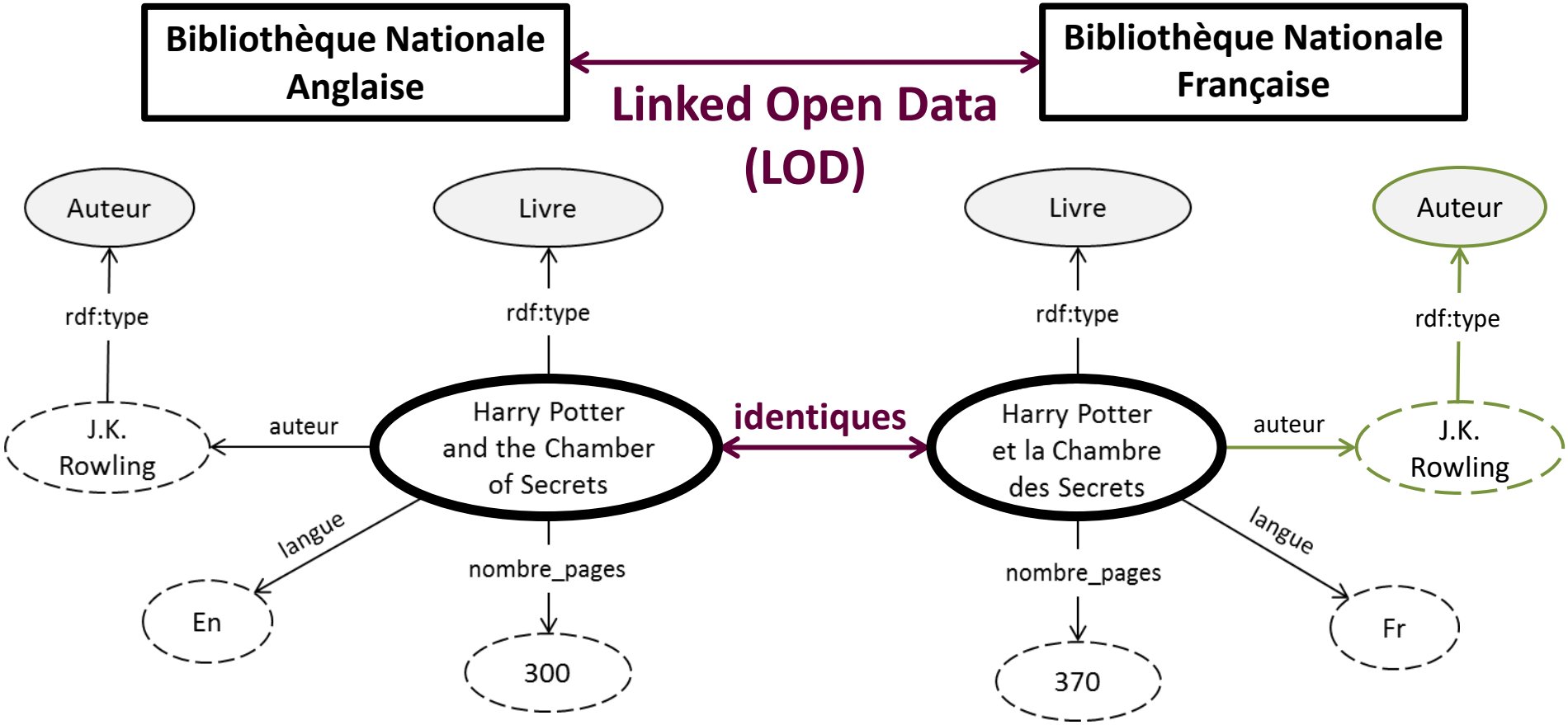
Détection de liens d'identité contextuels dans une base de connaissances

Joe Raad, Nathalie Pernelle, Fatiha Saïs
prenom.nom@lri.fr
LRI, Université Paris-Sud
Orsay, France

IC 2017
5 Juillet 2017

Identité dans le Web Sémantique

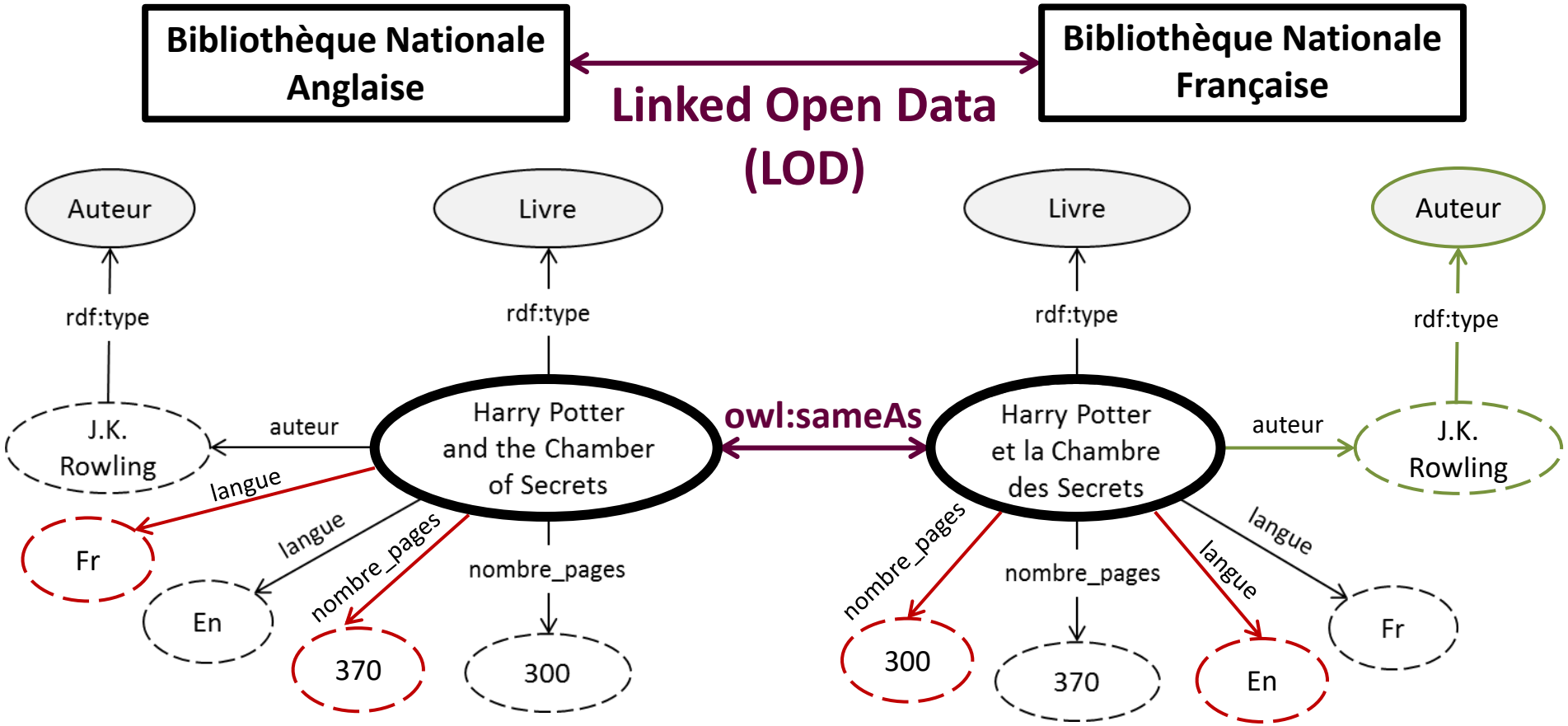
Pourquoi ?



- ✓ **Enrichir la bibliothèque nationale française**
- ✓ **Intégrer des informations issues de plusieurs sources**

Identité dans le Web Sémantique

Comment ?



Identité dans le Web Sémantique

owl:sameAs

- Relation qui indique que deux individus avec des URI différentes font référence à la même chose
- Transitive, symétrique, et réflexive
- Report de toutes les propriétés: *owl:sameAs*(i1, i2) ∧ *prop*(i1, v) → *prop*(i2, v)

Limites

- Représente seulement une identité complète
- En cas de mauvaise utilisation, peut aboutir lors d'inférence à des faits erronés et des incohérences par rapport à des axiomes définis dans une ontologie

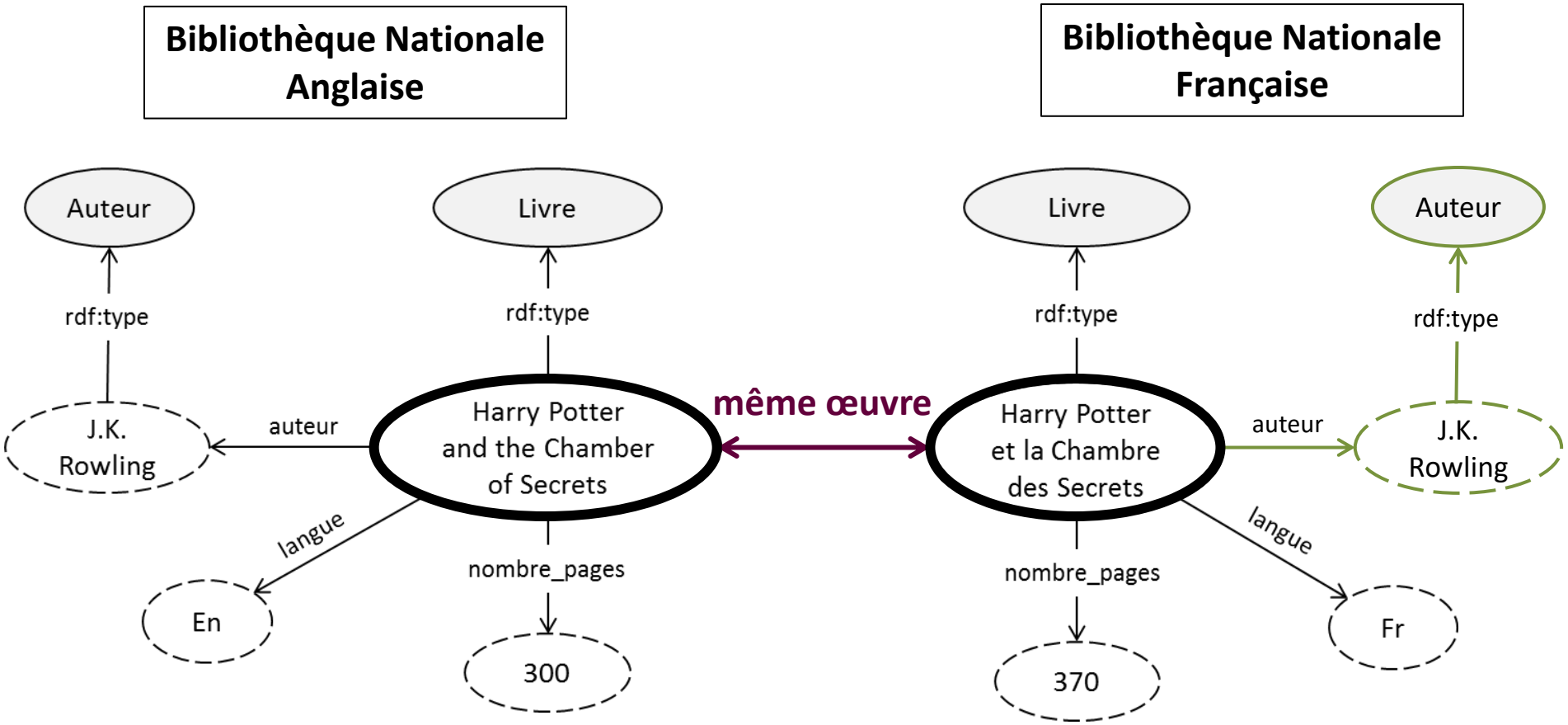
Liens owl:sameAs dans le LOD

≈ 58 millions **(OpenLink 2010)** → ≈ 558 millions **(LOD stat 2016)**

McGuinness et al. "OWL web ontology language overview." W3C recommendation 10.10. 2004.

Identité dans le Web Sémantique

Solution ?



Identité Contextuelle

Identité Contextuelle - Exemple

- 1992
- Bordeaux
- Rouge
- 13% alcool

?
identiques

- 2016
- Californie
- Rouge
- 13% alcool

Contexte 1
Différents

Contexte 2
Différents

Contexte 3
Identiques

Contexte 4
Identiques

Identité Contextuelle - Etat de l'Art

1. **skos:exactMatch :** indique un haut degré de confiance sur le fait que les éléments peuvent être utilisés de manière interchangeable dans un grand nombre de contextes mais sans pour autant préciser ces contextes

- Ne peut être utilisé que pour des URIs dont le type est un concept SKOS
- Contexte d'identité pas défini

Miles et al. "Skos simple knowledge organization system reference." 2009.

2. **The Similarity Ontology :** présente une hiérarchie de 13 prédicats (8 nouveaux)

Chaque prédicat est caractérisé par les propriétés de réflexivité, transitivité et de symétrie

- Difficile à utiliser car trop subjectif
- Contexte d'identité pas défini

Halpin et al. "When owl:sameAs isn't the same: An analysis of identity in linked data." ISWC 2010.

Identité Contextuelle - Etat de l'Art

3. Liens d'identité spécifiques au domaine :

taux_Alcool(vin1, a1) ∧ taux_Alcool(vin2, a1) → même_vin(vin1, vin2)

- Nécessité de l'intervention des experts

4. owl:sameAs contextualisé : définir des relations d'identités dans un contexte représenté par un ensemble de propriétés

Les contextes sont organisés dans un treillis de contextes (inclusion des propriétés)

- Le contexte est un ensemble de propriétés qui ne tient pas compte des classes de l'ontologie
- Identité est définie localement (sans propagation dans le graphe RDF)

Beek et al. "A Contextualised Semantics for owl: sameAs." ISWC 2016

Objectifs

- **Représenter un contexte en fonction du vocabulaire de l'ontologie**
  - Difficulté pour les experts de lister les contextes d'identités pertinents puisqu'ils peuvent varier en fonction des tâches
  - Plus facile de spécifier les contraintes sur les contextes

Contrainte 1 Propriétés Non-Pertinentes	Contrainte 2 Propriétés Essentielles	Contrainte 3 Propriétés Co-Occurrentes
Ne seront pas considérés dans les contextes d'identités	**Doivent être prises en compte dans les contextes d'identités**	**Doivent exister ensemble si elles sont prises en compte dans les contextes d'identités**
Valeurs non structurées (texte libre); variations non significatives; valeurs évolutives	*Ne comparer deux produits que s'ils partagent le même type*	*Une mesure n'a pas de sens sans son unité et réciproquement*

- **Proposer une approche qui est capable de calculer les contextes les plus spécifiques dans lequel deux instances sont identiques** (en prenant en considération les contraintes des experts quand elles sont disponibles)

Plan

1. **Contextes**
   a) Contexte Local
   b) Contexte Global

2. **Identité dans un contexte global**

3. **Détection de liens d'identités contextuels**

4. **Expérimentations**

Contexte Local

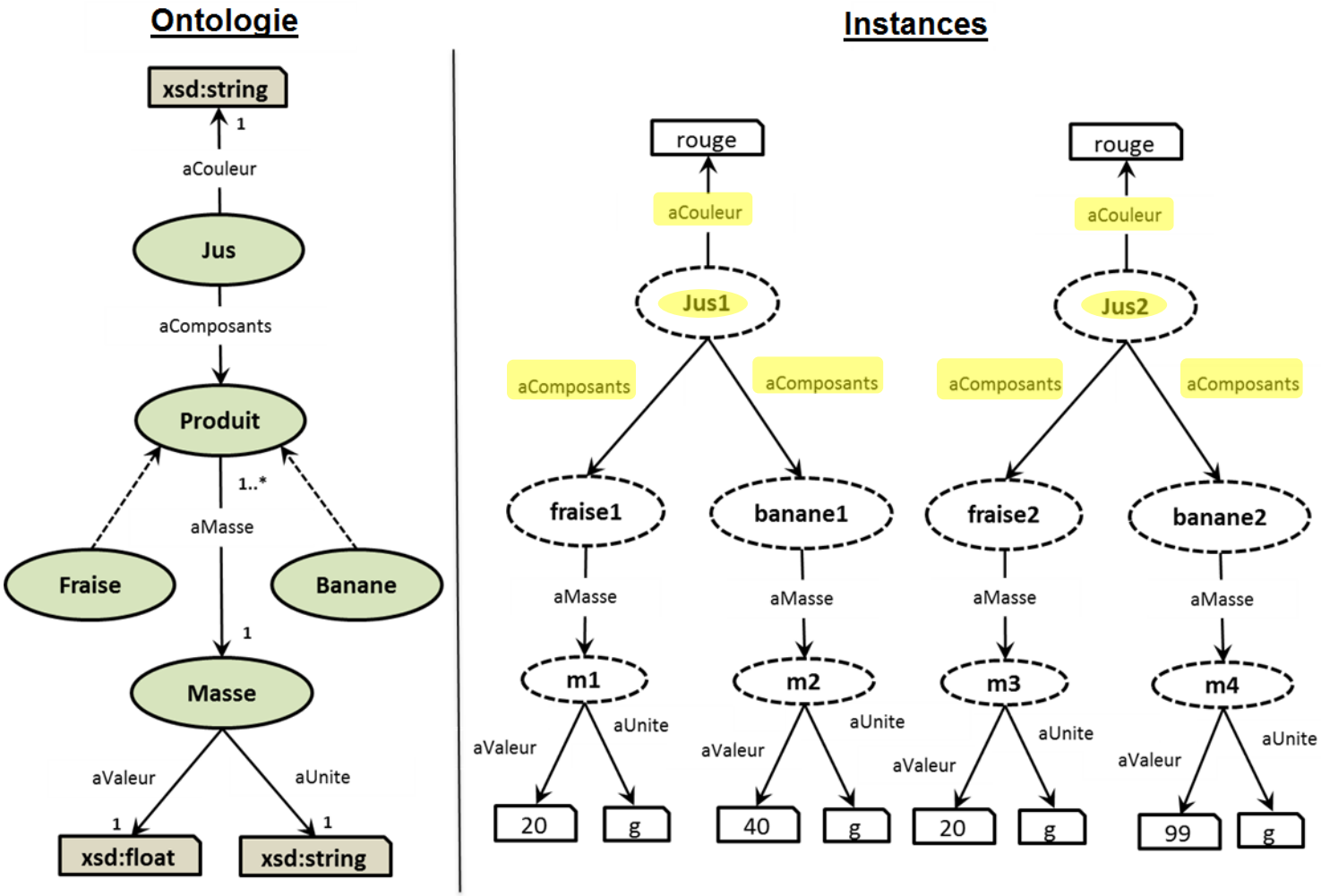
$$\pi(c) = (c, DP^c, OP^c)$$

- **c :** une classe donnée de l'ontologie

- **DP^c :** ensemble de propriétés qui peuvent être soit *rdf:Type*, *owl:DataTypeProperty* ou *owl:AnnotationProperty* dans lequel c est défini en domaine

- **OP^c :** ensemble de *owl:ObjectProperty* dans lequel *c* apparait en domaine ou en co-domaine (noté op^{-1})

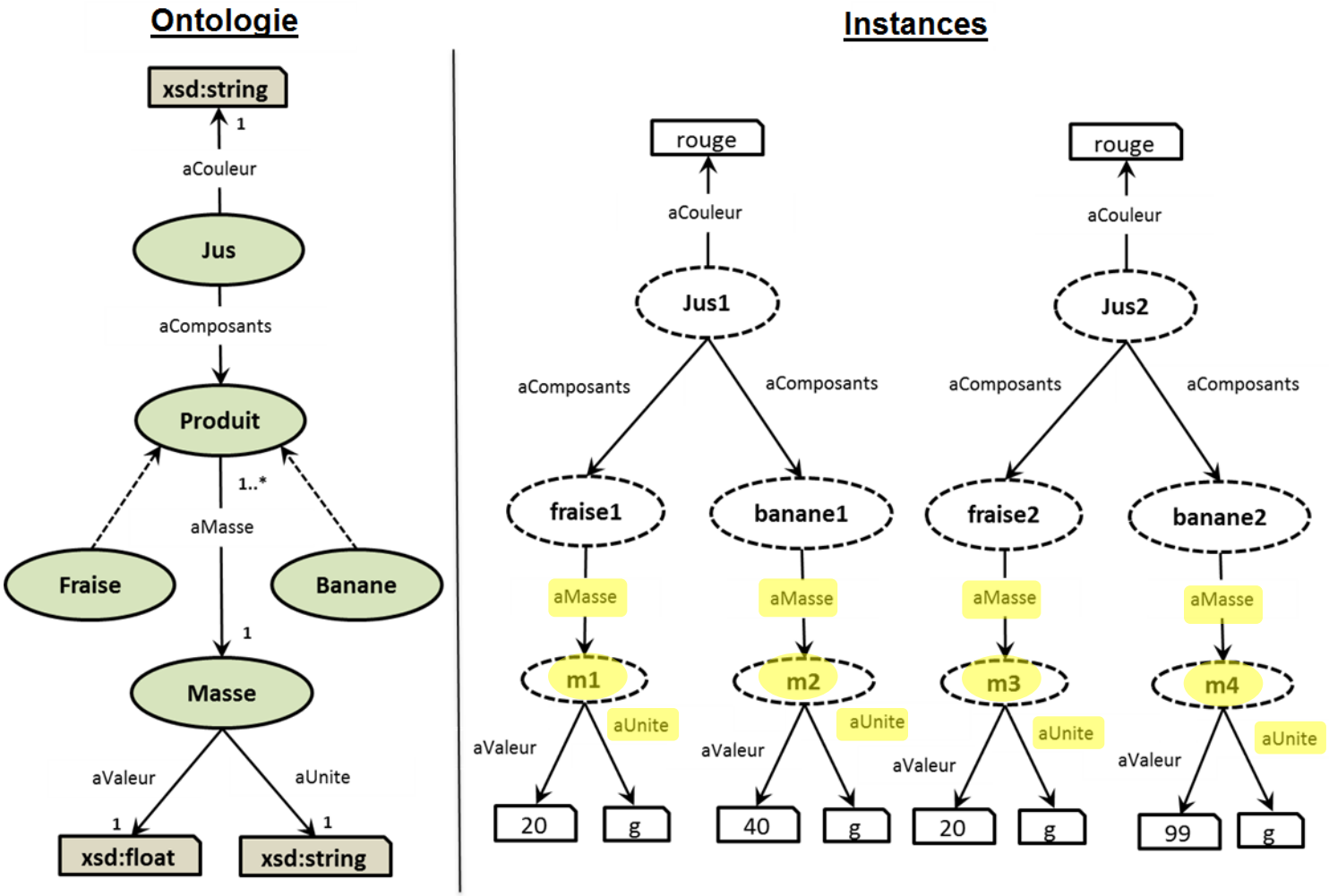
Contexte Local



c DP^c OP^c

Exemple: π_1**(Jus) =** (Jus, {rdf:Type, aCouleur}, {aComposants})

Contexte Local



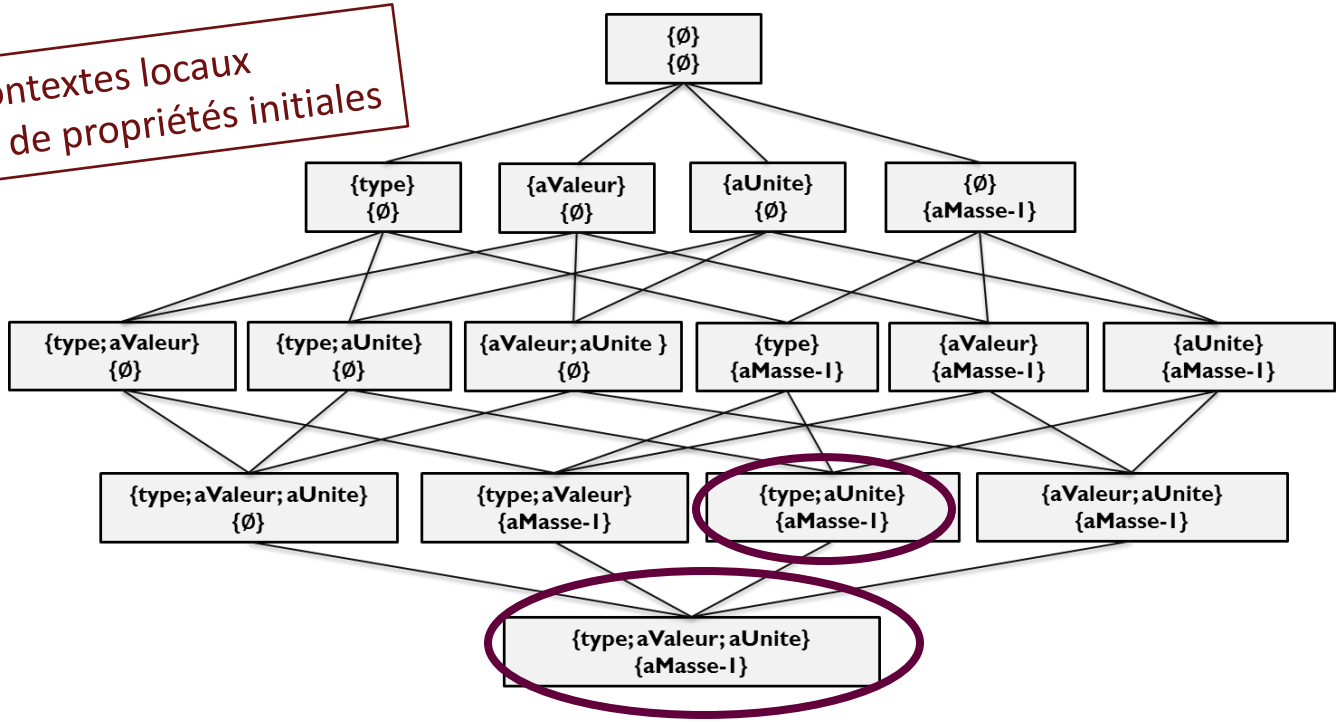
$$\pi_3(\text{Masse}) = (\underbrace{\text{Masse}}_{c},\ \underbrace{\{\text{rdf:Type}, \text{aUnite}\}}_{DP^c},\ \underbrace{\{\text{aMasse}^{-1}\}}_{OP^c})$$

Exemple: $\pi_3(\text{Masse})$ = (Masse, {rdf:Type, aUnite}, {aMasse^{-1}})

Contexte Local

Relations d'ordre pour la classe Masse

2^n contextes locaux
n = nombre de propriétés initiales



π_1**(Masse) =** (Masse, $\overbrace{\{rdf:Type, aValeur, aUnite\}}^{DP_1}$, $\overbrace{\{aMasse^{-1}\}}^{OP_1}$)

π_3**(Masse) =** (Masse, $\overbrace{\{rdf:Type, aUnite\}}^{DP_3}$, $\overbrace{\{aMasse^{-1}\}}^{OP_3}$)

$DP_3 \subseteq DP_1$ et
$OP_3 \subseteq OP_1$

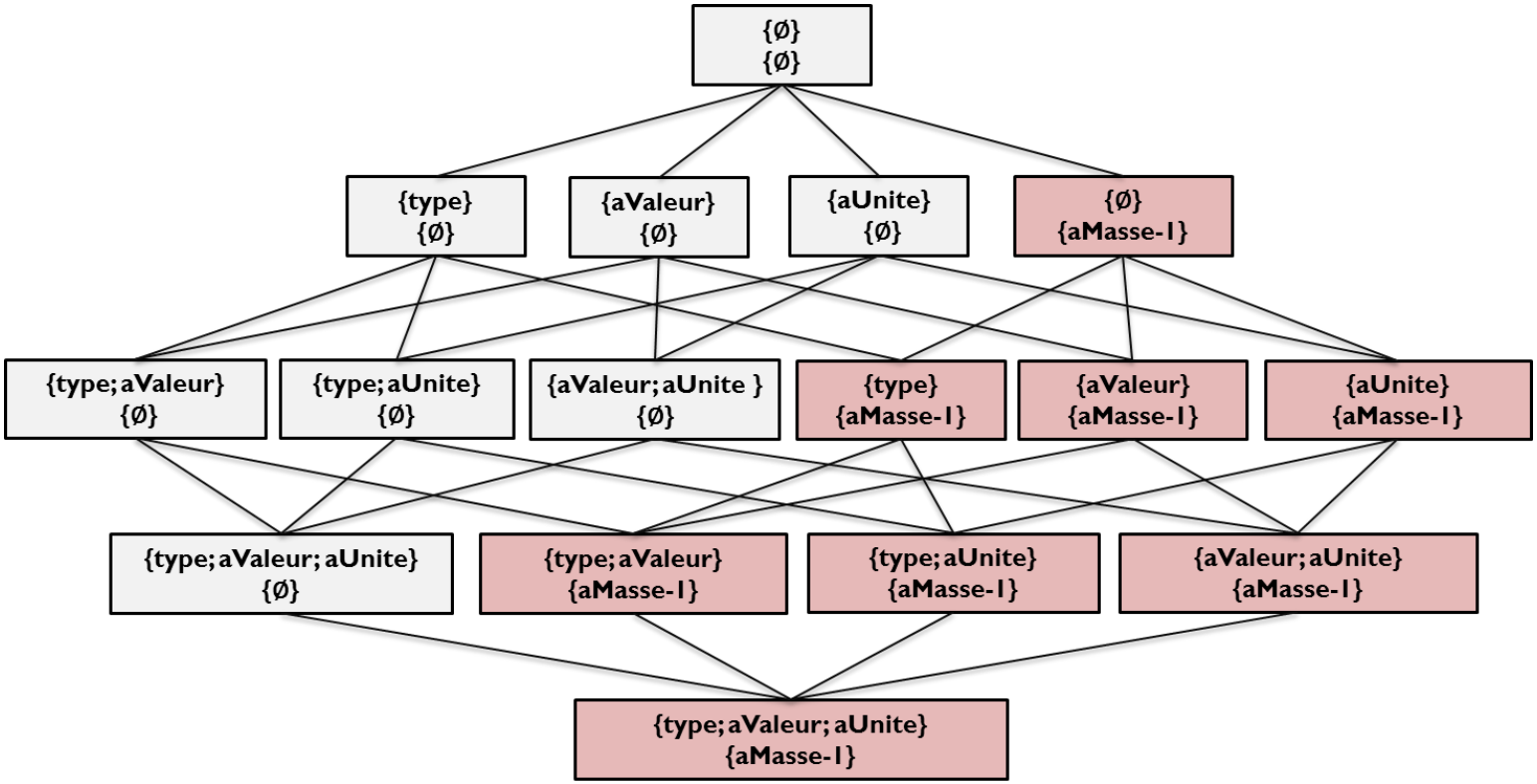
π_1**(Masse)** $\leq$ π_3**(Masse)**

Contexte Local

Contraintes d'élagage des contextes locaux

Liste de Propriétés Non Pertinentes

Contrainte Experte 1 : la propriété aMasse^{-1} est non pertinente et n'a pas de sens d'exister dans le calcul d'identité

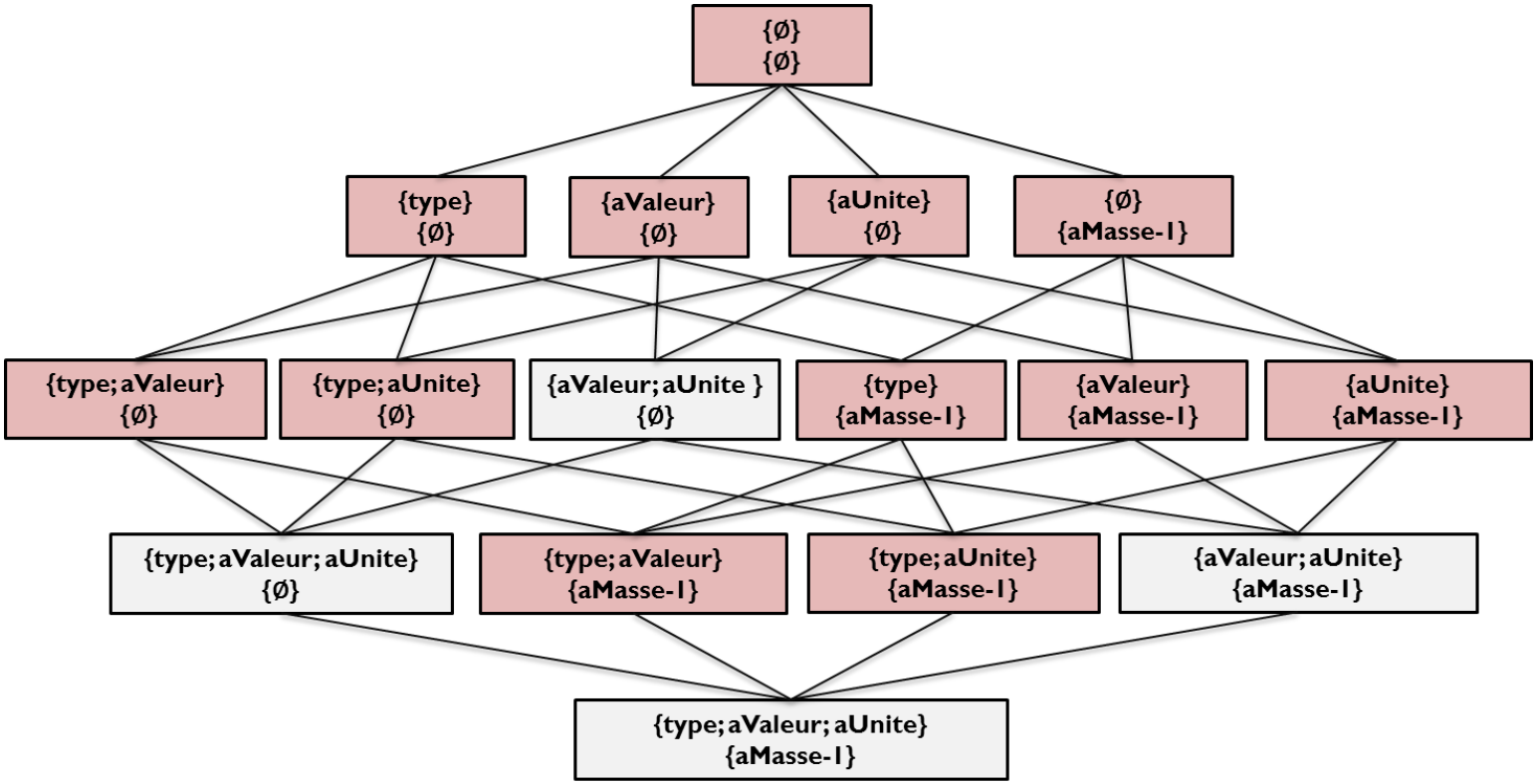


Contexte Local

Contraintes d'élagage des contextes locaux

Liste de Propriétés Essentielles

Contrainte Experte 2 : la valeur et son unité sont essentielles pour le calcul de l'identité

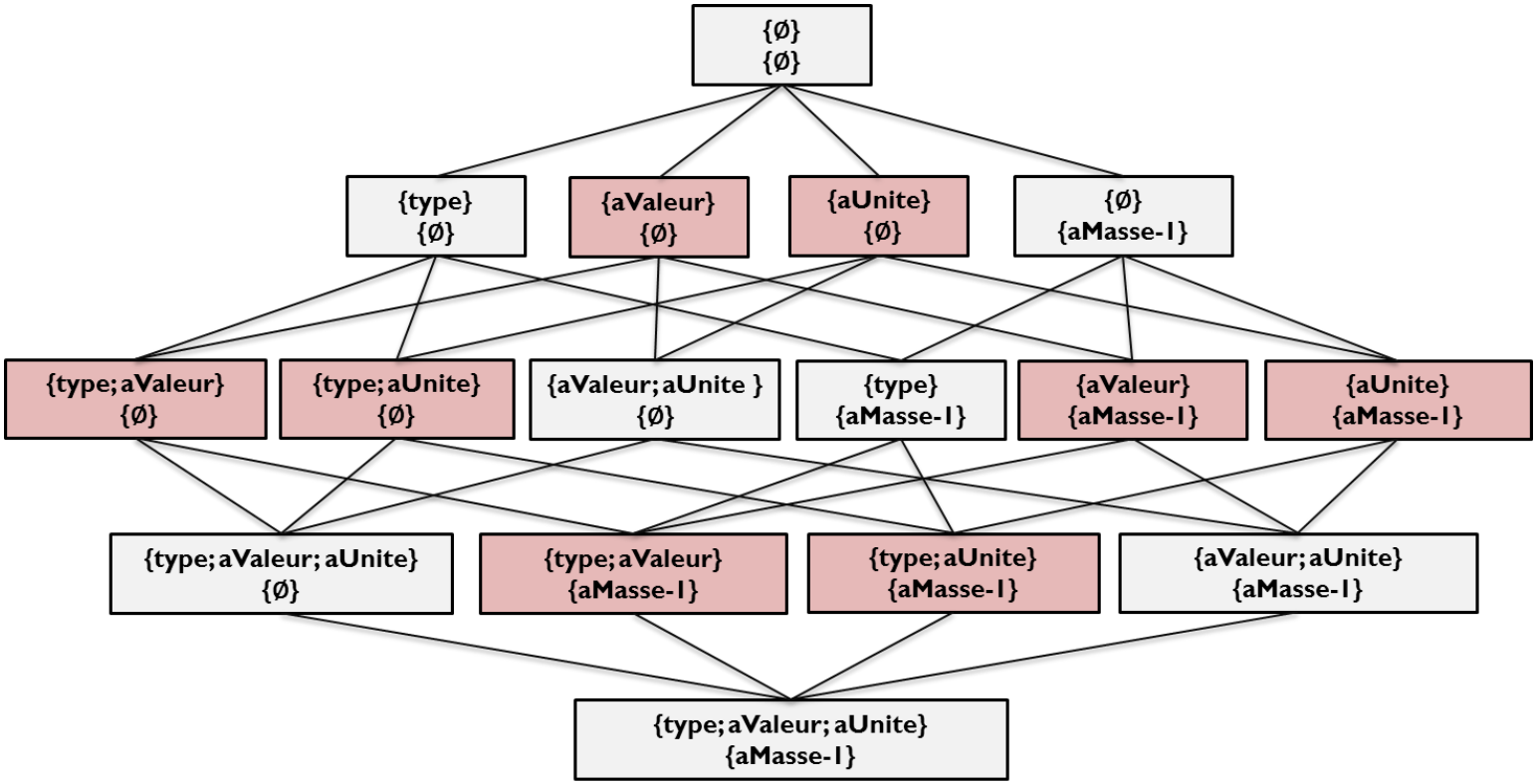


Contexte Local

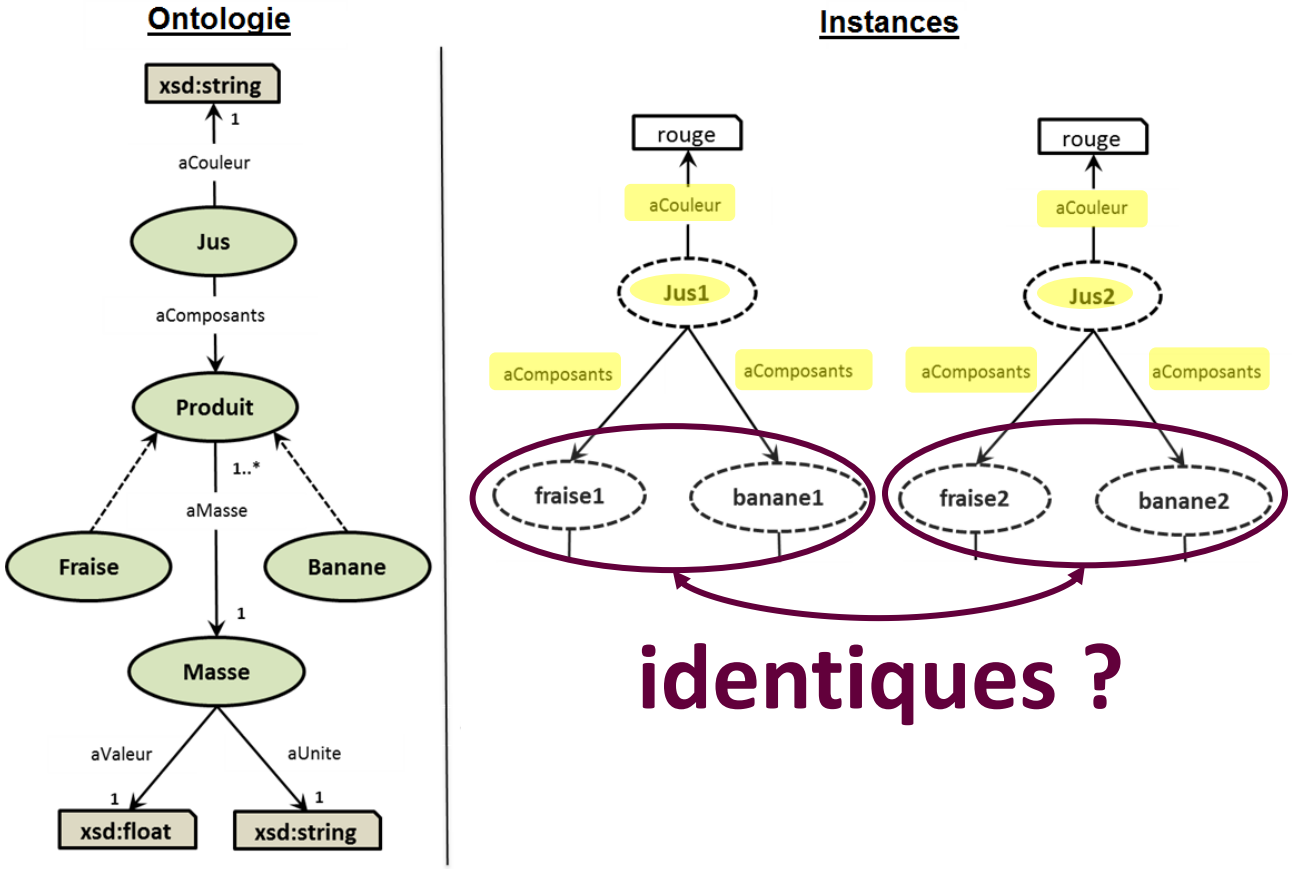
Contraintes d'élagage des contextes locaux

Liste de propriétés qui doivent être Co-Occurrentes

Contrainte Experte 3 : la valeur n'a pas de sens sans son unité et réciproquement



Identité dans un contexte local ?



c DP^c OP^c

Exemple: π_1**(Jus) =** (Jus, {rdf:Type, aCouleur}, {aComposants})

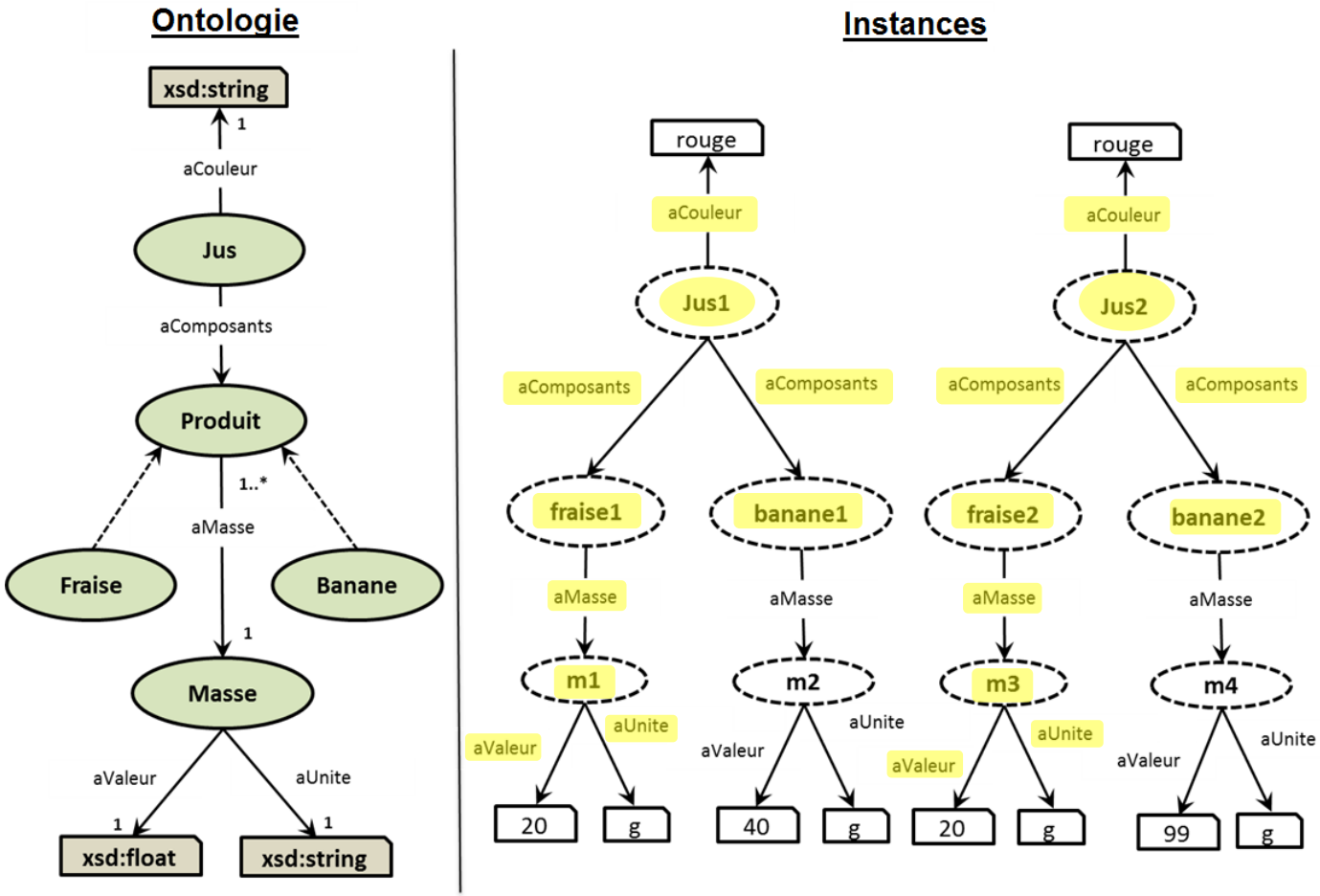
Contexte Global

$$\Pi(cbl) = \cup_{c^k \in C^G} \pi(c_k)$$

Etant donné une classe cible *cbl*, un contexte global *Π(cbl)* est un sous-graphe connexe *G* du graphe contenant *cbl*

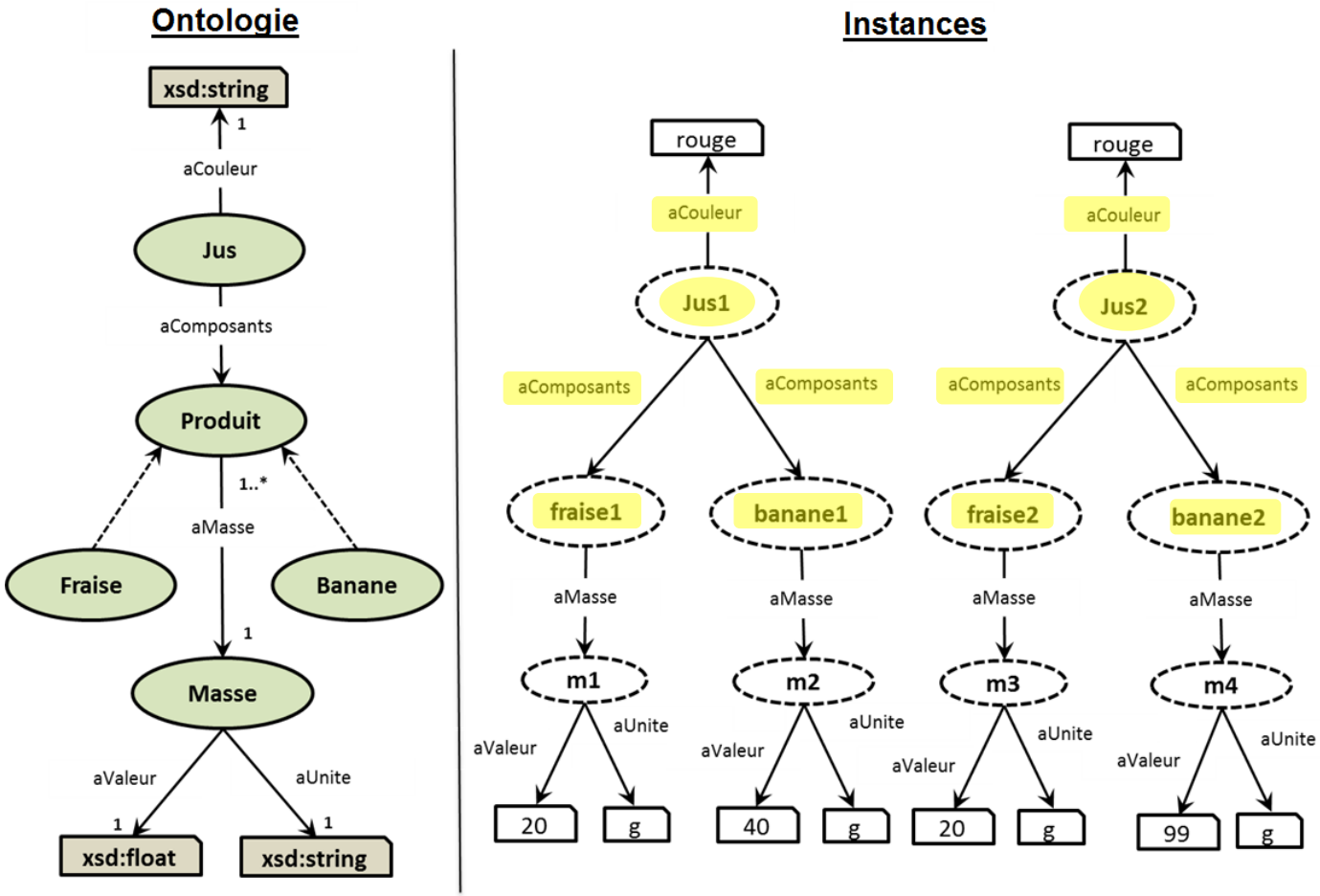
- **cbl :** une classe cible donnée de l'ontologie
- **G :** sous-graphe connexe contenant *cbl*
- $\mathbf{C_G}$ **:** ensemble des classes appartenant à *G*
- $\mathbf{c_k}$ **:** classe appartenant à C_G
- $\boldsymbol{\pi(c_k)}$ **:** contexte local de la classe c_k

Contexte Global



Exemple: Π_a(Jus) = { (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Banane, {rdf:Type}, {aComposants $^{-1}$}),
(Masse, {rdf:Type, aValeur, aUnite}, {aMasse^{-1}}) **}**

Contexte Global



Exemple: Π_c(Jus) = { (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aComposants^{-1}}),
(Banane, {rdf:Type}, {aComposants $^{-1}$}) **}**

Contexte Global

Relations d'ordre

Π_1(cbl) est plus spécifique à Π_2(cbl), noté $\Pi_1(cbl) \leq \Pi_2(cbl)$ ssi:

$\forall\ \pi_i(c) \in \Pi_2(cbl),\ \exists\ \pi_j(c) \in \Pi_1(cbl)$ tel que $\pi_j(c) \leq \pi_i(c)$

Π_c(Jus) = { (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aComposants^{-1}}),
(Banane, {rdf:Type}, {aComposants^{-1}}) **}**

Π_a(Jus) = { (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Banane, {rdf:Type}, {aComposants^{-1}}),
(Masse, {rdf:Type, aValeur, aUnite}, {aMasse^{-1}}) **}**

$$\Pi_a(Jus) \leq \Pi_c(Jus)$$

chaque contexte local dans Π_c(Jus) est moins spécifique ou égale à son contexte local dans Π_a(Jus)

Identité dans un Contexte Global

Relation d'identité contextuelle

$$\textit{identiConTo}_{<\Pi i(cbl)>}(i_1, i_2)$$

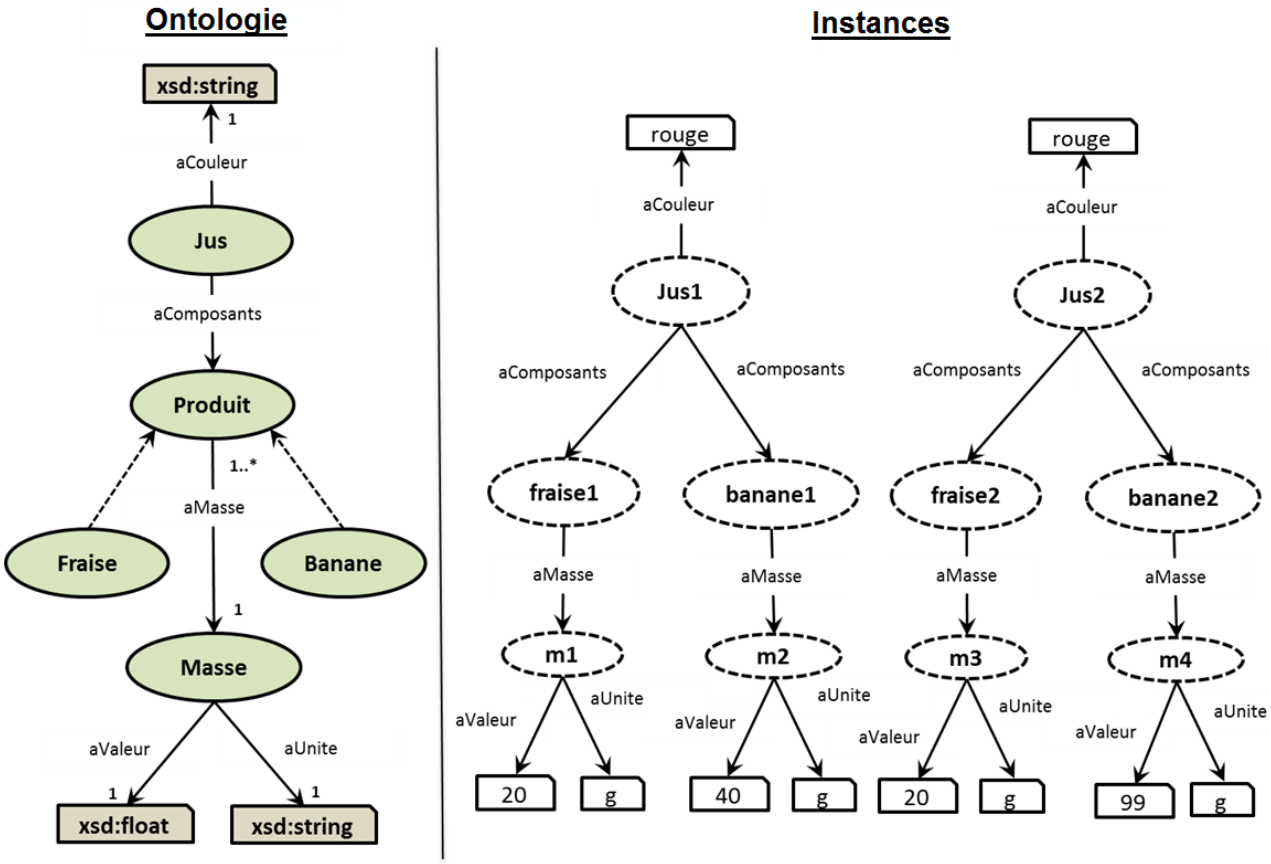
- **cbl :** une classe cible donnée de l'ontologie
- **Π_i(cbl) :** un contexte global de la classe *cbl*
- **i_1 et i_2 :** deux instances de la classe *cbl*

i_1 et i_2 sont identiques par rapport au contexte global Π_i(cbl) ssi :

les sous-graphes RDF décrivant i_1 et i_2, obtenus en utilisant la partie de l'ontologie représentée dans le contexte global, sont identiques :
- au renommage d'URI près
- et à une réécriture de valeurs littérales près (similarité)

✓ **Transitive** ✓ **Symétrique** ✓ **Réflexive**

Identité dans un Contexte Global



Π_a**(Jus) = {** (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Banane, {rdf:Type}, {aComposants^{-1}}),
(Masse, {rdf:Type, aValeur, aUnite}, {aMasse^{-1}}) **}**

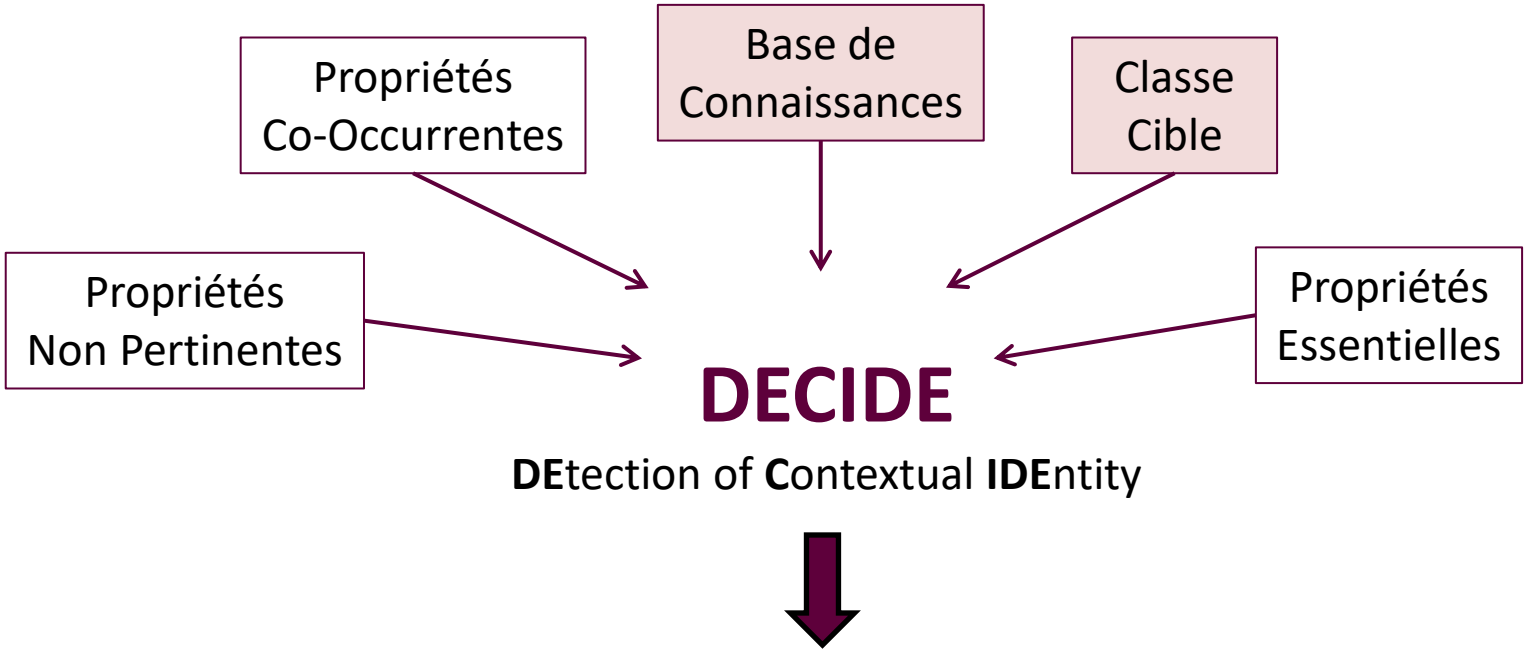
Π_b**(Jus) = {** (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Banane, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Masse, {rdf:Type, aUnite}, {aMasse^{-1}}) **}**

$identiConTo_{<\Pi a(Jus)>}$**(jus1, jus2)**

$identiConTo_{<\Pi b(Jus)>}$**(jus1, jus2)**

Détection de liens d'identités contextuels

Comment peut-on automatiquement détecter et ajouter ces liens d'identité dans une base de connaissances ?



Pour chaque couple d'individus (i_1, i_2) de la classe cible

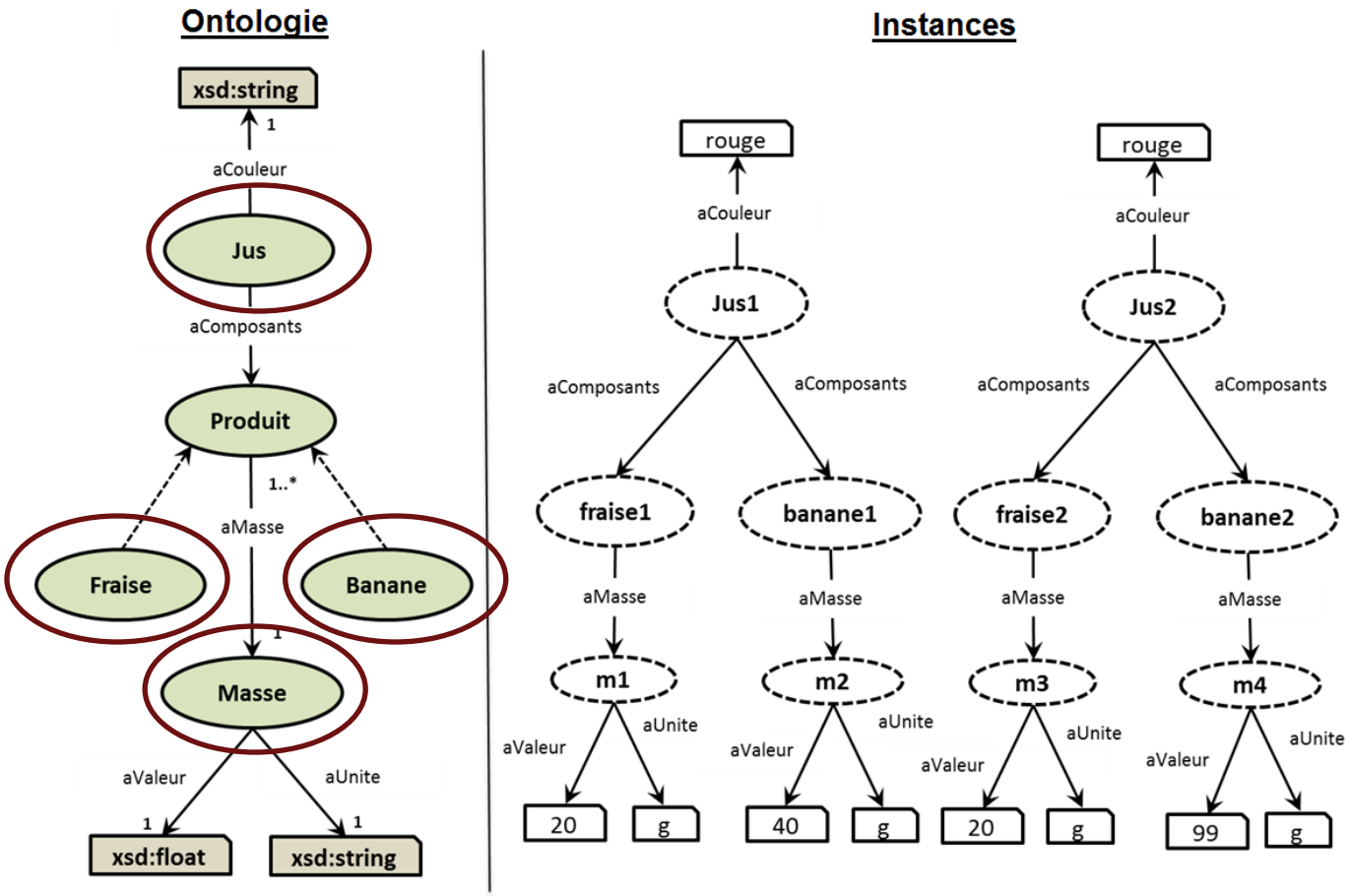
ensemble des contextes globaux les plus spécifiques dans lesquels (i_1, i_2) sont identiques

Détection de liens d'identités contextuels

DECIDE

DEtection of **C**ontextual **IDE**ntity

1. **Construction de la liste *cDep* des classes les plus générales du graphe connexe maximal de *cbl* qui comportent des instances directement typées par ces classes**

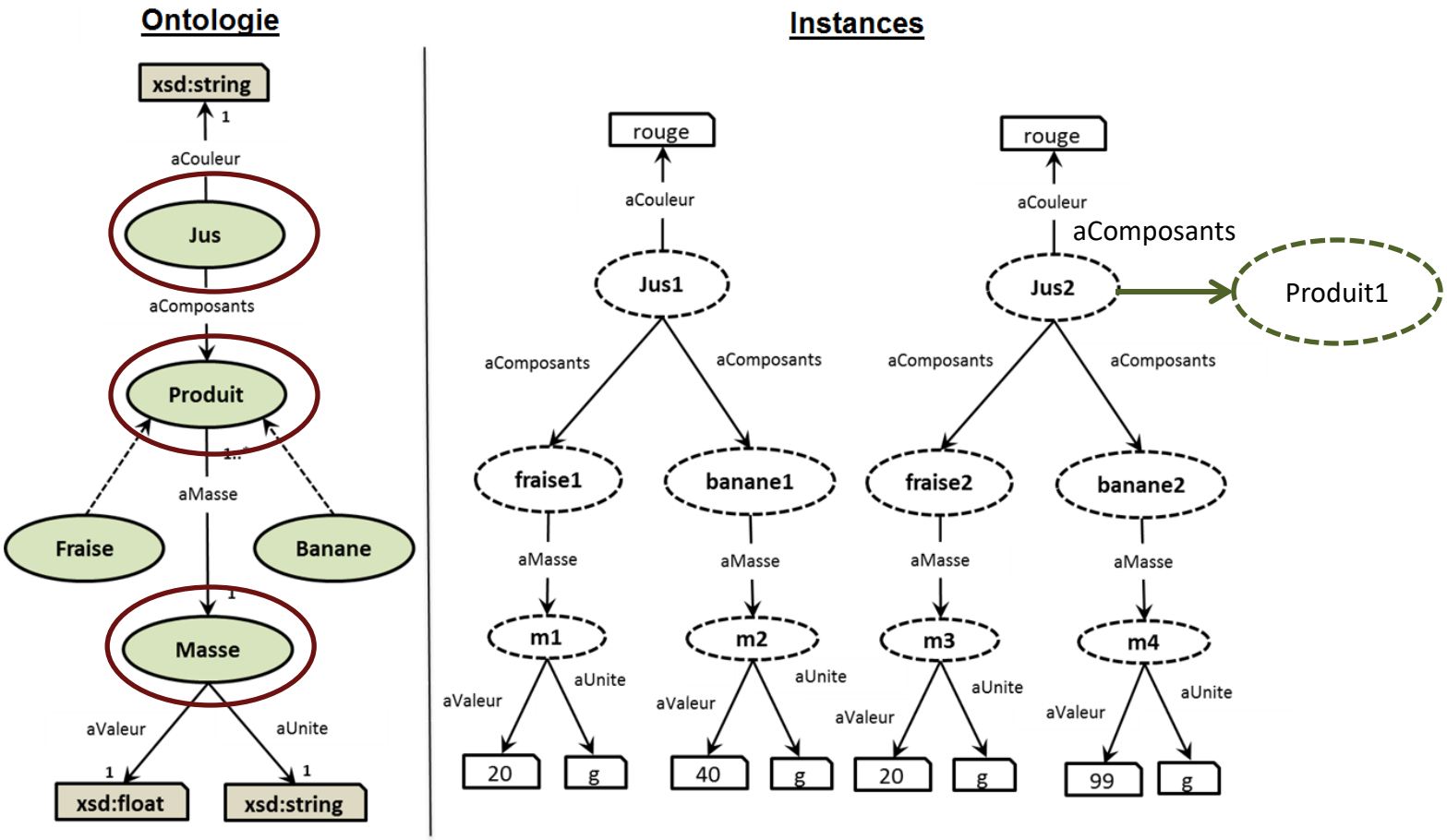


Détection de liens d'identités contextuels

DECIDE

DEtection of **C**ontextual **IDE**ntity

1. **Construction de la liste *cDep* des classes les plus générales du graphe connexe maximal de *cbl* qui comportent des instances directement typées par ces classes**



Détection de liens d'identités contextuels

DECIDE

DEtection of **C**ontextual **IDE**ntity

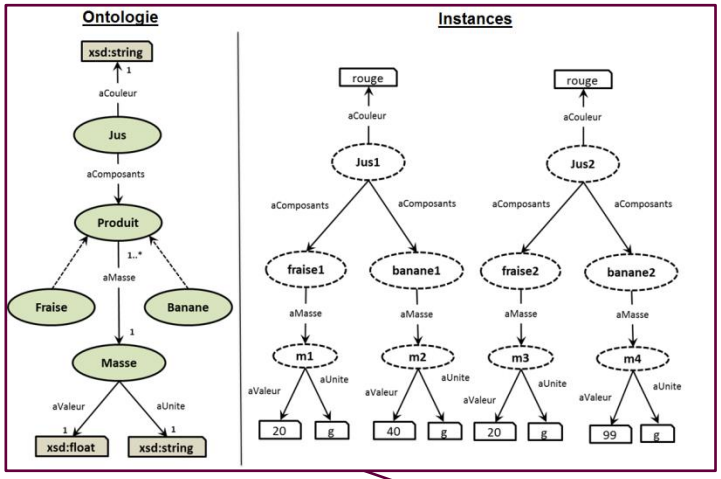
1. **Construction de la liste *cDep* des classes les plus générales du graphe connexe maximal de *cbl* qui comportent des instances directement typées par ces classes**

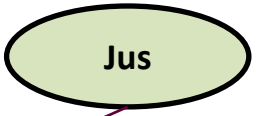
2. **Pour chaque classe *c* ∈ *cDep*, construction des treillis de contextes locaux *T(c)* pertinents** (en tenant compte des propriétés essentielles, non-pertinentes et *co-occurrentes* définies par les experts)

3. **Pour chaque couple d'instances (i_1,i_2) de *cbl*, appel de la fonction *IdentiConMax* qui calcule l'ensemble des contextes globaux les plus spécifiques**

Détection de liens d'identités contextuels

Exemple





DECIDE

DEtection of **C**ontextual **IDE**ntity

(jus1, jus2)

Π_a(Jus) = { (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Banane, {rdf:Type}, {aComposants^{-1}}),
(Masse, {rdf:Type, aValeur, aUnite}, {aMasse^{-1}}) **}**

Π_b(Jus) = { (Jus, {rdf:Type, aCouleur}, {aComposants}),
(Fraise, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Banane, {rdf:Type}, {aMasse, aComposants^{-1}}),
(Masse, {rdf:Type, aUnite}, {aMasse^{-1}}) **}**

Expérimentations

Fabrication et Stabilisation de micro-organismes

Processus de Transformation des gels laitiers

PO²
A Process and Observation Ontology

PO²
A Process and Observation Ontology
CellExtraDry

- **Classes :** ≈ 4 700
- **Individus :** ≈ 415 000
- **Faits :** ≈ 1 700 000

PO²
A Process and Observation Ontology
Carredas

- **Classes :** ≈ 5 000
- **Individus :** ≈ 42 000
- **Faits :** ≈ 237 000

Ibanescu et al. "PO² - A Process and Observation Ontology in Food Science. Application to Dairy Gels". *MTSR16*

Expérimentations

Expérimentation 1

A Process and Observation Ontology

CellExtraDry

- **Classes :** ≈ 4 700
- **Individus :** ≈ 415 000
- **Faits :** ≈ 1 700 000

Mixture

Mixture

Expérimentation 2

A Process and Observation Ontology

Carredas

- **Classes :** ≈ 5 000
- **Individus :** ≈ 42 000
- **Faits :** ≈ 237 000

DECIDE

DEtection of **C**ontextual **IDE**ntity

	CellExtraDry	Carredas
# Individus (type:Mixture)	220	36
# Paires d'individus	24 090	630
# Liens d'identité les plus spécifiques	25 189	1 035
# Liens d'identité les plus spécifiques par paire	1,04	1,64
# Classes Dépendantes (cDep)	180	102
# Contextes Globaux Différents	30	20
Temps d'exécution	102 secondes	16 secondes

Expérimentations

Exemples de Contextes Globaux détectés dans CellExtraDry

Π_9(Mixture) = { (Mixture, {rdf:type} {isComposedOf, hasForAttribute})
(Eau, {rdf:type} {Ø }) **}**

→ **Deux mixtures identiques dans Π_9(Mixture) ont de l'eau dans leurs composants (1620 / 25189 des liens d'identité les plus spécifiques)**

Π_{23}(Mixture) = { (Mixture, {rdf:type} {isComposedOf, hasForAttribute}), (Eau, {rdf:type} {hasForAttribute }), (Glucose, {rdf:type} {hasForAttribute }), (Antimousse, {rdf:type} {hasForAttribute }), (Levure, {rdf:type} {Ø}), (Cysteine, {rdf:type} {hasForAttribute }), (kh2po4, {rdf:type} {hasForAttribute }), (Masse, {rdf:type, valeur, unite} {Ø }) **}**

→ **Deux mixtures identiques dans Π_{23}(Mixture) ont les mêmes masses d'eau, glucose, antimousse, cystéine, kh2po4 et contiennent de la levures (2 / 25189 des liens d'identité les plus spécifiques)**

Conclusion et Perspectives

- **Proposition d'un nouveau lien d'identité contextuelle (identiConTo)**
  - Transitif, Symétrique, Réflexif
  - Basé sur les notions des contextes locaux et contextes globaux

- **Proposition d'un algorithme de détection des contextes globaux les plus spécifiques pour les paires d'individus d'une classe cible (DECIDE)**

- **Expérimentations sur un ensemble de données scientifiques réelles**
  - Fabrication et Stabilisation des micro-organismes
  - Processus de Transformation des gels laitiers

- **Utilisation des liens pour la prédiction à différents niveaux de confiances**

$$identiConTo_{<\Pi\alpha(Mixture)>}(x, y) \rightarrow \text{même_consommation_électrique}(x, y)$$

≈ **2000 règles** avec chaque règle ayant un certain niveau de précision et support

Détection de liens d'identité contextuels dans une base de connaissances

Merci pour votre attention…

Joe Raad, Nathalie Pernelle, Fatiha Saïs
prenom.nom@lri.fr
LRI, Université Paris-Sud
Orsay, France