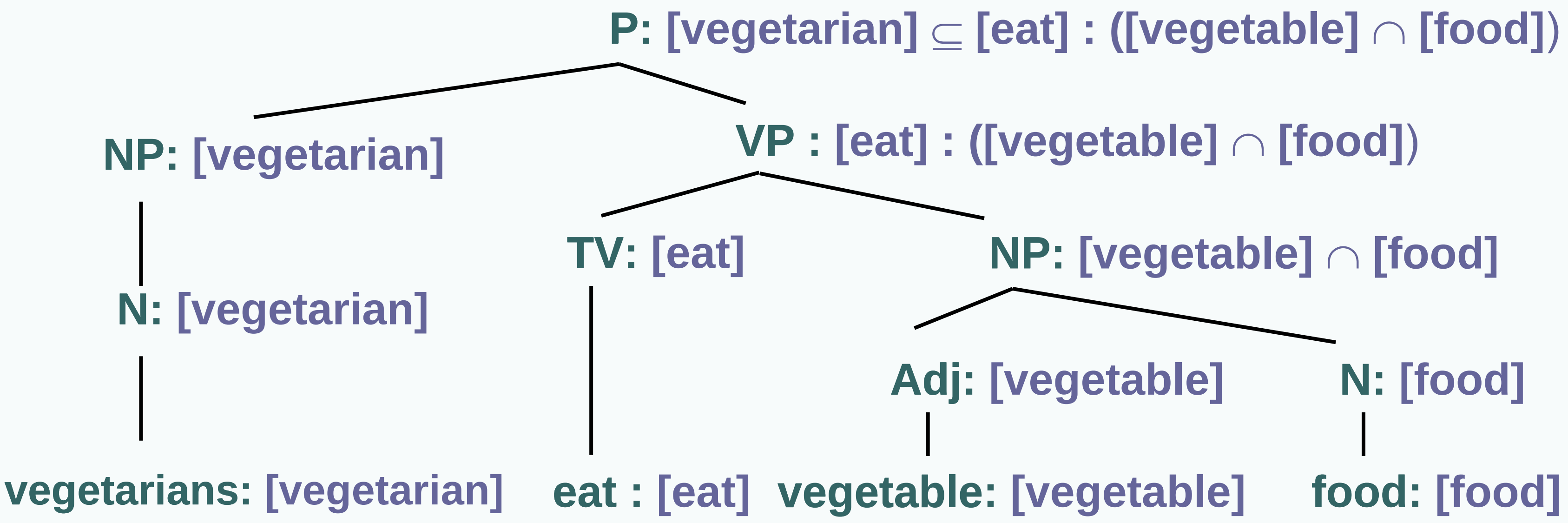
# The translation step – NL and Pierce algebras

- Semantics

  [.] : {N, Adj, NP} → sets

  {TV} → binary relation

| | Production rule | Semantic association |
|---|---|---|
| (i) | S → NP + VP | [NP] ⊆ [VP] |
| (ii) | VP → TV + NP | [TV]:[NP] |
| (iii) | NP → Adj + N | [Adj] ∩ [N] |
| (iv) | NP → N | [N] |



P: [vegetarian] ⊆ [eat] : ([vegetable] ∩ [food])

NP: [vegetarian]

VP : [eat] : ([vegetable] ∩ [food])

N: [vegetarian]

TV: [eat]

NP: [vegetable] ∩ [food]

Adj: [vegetable]

N: [food]

vegetarians: [vegetarian]

eat : [eat]

vegetable: [vegetable]

food: [food]

# The translation step – Pierce algebra and DL

| Pierce algebra | Description logics |
|---|---|
| set | concept |
| Binary relation | role |
| Subset relation ( $\subseteq$ ) | subsumption ( \| ) |
| intersection ( $\cap$ ) | conjunction ( ʃ ) |
| Pierce product ( : ) | Existential quantification ( $\exists$ ) |

$$[\text{Vegetarian}] \subseteq [\text{eat}] : ([\text{Vegetarian}] \cap [\text{Food}])$$

$$\text{Vegetarian} \sqsubseteq \exists \text{eat}.(\text{Vegetarian} \sqcap \text{Food})$$

# The translation step

- Restricted framework: sentences with complements, quantifiers (all, some, only), number restrictions, negation, passive form.

- Quantifiers
  - « Some persons eat fruit » Person ◆ ∃ eat.Fruit $\not{\text{B}}$ ⊥
  - « All persons eat fruit » Person ❖ ∃ eat.Fruit
  - « No persons eat fruit » Person ◆ ∃ eat.Fruit B ⊥
- Number restrictions
  - « John loves more than 3 girls » John ❖ ≥3 love. Girl
  - « John loves at most 2 girls » John ❖ ≤2 love. Girl
  - « John loves exactly 1 girl » John ❖ ≤1 love.Girl ◆ ≥1loves.Girl

## The translation step

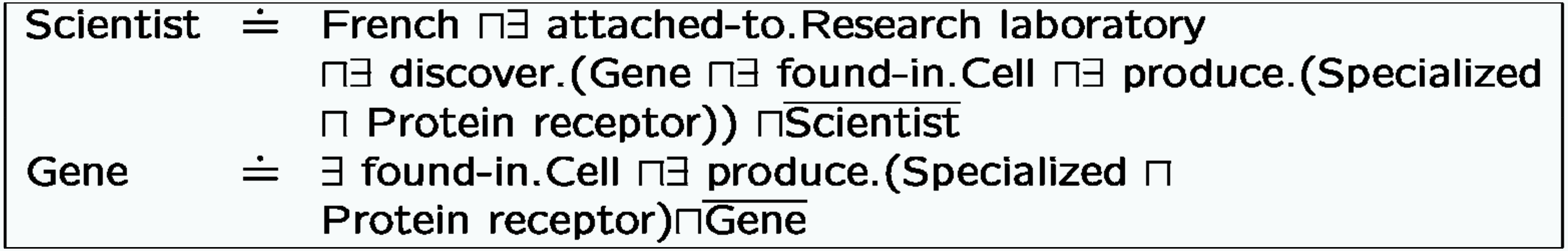
- Relational nouns
  - « A Father has sons »   Father ❖ ∃ son. T
- Negation
  - « is not comfortable »   ¬ comfortable
- Passive form
  - « is teached by »   ∃ teach $^{-1}$

# The translation step – An example

Q

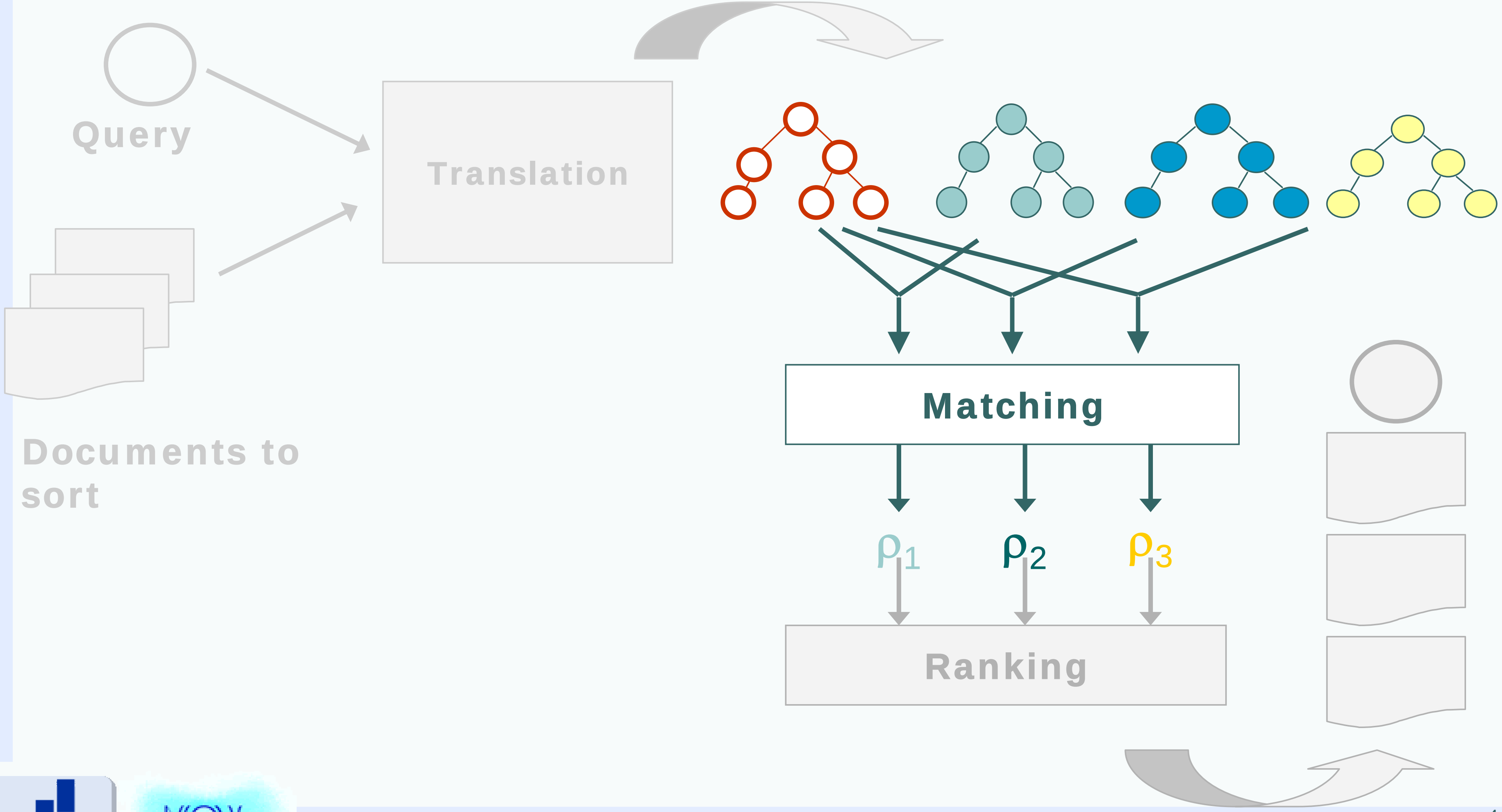
**French scientists are attached to a research laboratory. They discover genes. These genes are found in cells and produce specialized protein receptors.**

$T_Q$



$$\begin{aligned}
\text{Scientist} \quad &\doteq \quad \text{French} \sqcap \exists\, \text{attached-to.Research laboratory} \\
&\sqcap \exists\, \text{discover.(Gene} \sqcap \exists\, \text{found-in.Cell} \sqcap \exists\, \text{produce.(Specialized} \\
&\sqcap\ \text{Protein receptor))} \sqcap \overline{\text{Scientist}} \\
\text{Gene} \quad &\doteq \quad \exists\, \text{found-in.Cell} \sqcap \exists\, \text{produce.(Specialized} \sqcap \\
&\text{Protein receptor)} \sqcap \overline{\text{Gene}}
\end{aligned}$$

# The global process



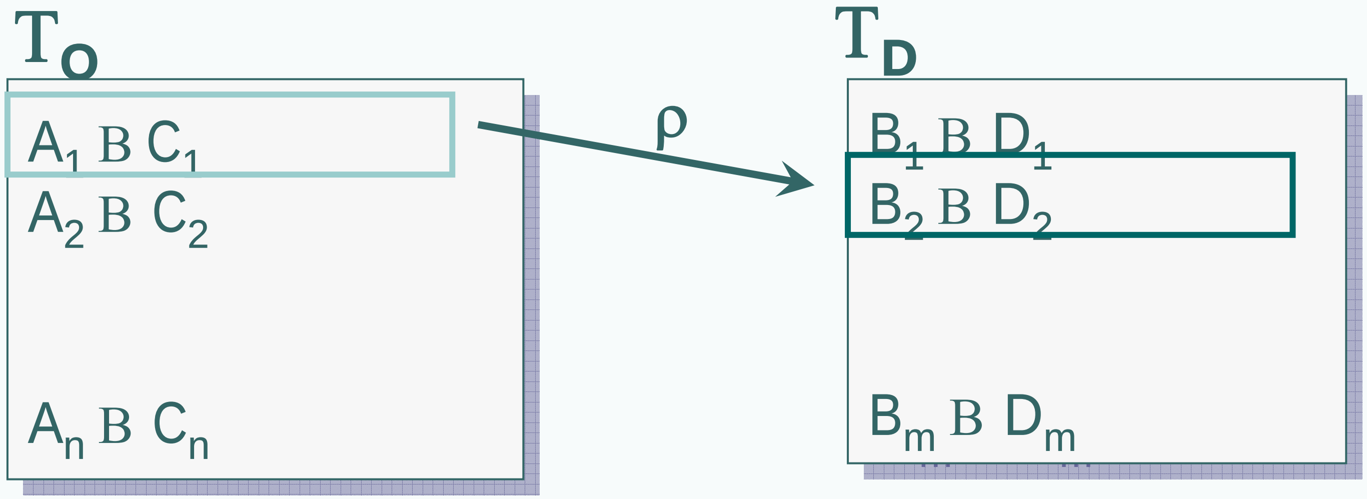
# The matching step

- Similar to schema matching problems (Databases, XML,…)
- Existing approaches: schema = tree structure
- Framework : description logics
  - Schema = Terminology

## Matching definition

- Operation that takes two schema as input and returns a correspondence between elements from the two schemas
- Correspondence is a pair of related elements
- Matching terminologies
  - Elements to relate: defined concepts in the terminologies

# The matching step



$T_O$

$A_1$ B $C_1$

$A_2$ B $C_2$

$A_n$ B $C_n$

$\rho$

$T_D$

$B_1$ B $D_1$

$B_2$ B $D_2$

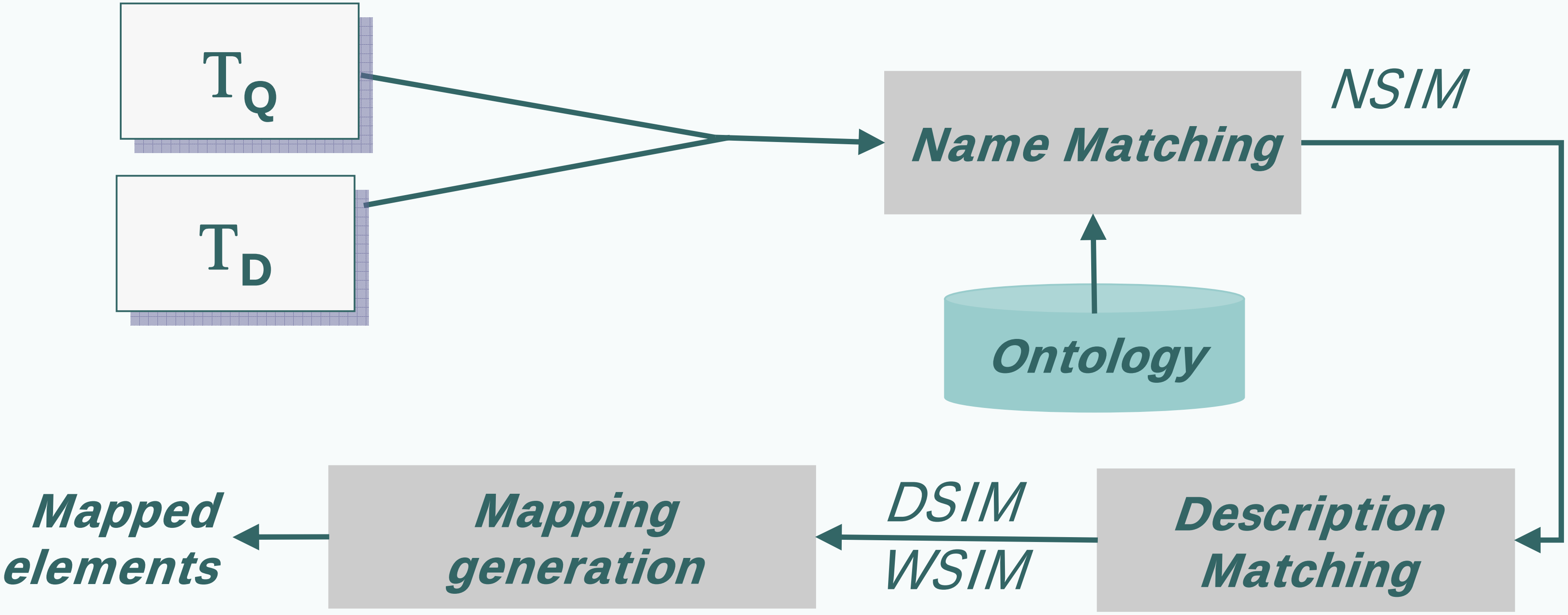
$B_m$ B $D_m$

Names $A_1$ and $B_2$ are similar

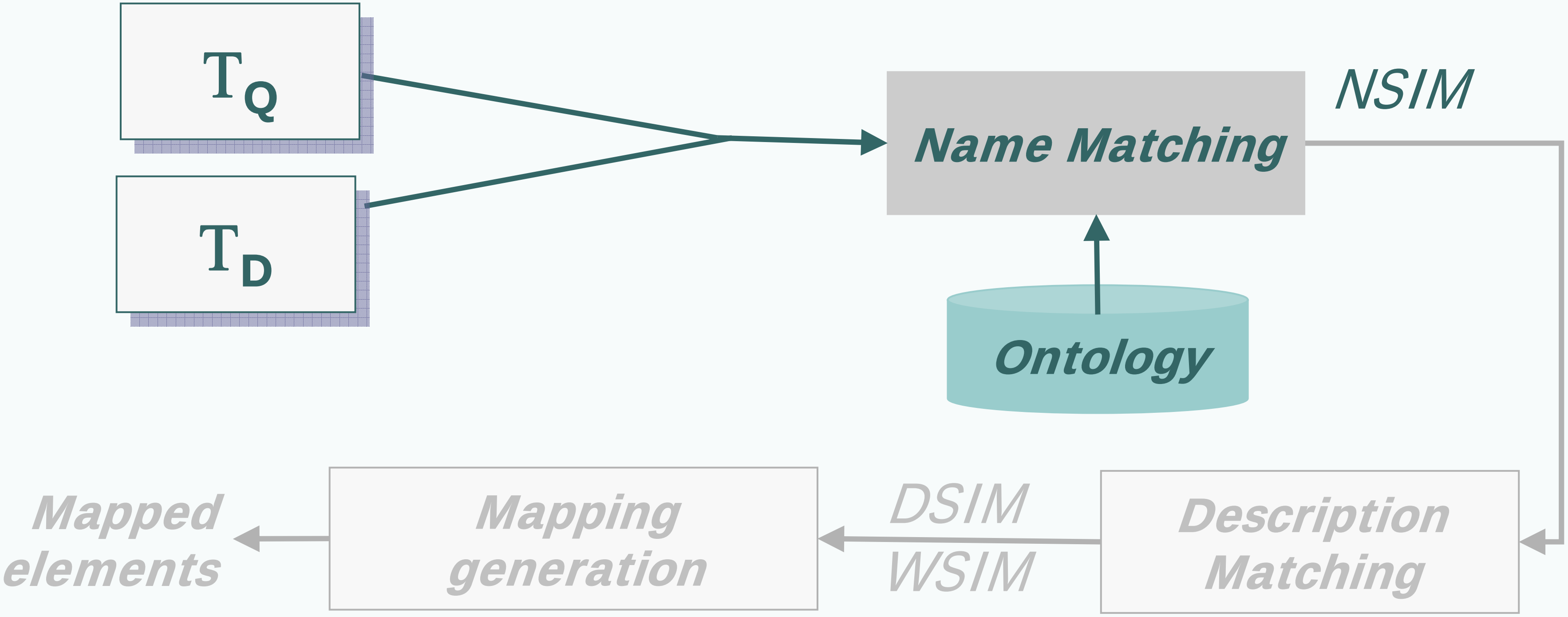
Descriptions $C_1$ et $D_2$ are similar

- 2 steps:
  - **Name matching**
  - **Description matching**

# The matching step



# The matching step

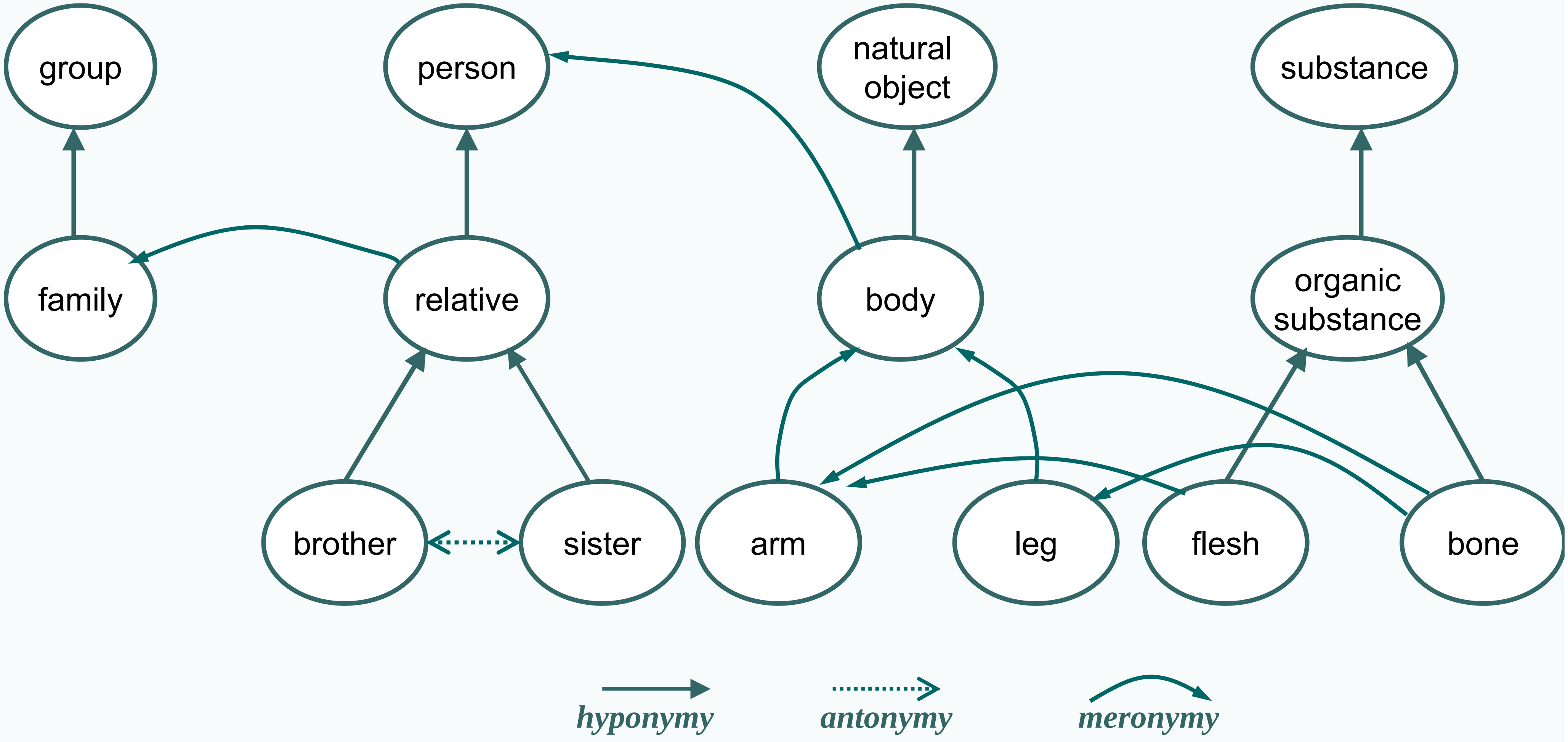


# The name matching

- Computes name similarity coefficients ***NSIM*** between concept names

- Based on the notion of "semantic relatedness" (**rel**)
  - Degree of semantic similarity between two lexically expressed concepts
  - Based on the semantic relations of **WorNet**

# The name matching - WordNet



# The name matching

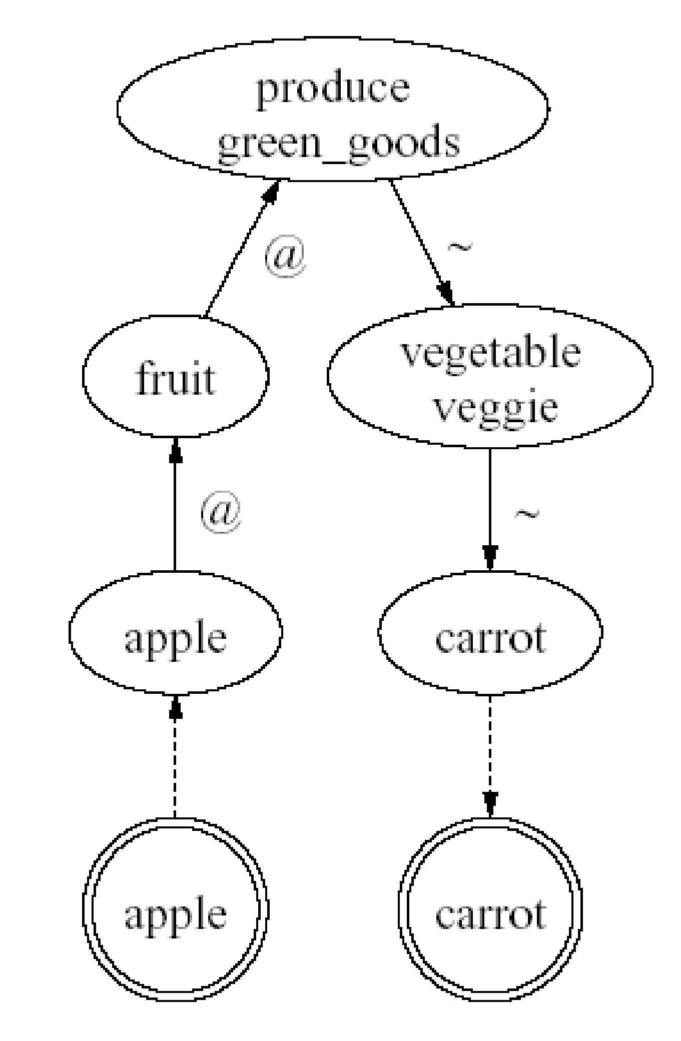
- Semantic relatedness
- 2 concepts are semantically close if:
  - Path not long
  - Path does not change direction too often

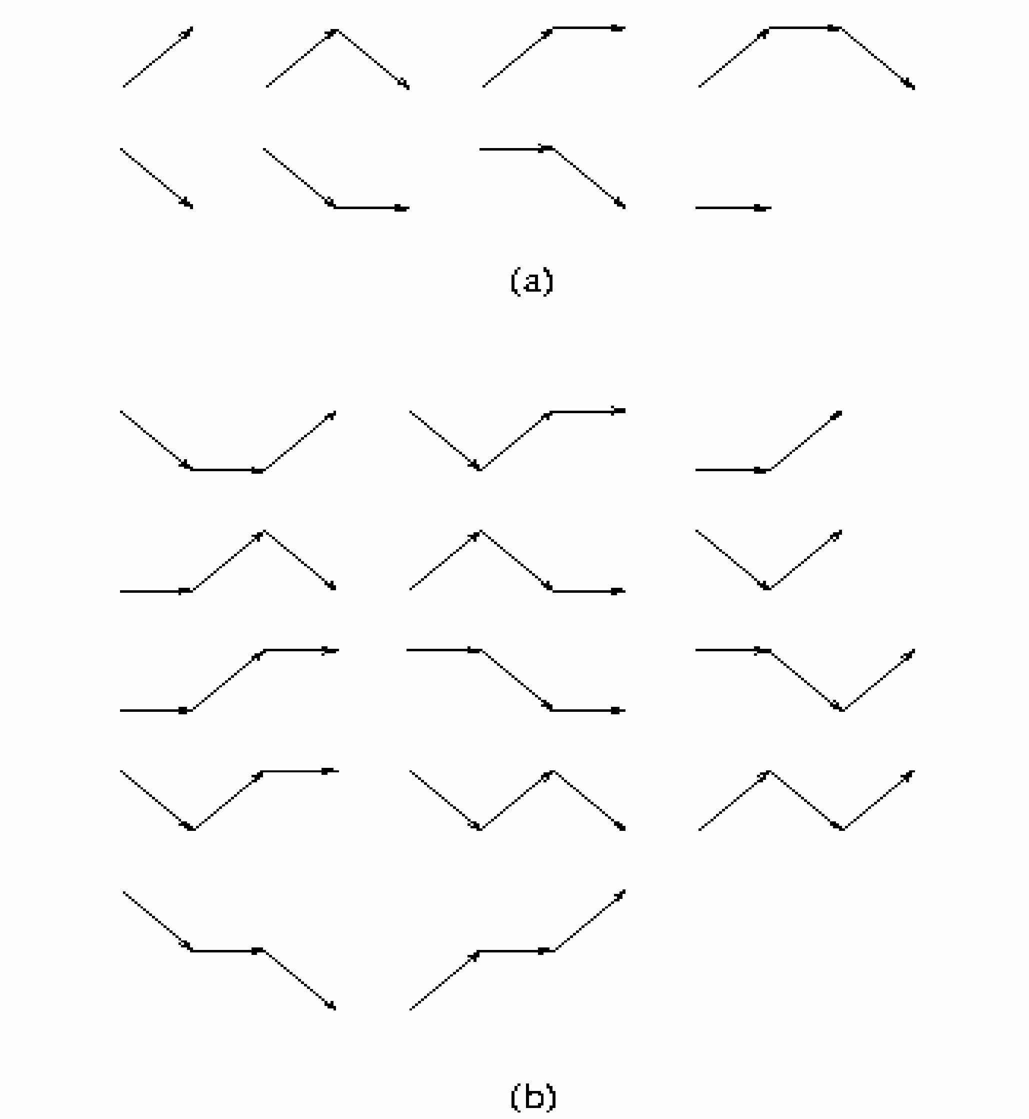
- **NSIM** ™ [0,1]

$$\text{rel}\,(C_1,C_2) = \lambda - \mathit{len}\,(C_1,C_2) - k * \mathit{turns}\,(C_1,C_2)$$

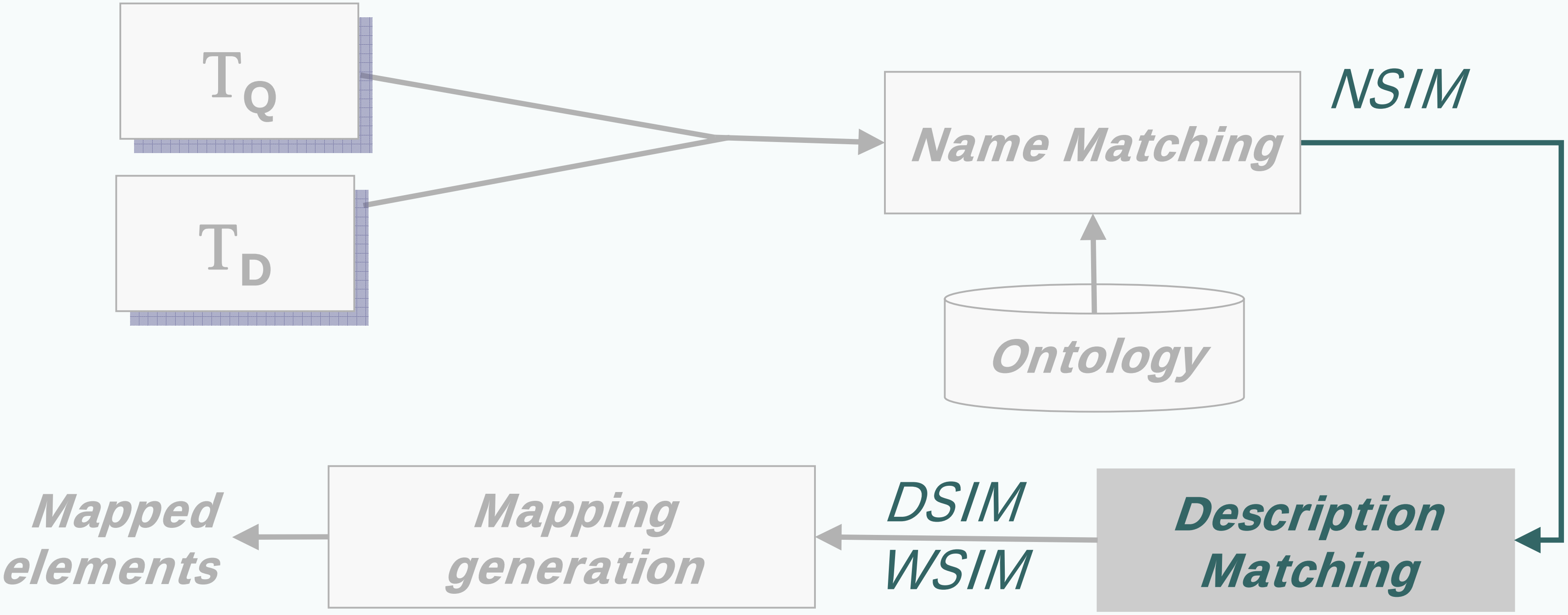
$$\text{NSIM}\,(C_1,C_2) = \text{rel}\,(C_1,C_2) \,/\, \lambda$$

# The name matching



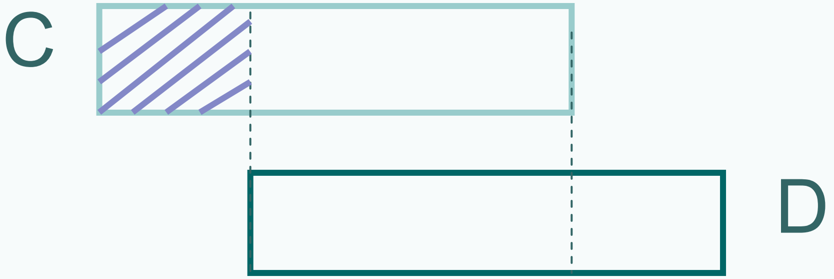


# The matching step



# The description matching

- Intuition



$$DSIM(C,D) = 1 - \frac{|\,C - D\,|}{|\,C\,|}$$

Difference operator

# The difference operator

- Allows to remove from a given description the information contained in another description
- Take into account linguistic relations (semantic relatedness) between concept and role names when computing the difference → “**Similarity difference**”

The difference algorithm based on the notion of **subsumption**

- **Goal :** define a subsumption taking into account linguistic relations between concept and role names

→ based on hierarchies

→ based on similarities “**Similarity subsumption**”

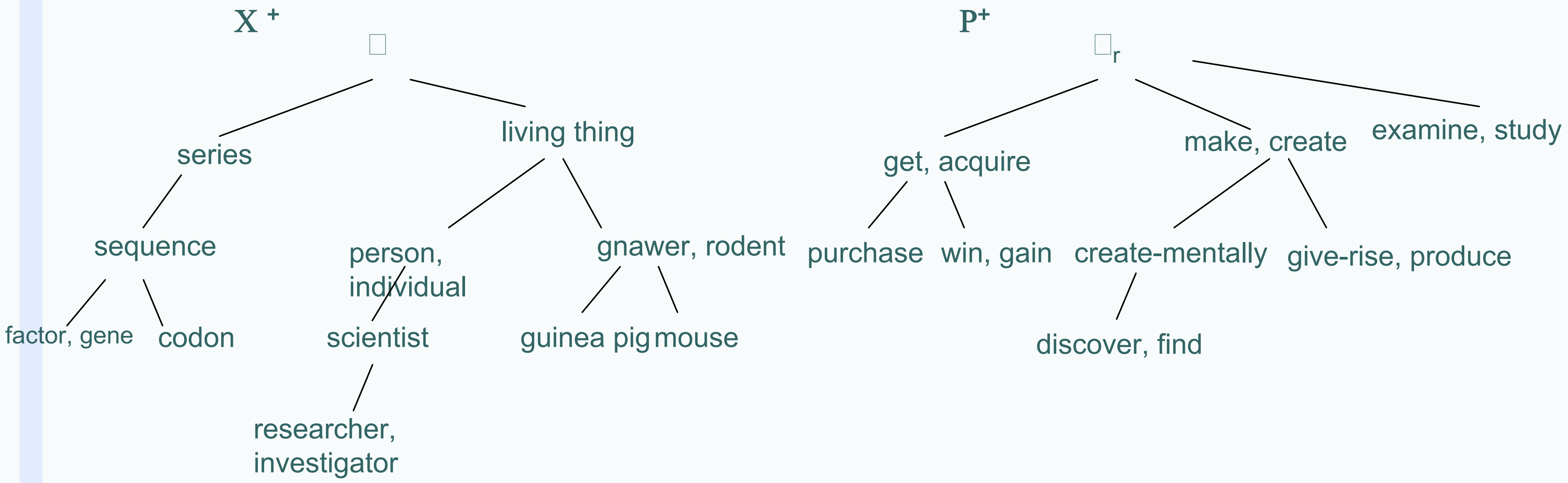
# Difference based on concept and role hierarchies

# Hierarchies

- A support $\mathcal{S} = (\mathcal{C}^+, \mathcal{R}^+)$

  $\mathcal{C}^+ = (\mathcal{N}_A, \leq_C)$ where $\leq_C$ is a partial order relation defined on $\mathcal{N}_A$.

  $\mathcal{R}^+ = (\mathcal{N}_R, \leq_R)$ where $\leq_R$ is a partial order relation defined on $\mathcal{N}_R$.



Logic under consideration: $\mathcal{ALEH}_\mathcal{S}$

# A structural subsumption algorithm for ALEH$_S$

- Based on graphs
- 3 steps
  - Concept descriptions are turned into a normal form
  - Normal forms represented by tree descriptions
  - Subsumption est caracterized in term of tree homomorphism

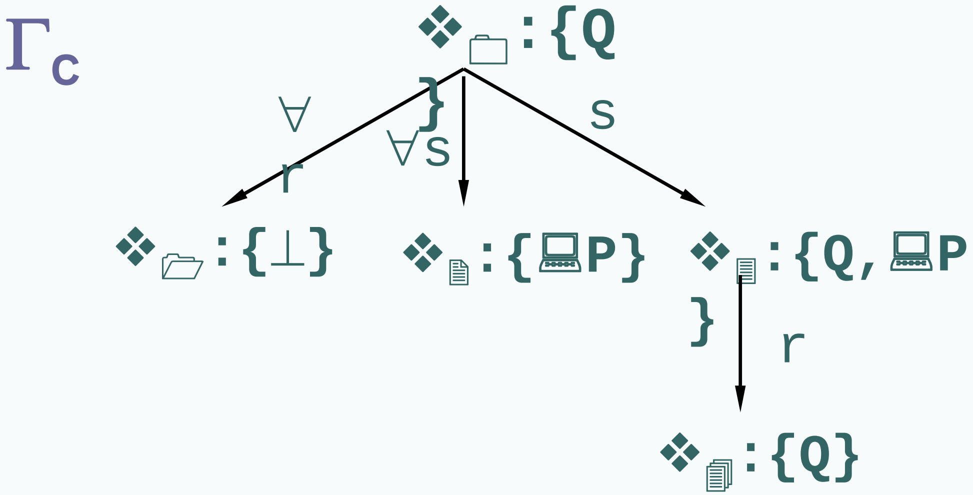
# Normalisation rules for ALEH$_S$

$$\begin{array}{rcl}
\forall r.C \sqcap \forall r.D & \rightarrow & \forall r.(C \sqcap D) \\
\forall s.C \sqcap \forall r.D & \rightarrow & \forall s.C \sqcap \forall r.(C \sqcap D), \text{si } r <_R s \\
\forall s.C \sqcap \exists r.D & \rightarrow & \forall s.C \sqcap \exists r.(C \sqcap D), \text{si } r \leq_R s \\
\forall r.\top & \rightarrow & \top \\
C \sqcap \top & \rightarrow & C \\
P \sqcap \neg Q & \rightarrow & \bot, \text{pour tout } P, Q \in N_C \text{ tel que } P \leq_C Q \\
\exists r.\bot & \rightarrow & \bot \\
C \sqcap \bot & \rightarrow & \bot
\end{array}$$

# Decription trees

$$C \doteq Q \sqcap \forall r.P \sqcap \forall s.\neg P \sqcap \exists s.(Q \sqcap \exists r.Q) \qquad r \leq_R s$$

$$C' \doteq Q \sqcap \forall r.\bot \sqcap \forall s.\neg P \sqcap \exists s.(Q \sqcap \neg P \sqcap \exists r.Q)$$

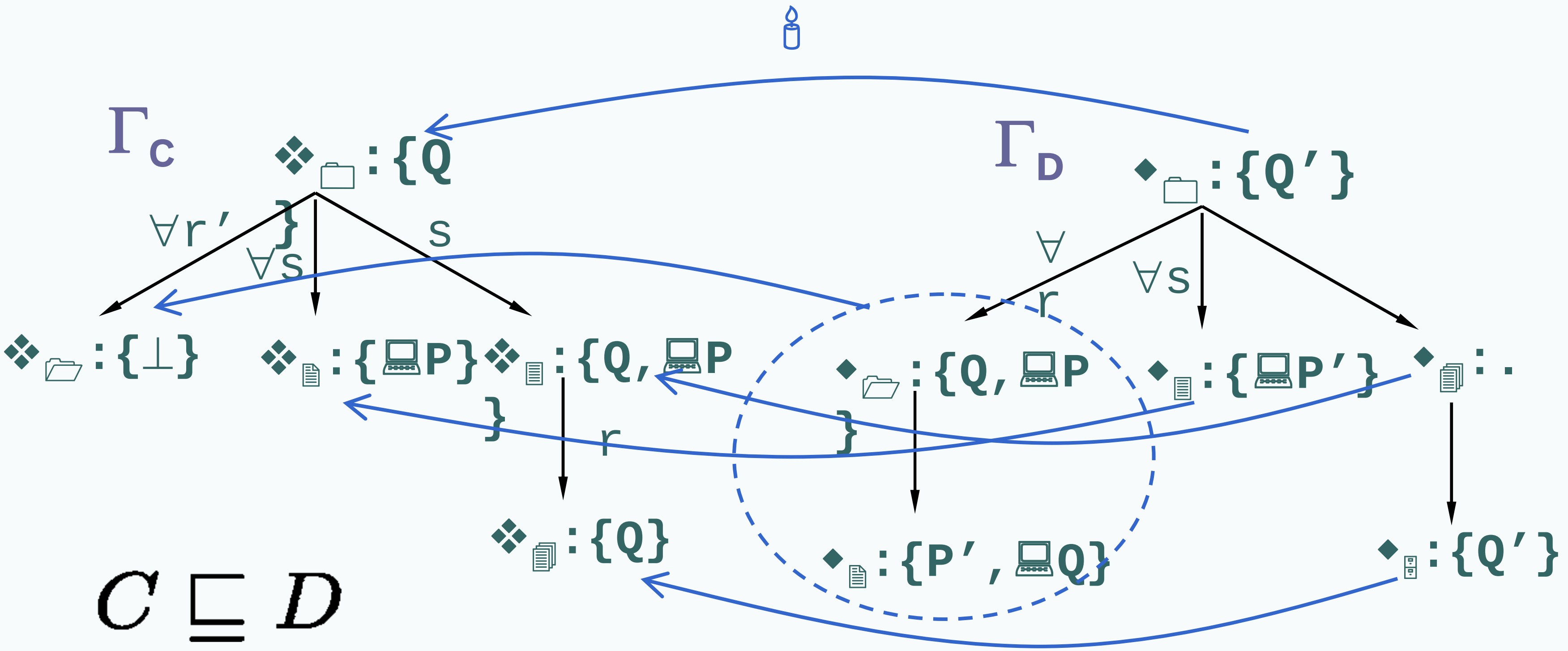


# Tree based caracterisation of subsumption in ALEH$_S$

$$C \doteq Q \sqcap \forall r'.\bot \sqcap \forall s.\neg P \sqcap \exists s.(Q \sqcap \neg P \sqcap \exists r.Q)$$

$$D \doteq Q' \sqcap \forall r.(Q \sqcap \exists s.(P' \sqcap \neg Q)) \sqcap \forall s.\neg P' \sqcap \exists s.(\exists s.Q')$$

$$P \leq_C P' \qquad r \leq_R r' \leq_R s$$

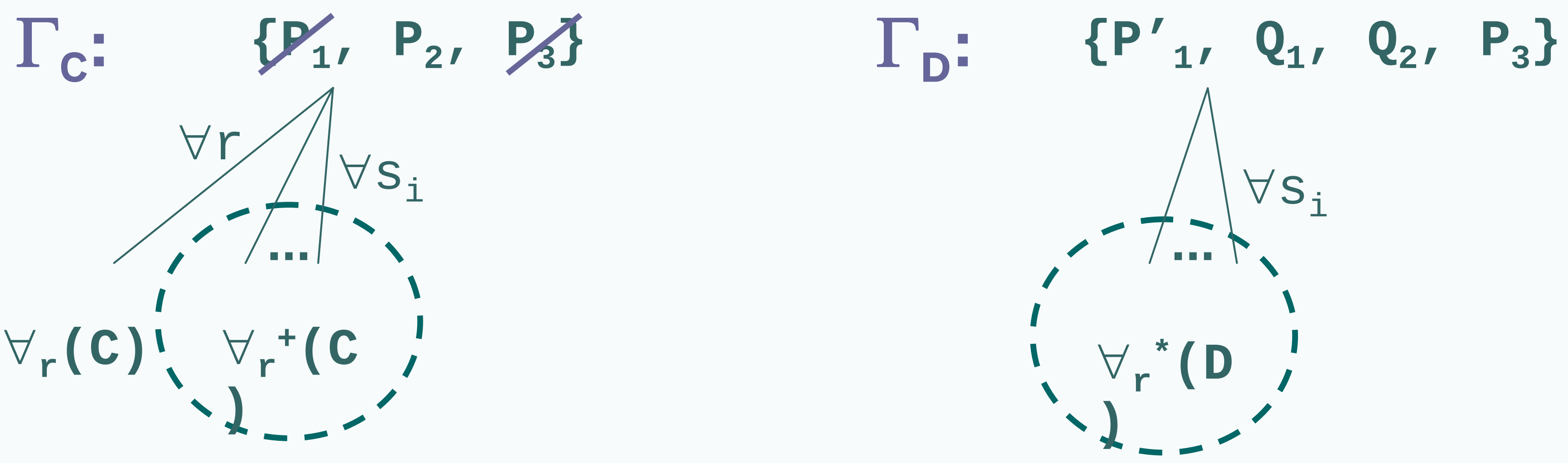


$$C \sqsubseteq D$$

# ALEH$_S$ Difference

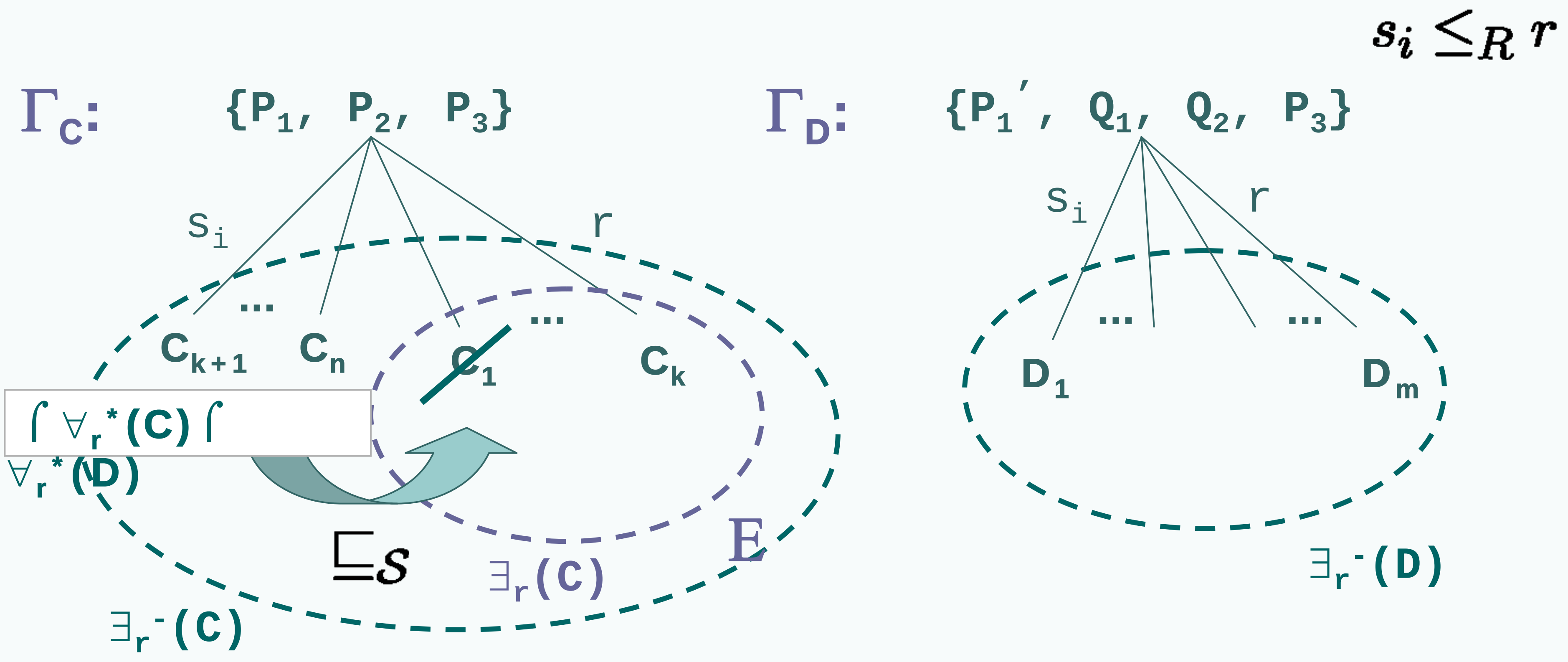
$P' \leq_C P$

$r \leq_R s_i$



$$diff_s(C,D) = P_2 \sqcap \forall r.\ diff_s(\forall_r(C), \forall_r^+(C) \sqcap \forall_r^*(D)) \sqcap \bigsqcap_{E \in \mathcal{E}} \exists r.E$$

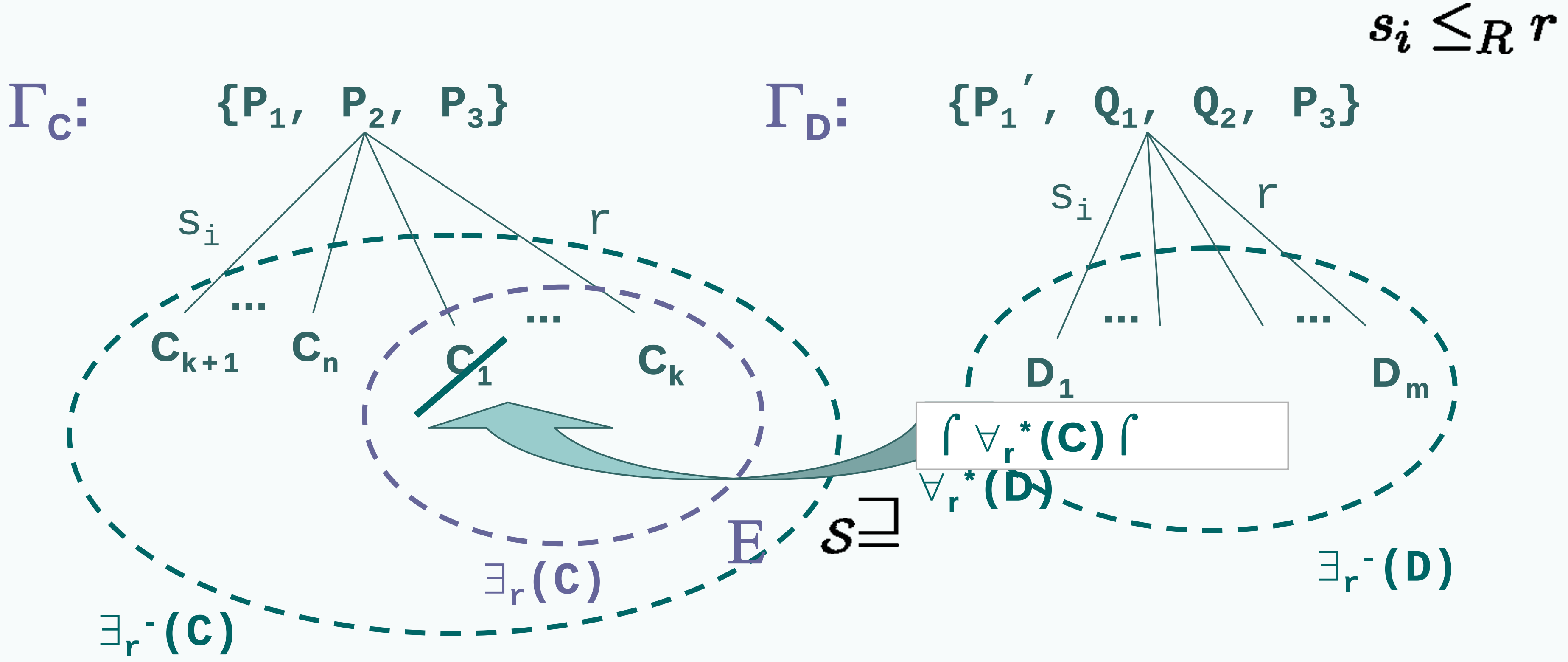
# ALEH$_S$ Difference



$diff_s(C,D) = P_2$ ⌠ $\forall r.\ diff_s(\forall_r(C), \forall_r^+(C)$ ⌠ $\forall_r^*(C))$

⌠ ⌠$_{E™E}$ $\exists r.E$

# ALEH$_S$ Difference



$$\mathit{diff}_s(C,D) = P_2 \;⌠\; \forall r.\ \mathit{diff}_s(\forall_r(C), \forall_r^+(C) ⌠ \forall_r^*(C))$$

$$⌠ \int_{E™E} \exists r.E$$

# Difference based on similarities between concept and role names

# Similarity subsumption

- $\Sigma_1$, $\Sigma_2$ the sets of symbols of two terminologies $T_1$ and $T_2$

$$\alpha(c) = \{c' \in \Sigma_2 \mid nsim(c, c') > TH\}$$
$$\alpha(\text{human}) = \{\text{person, individual, someone}\}$$

- $\sigma$ Substitution: can replace a symbol c by an element of $\alpha(c)$

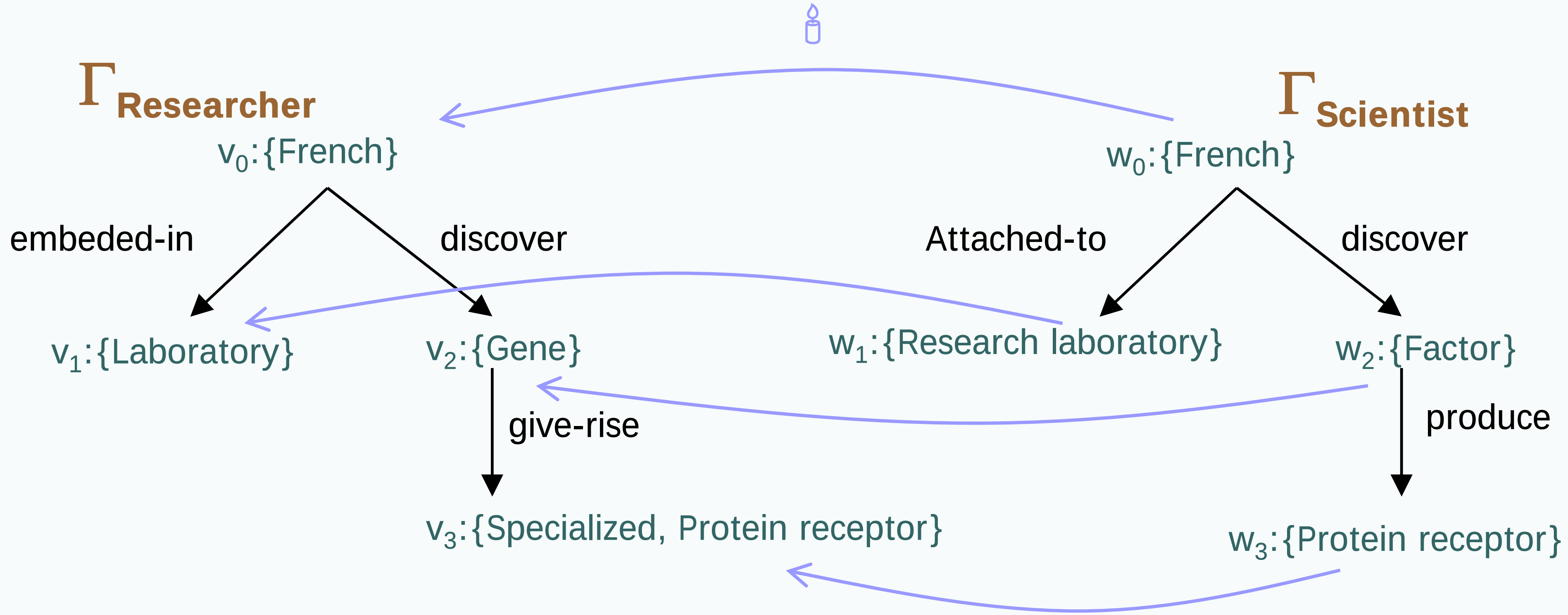
$$\sigma(\exists\text{has-child.human}) = \exists\ \text{has-offspring.person}$$

$$\sigma(\text{human}\sqcap\exists\text{has-child.human}) = \text{individual}\exists\text{has-offspring.person}$$

$C \sqsubseteq_\alpha D$, iff there exists a substitution $\sigma$ w.r.t. $\alpha$ such that $C \sqsubseteq \sigma(D)$.

# Similarity subsumption

- **Homomorphism**



Resercher $\sqsubseteq_\alpha$ Scientist

$\sigma$ over $\alpha$ such that: Researcher $\sqsubseteq \sigma$ (Scientist)

# Similarity difference

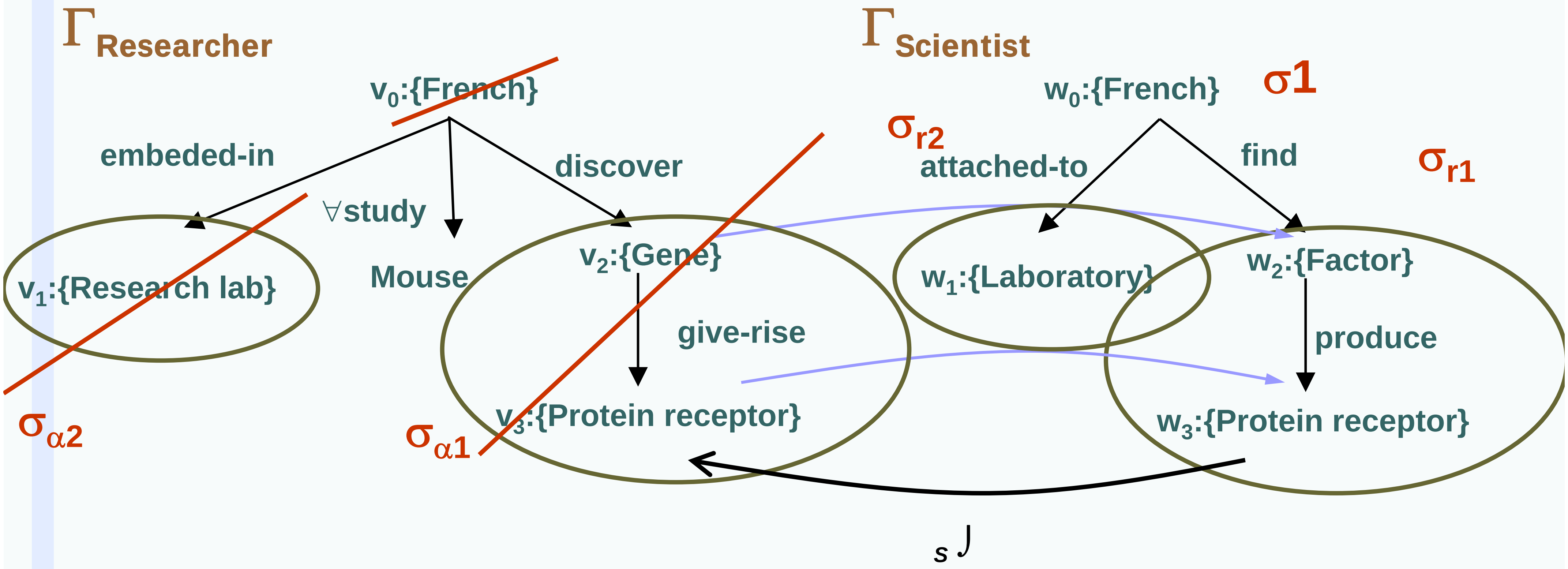
An expansion $\delta_{\mathcal{S}}$ w.r.t. $\mathcal{S}$

$$\delta_{\mathcal{S}}(C) = \prod_{\substack{\sigma_i \in \mathcal{S} \\ C_i \in C}} \sigma_i(C_i)$$

$E = C -_{\mathcal{S}} D$, iff there exists two sets of substitutions $\mathcal{S}_1$ and $\mathcal{S}_2$ w.r.t. $\alpha$ such that

$$\delta_{\mathcal{S}_1}(C) \sqcap \delta_{\mathcal{S}_2}(D) \equiv \delta_{\mathcal{S}_1}(E) \sqcap \delta_{\mathcal{S}_2}(D).$$

# Similarity difference



The sets of substitutions $\Sigma_1 = \{\sigma_{\alpha 1}, \sigma_{\alpha 2}\}$ $\Sigma_2 = \{\sigma_1, \sigma_{r1}, \sigma_{r2}\}$

s-diff (Resercher, Scientist) = ∀study. Mouse

# Similarity difference

$$\delta_{\mathcal{S}_1}(C) \sqcap \delta_{\mathcal{S}_2}(D) \equiv \delta_{\mathcal{S}_1}(E) \sqcap \delta_{\mathcal{S}_2}(D).$$

$$\begin{aligned}
&\mathsf{French} \sqcap \exists \mathsf{discover}.(\sigma_{\alpha_1}(\mathsf{Gene}) \sqcap \exists \sigma_{\alpha_1}(\mathsf{give\text{-}rise}).\sigma_{\alpha_1}(\mathsf{Protein\ \ receptor})) \sqcap \\
&\exists \mathsf{embedded\text{-}in}.\sigma_{\alpha_2}(\mathsf{Reaserach\ \ lab})) \sqcap \forall \mathsf{study.Mouse} \sqcap \sigma_1(\mathsf{French}) \sqcap \\
&\exists \sigma_{\alpha_{r1}}(\mathsf{find}).(\mathsf{Factor} \sqcap \exists \mathsf{produce.\ \ Protein\ \ receptor})) \sqcap \\
&\exists \sigma_{\alpha_{r1}}(\mathsf{attached\text{-}to}).\mathsf{laboratory} \\
&\equiv \\
&\forall \mathsf{study.Mouse} \sqcap \sigma_1(\mathsf{French}) \sqcap \exists \sigma_{\alpha_{r1}}(\mathsf{find}).(\mathsf{Factor} \sqcap \\
&\exists \mathsf{produce.\ \ Protein\ \ receptor})) \sqcap \exists \sigma_{\alpha_{r1}}(\mathsf{attached\text{-}to}).\mathsf{laboratory}
\end{aligned}$$

$$\textbf{\textit{DSIM}}\ (\text{Researcher, Scientist}) = 1 - \frac{|\ \text{s-diff (Researcher, Scientist)}\ |}{|\ \text{Researcher}\ |} = 0.77$$

# The description matching

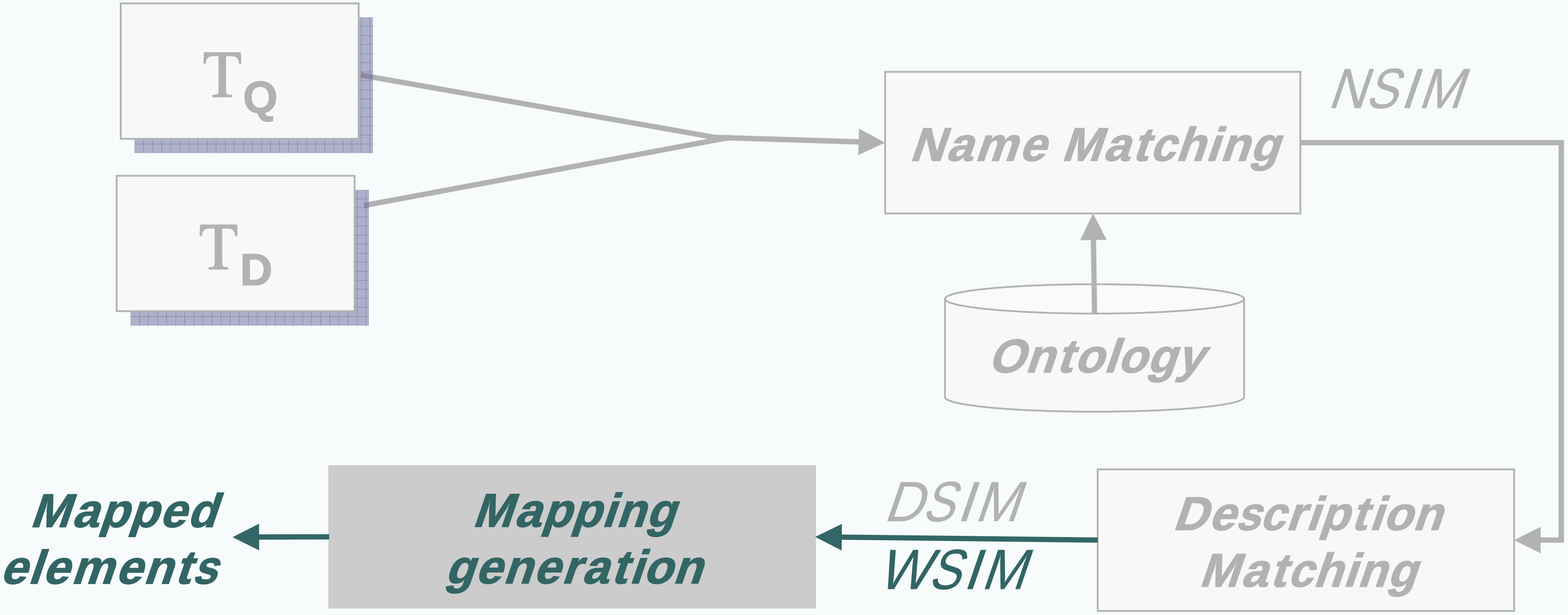
- ***WSIM*** is mean of ***NSIM*** and ***DSIM***

$$\boxed{WSIM = w * NSIM + (1 - w) * DSIM}$$

***NSIM*** (Researcher, Scientist) = 1
***DSIM*** (Reasearcher, Scientist) = 0.77

***=> WSIM*** (Reasearcher, Scientist) = 0.83
($w$= 0.3)

# The matching step



# The mapping generation

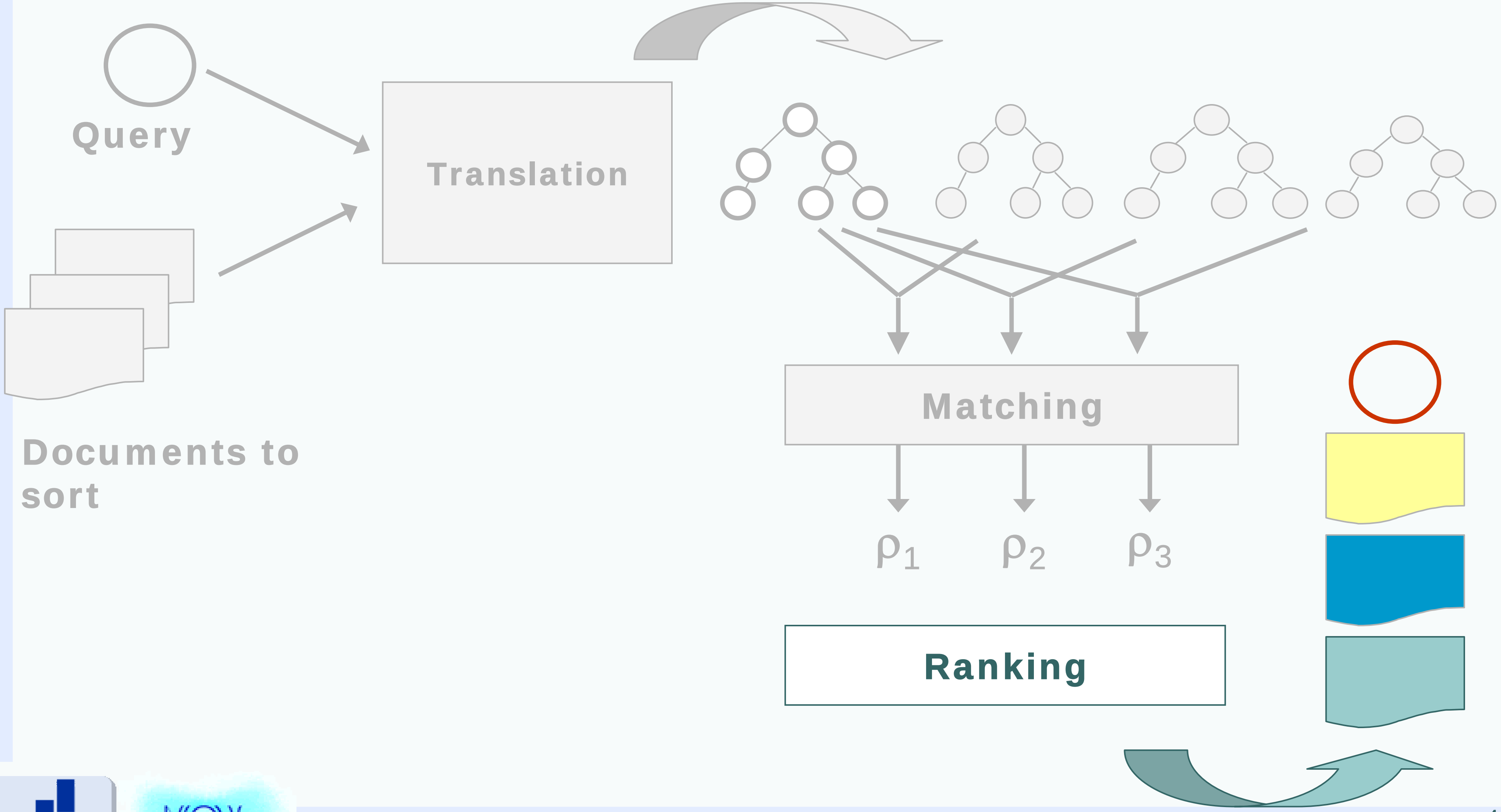
- A mapping is returned between elements having a weighted similarity greater than $th_{map}$

$th_{map}$ = 0.75

***WSIM*** (Reasearcher, Scientist) = 0.83 **>** $th_{map}$

ρ (Reasearcher) = Scientist

# The global process



# The ranking step

- The ranking function
  - Based on the matching result
  - Computes the non covered part of the query by each document
  - Ranks the documents according to the size of this part

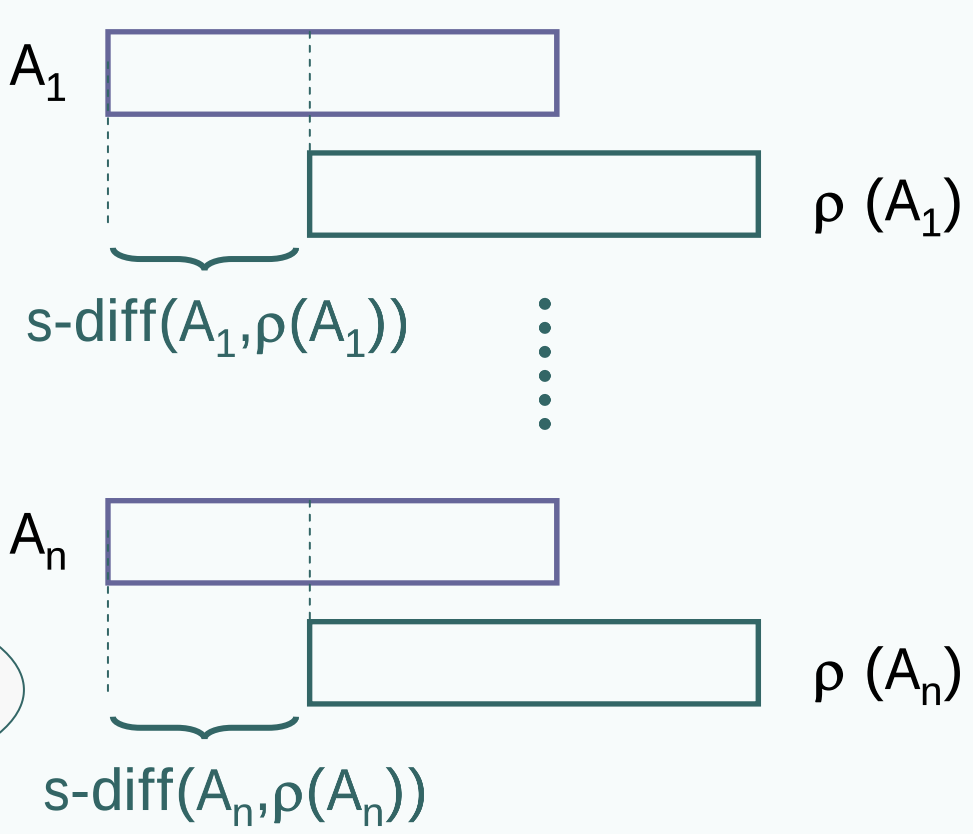
# The ranking step

$T_Q$

$A_1$ B ...
$A_2$ B ...
$A_n$ B ...

$T_D$

$B_1$ B ...
$B_2$ B ...
$B_m$ B ...


$A_1$
$\rho\ (A_1)$
$\text{s-diff}(A_1,\rho(A_1))$
$A_n$
$\rho\ (A_n)$
$\text{s-diff}(A_n,\rho(A_n))$

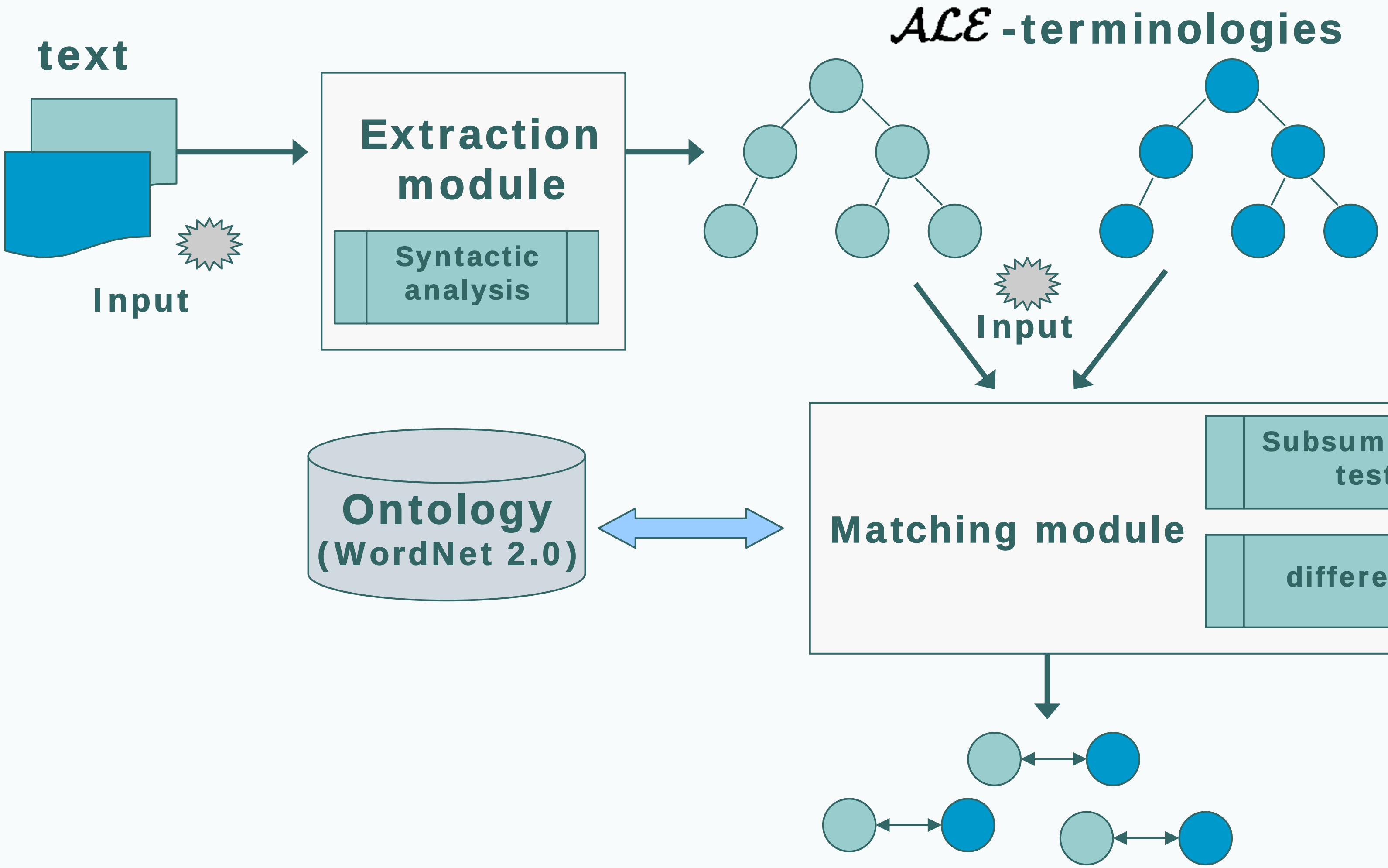
**Part of Q non covered by D**

$$\mathrm{diff}_s\ (\text{❄}_Q,\ \text{❄}_D) = \int_{i=1..n} \text{s-diff}\ (A_i\ ,\rho\ (A_i))$$

# Future work

- Approximate matching
- Application of matching to other type of data: web services
  - Representation / Adaptation to needs
- Extension of the method to
  - Structural subsumption algorithm $\mathcal{ALCEN}$

# Implementation


text
Input
Extraction module
Syntactic analysis
$\mathcal{ALE}$-terminologies
Input
Ontology (WordNet 2.0)
Matching module
Subsumption test
difference