Comment représenter et découvrir des liens d'identité contextuelle dans une base de connaissances : Application à des données expérimentales en sciences du vivant

Joe Raad, Nathalie Pernelle, Fatiha Saïs, Juliette Dibie-Barthelemy, Liliana Ibanescu, Stéphane Dervaux

Comment représenter et découvrir des liens d'identité contextuelle dans une base de connaissances

Application à des données expérimentales en sciences du vivant

Joe Raad [1,2], **Nathalie Pernelle** [2], **Fatiha Saïs** [2], **Juliette Dibie** [1], **Liliana Ibanescu** [1], **Stéphane Dervaux** [1]

1. *UMR MIA-PARIS, INRA, AgroParisTech, Université Paris-Saclay*
   Paris, France
   {joe.raad,juliette.dibie,liliana.ibanescu,stephane.dervaux}@agroparistech.fr
2. *LRI, Paris Sud University*
   Orsay, France
   {nathalie.pernelle,fatiha.sais}@lri.fr

RÉSUMÉ. *De nombreuses applications du web de données exploitent des liens d'identités déclarés à l'aide du constructeur* owl:sameAs. *Cependant, différentes études ont montré qu'une utilisation abusive de ces liens peut conduire à des résultats erronés ou contradictoires. Dans cet article nous proposons un nouveau lien d'identité contextuelle qui permet de représenter les contextes dans lesquels deux instances de classes sont identiques. Nous avons proposé et développé un algorithme nommé DECIDE pour détecter ces liens contextuels. Cette approche a été testée sur des données scientifiques décrivant des processus de transformations issus de projets menés par l'INRA.*

ABSTRACT. *Most of the Linked Data applications currently rely on the use of* `owl:sameAs` *for linking ontology instances. However, several studies have noticed multiple misuses of this identity link, which can lead to erroneous statements or inconsistencies. We propose in this paper a new contextual identity link, that could serve as a replacement in linking identical instances in a specified context. To detect these contextual links, we have defined an algorithm named DECIDE, which has been tested on scientific knowledge bases from several INRA projects.*

MOTS-CLÉS : *contexte, liens d'identités, base de connaissances, données scientifiques.*

KEYWORDS: *context, identity links, knowledge base, scientific data.*

DOI:10.3166/RIA.32.345-372

1. Introduction

Grâce à l'initiative du web des données, Linked Open Data (LOD) [1], de plus en plus de sources de données structurées et d'ontologies sont publiées. Le LOD comprenait 500 millions de triplets RDF (Resource Description Framework) [2] en 2007. Il comprend plus de 149 milliards de triplets RDF en 2017 [3]. Les sources de données peuvent être généralistes comme DBpedia [4], Yago [5] ou Wikidata [6], couvrir un domaine d'application spécifique comme GeoNames [7] pour la géographie ou Bio2RDF (Belleau *et al.*, 2008) pour les sciences du vivant.

Pour tirer profit de la richesse de toutes ces données et connaissances, il est important d'établir des liens sémantiques entre les instances. En particulier, les liens d'identité `owl:sameAs` sont utilisés pour exprimer que deux instances différentes réfèrent à la même entité du monde réel (e.g., même personne, même article, même gène). Ce type de lien est défini dans (Dean *et al.*, 2004) avec une sémantique très stricte : déclarer un lien `owl:sameAs` entre deux instances implique que toutes les valeurs de propriétés déclarées pour l'une peuvent aussi être déclarées pour l'autre. Aussi, si des liens `owl:sameAs` erronés sont déclarés dans une base de connaissances cela peut conduire à inférer des informations erronées voire même contradictoires.

Avec plus que 558 millions [8] `owl:sameAs` existants dans le LOD jusqu'à présent (Beek *et al.*, 2018), plusieurs études ont montré que ce lien d'identité était utilisé de manière abusive, liant dans certains cas des individus similaires, mais non identiques. Dans (Jaffri *et al.*, 2008), les auteurs ont évalué la qualité du résultat du liage de données obtenu entre des données DBLP [9] et des données de DBpedia. Pour cela, ils ont mesuré le niveau de correction des nouveaux faits inférés en exploitant la sémantique logique des liens `owl:sameAs`. En choisissant arbitrairement 49 noms parmi les 491 796 auteurs disponibles dans DBLP 2006, ils ont montré que 92 % de ces 49 auteurs ont eu des publications incorrectement affiliées. De même, les auteurs de (Halpin *et al.*, 2015) ont évalué 250 liens `owl:sameAs` parmi les 58 millions de liens présents dans OpenLink Data Explorer [10], provenant de différentes sources de données. Cette évaluation manuelle s'appuie sur la description de chacune des URI et montre qu'environ 21% des liens d'identité existants devraient être considérés comme des liens de similarité ou comme des liens "related-to".

---

1. https://www.w3.org/DesignIssues/LinkedData.html
2. https://www.w3.org/RDF/
3. http://stats.lod2.eu/
4. http://wiki.dbpedia.org
5. http://mpi-inf.mpg.de
6. https://www.wikidata.org/
7. http://www.geonames.org
8. accessibles sur http://sameas.cc
9. https://dblp.uni-trier.de/
10. http://ode.openlinksw.com/

Dans les bases de connaissances scientifiques, l'utilisation des liens d'identité pose de nombreux problèmes. En effet, les données sont souvent collectées par plusieurs scientifiques et les conditions expérimentales changent, même légèrement, d'une expérience à une autre. Il est ainsi rarement pertinent de déclarer que deux instances sont strictement identiques. De plus, la notion d'identité peut dépendre du contexte dans lequel elle est utilisée. Par exemple, dans certaines applications, le fait que deux médicaments partagent les mêmes noms suffit pour les considérer comme identiques, tandis que dans d'autres applications il est nécessaire que ces deux médicaments partagent les mêmes structures chimiques (Batchelor *et al*., 2014). De même, deux jus de fruits avec différentes quantités d'ingrédients, mais avec des proportions égales peuvent être considérées comme étant les mêmes dans un contexte gustatif, mais considérés comme différents dans le contexte d'une étude nutritionnelle et énergétique. Enfin, nos discussions avec les experts de l'INRA [11] dans le cadre de deux projets, CellExtraDry sur la stabilisation des micro-organismes et CARéDAS sur les gels laitiers, montrent qu'il est difficile de spécifier les contextes qui pourraient être d'intérêt pour leurs applications afin d'établir des relations entre les différentes expériences scientifiques menées selon différentes conditions expérimentales avec différents paramètres.

Dans cet article, qui est une version étendue du travail présenté dans (Raad, Pernelle, Saïs, 2017 ; Raad, Pernelle, Saïs, 2017), nous proposons un nouveau type de lien d'identité nommé `identiConTo`. Ce lien exprime une identité entre deux instances, qui est valide dans un contexte défini par rapport à une ontologie de domaine. Nous avons proposé et développé un algorithme nommé *DECIDE*, qui permet de détecter l'ensemble des contextes les plus spécifiques dans lesquels deux instances sont identiques. Cet algorithme peut être guidé par un ensemble de contraintes sémantiques définies par les experts du domaine. Nous avons évalué notre approche sur des données scientifiques provenant des deux projets, CellExtraDry et CARéDAS, dont les données sont décrites dans le vocabulaire de l'ontologie PO^2 (Ibanescu *et al*., 2016).

Nous présentons en section 2 les objectifs ainsi que les définitions utilisées dans notre approche. En section 3, nous décrivons notre algorithme *DECIDE* qui calcule les liens d'identité contextuelle. En section 4, nous présentons les résultats des expérimentations conduites sur des données scientifiques collectées dans le cadre de plusieurs projets INRA. Enfin, section 5 positionne notre travail par rapport à l'état de l'art.

2. Identité contextuelle

Dans cet article nous présentons une nouvelle approche pour la détection de liens d'identité dans des bases de connaissances RDF. Cette approche vise a détecter des liens d'identité qui sont valides dans des contextes pouvant être définis comme des sous-parties d'une ontologie de domaine. Dans cette section nous introduisons les notions préliminaires à notre approche et nous définissons notre problèmatique.

11. Institut National de Recherche Agronomique.

2.1. Base de connaissances

L'approche de détection de liens d'identité contextuelle s'appuie sur une base de connaissances RDF où l'ontologie est représentée en RDFS (Resource Description Framework Schema) [12] et les données en RDF.

DÉFINITION 1. — **Base de connaissances**. *Une base de connaissances $\mathcal{B}$ est définie par un couple $(\mathcal{O}, \mathcal{F})$ où :*

- *$\mathcal{O} = (\mathcal{C}, \mathcal{P}, \mathcal{A})$ représente la partie conceptuelle de la base de connaissances définie par un ensemble de classes $\mathcal{C}$, un ensemble $\mathcal{P}$ de propriétés, et un ensemble d'axiomes $\mathcal{A}$ tels que la relation de subsomption entre classes et le typage des domaines et co-domaines des propriétés. Dans cet article, nous utiliserons notamment les axiomes de subsomption avec la notation suivante : $C_2 \sqsubseteq C_1$ pour exprimer que la classe C_2 est subsumée par la classe C_1 i.e., la classe C_2 est plus spécifique que la classe C_1.*
- *$\mathcal{F} = \{(s, p, o)\}$ est une collection de triplets (i.e. faits) de la forme* (sujet, propriété, objet)*, exprimant des liens entre deux instances de classe ou entre une instance d'une classe et une valeur littérale* [13]*. On notera $\mathcal{I}^{\mathcal{C}}$ l'ensemble des instances i de $\mathcal{C}$ telles que $\exists(i, p, \_) \in \mathcal{F}$ ou $\exists(\_, p, i) \in \mathcal{F}$.*

Une ontologie RDFS peut être représentée par un graphe $\mathcal{G} = (V, E)$ où l'ensemble de sommets V représente les classes et les types de valeurs littérales (e.g. String, Date, Integer), et l'ensemble des arcs E correspond aux propriétés liant les classes entre elles ou liant les classes à des types de valeurs littérales.

2.2. Problème de détection de liens d'identité contextuelle

Un contexte est un sous-graphe de l'ontologie $\mathcal{O}$ représentant le vocabulaire pour lequel deux instances sont identiques. Il s'agit plus précisément d'un sous-ensemble de classes, de propriétés et d'axiomes de l'ontologie (domaines et co-domaines des propriétés).

Le problème de détection de liens d'identité contextuelle auquel nous nous intéressons dans cet article peut être défini comme suit : étant donnés une base de connaissances $\mathcal{B} = (\mathcal{O}, \mathcal{F})$ et une classe cible c de l'ontologie $\mathcal{O}$ choisie par l'expert, il s'agit de découvrir pour chaque paire d'instances $(i_1, i_2) \in (\mathcal{I}^c \times \mathcal{I}^c)$ les contextes les plus spécifiques pour lesquels (i_1, i_2) sont identiques.

Par exemple, dans la figure 2.2, les deux instances pr_3 et pr_4 de la classe cible $Processus$ sont identiques lorsqu'on considère dans leur description toutes les propriétés et toutes les classes de l'ontologie. En revanche, les deux instances pr_1 et pr_2 sont considérées comme étant identiques dans deux contextes plus spécifiques. Le premier contexte est celui dans lequel la description des volumes décrivant les matériels

12. https://www.w3.org/TR/rdf-schema
13. Nous ne considérons pas les *blank-nodes* dans ce travail.

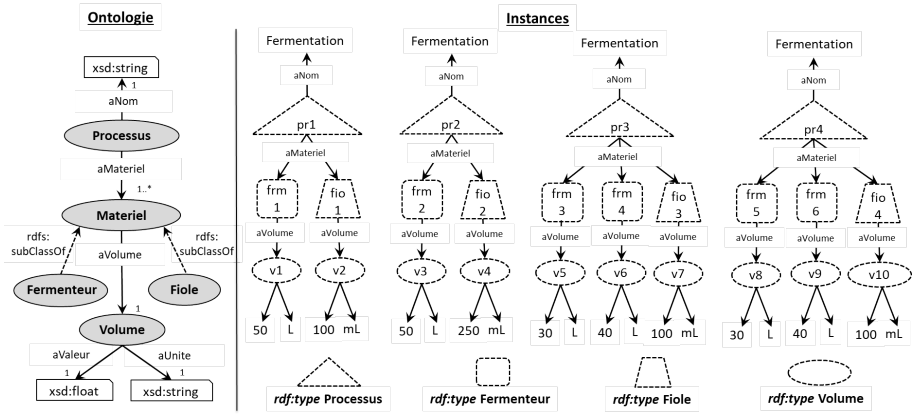


Figure 1. Un extrait d'ontologie O, quatre instances de la classe cible Processus

(i.e. fioles et fermenteurs) est restreinte à leur unité de mesure, sans considérer leur quantité.

Le second contexte est celui dans lequel seul le volume du Fermenteur est considéré (sans celui des fioles) ; et pour chaque volume leur quantité et unité de mesure.

On remarquera que les propriétés prises en compte pour comparer les instances de la classe $Volume$ ne doivent pas varier selon que l'on compare le volume du fermenteur ou celui de la fiole (i.e. dans le deuxième contexte, le volume des fioles n'est pas considéré). Il ne s'agit donc pas ici de calculer le graphe le plus spécifique partagé par deux instances de $Processus$ dans lequel les descriptions de classes pourraient varier selon les instances considérées. Pour garantir une certaine uniformité sémantique, il s'agira plutôt de calculer les contextes globaux les spécifiques de telle sorte que si une propriété p d'une classe C apparaît dans un contexte, alors elle doit être instanciée et identique pour toutes les instances de la classe C indépendamment du rôle de l'entité représentée par la classe dans les données (e.g. une personne peut être un chercheur, un directeur de laboratoire, etc.).

Afin d'améliorer l'efficacité de notre approche, nous proposons de prendre en compte certaines connaissances expertes durant le calcul des contextes globaux. Les liens d'identité contextuelle ne sont en effet pas forcément d'intérêt pour toutes les classes (e.g. instances de la classes Volume), mais pour seulement une ou plusieurs classes cibles dont les liens d'identité sont d'intérêt pour l'application considérée (e.g. Processus impliqué dans une expérience). Nous considérons ainsi des connaissances sur le fait qu'une propriété ou une classe puisse être ignorée, que deux propriétés doivent apparaître ensemble (e.g. $aUnite$, $aValue$) ou encore qu'une propriété doit nécessairement apparaître dans un contexte global.

2.3. Contextes

Un contexte global est représenté comme une sous-ontologie connexe de l'ontologie $\mathcal{O}$, composé d'un ensemble de classes et de propriétés de $\mathcal{O}$ et d'un ensemble d'axiomes de typage des propriétés par leur domaines et co-domaines dans l'ontologie.

Dans ce qui suit nous introduisons l'ensemble de classes $CDep$ pouvant être impliquées dans les contextes. Puis nous définissons formellement les contextes globaux ainsi que la relation d'identité contextuelle proposée.

Afin de limiter le nombre de classes et donc le vocabulaire dans lequel peuvent être exprimés les contextes globaux et pour simplifier leur calcul, nous avons introduit la notion de classes descriptives.

Classes descriptives. L'ensemble des classes descriptives d'un contexte global, noté $CDep$, est composé des classes les plus générales (au sens de la relation de subsomption) de l'ontologie $\mathcal{O}$ parmi les classes explicitement instanciées dans $\mathcal{F}$, i.e. dont le type n'est pas inféré. Dans la suite nous notons $directType(i, c)$ la classe c explicitée pour l'instance i dans la base de connaissances.

DÉFINITION 2. — *L'ensemble des classes $CDep$ pouvant appartenir à un contexte global est le sous-ensemble des classes directement instanciées c_i de $\mathcal{B}$ telles que :*

$$CDep = \{c_i \in \mathcal{C} \mid \nexists c_j \in \mathcal{C}\ t.q.\ \exists x, directType(x, c_j)\ et\ c_i \sqsubset c_j\}$$

EXEMPLE 3. — *Dans les graphes exemples de la figure 2.2, $CDep$ correspond à toutes les classes de l'ontologie à l'exception de la classe $Materiel$, qui n'est pas directement instanciée. Ainsi, frm_1 et fio_1 seront uniquement considérées comme étant de type $Fermenteur$ et $Fiole$ respectivement.*

Contexte global. Un contexte global est une sous-partie de l'ontologie $\mathcal{O}$. Il est composé d'un ensemble de classes et de propriétés de $\mathcal{O}$ et d'un ensemble d'axiomes de typage des propriétés par leur domaines et co-domaines dans l'ontologie.

DÉFINITION 4. — *Un contexte global est une sous-partie CG_u=(C_u, P_u, A_u) de l'ontologie $\mathcal{O}$ telle que $C_u \subseteq CDep$, $P_u \subseteq P$, et A_u est un ensemble d'axiomes de typage des domaines et co-domaines des propriétés qui sont plus spécifiques que celles décrites dans A : $\forall p \in P_u$, $domaine_u(p) \sqsubseteq domaine_{\mathcal{O}}(p)$ et co-$domaine_u(p) \sqsubseteq$ co-$domaine_{\mathcal{O}}(p)$*

EXEMPLE 5. — *Si l'on considère l'ontologie de la figure 2.2, un contexte global possible est :*

CG_1=$(C = \{Processus, Fermenteur, Fiole, Volume\}$,
$P = \{aMateriel, aVolume, aUnite\}$,
$A = \{domain(aMateriel) = Processus,$
$co\text{-}domaine(aMateriel) = Fermenteur \sqcup Fiole,$
$domain(aVolume) = Fermenteur \sqcup Fiole$, $co\text{-}domaine(aVolume) = Volume,$
$domain(aUnite) = Volume$, $co\text{-}domaine(aUnite) = xsd : string\}$)

Relation d'ordre sur les contextes globaux. Nous définissons ici la relation d'ordre sur les contextes globaux en s'appuyant sur l'inclusion d'ensembles de propriétés et de classes ainsi que sur la relation de subsomption entre classes.

DÉFINITION 6. — *Soit $CG_u = (C_u, P_u, A_u)$ et $CG_v = (C_v, P_v, A_v)$ deux contextes globaux. Le contexte CG_u est dit plus spécifique que CG_v, noté $CG_u \leq CG_v$, si et seulement si $C_v \subseteq C_u$, $P_v \subseteq P_u$, et $\forall p \in P_v$, $domaine_v(p) \sqsubseteq domaine_u(p)$ et $co\text{-}domaine_v(p) \sqsubseteq co\text{-}domaine_u(p)$.*

EXEMPLE 7. — $CG_1 \leq CG_2$, *avec* CG_2=
$(C = \{Processus, Fermenteur\}, P = \{aMateriel\},$
$A = \{domaine(aMateriel) = Processus,\ co\text{-}domaine(aMateriel) = Fermenteur\})$

Description contextuelle d'une instance dans un contexte global. Dans notre approche, deux instances sont considérées comme identiques dans un contexte global donné, lorsque toutes les informations décrites dans ce contexte sont instanciées pour les deux instances et que ces informations sont les mêmes. Nous définissons tout d'abord la notion de description contextuelle d'une instance de la classe cible.

DÉFINITION 8. — *Soient $\mathcal{F}$ un ensemble de triplets RDF, un contexte global $CG_u = (C_u, P_u, A_u)$, et une instance i. Une description contextuelle G_i de i dans CG_u est l'ensemble maximal de triplets décrivant i dans $\mathcal{F}$ tel que :*

- *G_i forme un graphe connecté contenant au moins un triplet dans lequel i est le sujet ou l'objet.*
- *$\forall\ t = (s,\ p,\ o) \in G_i$, $p \in P_u$, $type(s) \sqsubseteq domaine_u(p)$ et $type(o) \sqsubseteq co\text{-}domaine_u(p)$.*
- *$\forall\ j$ une instance de classe dans G_i, et $\forall p \in P_u$ tels que $type(j) \sqsubseteq domain_u(p)$, alors $\exists\ t_a = (j, p, k) \in G_i$, t.q. $type(k) \sqsubseteq co\text{-}domaine_u(p)$.*

EXEMPLE 9. — *La figure 2.3 présente le contexte global CG_1 et les descriptions contextuelles G_{pr_1} et G_{pr_2} de pr_1 et pr_2 respectivement dans CG_1.*

2.4. Identité de deux instances dans un contexte global

A partir de deux descriptions contextuelles de deux instances de classes, nous cherchons à définir sous quelles conditions ces instances peuvent être considérées comme identiques. Pour le cas des propriétés multi-valuées, nous considérons que les propriétés sont localement complètes : si une propriété est instanciée pour une instance alors nous supposons que toutes les valeurs de cette propriété sont déclarées pour cette instance dans la base de connaissance.

DÉFINITION 10. — *Soit CG_u un contexte global, deux instances i_1 et i_2 sont identiques dans CG_u, noté $identiConTo_{<CG_u>}(i_1, i_2)$, si et seulement si les deux graphes G_{i_1} et G_{i_2} qui représentent la description contextuelle de i_1 et i_2 respec-*

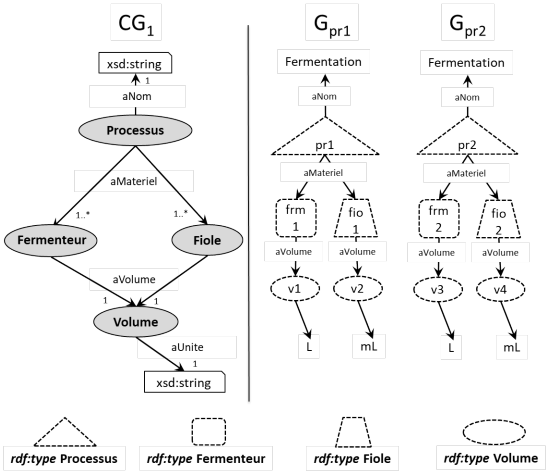


Figure 2. Le contexte global CG_1 et les descriptions contextuelles G_{pr_1} et G_{pr_2} de pr_1 et pr_2, respectivement dans CG_1

tivement, sont isomorphes à un renommage près des URI[14] *des instances de classes et en considérant l'égalité pour les littéraux et les propriétés.*

EXEMPLE 11. — *Soit le contexte CG_3 suivant :*

CG_3=($C = \{Processus, Fermenteur, Fiole, Volume\}$,
$P = \{aMateriel, aVolume, aValeur, aUnite\}$,
$A = \{domain(aMateriel) = Processus,$
$co\text{-}domaine(aMateriel) = Fermenteur \sqcup Fiole,$
$domain(aVolume) = Fermenteur$, $co\text{-}domaine(aVolume) = Volume,$
$domain(aValeur) = Volume$, $co\text{-}domaine(aValeur) = xsd : float,$
$domain(aUnite) = Volume$, $co\text{-}domaine(aUnite) = xsd : string\}$)

Les deux liens d'identité contextuelle exprimant que les instances pr_1 et pr_2 sont identiques dans les contextes globaux CG_1, défini dans l'exemple 5, et CG_3, sont notés $identiConTo_{<CG_1>}(pr_1, pr_2)$, $identiConTo_{<CG_3>}(pr_1, pr_2)$.

Ces contextes représentent aussi les contextes les plus spécifiques dans lesquels pr_1 et pr_2 sont considérés comme étant identiques.

Les liens d'identité ne sont déclarés que pour les contextes les plus spécifiques, mais d'autres liens d'identité contextuelle pourront être inférés pour des contextes plus généraux si besoin, en exploitant la relation d'ordre entre contextes globaux (voir définition 6). Soient CG_u et CG_v deux contextes globaux, tels que $CG_u \leqslant CG_v$, on a alors : $identiConTo_{<CG_u>}(i_1, i_2) \Rightarrow identiConTo_{<CG_v>}(i_1, i_2)$.

14. Uniform Resource Identifier.

3. DECIDE – Méthode de détection des liens d'identité contextuelle

L'objectif de notre approche de détection de liens d'identité contextuelle est de découvrir pour chaque paire d'instances les contextes, c'est-à-dire la ou les sous-partie(s) de l'ontologie de domaine, dans lesquelles ces instances sont identiques. Nous proposons la méthode $DECIDE$ (DEtection of Contextual IDEntity) qui prend en entrée une classe cible *cbl* de l'ontologie et vise à déterminer pour chaque couple d'instances $(i_1, i_2) \in I^{cbl} \times I^{cbl}$ de *cbl*, l'ensemble des contextes globaux les plus spécifiques dans lesquels la relation d'identité `identiConTo` est valide. Elle opère en trois étapes principales : (i) sélection de l'ensemble de classes candidates $CDep$ (voir définition 2), (ii) construction des graphes d'identité et enfin (iii) calcul des contextes globaux les plus spécifiques.

Notre algorithme de détection de liens d'identité contextuelle s'appuie sur la notion de contexte local composant les contextes globaux.

DÉFINITION 12. — **Contexte local**. *Un contexte local d'une classe c est un contexte global où l'on considère uniquement les propriétés pour lesquelles la classe c apparaît en domaine ou en co-domaine.*

Nous considérons le *contexte local sortant* capturant les propriétés pour lesquelles c est en domaine et le *contexte local entrant* capturant les propriétés pour lesquelles c est en co-domaine. On notera :

- $CL_k^{out}(c) = (C_k^{out}, P_k^{out}, A_k^{out})$, le contexte local sortant tel que $\forall p \in P_k^{out}$, $domaine(p) = c$
- $CL_k^{in}(c) = (C_k^{in}, P_k^{in}, A_k^{in})$ le contexte local entrant tel que $\forall p \in P_k^{in}$, $co\text{-}domaine(p) = c$.

3.1. Connaissances expertes

Afin de filtrer certains contextes globaux non pertinents, nous exploitons des connaissances expertes, quand elles sont disponibles, concernant l'inutilité ou la nécessité de certaines propriétés ainsi que des informations sur l'importance de la co-occurrence de certaines propriétés. Plus précisément, un expert peut spécifier trois types de contraintes :

- *Propriétés Non-Pertinentes (ensemble des contraintes PN)* : il s'agit de propriétés qui ne doivent pas être considérées dans le calcul des liens d'identité, soit parce qu'il s'agit de propriétés dont les valeurs sont non structurées (textuelles), soit parce que leurs valeurs sont trop hétérogènes, soit parce que leurs variations ne sont pas significatives dans un contexte donné (e.g. pour une étude gustative la quantité de produit testé peut ne pas être considérée, mais elle peut l'être dans une étude sur les apports énergétiques). Un expert peut déclarer une propriété p comme non-pertinente pour un certain domaine c_i (ou un co-domaine c_j) en ajoutant la contrainte $pn = (c_i, p, *)$ (resp. $pn = (*, p, c_j)$ à l'ensemble des contraintes PN. Pour déclarer une propriété non-pertinente pour tous les domaines et les co-domaines de la propriété p, on utilisera

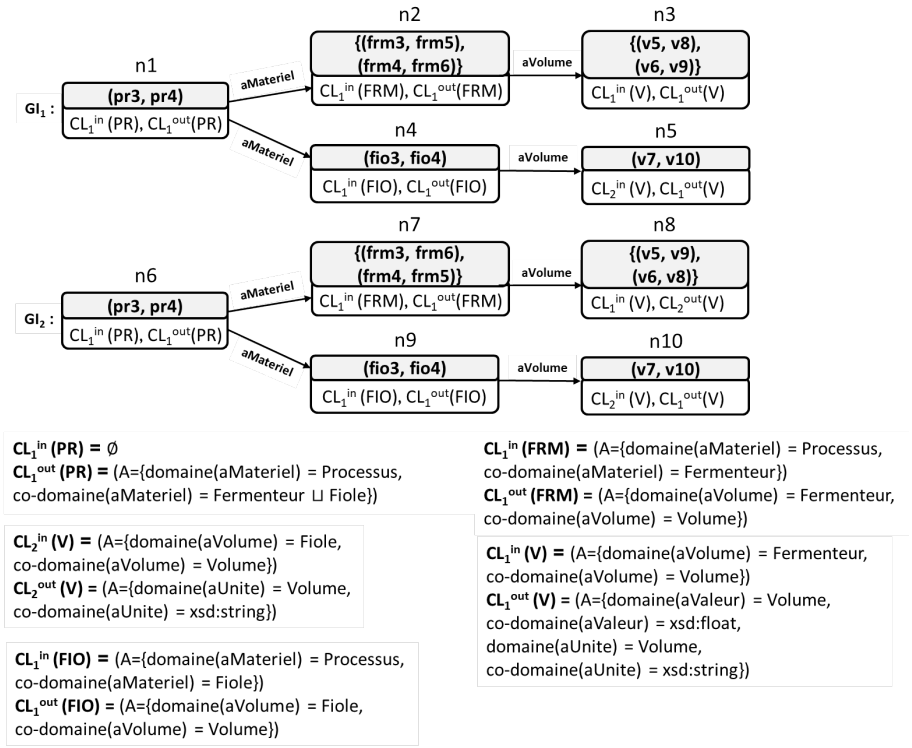


Figure 3. Deux graphes d'identité possibles pour la paire $(pr3, pr4)$. Pour des raisons de simplicité, C et P ne sont pas présentés dans les contextes locaux

la contrainte $(*, p, *)$. Dans ce cas, nous avons $p \notin P$ quel que soit le contexte.

- *Propriétés Essentielles (ensemble des contraintes PE)* : une contrainte notée pe = $(c_i, p, *)$ (ou pe = $(*, p, c_j)$) permet de déclarer que la propriété p est essentielle. En ajoutant une telle contrainte à l'ensemble des contraintes PE, on considérera seulement les contextes globaux dans lesquels apparait la propriété $p \in P$ et tels que $c_i \in domaine(p)$ (resp. $c_j \in co\text{-}domaine(p)$).
- *Propriétés Co-Occurrentes (ensemble des contraintes PC)* : une contrainte de co-occurrence $pc = \{(c_i, p_1, *), ..., (c_i, p_n, *)\}$ permet de déclarer que si la classe c_i apparaît comme domaine (ou co-domaine) d'une des propriétés p_i de la contrainte, alors elle doit apparaître comme domaine (ou co-domaine) de toutes les autres propriétés p_j, $j \in [1, n], j \neq i$, de la contrainte. Par exemple, pour déclarer que la valeur du volume n'a pas de sens sans son unité de mesure, un expert peut ajouter la contrainte $pc_1 = \{(Volume, aValeur, *), (Volume, aUnite, *)\}$.

3.2. La méthode DECIDE

Le but de la méthode $DECIDE$ (DEtection of Contextual IDEntity) est de déterminer pour chaque paire d'instances $(i_1, i_2) \in I^{cbl} \times I^{cbl}$ d'une classe cible *cbl* donnée par l'utilisateur, l'ensemble des contextes globaux les plus spécifiques dans lesquels la relation d'identité `identiConTo` est valide. $DECIDE$ prend en paramètres la base de connaissance $\mathcal{B}$, une classe cible *cbl* de l'ontologie $\mathcal{O}$, et les trois listes de contraintes spécifiées par les experts PN, PE, et PC, lorsqu'elles existent.

Algorithme 1 : DECIDE

Input :
– $\mathcal{B}$: la base de connaissance considérée
– *cbl* : classe cible
– $X(PN, PE, PC)$: contraintes expertes
Output : $PScontextes$: ensemble des contextes globaux les plus spécifiques pour chaque paire d'instances

$CDep \leftarrow getCDep(\mathcal{B})$;
$I^{cbl} \leftarrow$ liste des instances de la classe *cbl* dans $\mathcal{B}$;
foreach ($(i_1, i_2) \in I^{cbl} \times I^{cbl}$ *tel que* $i_1 \neq i_2$) **do**
  $CGset \leftarrow \emptyset$;
  $GIset \leftarrow construireGraphesIdentite(i_1, i_2, CDep, X, \mathcal{B})$;
  foreach $(GI \in GIset)$ **do**
    $n_0 \leftarrow GI.getNoeud(i_1, i_2)$;
    $N \leftarrow \emptyset$; $a \leftarrow \emptyset$; $CG \leftarrow \emptyset$; $CLset \leftarrow \emptyset$;
    $CG \leftarrow genererCG(n_0, a, CG, CLset, N, GI, X)$;
    if ($\nexists\, CG_1 \in CGset$, *tels que* $CG_1 \leq CG$) **then**
      $CGset.add(CG)$;
    if ($\exists\, CG_2 \in CGset$, *tels que* $CG \leq CG_2$) **then**
      $CGset.remove(CG_2)$;
    foreach $(CL \in CLset)$ **do**
      $CG \leftarrow \emptyset$;
      $CG.add(CL)$;
      $CG \leftarrow genererCG(n_0, a, CG, CLset, N, GI, X)$;
      if ($\nexists\, CG_1 \in CGset$, *tels que* $CG_1 \leq CG$) **then**
        $CGset.add(CG)$;
      if ($\exists\, CG_2 \in CGset$, *tels que* $CG \leq CG_2$) **then**
        $CGset.remove(CG_2)$;
  $PScontextes.add(CGSet, (i_1, i_2))$;
retourner $PScontextes$;

L'algorithme 1 détaille le déroulement de la méthode $DECIDE$ de détection des liens d'identité contextuelle. Il se déroule en trois étapes principales que nous détaillons ci-dessous.

1) Collecter l'ensemble des classes descriptives. L'ensemble $CDep$ (voir définition 2), est collecté afin d'indiquer le niveau d'abstraction (au sens de la relation de subsomption) des classes à considérer dans la construction des graphes d'identité et des contextes globaux.

Par exemple, la liste $CDep$ pour la base de connaissances de la figure 2.2 contiendrait toutes les classes de l'ontologie sauf la classe $Materiel$ qui n'est pas instanciée. $CDep$ pour la base de connaissances de la figure 30 contiendrait toutes les classes de l'ontologie.

2) Construire le(s) graphe(s) d'identité. Pour chaque paire d'instances de la classe cible, un ou plusieurs graphes d'identité sont construits. Un graphe d'identité représente un ensemble possible d'appariements des instances de classes (représentées par des URIs) pour chacune des propriétés apparaissant dans les descriptions RDF des deux instances. Un nœud n_i du graphe d'identité représente l'ensemble de paires d'instances d'une classe c dans $I^c \times I^c$ qui sont appariées et le contexte local le plus spécifique commun pour l'ensemble des paires du nœud. Chaque contexte local le plus spécifique est composé du contexte entrant $CL^{in}_{n_i}(c)$ et du contexte sortant $CL^{out}_{n_i}(c)$ dans lequel elles sont identiques (cf. définition 10). Ces contextes locaux vérifient l'ensemble des contraintes X spécifiées par les experts. La construction du graphe d'identité est dirigé par un nœud principal représentant le couple d'instances de la classe cible. L'orientation des arcs du graphe indique les domaines et les co-domaines des propriétés considérés dans les axiomes des contextes locaux, et n'influence pas le parcours du graphe.

Par exemple, les graphes d'identité correspondant au couple ($pr3$, $pr4$) de la classe cible $Processus$ décrits dans la figure 2.2 sont présentés dans la figure 3. Dans cet exemple, la propriété $aMateriel$ étant multi-valuée pour la classe $Fermenteur$, cela conduit à la construction de deux graphes d'identité : GI_1 considère l'appariement des instances $frm3$ avec $frm5$ et $frm4$ avec $frm6$, tandis que GI_2 considère l'appariement des instances $frm3$ avec $frm6$ et $frm4$ avec $frm5$. Les nœuds des graphes d'identité correspondant aux volumes des fermenteurs sont associés à des contextes locaux qui diffèrent selon le choix des appariements. Dans l'exemple de la figure 30, les graphes d'identité obtenus sont présentés dans la figure 30. Ces graphes sont construits pour les trois couples ($med1$, $med2$), ($med1$, $med3$) et ($med2$, $med3$). Par exemple, le graphe d'identité GI_2 correspondant à ($med1$, $med3$) contient trois nœuds. Les instances du premier nœud, correspondant aux instances de la classe cible ($med1$, $med3$), ne partagent aucune valeur des propriétés sortantes, par conséquent, le contexte local sortant ($CL^{out}_2(M)$) est vide. En revanche, les propriétés entrantes sont identiques dans le contexte local entrant $CL^{in}_2(M)$, puisque $med1$ apparaît en co-domaine de la propriété $vend$ autant de fois que $med3$, et $e1$ apparaît en domaine de cette propriété autant de fois que $e2$. Le couple ($e1$, $e2$) est lié au couple d'instances ($med2$, $med2$)

Algorithme 2 : Générer CG

Input :
– n : un nœud du graphe d'identité
– a_s : un axiome indiquant le type du nœud source et la propriété source
– CG : le contexte global courant
– $CLset$: ensembles des contextes locaux non utilisés
– N : ensemble des nœuds visités
– GI : le graphe d'identité
– $X(PN, PE, PC)$: contraintes expertes
Output : CG : le contexte global courant

if $(n \notin N)$ **then**
  $N.add(n)$;
  $CL_n(c) \leftarrow getOutgoingLocalContext(n)$;
  $CL_{ex}(c) \leftarrow CG.getExistingLocalContext(c)$;
  if $(CL_{ex}(c) == null$ *or* $CL_{ex}(c) == CL_n(c))$ **then**
    $CG.add(CL_n(c))$; // si le contexte n'existe pas déjà
    $E^n \leftarrow GI.getEdges(n)$;
    foreach $(e_{out} = p(n, n_{dst}) \in E^n)$ **do**
      $a \leftarrow \{domaine(p) = c, co\text{-}domaine(p) = type(n_{dst})\}$;
      $CG \leftarrow genererCG(n_{dst}, a, CG, CLset, N, GI, X)$;
    foreach $(e_{in} = p(n_{dst}, n) \in E^n)$ **do**
      $a \leftarrow \{domaine(p) = type(n_{dst}), co\text{-}domaine(p) = c\}$;
      $CG \leftarrow genererCG(n_{dst}, a, CG, CLset, N, GI, X)$;
  else
    if $(CL_n(c) \leqslant CL_{ex}(c))$ **then**
      $E^n \leftarrow GI.getEdges(n)$;
      foreach $(e_{out} = p(n, n_{dst}) \in E^n)$ **do**
        $a \leftarrow \{domaine(p) = c, co\text{-}domaine(p) = type(n_{dst})\}$;
        if $(a \in CG)$ **then**
          $CG \leftarrow genererCG(n_{dst}, a, CG, CLset, N, GI, X)$;
      foreach $(e_{in} = p(n_{dst}, n) \in E^n)$ **do**
        $a \leftarrow \{domaine(p) = type(n_{dst}), co\text{-}domaine(p) = c\}$;
        if $(a \in CG)$ **then**
          $CG \leftarrow genererCG(n_{dst}, a, CG, CLset, N, GI, X)$;
    else
      $CG.remove(a_s)$; // supprimer l'axiome source du CG
      $CG \leftarrow updateCG(X, GI)$; // vérifier connexité du CG et les contraintes expertes
    $CLset.add(CL_n(c))$; // si le contexte n'existe pas déjà
    $CLset.add(intersect(CL_n(c), CL_{ex}(c)))$; // si le contexte n'existe pas
retourner CG ;

qui sont identiques dans tous les contextes possibles, et au couple ($med1$, $med3$) qui a déjà été exploré.

3) Générer le(s) contexte(s) global(aux) le(s) plus spécifique(s). En s'appuyant sur le(s) graphe(s) d'identité construit(s), un contexte global CG est construit en utilisant l'ensemble des contextes locaux tout en garantissant la présence d'au plus un contexte local par classe dans le même contexte global. Les contextes globaux les plus spécifiques sont générés par la fonction $genererGC$ en parcourant le graphe d'identité GI et en utilisant un parcours en profondeur du graphe d'identité. Cette fonction, décrite dans l'algorithme 2, vise à ajouter pour chaque nœud du graphe d'identité, son contexte local sortant $CL_n(c)$ dans le contexte global courant CG (i.e. le contexte global le plus spécifique). Nous distinguons trois cas :

1. Si CG ne contient pas de contexte local $CL_{ex}(c)$ pour la classe c, ou si CG contient un contexte local $CL_{ex}(c)$ égal au contexte local $CL_n(c)$ du nœud n, alors $CL_n(c)$ est ajouté à CG. La fonction $genererGC$ est ensuite récursivement appelée pour chaque nœud n_{dst} dans GI pour qu'il existe un arc entre n et n_{dst}.

2. Si CG contient un contexte local $CL_{ex}(c)$ pour la classe c et $CL_n(c)$ est plus spécifique que $CL_{ex}(c)$, alors la fonction $genererGC$ est récursivement appelée pour chaque nœud destinataire n_{dst} dans GI pour qu'il existe un arc entre n et n_{dst} étiquetée par p et pour qu'il existe un axiome a dans CG, avec $a = \{domaine(p) = c$ et $type(n_{dst}) \sqsubseteq$ *co-domaine*$(p)\}$ ou $a = \{co - domaine(p) = c$ et $type(n_{dst}) \sqsubseteq domaine(p)\}$.

3. Si CG contient un contexte local $CL_{ex}(c)$ pour la classe c, et $CL_n(c)$ n'est pas plus spécifique que $CL_{ex}(c)$, alors la fonction $genererGC$ n'est pas appelée pour ce nœud du graphe. Par ailleurs, le domaine représentant le type du nœud source et le co-domaine représentant la classe c de la propriété p qui a conduit à ce nœud du graphe sont supprimés du contexte global. Finalement CG est mis à jour, vérifiant que la connexité du graphe et les contraintes expertes sont encore respectés.

Dans les cas (2) et (3), $CL_n(c)$ et le contexte local le plus spécifique qui généralise $CL_n(c)$ et $CL_{ex}(c)$ sont ajoutés à la liste $CLset$ pour garantir leur présence dans d'autres contextes globaux. Ainsi, nous obtenons plusieurs contextes globaux les plus spécifiques non comparables pour un même couple d'instances.

Par exemple, l'application de $DECIDE$ sur le couple ($pr1$, $pr2$) permet d'obtenir les deux contextes globaux CG_1 et CG_2 (voir exemple 5), représentant les contextes les plus spécifiques et non comparables, dont lesquels les deux processus $pr1$ et $pr2$ sont identiques.

$DECIDE$ est implémenté en $Java$ en utilisant le Framework $SESAME$. Il est disponible à l'adresse suivante : http://github.com/raadjoe/DECIDE_v2.

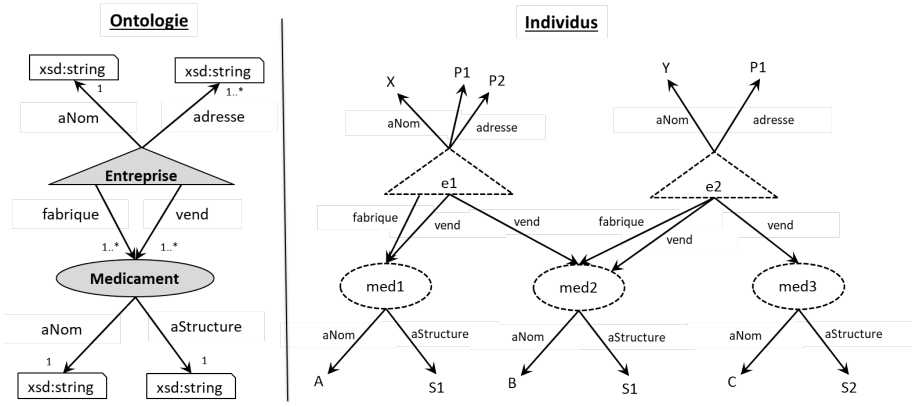
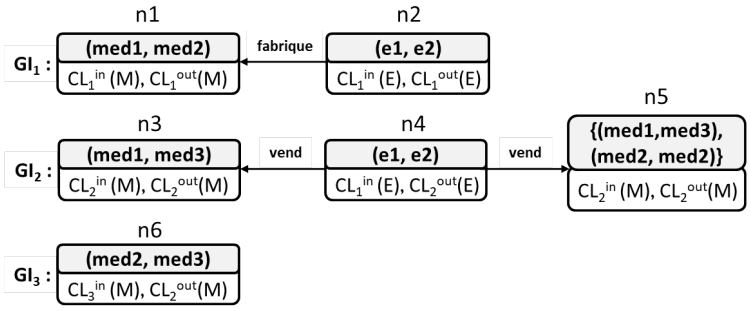


Figure 4. Un extrait d'une ontologie $\mathcal{O}$ et la description de trois instances de la classe cible Medicament



CL_1^{out} (M) = (A={domaine(aStructure) = Medicament, co-domaine(aStructure) = xsd:string})
CL_2^{out} (M) = ∅

CL_1^{in} (M) = (A={domaine(fabrique) = Entreprise, co-domaine(fabrique) = Medicament})
CL_2^{in} (M) = (A={domaine(vend) = Entreprise, co-domaine(vend) = Medicament})
CL_3^{in} (M) = ∅

CL_1^{in} (E) = ∅
CL_1^{out} (E) = (A={domaine(fabrique) = Entreprise, co-domaine(fabrique) = Medicament})
CL_2^{out} (E) = (A={domaine(vend) = Entreprise, co-domaine(vend) = Medicament})

Figure 5. Les trois graphes d'identité construits pour les instances de la classe cible $Medicament$. Pour des raisons de simplicité, C et P ne sont pas présentés dans les contextes locaux

4. Expérimentations

Avant de présenter les résultats des expérimentations conduites, nous décrivons tout d'abord la base de connaissances utilisée pour évaluer notre approche.

4.1. Description des données et de l'ontologie

Notre approche a été évaluée sur des données scientifiques relatives au domaine des gels laitiers. Ces données ont été collectées dans le cadre de plusieurs projets INRA. L'un de ces projets s'intéresse par exemple aux effets d'un changement de composition des modèles fromagers sur la mobilité, la libération et la perception des molécules déterminant les aspects gustatifs du modèle (e.g. sel, composés d'arôme). Pour modéliser ces données, nous avons utilisé la version 1.5 [15] de l'ontologie PO^2 - Process and Observation Ontology (Ibanescu *et al.*, 2016).
L'ontologie PO^2 permet de représenter des processus des transformations. Ses concepts principaux qui sont illustrés dans la figure 4.1 se répartissent en 5 parties :

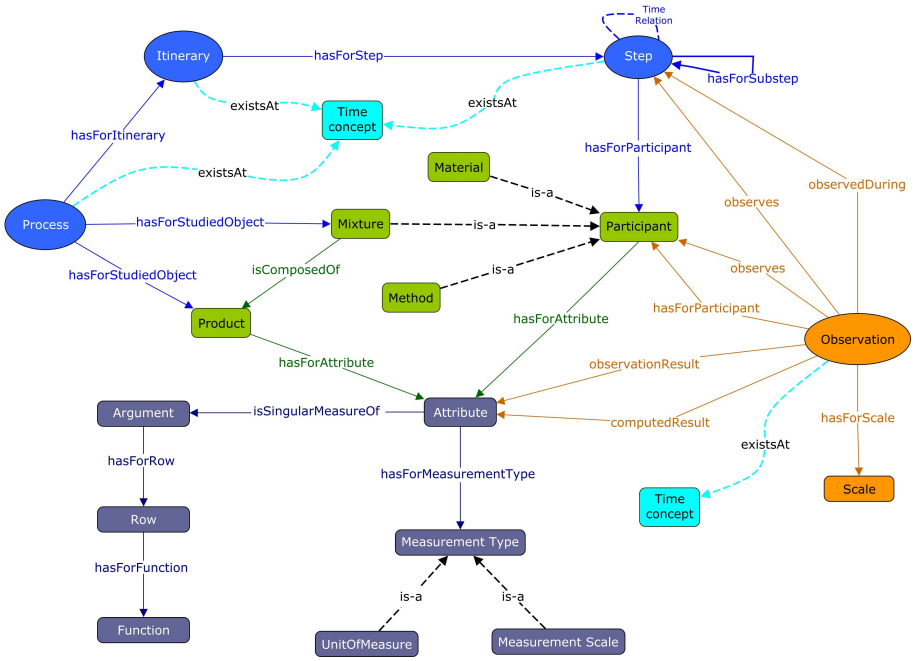


Figure 6. Les concepts principaux de l'ontologie PO^2

– *Processus* : cette partie de l'ontologie concerne les processus, les itinéraires possibles et les étapes composant chaque itinéraire. Par exemple, un processus de

15. http://agroportal.lirmm.fr/ontologies/PO2

fabrication de fromage peut être conduit suivant différents itinéraires, et chaque itinéraire correspond à une succession d'étapes de transformations (e.g. égouttage, salage).
– *Participants* : cette partie représente les participants à un processus, i.e. les mixtures qui se transforment au cours des étapes d'un processus, les matériels et les méthodes qui sont utilisés pour conduire un processus. Cette partie a été enrichie semi-automatiquement par des termes du vocabulaire du domaine des gels laitiers en s'appuyant notament sur une partie du thésaurus d'Agrovoc [16].
– *Observations* : cette partie décrit les observations qui peuvent être réalisées tout au long des étapes d'un processus de transformation. Une observation peut être réalisée sur une étape ou sur l'un des participants de cette étape. Elle peut également être faite à différentes échelles (e.g. au niveau cellulaire, moléculaire).
– *Attributs* : cette partie décrit les attributs des participants et les mesures des observations.

– *Aspect Temporel* : cette partie décrit les relations temporelles du processus et des observations. Elle a été construite à partir de la Time ontology [17].

La base de connaissances sur laquelle nous avons testé notre approche comporte 1 491 255 triplets. Ces données instancient 950 classes et 125 propriétés de PO^2, propriétés parmi lesquelles 64 sont des propriétés d'objets. Ces données décrivent 284 processus de transformation issus de 11 projets différents, où chaque processus est généralement composé de quatre à cinq étapes. Durant ces processus de transformations, 2 603 observations ont été réalisées qui portent sur 509 attributs différents (e.g. température, viscosité) mesurés à 6 échelles différentes. La base de connaissance, qui est régulièrement mise à jour est accessible sur http://sonorus.agroparistech.fr:7200/sparql.

4.2. Détection de liens d'identité contextuelle

L'objectif de ces expérimentations est de détecter pour chaque couple de mixtures, et chaque couple d'étapes, les contextes dans lesquels ils sont identiques. Pour comparer deux instances, nous n'avons pas pris en compte les observations. D'une part, elles sont peu pertinentes pour comparer deux étapes ou mixtures d'un processus et d'autre part un grand nombre d'observations sont manquantes (e.g. la $Temperature$ n'est pas nécessairement mesurée pour chaque mixture et chaque étape d'expérimentation).

16. http://aims.fao.org/fr/agrovoc
17. https://www.w3.org/TR/owl-time/

Nombre de liens et de contextes détectés. Le tableau 1 présente les résultats de $DECIDE$ quand la classe $Mixture$ et la classe $Step$ sont considérées comme classe cible. Il existe dans ce jeu de données, 1 187 instances de type $Mixture$, et 581 de type $Step$, formant respectivement 703 891 et 168 490 couples d'instances à considérer. Parmi les 950 classes de l'ontologie, seules les 784 classes instanciées les plus générales sont utilisées pour expliciter les contextes (définition 2). En moyenne, un seul graphe d'identité par couple d'instances est généré pour détecter les liens d'identité contextuelle. En effet, dans ce jeu de données, peu de propriétés multivaluées ayant des valeurs de même type peuvent amener à construire différents graphes d'identité (comme dans le cas de la propriété liant les processus $pr3$ et $pr4$ aux fermenteurs présenté en figure 2.2). Un graphe d'identité est en moyenne composé de 5,26 nœuds pour la classe $Mixture$, et de 8,25 nœuds pour la classe $Step$. Ces graphes d'identité permettent de générer 1 279 376 liens d'identité valides dans 2 232 contextes globaux différents pour les mixtures et 348 017 liens valides dans 718 contextes pour les étapes. Ces résultats montrent que deux instances de ces classes cibles peuvent être identiques dans plus d'un contexte global plus spécifique (1,81 et 2,06 respectivement).

Tableau 1. Résultats de $DECIDE$ pour les deux classes cibles $Mixture$ et $Etape$

	Mixture	***Etape***
# Instances de la Classe Cible	1 187	581
# Couples Possibles	703 891	168 490
# Classes Sélectionnées	784	784
# Graphes d'Identité par Couple	1,003	1,785
# Nœuds par Graphe d'identité	5,26	8,25
# Contextes Globaux Différents	2 232	718
# Liens d'Identité	1 279 376	348 017
# Liens d'Identité par Couple	1,81	2,06

Ajout de Contraintes. Nous avons également étudié comment l'ajout de contraintes expertes pouvait impacter les résultats de l'approche. Pour cela, nous avons utilisé un échantillon de données représentant un seul projet des 11 existants, et contenant 153 couples de la classe cible $Mixture$. Sans contrainte experte, $DECIDE$ détecte 502 liens d'identité contextuelle valides dans 37 contextes globaux différents (3,28 liens par couple). Une première contrainte experte impose que la valeur d'un attribut (instance de la classe $Attribute$) ne peut exister sans son unité de mesure. En ajoutant cette contrainte de cooccurrence, $DECIDE$ découvre alors 377 liens d'identité valides dans 24 contextes globaux différents (2,46 liens par couple). Les experts nous ont également indiqué que si la présence d'eau dans les mixtures est considérée, il est nécessaire que la quantité d'eau soit la même pour que le contexte d'identité soit significatif. En ajoutant cette deuxième contrainte de cooccurrence $pc_2 = \{(Mixture, isComposedOf, Water), (Water, hasAttribute, Weight)\}$, le nombre de contextes globaux diminue de 37 à 35. Ces expérimentations indiquent que l'ajout de contraintes peut permettre de diminuer sensiblement le nombre de contextes et donc le nombre de liens d'identité contextuelle détectés. Bien sûr, toutes les contraintes n'ont pas le même impact sur les résultats. Ainsi, si l'expert indiquait

que la relation $isComposedOf$ reliant une $Mixture$ aux instances de la classe $Product$ est une propriété non pertinente, il n'existerait plus que 4 contextes globaux différents et le nombre de liens d'identité contextuelle serait seulement de 198 (1,29 liens par couple). En effet, la suppression d'une telle propriété diminue fortement la taille des graphes décrivant les instances à comparer.

Représentation des liens. Un contexte global est représenté par un graphe nommé (Carroll *et al*., 2005). Chaque graphe nommé contient les axiomes de l'ontologie ainsi que les liens d'identité valides dans ce contexte. Parmi les contextes globaux détectés, le nombre d'axiomes varie entre 2 axiomes dans le contexte (i.e. un seul domaine et un seul co-domaine) et 88 axiomes, selon la spécificité du contexte global. Un lien d'identité `identiConTo` dans un contexte global entre deux instances i_1, i_2 d'une classe cible indique que ce contexte représente le contexte ou un des contextes les plus spécifiques dans lesquels i_1 et i_2 sont identiques (voir définition 10) et chaque lien d'identité contextuelle est symétrique, transitif et réflexif dans ce contexte. La relation d'ordre entre les contextes globaux est représentée, dans le graphe par défaut, par une relation transitive $moreSpecificThan$. Les sorties de l'algorithme $DECIDE$ sur les exemples des Figure 2.2 et 30 sont présentées à l'adresse suivante : https://github.com/raadjoe/DECIDE_v2/tree/master/Example.

4.3. Utilisation des liens d'identité contextuelle pour la découverte de règles

Le but de cette expérimentation est d'évaluer si les liens d'identité contextuelle peuvent être utilisés pour découvrir des règles. Plus précisément, et puisque nous n'avons pas considéré les observations dans les contextes d'identité, nous cherchons à déterminer quelle est la probabilité que deux expérimentations identiques dans un certain contexte, aient des observations similaires. Il serait alors éventuellement possible de prédire, avec un certain degré de confiance, des mesures non observées dans une expérimentation.

Selon le principe d'"Indiscernabilité des Identiques" de Leibniz (Forrest, 2016), une identité réelle entre deux objets (e.g. mixtures), indique que chaque propriété (e.g. une mesure observée) déclarée pour l'un des objets est forcément déclarée pour l'autre : $x = y \wedge p(x, z) \rightarrow p(y, z)$ avec $p \in P$. Dans cette tâche, nous cherchons à découvrir pour chaque contexte CG_i, l'ensemble Ψ des propriétés $\{p_1, ..., p_n\}$ (ou chemins de propriétés), telles que $identiConTo_{<CG_i>}(x, y) \wedge p(x, z_1) \rightarrow p(y, z_2)$ avec $z_1 \simeq z_2$ et $\Psi \wedge (P^{GC_i}) = \emptyset$. De telles règles seront notées plus simplement: r = $identiConTo_{<GC_i>}(x, y) \rightarrow$ same(m), où m représente une certaine mesure (e.g. la mesure du pH). Puisque les liens d'identité contextuelle sont seulement déclarés pour les contextes les plus spécifiques, nous avons exploité la relation d'ordre des contextes globaux (voir définition 6) pour obtenir l'ensemble complet des liens d'identité pour chaque contexte.

Pour évaluer la qualité d'une règle r, nous avons calculé :

- **le taux d'erreur :** pour chaque couple identique dans CG_i pour lequel une mesure m est observée, nous calculons un taux d'erreur er. Le taux d'erreur pour une mesure m entre deux instances x et y est calculé de la manière suivante : $er_m(x,\ y) = \frac{|m(x)-m(y)|\times 100}{|m(max)-m(min)|}$ où $m(max)$ et $m(min)$ représentent la valeur maximale (resp. minimale) prise pour la mesure m dans tout le jeu de données. Le taux d'erreur d'une règle pour un contexte global CG_i est la moyenne des taux d'erreur de chaque paire déclarée comme identique dans ce contexte.

- **le support :** il représente le nombre de couples identiques dans CG_i pour lesquels la mesure m a été observée.

Tableau 2. Evaluation de 20 règles par les experts

Impossible	**Peu Probable**	**Je ne sais pas**	**Pourquoi pas**	**Très Plausible**
3	5	4	5	3

En nous appuyant sur les résultats de $DECIDE$ sur la classe cible $Mixture$, nous avons généré 38 844 règles. Le nombre de règles varie entre une seule règle et 313 règles par contexte. En moyenne, le support d'une règle varie de 1 (e.g. un seul couple ayant en commun la mesure d'observation "Amer" dans un contexte) à 15 075. Le taux d'erreur de ces règles varie de 0 à 100 % et 1 005 règles ont un taux d'erreur inférieur à 1 %. En moyenne, le taux d'erreur d'une règle est de 34,86 %. En moyenne le taux d'erreur d'une règle diminue de 12 % quand un contexte global est remplacé par un contexte plus spécifique. Ce qui montre que plus le contexte est spécifique, plus les règles découvertes sont précises. Ainsi, les liens d'identité contextuelle pourraient être exploités pour prédire des observations manquantes avec différents niveaux de confiance.

Nous avons présenté aux experts les 20 meilleures règles, en s'appuyant sur le taux d'erreur et le support. Plus précisément, nous avons choisi les règles les plus simples à comprendre par les experts (i.e. comportant peu d'axiomes) telles que le taux d'erreur est inférieur à 15 % et ayant le plus grand support. La plausibilité des 20 règles données aux experts a été évaluée en utilisant une échelle de 5 appréciations : "impossible", "peu probable", "je ne sais pas", "pourquoi pas", et "très plausible". Le tableau 2, qui présente l'évaluation des experts, montre que parmi ces 20 règles, 3 sont très plausibles. Le détail de ces 3 règles est présenté dans le tableau 3. Il s'agit de règles pour lesquelles l'expert est sûr de l'impact des attributs du contexte sur la valeur de la mesure observée. Par exemple, l'expert a jugé qu'il était très plausible que deux mixtures ayant la même teneur en acide citrique, aient la même valeur observée de pH (première règle). Nous avons pu fournir aux experts 14 règles qui pourraient faire l'objet de plus d'études. Il s'agit des règles ayant été évaluées comme "plausible", "je ne sais pas", et "peu probable". Par exemple, l'expert a jugé qu'il était

possible que lorsque deux mixtures partagent la même teneur en eau, elles partageront aussi la même mesure observée de viscosité (règle jugée comme plausible). En revanche, 3 des règles fournies lui apparaissent comme impossibles : l'expert est sûr qu'il n'existe aucune dépendance entre le contexte d'identité et la mesure observée.

Tableau 3. Le taux d'erreur et le support des règles considérées les plus plausibles

Règle	***Taux d'erreur***	***Support***
$identiConTo_{<GC_1>}(x, y)$ $\rightarrow$ same(pH)	6.19 %	57
$identiConTo_{<GC_3>}(x, y)$ $\rightarrow$ same(Dureté)	1.86 %	66
$identiConTo_{<GC_2>}(x, y)$ $\rightarrow$ same(Friabilité)	4.52 %	647

Nous avons également utilisé les liens d'identité contextuelle et les règles pour répondre à des questions de compétences intéressant les experts. Par exemple, les experts se demandent s'il existe une dépendance entre la teneur en lipide des mixtures et les notes de rhéologie observées (trois types d'attributs). Pour répondre à cette question, nous avons sélectionné tous les contextes globaux les moins spécifiques $GCset$ contenant la propriété "isComposedOf" ayant comme domaine la classe "Mixture", et comme co-domaine la classe "Lipide", et contenant la propriété "hasWeight" ayant comme domaine la classe "Lipide", et comme co-domaine la classe "Teneur". Donc pour répondre à cette question, nous avons pu fournir aux experts la moyenne des taux d'erreur et des supports de chaque règle de type : $identiConTo_{<GC_i>}(x, y) \rightarrow$ same($note\_rheologie$), avec $GC_i \in GCset$. Les taux d'erreurs obtenus pour ces trois types d'attributs varient entre 5,19 % et 13,85 %, indiquant en effet une dépendance entre la teneur en lipide et les notes de rhéologie.

4.4. Discussion

Notre collaboration avec les experts du domaine et les résultats de nos expérimentations conduites sur ce jeu de données scientifique nous ont montré que :

- l'utilisation d'un lien d'identité stricte comme le lien *owl:sameAs* n'est pas pertinent pour les données scientifiques, puisque les conditions expérimentales et les paramètres étudiés changent, même légèrement, d'une expérience à l'autre.
- les experts du domaine peuvent difficilement expliciter les contextes dans lesquels ils considéreraient que deux objets sont identiques, ces contextes variant en fonction des objectifs de l'étude. En revanche, les experts spécifient plus facilement des contraintes qui peuvent être utilisées pour limiter la génération des contextes d'identité non-pertinents. Quand l'ontologie est complexe, la spécification de ces contraintes peut faire partie d'un processus itératif, dans lequel les experts ajoutent des contraintes au fur et à mesure en s'appuyant sur la sémantique des contextes d'identité et les règles générés précédemment.

- les liens d'identité contextuelle permettent de stocker les similarités des instances et faciliter leur interrogation.
- les liens d'identité contextuelle générés peuvent être utilisés pour la prédiction de certaines mesures d'observations manquantes. Comme les règles détectées dans les contextes les plus spécifiques ont un meilleur taux d'erreur que celles détectées dans des contextes moins spécifiques, la spécificité d'un contexte peut servir comme indicateur de confiance d'une règle.
- la pertinence d'un contexte d'identité peut varier selon les observations et l'étude. Par exemple, l'identité de la composition des mixtures est nécessaire dans les études où l'on s'intéresse à l'acidité de la mixture, tandis que l'identité des étapes dans lesquelles les mixtures apparaissent est nécessaire quand on s'intéresse à l'étude de l'impact environnemental du processus.

5. Travaux connexes

Dans cette section, nous présentons dans un premier temps les approches de liage de données permettant de détecter que deux descriptions différentes réfèrent à la même entité du monde réel (e.g. la même personne, le même lieu, le même gène). Nous présentons ensuite les travaux s'intéressant à la définition de relations sémantiques exprimant différents niveaux d'identité (plus ou moins stricte) et/ou de similarité entre entités. Nous nous intéressons ensuite aux langages de représentation de méta-données qui peuvent permettre de représenter des contextes et de les associer à des triplets. Enfin, nous présentons les approches existantes de détection des liens d'identité contextuelle.

Liage de données. Avec l'initiative du Linked Open Data (LOD) proposée par Tim-Berners Lee en 2006, un fort engouement a été constaté pour le développement d'approches de liage de données RDF, dans le domaine du web sémantique (voir Ferrara *et al.*, (2011) pour un état de l'art). Les approches de liage de données développées peuvent être classées selon trois critères.

Premièrement, les approches peuvent être qualifiées de (semi-)supervisées (Dong *et al.*, 2005 ; Hu *et al.*, 2011) ou non-supervisées (Nikolov *et al.*, 2012 ; Saïs *et al.*, 2009 ; Al-Bakri *et al.*, 2016 ; 2015), selon que celles-ci utilisent ou non un ensemble de données étiquetées pour apprendre certains paramètres (e.g., poids sur les propriétés, seuils de similarité) et/ou des fonctions (e.g., fonctions d'agrégation, mesures de similarité).

Deuxièmement, les approches peuvent être qualifiées d'approches non-collectives (Volz *et al.*, 2009) ou collectives (Dong *et al.*, 2005 ; Saïs *et al.*, 2009 ; Al-Bakri *et al.*, 2016 ; 2015), selon que celles-ci exploitent ou non les relations (i.e. les propriétés de type *owl*:*ObjectProperty*) ou la sémantique du `owl:sameAs` (i.e. transitivité), pour propager des décisions de liage de données (e.g., si deux entités réfèrent au même musée alors les villes dans lesquelles ces musées sont localisés sont également identiques).

Enfin, des approches dites informées (Saïs *et al.*, 2009 ; Al-Bakri *et al.*, 2016 ; 2015) ou non-informées selon que les méthodes considèrent ou non des connaissances déclarées dans une ontologie, ou définies par un expert du domaine (e.g. clés, fonctionnalité sur les propriétés, règles de liage).

L'approche que nous présentons dans cet article n'est pas collective : elle exploite tout le graphe RDF décrivant les instances mais un lien d'identité contextuelle généré pour un couple d'instances n'influence pas la génération de liens d'identité pour d'autres couples. Elle est non supervisée dans le sens où elle ne nécessite aucun ensemble de données étiquetées. Enfin, elle est informée car elle peut prendre en compte des contraintes spécifiques au domaine qui permettent d'éliminer des contextes d'identité non pertinents pour l'expert.

Représentation de la relation d'identité ou de similarité. Il existe quelques propositions pour la représentation de liens d'identité faible tels que les prédicats SKOS (Miles, Bechhofer, 2009) `skos:exactMatch`, `skos:closeMatch`, `skos:broadMatch` et `skos:narrowMatch`. Par exemple, le prédicat `skos:closeMatch` indique que deux concepts sont assez similaires pour être utilisés de manière interchangeable dans certaines applications. Cependant, ces contextes d'applications ne sont pas spécifiés et ces prédicats ne peuvent être utilisés que pour des URI dont le type est un concept SKOS, ce qui limite les cas d'utilisation possibles dans le LOD.

(Halpin *et al.*, 2010) propose l'Ontologie de Similarité (SO) dans laquelle 13 relations d'identités dont prédicats SKOS, `rdfs:seeAlso` et `owl:sameAs`, prédicats qui sont hiérarchisées par la relation `rdfs:subPropertyOf`. Les prédicats préfixés par le mot `claims`, tels que `so:similar` et `so:claimsIdentical`, expriment une relation d'identité ou de similarité subjective, dont la véracité dépend du contexte et/ou de l'interprétation d'un expert humain. Dans cette hiérarchie, chaque propriété est caractérisée par les propriétés de réflexivité, de transitivité et de symétrie. Cette formalisation ne permet cependant pas d'expliciter les contextes dans lesquels une relation d'identité serait valide.

D'autres vocabulaires ont introduit des liens d'identités alternatifs. Dans le vocabulaire UMBEL [18], le prédicat `umbel:isLike` définit une relation associative entre des individus similaires, qui peuvent être identiques ou non. De même, vocab.org [19] a introduit le prédicat `similarTo` pour lier deux individus qui ne sont pas `owl:sameAs`, mais similaires dans une certaine mesure. (Melo, 2013) propose d'utiliser trois prédicats : le prédicat `lvont:strictlySameAs` qui est équivalent à `owl:sameAs` et introduit dans le but de distinguer les liens correspondant réellement à une identité stricte, les prédicats `lvont:nearlySameAs` et

18. http://umbel.org
19. http://vocab.org

`lvont:someWhatSameAs` qui expriment des relations de similarité [20]. Cependant, la sémantique de ces deux derniers prédicats reste volontairement vague.

Aucun des prédicats proposés ne permet de spécifier les contextes dans lesquels un lien d'identité est valide contrairement à ce que nous proposons dans cet article.

Représentation d'un contexte. Le prédicat `owl:sameAs` représente une identité entre deux instances dans un contexte implicite correspondant à toutes les propriétés qui peuvent être décrites pour l'une ou l'autre des deux instances liées. Pour expliciter les contextes associés aux liens d'identité il est important de disposer d'un langage permettant d'associer des méta-données aux triplets RDF représentant un lien d'identité contextuelle. Il est possible d'utiliser un mécanisme de réification [21] qui utilise une nouvelle ressource de type `rdf:Statement` à laquelle des métadonnées (un contexte) peuvent être associées. Cependant, la réification génère un nombre important de faits RDF et ne permet pas de raisonner sur les données réifiées. Il est également possible d'utiliser des relations n-aires [22], qui permettent de représenter une propriété comme une classe. Dans (Nguyen *et al.*, 2014), les auteurs proposent une solution moins coûteuse au niveau du nombre de triplets ajoutés pour représenter un contexte, par l'utilisation de propriétés dites singletons qui représentent une relation entre deux entités dans un certain contexte. Une propriété singleton associée à des métadonnées est reliée à la propriété plus générique correspondante par la relation `rdf:singletonPropertyOf`. Par exemple, la propriété singleton `estMariéà` pour laquelle on peut spécifier le contexte (e.g. la provenance, la date, etc.) est `rdf:singletonPropertyOf` de la propriété générique `estMariéà`. Finalement, il est possible d'utiliser les graphes nommés [23] pour associer des métadonnées à un ensemble de triplets.

Dans l'approche que nous présentons dans cet article, nous utilisons des graphes nommés pour représenter les contextes d'identité. Un graphe nommé permet d'associer un contexte à un ensemble de liens d'identité contextuelle, tout en minimisant le nombre de triplets créés, contrairement à un processus de réification (Dodds, Davis, 2012).

Découverte de liens sémantiques contextuels. Dans (Beek *et al.*, 2016), les auteurs proposent une approche permettant de calculer et de représenter tous les contextes dans lesquels un lien d'identité est valide. Un contexte est représenté par un sous-ensemble de propriétés pour lesquelles deux instances sont identiques (i.e. ont les mêmes valeurs). Ces ensemble de propriétés sont hiérarchisés dans un treillis de contextes en utilisant une relation d'inclusion. Néanmoins, cette approche ne considère que les propriétés décrivant une instance localement (chemin de longueur 1) dans le graphe RDF. De plus, cette représentation des contextes ne considère pas les classes

20. http://lexvo.org
21. https://www.w3.org/TR/rdf11-mt/
22. https://www.w3.org/TR/swbp-n-aryRelations
23. https://www.w3.org/2004/03/trix/

de l'ontologie, et par conséquent ne permet pas de sélectionner des propriétés différemment, en fonction de chaque classe de l'ontologie. Aussi, les contextes découverts sont moins précis que ceux définis dans l'approche proposée dans ce papier.

Les approches d'analyse relationnelle de concepts (ARC) et d'analyse formelle de concepts dans des Graphes (Graph-FCA) ont été récemment introduites pour découvrir des concepts dans des jeux de données graphe et les ordonner hiérarchiquement dans une structure de treillis. Dans (Hacene *et al.*, 2013), un processus itératif génère de nouveaux attributs à partir de propriétés explorées à différents niveaux de profondeur dans le graphe RDF. L'intention d'un concept formel est exprimée en Logique de description (DL) (Baader *et al.*, 2017) et est composée de rôles originaux et de restrictions de rôles (i.e. restrictions existentielles ou universelles) définis pour des concepts formels calculés dans une étape précédente. Dans (Ferré, Cellier, 2016), l'intention des concepts formels construits sont des motifs de graphes projetés. Cependant, ces approches ne considèrent pas les classes de l'ontologie dans la construction des intentions partagées, décrites dans les concepts formels.

Discussion. Il est maintenant largement reconnu que des liens `owl:sameAs` erronés existent dans le web des données. La présence de ces liens d'identités erronés peut aboutir à la génération de faits incorrects voire même incohérents en cas d'inférence. Ce problème est principalement causé par le manque de liens alternatifs dont la sémantique est bien définie. Nous nous sommes intéressés aux applications pour lesquelles un contexte peut être défini par un sous-ensemble de l'ontologie de domaine. Nous proposons un nouveau prédicat pour représenter les liens d'identité contextuelle et une approche de détection de ces liens. Cette approche peut permettre de limiter l'utilisation incorrecte du `owl:sameAs`, en particulier pour le cas de données scientifiques où l'existence de liens d'identité permet d'exploiter les résultats issus d'expérimentations qui peuvent différer sur des aspects qui ne sont pas pertinents pour une étude donnée. L'explicitation des contextes permet d'associer une explication aux liens d'identité nécessaires à leur validité.

6. Conclusion

Dans cet article, nous avons présenté une approche de détection de liens d'identité contextuelle ($DECIDE$) dans une base de connaissances RDF. L'approche est fondée sur la notion de contexte global exprimant une sous-partie de l'ontologie dans laquelle deux instances de classe sont identiques. En effet, la méthode $DECIDE$ permet de détecter pour chaque couple d'instances d'une classe cible donnée, les contextes globaux les plus spécifiques dans lesquels ces instances sont identiques. Les contextes globaux les plus généraux peuvent être inférés à partir des contextes globaux détectés, grâce à la relation d'ordre entre les contextes globaux que nous avons définis. De plus, cette approche peut prendre en compte des contraintes expertes, spécifiées sous forme de listes de propriétés nécessaires, propriétés non pertinentes, et propriétés co-occurrentes qu'un contexte d'identité doit respecter.

Une première évaluation expérimentale de cette approche a été réalisée sur un jeu de données scientifiques relevant du domaine des gels laitiers, collectées dans le cadre de 11 projets scientifiques menés à l'INRA. Dans ce cadre applicatif, les liens d'identité contextuelle ont été utilisés pour générer des règles permettant de prédire certaines mesures d'observations manquantes. Nous avons constaté que plus le contexte d'identité est spécifique, plus le taux d'erreur est faible.

Comme perspectives, nous envisageons de réaliser d'autres évaluations expérimentales où les contextes d'identité obtenus par $DECIDE$ seraient exploités dans d'autres tâches. En particulier, nous envisageons de découvrir des relations de causalité pour lesquelles les liens d'identité contextuelle pourront permettre de comparer les expériences et de détecter les causes de variations dans les résultats d'observations. Nous souhaitons également nous intéresser à la détection et à la représentation des contextes dans lesquels deux instances sont différentes. Enfin l'approche que nous proposons dans cet article, peut également être utilisée pour compléter des approches d'invalidation de liens d'identité (Guéret *et al.*, 2012 ; Melo, 2013 ; Papaleo *et al.*, 2014), en permettant la requalification d'un lien d'identité erroné par un lien d'identité contextuel.

Remerciements

Ce travail est soutenu par le Center for Data Science (CDS), financé par IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

Bibliographie

Al-Bakri M., Atencia M., David J., Lalande S., Rousset M. (2016). Uncertainty-sensitive reasoning for inferring sameas facts in linked data. In G. A. Kaminka *et al.* (Eds.), *ECAI 2016 - 22nd european conference on artificial intelligence, 29 august-2 september 2016, the hague, the netherlands - including prestigious applications of artificial intelligence (PAIS 2016)*, vol. 285, p. 698–706. IOS Press.

Al-Bakri M., Atencia M., Lalande S., Rousset M. (2015). Inferring same-as facts from linked data: An iterative import-by-query approach. In B. Bonet, S. Koenig (Eds.), *Proceedings of the twenty-ninth AAAI conference on artificial intelligence, january 25-30, 2015, austin, texas, USA.*, p. 9–15. AAAI Press.

Baader F., Horrocks I., Lutz C., Sattler U. (2017). *An introduction to description logic*. Cambridge University Press. Consulté sur http://www.cambridge.org/de/academic/subjects/computer-science/knowledge-management-databases-and-data-mining/introduction-description-logic?format=PB#17zVGeWD2TZUeu6s.97

Batchelor C. R., Brenninkmeijer C. Y. A., Chichester C., Davies M., Digles D., Dunlop I. *et al.* (2014). Scientific lenses to support multiple views over linked chemistry data. In *The semantic web - ISWC 2014 - 13th international semantic web conference, riva del garda, italy, october 19-23, 2014. proceedings, part I*, p. 98–113.

Beek W., Raad J., Wielemaker J., Harmelen F. van. (2018). sameas.cc: The closure of 500m owl:sameas statements. In *International eswc conference.*

Beek W., Schlobach S., Harmelen F. van. (2016). A contextualised semantics for owl: sameas. In H. Sack, E. Blomqvist, M. d'Aquin, C. Ghidini, S. P. Ponzetto, C. Lange (Eds.), *The semantic web. latest advances and new domains - 13th international conference, ESWC 2016, heraklion, crete, greece, may 29 - june 2, 2016, proceedings*, vol. 9678, p. 405–419. Springer. Consulté sur https://doi.org/10.1007/978-3-319-34129-3_25

Belleau F., Nolin M.-A., Tourigny N., Rigault P., Morissette J. (2008). Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, vol. 41, n° 5, p. 706 - 716. Consulté sur http://www.sciencedirect.com/science/article/pii/S1532046408000415 (Semantic Mashup of Biomedical Data)

Carroll J. J., Bizer C., Hayes P., Stickler P. (2005). Named graphs, provenance and trust. In *Proceedings of the 14th international conference on world wide web*, p. 613–622.

Dean M., Schreiber G., Bechhofer S., Harmelen F. van, Hendler J., Horrocks I. *et al*. (2004). OWL web ontology language reference. *W3C Recommendation February*, vol. 10.

Dodds L., Davis I. (2012). *Linked data patterns: A pattern catalogue for modelling, publishing, and consuming linked data*. web. Consulté sur http://patterns.dataincubator.org/book/

Dong X., Halevy A., Madhavan J. (2005). Reference reconciliation in complex information spaces. In *Special interest group on management of data(acm sigmod)*, p. 85–96. NY, USA.

Ferrara A., Nikolov A., Scharffe F. (2011). Data linking for the semantic web. *Int. J. Semantic Web Inf. Syst.*, vol. 7, n° 3, p. 46–76.

Ferré S., Cellier P. (2016). Graph-FCA in practice. In *Graph-based representation and reasoning - 22nd international conference on conceptual structures, ICCS 2016, annecy, france, july 5-7, 2016, proceedings*, p. 107–121.

Forrest P. (2016). The identity of indiscernibles. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy*, Winter 2016 éd.. Metaphysics Research Lab, Stanford University.

Guéret C., Groth P., Stadler C., Lehmann J. (2012). Assessing linked data mappings using network measures. In *Extended semantic web conference*, p. 87–102.

Hacene M. R., Huchard M., Napoli A., Valtchev P. (2013). Relational concept analysis: mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell.*, vol. 67, n° 1, p. 81–108.

Halpin H., Hayes P. J., McCusker J. P., McGuinness D. L., Thompson H. S. (2010). When owl:sameAs isn't the same: An analysis of identity in linked data. In P. F. Patel-Schneider *et al*. (Eds.), *The semantic web – iswc 2010: 9th international semantic web conference, iswc 2010, shanghai, china, november 7-11, 2010, revised selected papers, part i*, p. 305–320. Berlin, Heidelberg, Springer Berlin Heidelberg. Consulté sur http://dx.doi.org/10.1007/978-3-642-17746-0_20

Halpin H., Hayes P. J., Thompson H. S. (2015). When owl: sameas isn't the same redux: towards a theory of identity, context, and inference on the semantic web. In *International and interdisciplinary conference on modeling and using context*, p. 47–60.

Hu W., Chen J., Qu Y. (2011). A self-training approach for resolving object coreference on the semantic web. In *Www*, p. 87–96.

Ibanescu L., Dibie J., Dervaux S., Guichard E., Raad J. (2016). po^2 - a process and observation ontology in food science. application to dairy gels. In *Metadata and semantics research:*

10th international conference, mtsr 2016, göttingen, germany, november 22-25, 2016, proceedings, p. 155–165.

Jaffri A., Glaser H., Millard I. (2008). URI disambiguation in the context of linked data. In C. Bizer, T. Heath, K. Idehen, T. Berners-Lee (Eds.), *Linked data on the web - ldow*, vol. 369. CEUR-WS.org.

Melo G. de. (2013). Not quite the same: Identity constraints for the web of linked data. In M. desJardins, M. L. Littman (Eds.), *Aaai*. AAAI Press. Consulté sur http://dblp.uni-trier.de/db/conf/aaai/aaai2013.html#Melo13

Miles A., Bechhofer S. (2009). *SKOS simple knowledge organization system reference. w3c recommendation 18 august 2009*. Consulté sur http://www.w3.org/TR/2009/REC-skos-reference-20090818/

Nguyen V., Bodenreider O., Sheth A. (2014). Don't like RDF reification?: making statements about statements using singleton property. In *Proceedings of the 23rd international conference on world wide web*, p. 759–770.

Nikolov A., d'Aquin M., Motta E. (2012). Unsupervised learning of link discovery configuration. In *9th extended semantic web conference (eswc)*, p. 119–133. Berlin, Heidelberg, Springer-Verlag. Consulté sur http://dx.doi.org/10.1007/978-3-642-30284-8_15

Papaleo L., Pernelle N., Saïs F., Dumont C. (2014). Logical detection of invalid sameas statements in RDF data. In *Knowledge engineering and knowledge management - 19th international conference, EKAW 2014, linköping, sweden, november 24-28, 2014. proceedings*, p. 373–384. Consulté sur http://dx.doi.org/10.1007/978-3-319-13704-9_29

Raad J., Pernelle N., Saïs F. (2017). Détection de liens d'identité contextuels dans une base de connaissances. In *Ic 2017-28es journées francophones d'ingénierie des connaissances*, p. 56–67.

Raad J., Pernelle N., Saïs F. (2017). Detection of contextual identity links in a knowledge base. In Ó. Corcho, K. Janowicz, G. Rizzo, I. Tiddi, D. Garijo (Eds.), *Proceedings of the knowledge capture conference, K-CAP 2017, austin, tx, usa, december 4-6, 2017*, p. 8:1–8:8. ACM.

Saïs F., Pernelle N., Rousset M.-C. (2009). Combining a logical and a numerical method for data reconciliation. *Journal on Data Semantics*, vol. 12, p. 66–94.

Volz J., Bizer C., Gaedke M., Kobilarov G. (2009). Discovering and maintaining links on the web of data. In *Proceedings of the 8th international semantic web conference(iswc)*, p. 650-665. Berlin, Heidelberg, Springer-Verlag. Consulté sur http://dx.doi.org/10.1007/978-3-642-04930-9_41