Towards the extraction of partial instances of N-Ary relations in textual data

Atelier IN'OVIVE - PFIA 2019

Martin Lentschat, Patrice Buche, Juliette Dibie-Barthelemy, Mathieu Roche

01/07/2019

L'équipe

LENTSCHAT Martin
thèse débutée le 01/10/2018

Encadrement :

- Patrice BUCHE
- Juliette DIBIE-BARTHÉLEMY
- Mathieu ROCHE

Table des matières

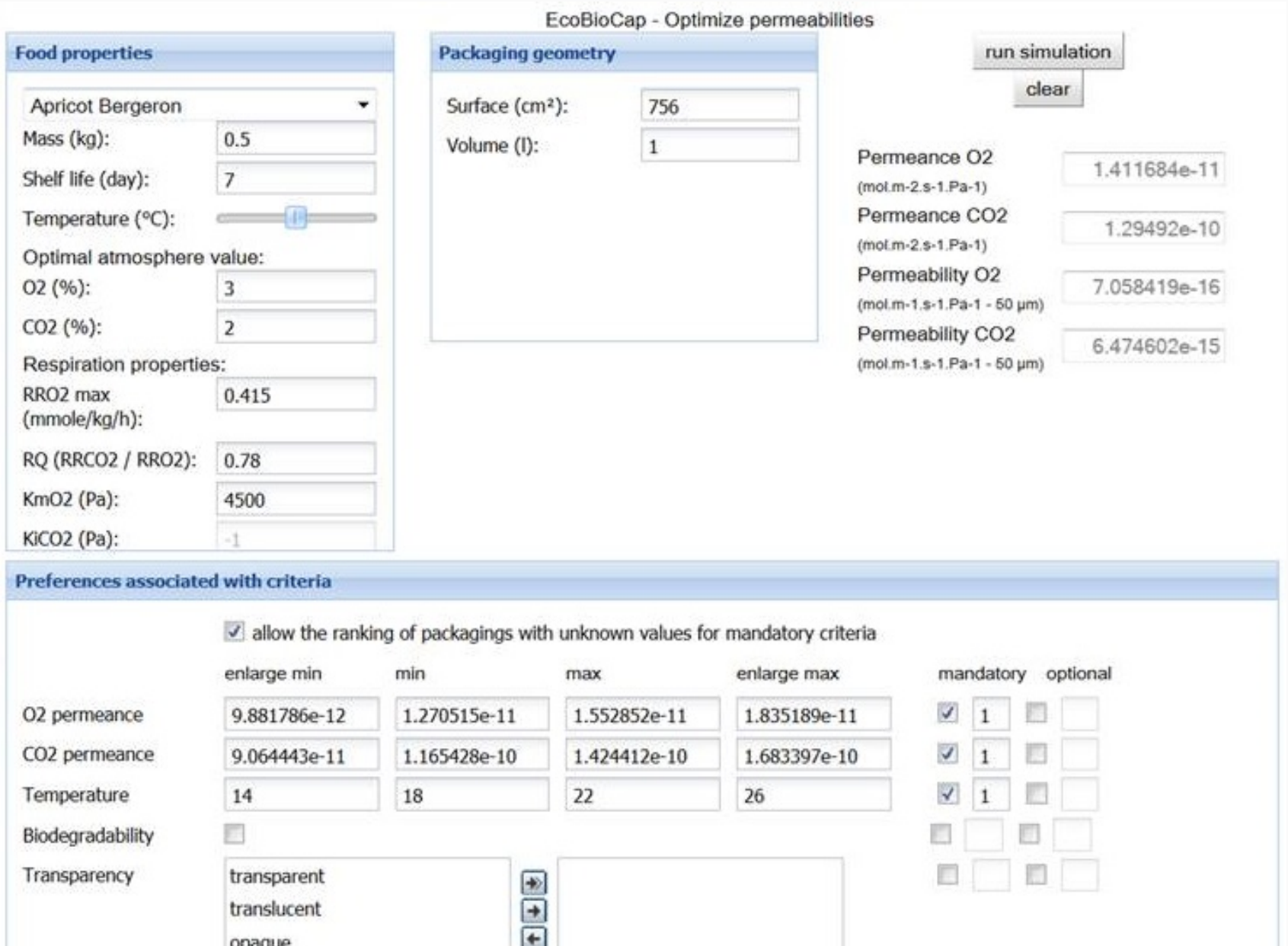
Contexte

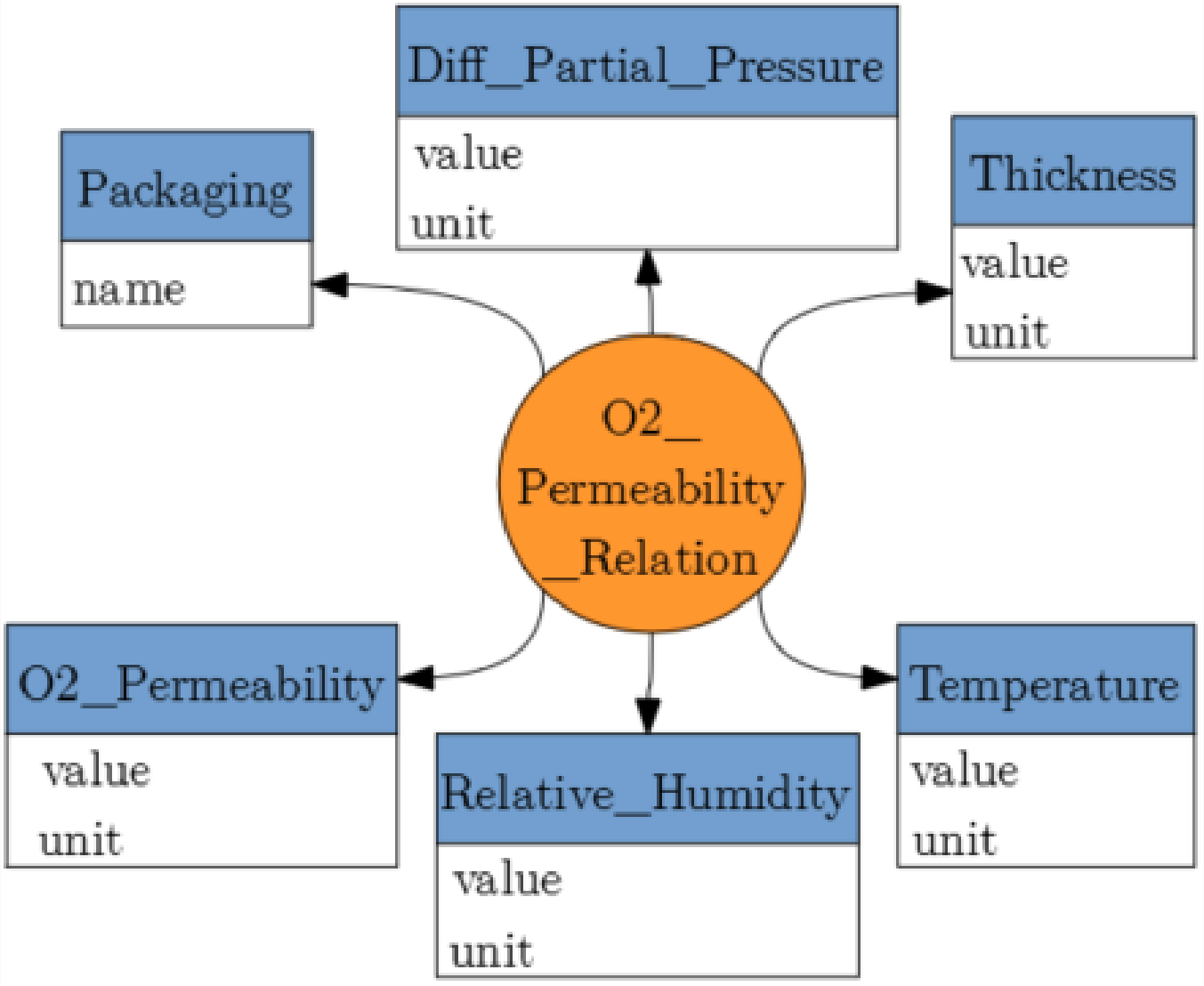
Objectif

n-ARy relaTions EXTraction for Linked Open Data

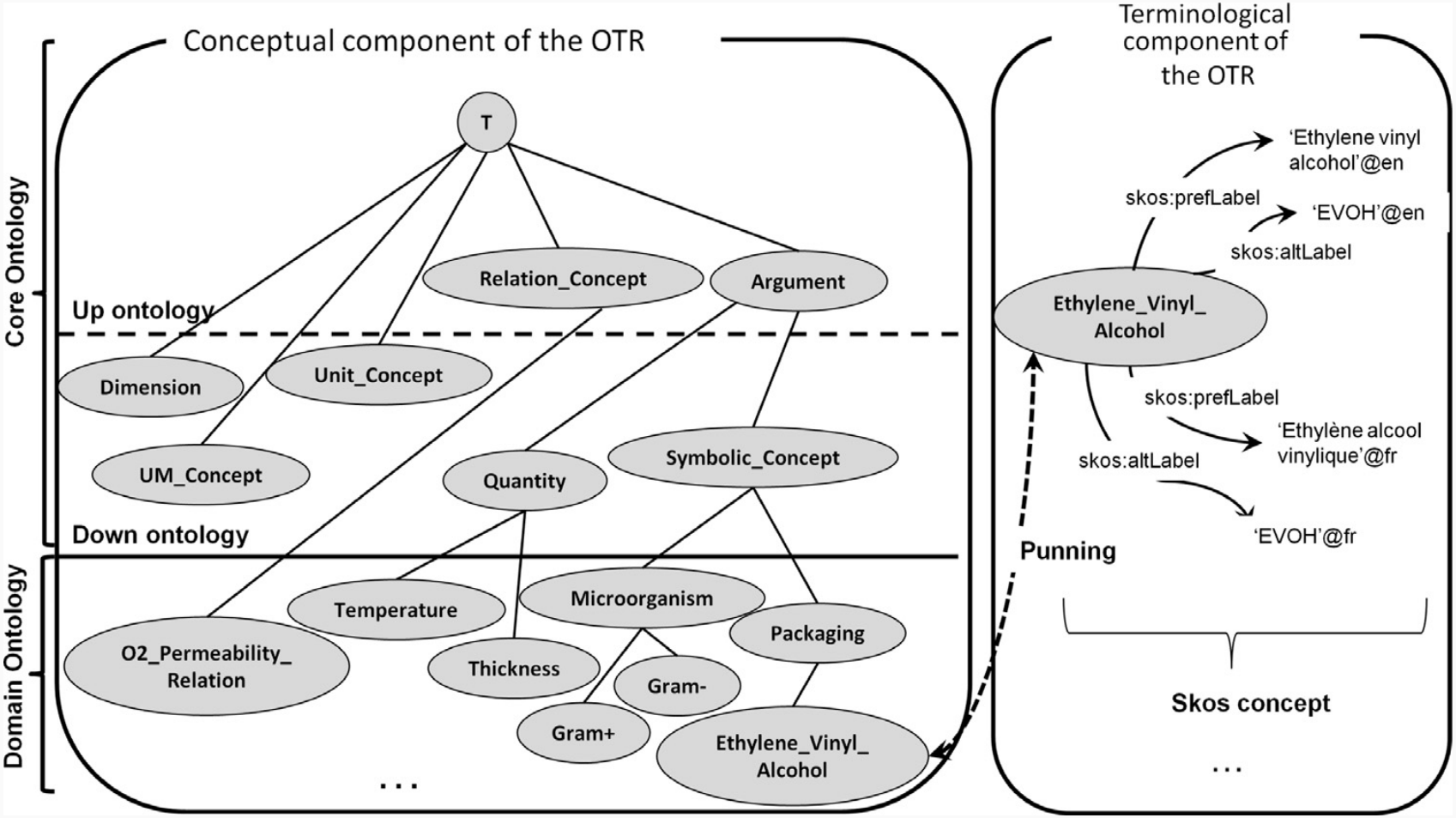
n-ARy relaTions EXTraction for Linked Open Data

Une application à un besoin : le cas ***EcoBioCap***

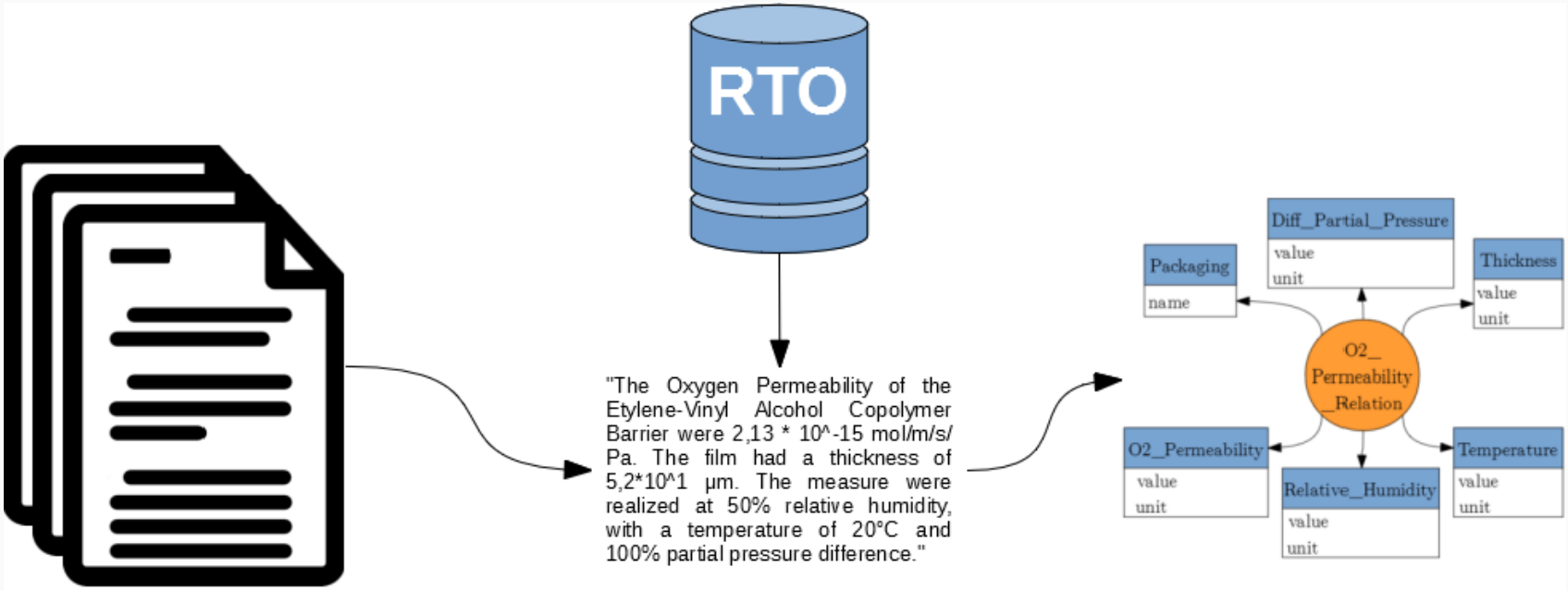



Diff_Partial_Pressure
value
unit
Packaging
name
Thickness
value
unit
O2_
Permeability
_Relation
O2_Permeability
value
unit
Relative_Humidity
value
unit
Temperature
value
unit

Notre *Ressource Termino-Ontologique* **TRANSMATT**



En bref


RTO
"The Oxygen Permeability of the Etylene-Vinyl Alcohol Copolymer Barrier were 2,13 * 10^-15 mol/m/s/Pa. The film had a thickness of 5,2*10^1 µm. The measure were realized at 50% relative humidity, with a temperature of 20°C and 100% partial pressure difference."
Diff_Partial_Pressure
value
unit
Packaging
name
Thickness
value
unit
O2_ Permeability _Relation
O2_Permeability
value
unit
Relative_Humidity
value
unit
Temperature
value
unit

Approche générale

Approche

Comment se manifeste l'information ?

Exemple :

"The Oxygen Permeability of the Etylene-Vinyl Alcohol Copolymer Barrier were $2.13 * 10^{-15} mol/m/s/Pa$. *The film had a thickness of* $5.2 * 10^{1} \mu m$. *The measure were realized at* 50% *relative humidity, with a temperature of* $20° C$ *and* 100% *partial pressure difference."*

Approche

Comment se manifeste l'information ?

Exemple :

"The Oxygen Permeability of the Etylene-Vinyl Alcohol Copolymer Barrier were $2.13 * 10^{-15}$ *mol/m/s/Pa. The film had a thickness of* $5.2 * 10^{1}$ μm*. The measure were realized at* 50 % *relative humidity, with a temperature of* 20 °*C and* 100 % *partial pressure difference."*

Approche

Comment se manifeste l'information ?

Exemple :

"The Oxygen Permeability of the Etylene-Vinyl Alcohol Copolymer Barrier were $2.13 * 10^{-15}$ *mol/m/s/Pa. The film had a thickness of* $5.2 * 10^{1}$ μm. *The measure were realized at* 50 % *relative humidity, with a temperature of* 20 °C *and* 100 % *partial pressure difference."*

Approche

Comment se manifeste l'information ?

Exemple :

"The Oxygen Permeability of the Etylene-Vinyl Alcohol Copolymer Barrier were $2.13 * 10^{-15}$ $mol/m/s/Pa$. *The film had a thickness of* $5.2 * 10^{1}$ μm. *The measure were realized at* 50 % *relative humidity, with a temperature of* 20 $^{\circ}C$ *and* 100 % *partial pressure difference."*

Travaux précédents

Une thèse en 2015 : **Extraction d'arguments de relations N-Aires dans les textes guidée par une RTO de domaine** [Ber15] [BBDR17]

Extraction de motifs séquentiels

Ancrage autour des Unités de Mesure

[...] The film had a thickness of $5.210^1 \mu m$ [...]

$\downarrow$

$films < thickness > [prep](numvalthick) < um >$

Besoins dans l'extraction d'arguments

Verrous identifiés :

1. variations dans les textes
2. dispersion des arguments

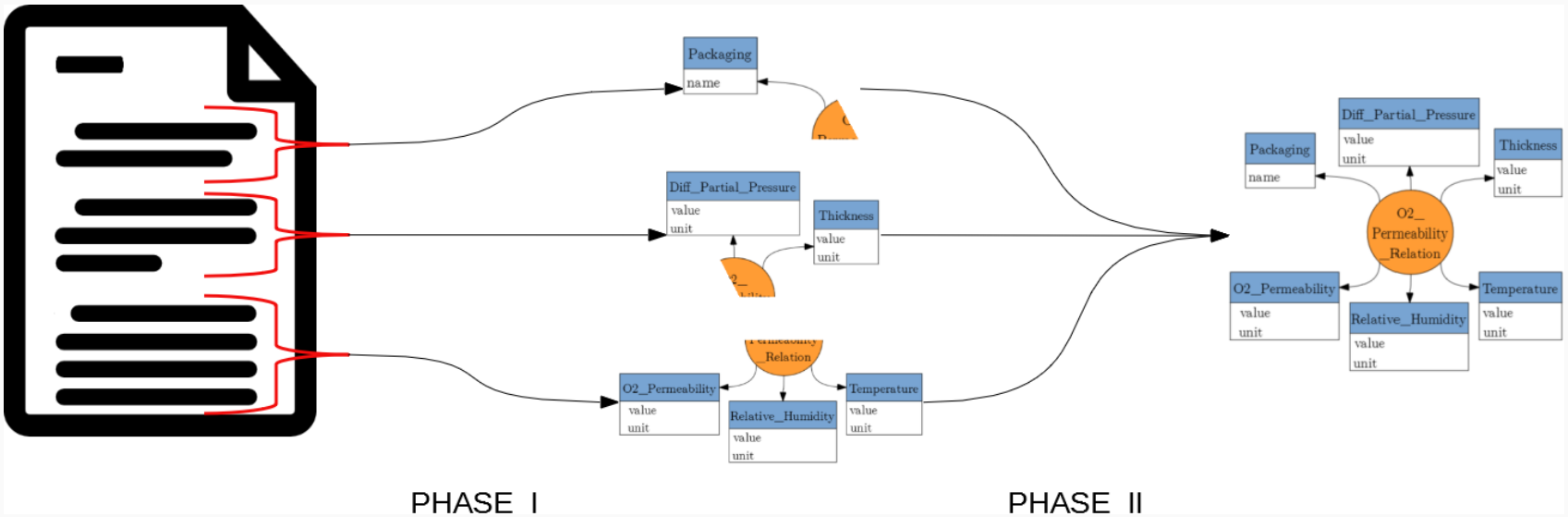
Propositions :

1. Augmentation de la couverture
   - nouvelles Unités de Mesure [Ber15]
   - reconnaissance des Variants Terminologiques
   - extraction des Acronymes
2. caractérisation des arguments avant reconstitution de relations N-Aires

Processus

Processus en deux temps :

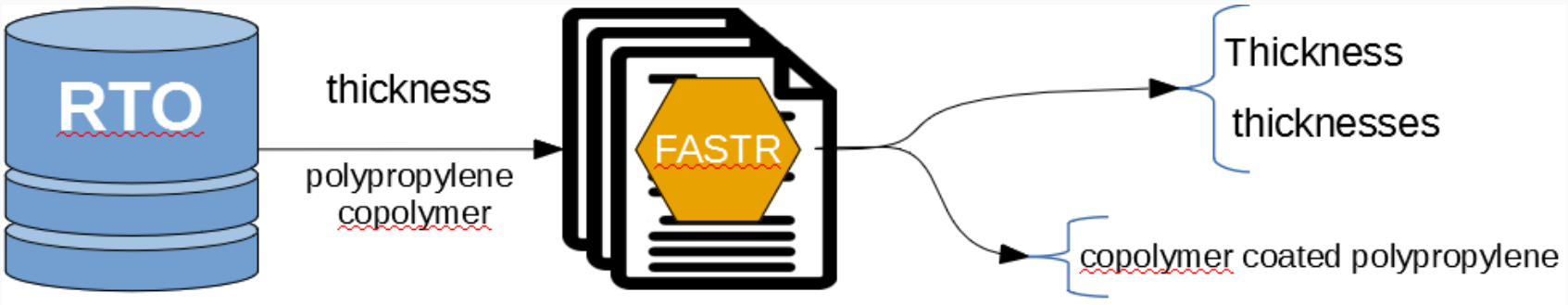
1. Extraction des Arguments
2. Reconstitution des relations N-Aires



Extraction des Arguments de relations N-Aires

Reconnaissance des Variations terminologiques

Intégration d'un outil de recherche des Variants Terminologique : **FASTR** [BJ99]



Évaluation

Corpus	60 documents
Entrée	termes relatifs au concept *packaging* de la *RTO*
Résultats	89 variations pour 35 termes
Précision	0.85

Acronymes

Importants à identifier car représentant librement des arguments de relations N-Aires

Des formes parfois complexes :

methyl cellulose $\rightarrow$ *MC*

low density polyethylene $\rightarrow$ *LDPE*

ethylene-vinyl alcohol copolymer with 44% *ethylene molar content* $\rightarrow$ *EVOH44*

potato starch film $\rightarrow$ *NS450*

Reconnaissance des Acronymes - méthodes possibles

Différentes approches [SV19]

- dictionnaires
- apprentissage
- heuristiques
- ...

Une méthode basée sur la fréquence d'apparition : $AcroRec_{FREQ}$, évaluation pour comparaison

Reconnaissance des Acronymes - notre méthode originale

Extraction d'Acronymes en domaine de spécialité

Particularités

- guidage par la RTO
- analyse syntaxique [OA06], extraction de *NounPhrases*
- recherche d'acronymes à proximité via un pattern
- mesure de similarité par *indice de Dice*

Reconnaissance des Acronymes - évaluation

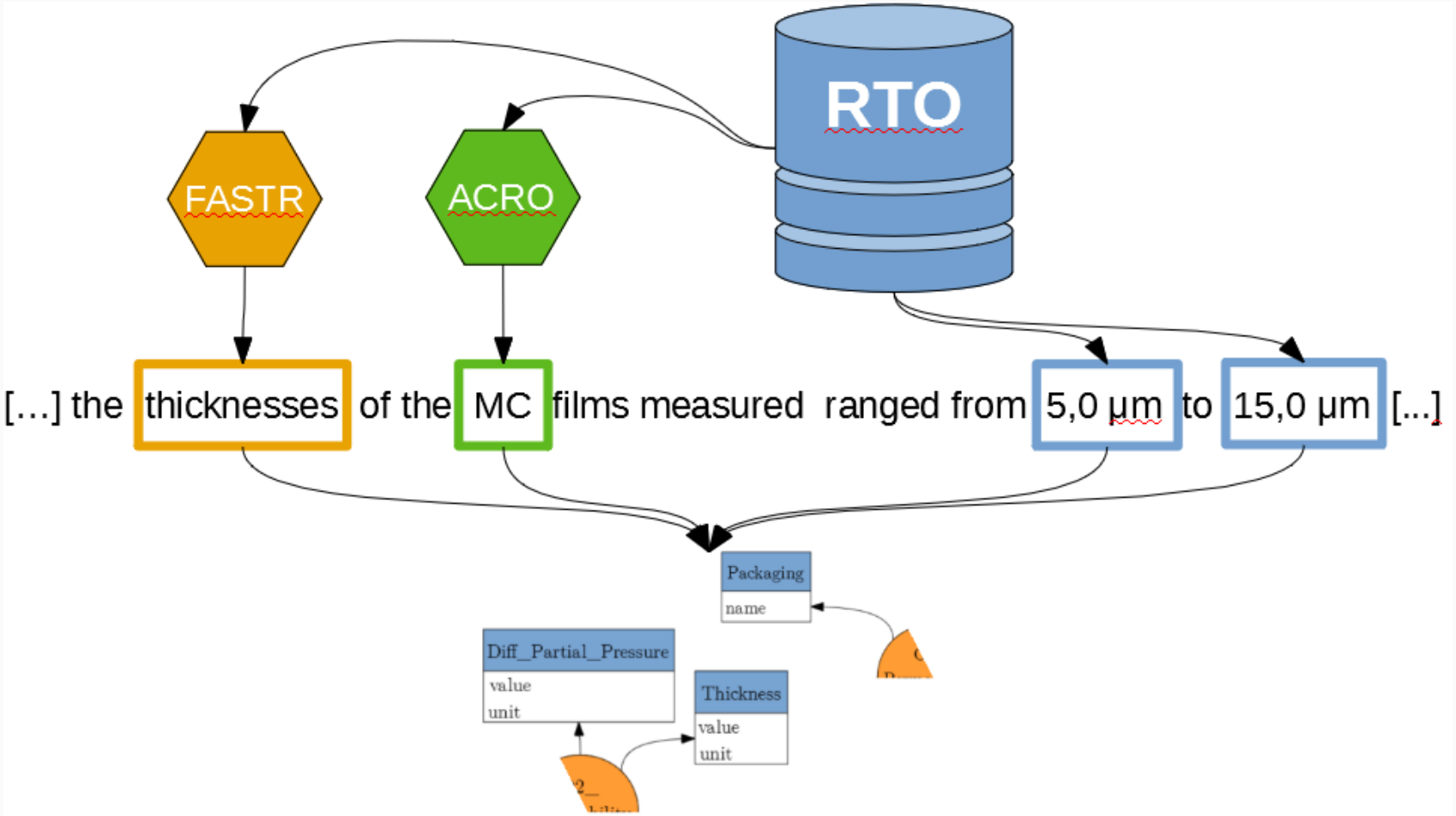
Évaluation

Corpus 60 documents contenant 53 acronymes d'intérêt

Entrée termes relatifs au concept *packaging* de la *RTO*

	$AcroRec_{FREQ}$ *VS* $AcroRec_{RTO}$	
Précision	0.49	**0.85**
Rappel	**0.58**	0.53
F1-score	0.53	**0.65**

Reconnaissance d'arguments



Nouveau besoin : contextualiser et caractériser les arguments avant la **PHASE II**

Travaux en Cours

Contextualisation des Arguments

Utilisation des Segments textuels (ie. parties, sous-parties ...)

classification des Segments

+

méthode de pondération TF et ICF

=

Description contextuelle et statistique

Perspectives et Développements

Perspectives et Développements

À moyen terme

- finalisation de la PHASE I
- évaluation générale des contributions

Perspectives Futures

- Mise en place de la PHASE II
  - heuristiques pilotées par la RTO
  - utilisation de la syntaxe
- application à un autre domaine

Bibliographie

Soumia Lilia Berrahou, Patrice Buche, Juliette Dibie, and Mathieu Roche.
Xart : Discovery of correlated arguments of n-ary relations in text.
Expert Systems with Applications, 73 :115–124, 2017.

Soumia Lilia Berrahou.
Extraction d'arguments de relations n-aires dans les textes guidée par une RTO de domaine.
PhD thesis, Université de Montpellier, 2015.

Didier Bourigault and Christian Jacquemin.
Term extraction-i-term clustering : An integrated platform for computer-aided terminology.
In Ninth Conference of the European Chapter of the Association for Computational Linguistics, 1999.

Naoaki Okazaki and Sophia Ananiadou.
A term recognition approach to acronym recognition.
In Proceedings of the COLING/ACL on Main conference poster sessions, pages 643–650. Association for Computational Linguistics, 2006.

R.Menaha Senthilkumar and Jayanthi VE.
A Survey on Acronym–Expansion Mining Approaches from Text and Web : Proceedings of the Second International Conference on SCI 2018, Volume 1, pages 121–133.
01 2019.