# Detection of Contextual Identity Links in a Knowledge Base

Joe Raad, Nathalie Pernelle, Fatiha Saïs

# Detection of Contextual Identity Links in a Knowledge Base

Joe Raad
INRA, UMR 518
Paris, France
joe.raad@agroparistech.fr

Nathalie Pernelle
LRI, Paris Sud University
Orsay, France
nathalie.pernelle@lri.fr

Fatiha Saïs
LRI, Paris Sud University
Orsay, France
fatiha.sais@lri.fr

## 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. These misuses, which are mainly caused by the lack of other well-defined linking alternatives, can lead to erroneous statements or inconsistencies. We propose in this paper a new contextual identity link: *identiConTo* that could serve as a replacement for *owl* : *sameAs* in linking identical instances in a specified context. To detect these contextual links, we have defined an algorithm named DECIDE that has been tested on scientific knowledge bases describing transformation processes.

## KEYWORDS

context, identity link discovery, scientific data

## 1 INTRODUCTION

Over the recent years, scientists have increasingly started to use ontologies in order to formalize the knowledge about their data. This formalization allows scientists to browse data collected and processed by other scientists for the purpose of facilitating data integration tasks and scientific knowledge discoveries. Identity links that can be declared between class instances are of importance since they can be used to fusion data described in different data sources, to predict missing values [24] or to evaluate the reliability of a scientific result based on its frequency [11]. To represent identity links, people are increasingly relying on the use of *owl* : *sameAs*. This relationship, defined in Dean et al. [6], has very strict semantics: "an *owl* : *sameAs* statement indicates that two URI references actually refer to the same thing". I.e. a statement of $material_1$ *owl* : *sameAs* $material_2$ indicates that every property asserted for $material_1$ can be inferred for $material_2$ and vice versa. In many situations, *owl* : *sameAs* is used to link two similar but distinct individuals. Jaffri et al. [18] study the implications of such erroneous use of *owl* : *sameAs* in linking authors of the DBLP dataset with the ones present in DBpedia. To conduct the study, they have chosen 49 names with common forenames and surnames from the 491 796 authors available in the 2006 DBLP dataset. This study shows that 92% of the 49 chosen names have incorrect publications affiliated to them, caused by erroneous inferences. In datasets that describe scientific experiments, data are collected by different scientists, and the experiment's circumstances and participants (e.g. products, materials, etc.) tend to change, even slightly, from one experiment to another. Therefore, individuals can rarely be declared to be the same. Furthermore, this type of genuine identity is not always required, as the notion of identity might change depending on the context. For instance, in some applications, the fact that two drugs share the same name is sufficient to consider them as equivalent, while in other applications it is also necessary that these drugs share the same chemical structure [2]. Likewise, two lemonades with different quantity but equal proportions of lemon, water and sugar can be considered the same in a gustatory context, and different in the context of an energetic and nutritional study. However, it is not easy for scientific experts to enumerate all the contexts of interest that can be relevant for a given task. Our discussions with the INRA[1] experts have shown that it is easier for them to declare constraints that should be respected by a semantic context in order to be considered relevant. For instance, an expert can declare that if the quantity of sugar is considered, the corresponding measure unit must also be considered. Then, when identity links are detected for all the contexts that respect the experts' constraints, it will be possible to focus on different links depending on the considered task.

In this paper, we propose a new contextual identity link named *identiConTo*. This link expresses an identity between two class instances, that is valid in a context defined regarding a domain ontology. We have defined an algorithm for DEtecting Contextual IDEntity links (DECIDE) that detects the most specific global contexts in which a couple of instances are identical. This algorithm can also be guided by a set of semantic constraints provided by experts. We have tested our approach on scientific data issued from two different projects related to the stabilization of micro-organisms and the transformation of dairy gels.

The rest of the paper is structured as follows. In the next section, we present the related work. In section 3, we present our objectives and the preliminaries. In section 4, we present the algorithm *DECIDE* that detects contextual identity links in a knowledge base. In section 5, and before concluding, we present the experiments we have conducted on two scientific datasets to test our approach.

## 2 RELATED WORK

**Data Linking.** After the Linked Open Data cloud (LOD) initiative, there has been a great interest in the development of RDF data-linking approaches (see [8] for a survey). Existing data-linking approaches can be classed into different categories. Firstly, the supervised and unsupervised approaches [16, 19], depending on whether the approaches use a set of labelled data to learn the parameters (e.g. weight of a property, similarity thresholds) and/or functions (e.g. aggregation functions, elementary similarity measures). Secondly, the local [16, 25] and global approaches [1, 23], depending on whether or not the approaches explore the properties of type *owl:ObjectProperty* while measuring similarity. Finally, the informed [1, 16, 23] and uninformed approaches, depending on whether the approaches consider the experts' knowledge declared

[1] French National Institute for Agricultural Research

as ontology axioms (e.g. keys, functionality constraints on properties) or as data-linking rules [25].

According to this classification, the approach we propose is unsupervised, global, and informed. However, our aim is not to detect *owl* : *sameAs* links but to discover identity links that are only valid in specific and explicit contexts.

**Identity link assessment.** Identity links generated by automatic approaches are mainly represented by the *owl*: *sameAs* constructor. This relationship [21], has a strict semantics and requires in particular the identity of all the properties of the related individuals (i.e. *owl*: $sameAs(i_1, i_2) \wedge p(i_1, v) \Rightarrow p(i_2, v)$). Several approaches focus on detecting existing erroneous *owl* : *sameAs* statements, such as [5, 7, 20]. Some approaches are based on the structural properties of large graphs of identity links [7, 12]. Other approaches are constraint-based [5], or logical-based [20].

These approaches aims to invalidate *owl* : *sameAs* links, while our proposed approach aims to qualify the specific contexts where two objects can be considered as identical.

**Weak-identity and similarity representation.** Some approaches have focused on the representation of weak identity links. Halpin et al. [14] propose the Similarity Ontology (SO) which introduces eight new relations such as *so:similar* and *so:claimsIdentical*. Predicates prefixed with the word *claims* express a subjective identity or similarity relation. Their veracity depends on the (contextual) interpretation of the user. These newly introduced relations are organized in a hierarchy where existing identity properties such as *rdfs:seeAlso*, *owl* : *sameAs*, and SKOS predicates are also described. In this hierarchy, each predicate is characterized by the reflexivity, transitivity, and the symmetry properties. In addition, this similarity ontology can be extended with domain-specific relations. However it may be difficult to reliably deploy these distinctions in open-ended domains, and this representation does not allow to explicit the contexts in which an identity link is valid. Therefore the authors of [15] have proposed the use of named graphs to represent contexts, and identity links that are valid in these contexts.

In Melo et al. [5], the authors have defined a new predicate for genuine identity: *lvont:strictlySameAs*. The aim is to distinguish correct identity links from the existing and possibly erroneous *owl* : *sameAs* statements: whenever *lvont:strictlySameAs* is used, the user will know that this link is intended in the strict sense of identity. Additionally, this ontology provides two near-identity predicates: *lvont:nearlySameAs* and *lvont:somewhatSameAs*, which are intentionally left vague (e.g. the relation somewhatSameAs is defined as 'the property of being at least somewhat the same as something else, the City of Los Angeles is somewhat the same as the Greater Los Angeles area').

**Contextual identity links discovery.** Beek et al. [3] propose an approach that allows to represent the possible contexts in which an identity link can be valid. A context is represented by a subset of properties for which two individuals have the same values. All the possible subsets of properties are organised in a lattice using the set inclusion relation. However, the proposed representation does not rely on ontology classes and does not allow selection of a property depending on the considered ontology classes.

To represent sets of instances described in RDF and their corresponding shared description, extensions of Formal Concept Analysis (FCA) framework have been recently introduced to handle graph descriptions [9, 13]. In Hacene et al. [13], an iterative process infers new attributes (propositionalized relations between individuals) from relations that are explored at several levels of depth in the RDF graph. A formal concept intent is made of original attributes and DL role restrictions (existential or universal restrictions) that exploits concepts that have been computed at the previous step ($\exists$ *haspublished*.*C*2 where C2 belongs to the concept lattice). In Ferré et al. [9], the intents of the constructed formal concepts are projected graph patterns. However, these approaches do not consider the ontology classes that can pre-exist and guide the construction of the shared intent described in the formal concepts.

## 3 CONTEXTUAL IDENTITY

In this paper we present a new approach for discovering contextual identity relationships in RDF knowledge bases. The approach aims at detecting identity links that are valid in contexts that can be defined as sub-ontologies of the domain ontology. In this section, we introduce the basic notions and the definitions that are needed to define a contextual identity link. We first present the considered data model and the problem statement. Then, we define the notion of global context and the contextual identity relationship *identiConTo*.

### 3.1 Knowledge Base

We consider a knowledge base where the ontology is represented in OWL[2] and the data represented in RDF[3]. A knowledge base $\mathcal{B}$ is defined by a couple $(\mathcal{O}, \mathcal{F})$ where:

– the ontology $\mathcal{O} = (\mathcal{C}, \mathcal{DP}, \mathcal{OP}, \mathcal{A})$ is defined by a set of classes $\mathcal{C}$, a set of *owl:DataTypeProperty* $\mathcal{DP}$, a set of *owl:ObjectProperty* $\mathcal{OP}$, and a set of axioms $\mathcal{A}$ (e.g property domains and ranges, subsumption).

– $\mathcal{F}$ is a collection of triples *(subject, property, object)*, that expresses that some relationship, indicated by the property, holds between the subject and object of the triple (between two resources or between a resource and a literal value) [4].

### 3.2 Problem statement

The problem of detecting contextual identity links can be defined as follows: given a knowledge base $\mathcal{B} = (\mathcal{O}, \mathcal{F})$ and a set $\mathcal{I}^{tc}$ of instances of a target class *tc* of the ontology $\mathcal{O}$, find for the set of all instance pairs $(i_1, i_2) \in (\mathcal{I}^{tc} \times \mathcal{I}^{tc})$ the most specific global contexts in which $(i_1, i_2)$ are identical.

A global context is a sub-ontology of $\mathcal{O}$ which represents the vocabulary on which two instances are considered as identical. For instance, in the example depicted in Figure 1, the two instances *drug*3 and *drug*4 of the target class *Drug* can be seen as identical when all the ontology's properties and classes are considered with the exception of the property *name* for the drugs. Similarly, the two instances *drug*1 and *drug*2 can be considered as identical in two distinct contexts. In a first context, we can consider all the

[2]https://www.w3.org/OWL/

[3]https://www.w3.org/RDF/

[4]We do not consider blank nodes in this work

products composing the drugs and for every product we consider its weight. However, in this context, the description of a weight is reduced to the measure unit: we do not consider the quantity (property *hasValue*). A second context in which these instances are identical is the context where we take into account the weight of *Paracetamol* described by its value and its measure unit, but we only consider the presence of *Lactose* in the drugs without considering its weight. Some contexts can be more relevant than others (e.g. a value of the weight without its measure unit does not have sense). Hence, we also aim to take into account some expert knowledge that can be represented as a set of constraints on the classes and/or properties that should or should not be involved in the considered contexts.

## 3.3 Contexts

A global context is represented as a connected sub-ontology of the ontology $O$ that is composed of a set of classes and properties of $O$, and a set of axioms which is limited to constraints on property domains and ranges. The set of classes that can be involved in a global context is the subset of classes, denoted by *DepC*, that are instantiated in $\mathcal{B}$ (see Definition 3.1 ). Moreover, we automatically choose the abstraction level of the classes involved in a global context by selecting, from the instantiated classes (direct types), the most general ones.

In what follows, we first introduce the set of classes *DepC* that can be involved in the contexts. Then, we formally define the global contexts and the contextual identity relation, named *identiConTo*, that expresses that two instances are identical in a given global context.

*Definition 3.1.* **Selected classes *DepC***. The set of selected classes *DepC* that can be involved in the contextual identity links is the subset of instantiated classes $c_i$ of $\mathcal{B}$ such that:

$$DepC = \{c_i \in C \mid \nexists c_j \in C \; s.t. \; \exists x, directType(x, c_j) \; and \; c_i \sqsubset c_j\}$$

**Example 1.** In Figure 1, *DepC* will contain all the classes of the graph except *Product* which is not instantiated. Therefore, $par1$ and $lac1$ will be uniquely considered as of type *Paracetamol* and *Lactose* respectively.

*Definition 3.2.* **Global Context**. A global context is a sub-ontology $GC_u$=$(C_u, DP_u, OP_u, A_u)$ of $O$ such that $C_u \subseteq DepC$, $DP_u \subseteq DP$, $OP_u \subseteq OP$ and $A_u$ is a set of domain and range constraints that are more specific than those described in $A$: $\forall op \in OP_u$, $domain_u(op) \sqsubseteq domain_O(op)$ and $range_u(op) \sqsubseteq range_O(op)$, and $\forall dp \in DP_u$, $domain_u(dp) \sqsubseteq domain_O(dp)$.

**Example 2.** In Figure 1, there exists many possible global contexts. We present one:
$GC_1$=($C = \{Drug, Paracetamol, Lactose, Weight\}$,
$OP = \{isComposedOf, hasWeight\}$, $DP = \emptyset$,
$A = \{domain(isComposedOf) = Drug$,
$range(isComposedOf) = Lactose \sqcup Paracetamol$,
$domain(hasWeight) = Lactose \sqcup Paracetamol$,
$range(hasWeight) = Weight\}$)

*Definition 3.3.* **Order relation between global contexts**. Let $GC_u = (C_u, OP_u, DP_u, A_u)$ and $GC_v = (C_v, OP_v, DP_v, A_v)$ be two global contexts. The context $GC_u$ is more specific than $GC_v$, noted $GC_u \leq GC_v$, if $C_v \subseteq C_u$, $OP_v \subseteq OP_u$, $DP_v \subseteq DP_u$ and $\forall op \in OP_v$, $domain_v(op) \sqsubseteq domain_u(op)$ and $range_v(op) \sqsubseteq range_u(op)$, and $\forall dp \in DP_v$, $domain_v(dp) \sqsubseteq domain_u(dp)$, and $range_v(dp)$=$range_u(dp)$.

In order to filter out the irrelevant contexts to consider, we take in consideration the experts' knowledge when it is available. An expert can supply three types of constraints:
– *Unwanted properties* ($UP$): this refers to properties that an expert wants to discard in the detection of contextual identity links. Such constraints can be used when property values correspond to unstructured (free) text, or are known to be particularly heterogeneous, or when the property subjects or objects are evolutive or insignificant to compare two instances for a given task. In such cases, an expert can declare that a property p is unwanted for a given domain $c_i$ (or a particular range $c_j$) by adding a constraint $up = (c_i, p, *)$ (resp. $up = (*, p, c_j)$) in $UP$. When a property is unwanted in all domains and ranges, the constraint $(*, p, *)$ can be used. In such cases, $p \notin OP \cup DP$.
– *Necessary properties* ($NP$): a necessary property is a constraint noted $np = (c_i, p, *)$ or $(*, p, c_j)$. When such constraints are added to $NP$, we will only consider global contexts where the property $p \in OP$ or $p \in DP$, and such that $c_i \in domain(p)$ (resp. $c_j \in range(p)$).
– *Co-occurring properties* ($CP$): a co-occurrence constraint $cp = \{(c_i, p_1, *), ..., (c_i, p_n, *)\}$ can be declared to guarantee that a certain class $c_i$ will be either declared as the domain (or range) of *all* the properties indicated in the constraint, or *none* of them. For instance, to declare that the weight's value has no meaning without its measure unit, an expert can add the constraint $cp_1 = \{(Weight, hasValue, *), (Weight, hasUnit, *)\}$.

## 3.4 Contextual identity links

In our approach, two instances are considered as identical in a given global context, when all the properties contained in the global context are instantiated and when their instances (values) are equal. Therefore, we firstly define the contextual description that is considered for one instance in one context. Then we will define the conditions that must hold to consider that two instance descriptions refer to the same entity.

*Definition 3.4.* **Contextual instance description according to a global context**. Given a set of RDF triples $\mathcal{F}$, a global context $GC_u = (C_u, OP_u, DP_u, A_u)$ and an instance $i$, a contextual description $G_i$ of $i$ in $GC_u$ is the maximal set of triples that describe $i$ in $\mathcal{F}$ such that:
– $G_i$ forms a connected graph that contains at least one triple where $i$ is a subject or an object
– $\forall \; t = < s, p, o > \in G_i$ then $p \in OP_u \cup DP_u$ and $type(s) \sqsubseteq domain_u(p)$ and $type(o) \sqsubseteq range_u(p)$
– $\forall \; j$ a class instance of $G_i$, and $\forall dp \in DP_u$ such as $type(j) \sqsubseteq domain(dp)$, then $\exists \; t_a = < j, p, v > \in G_i$, with $v$ of type literal
– $\forall \; j$ a class instance of $G_i$, and $\forall op \in OP_u$ such as $type(j) \sqsubseteq domain(op)$, and $c_1 \cup c_2 \sqsubseteq range(op)$ then $\exists \; t_a = < j, op, k >$ and $t_b = < j, op, l > \in G_i$ with $type(k) = c_1$ and $type(l) = c_2$

From two contextual descriptions of two class instances, defined in a given context, we can define if they can be considered as

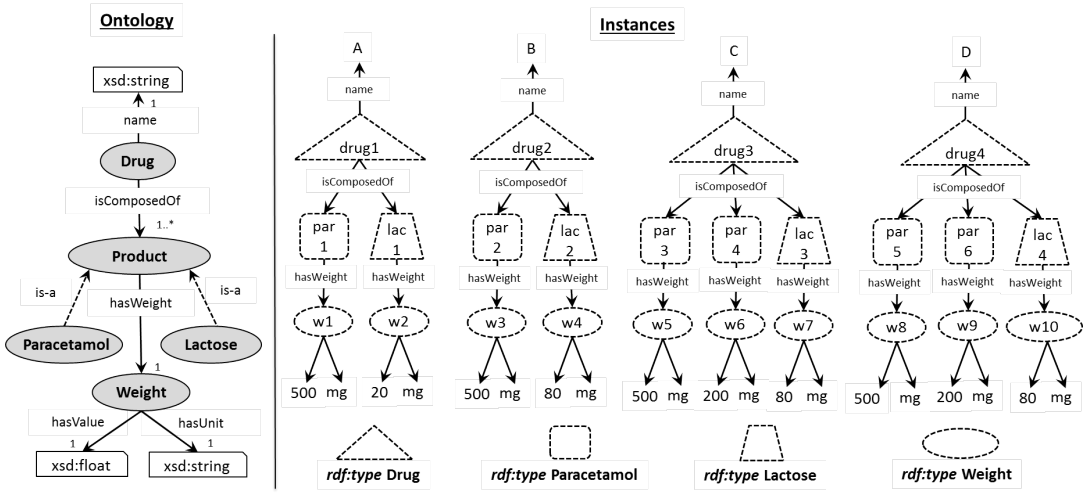


**Figure 1: An extract of ontology $O$, four instances $drug1$, $drug2$, $drug3$ and $drug4$ of the target class $Drug$.**

identical. In this work we will consider that properties are local complete: if a property $p$ is instantiated for a given class instance $i$, we consider that all the property instances are known for $i$. Since a local completness is assumed, two instances can be considered as identical when the contextual graphs, formed by the contextual descriptions, are isomorphic up to a renaming of the instance URI. Note that since some classes can be removed from the global context, this constraint can in fact be considered class by class.

*Definition 3.5.* **Identity in a global context**. Given a global context $GC_u$, a pair of instances $(i_1, i_2)$ are identical in $GC_u$, noted $identiConTo_{<GC_u>}(i_1, i_2)$, only if the two labelled graphs $G_{i_1}$ and $G_{i_2}$, that represent the contextual descriptions of $i_1$ and $i_2$ respectively, are isomorphic up to a rewriting of the URI of the class instances (literals must be equal).

**Example 3.** $drug1$ and $drug2$ are considered as identical according to the global context $GC_1$ defined in Example 2.
(i.e. $identiConTo_{<GC_1>}(drug1, drug2)$).

The contextual identity relations will only be specified for the most specific global context(s), but can be inferred for the more general ones using the order relation between global contexts: given $GC_u$ and $GC_v$ two global contexts, with $GC_u \leqslant GC_v$, then $identiConTo_{<GC_u>}(i_1, i_2) \Rightarrow identiConTo_{<GC_v>}(i_1, i_2)$.

# 4 DECIDE – DETECTING CONTEXTUAL IDENTITY

Before we present the algorithm in sub-section 4.2, we introduce in 4.1 the terminologies that are used throughout the algorithm.

## 4.1 Preliminaries

*Definition 4.1.* **Local Contexts**. A local context of a class $c$ is a context that is limited to datatype and object properties that are defined for $c$. In the algorithm, we will note:
– $LC_u^{out} = (C_u^{out}, OP_u^{out}, DP_u^{out}, A_u^{out})$, a local context where $\forall p \in OP_u^{out} \cup DP_u^{out}$, $domain(p) = c$ and
–$LC_u^{in} = (C_u^{in}, OP_u^{in}, DP_u^{in}, A_u^{in})$ a local context where $DP_u^{in} = \emptyset$ and $\forall op \in OP_u^{in}$, $range(op) = c$.

*Definition 4.2.* **Identity Graph**. An identity graph $IG_{<i_1,i_2>} = (V, E)$ for a pair of individuals $(i_1, i_2)$, is a connected labelled undirected graph, where $V$ is a set of nodes and $E$ is a set of edges. Each node $n_i$ represents a set of pairs $I_1 \times I_2$, and the local contexts $LC_n^{in}(c)$ and $LC_n^{out}(c)$ that generalize all the most specific local contexts $LC_n^{in}(c)$ and $LC_n^{out}(c)$ for which the pairs are considered as identical. A node $n_1$ representing a set of pairs $I_1 \times I_2$ is linked to a node $n_2$ representing the set of pairs $J_1 \times J_2$ by an edge $e(n1, n2)$ labelled as $p$, if $\forall\ (i_1, i_2) \in I_1 \times I_2$, $\exists\ j_1 \in J_1$ and $j_2 \in J_2$ such that:
– $\exists < i_1, p, j_1 >$ and $< i_2, p, j_2 > \in \mathcal{F}$ if $p \in LC_{n_1}^{out}(c)$
– $\exists < j_1, p, i_1 >$ and $< j_2, p, i_2 > \in \mathcal{F}$ if $p \in LC_{n_1}^{in}(c)$.
In an identity graph $IG_{<i_1,i_2>}$, a graph path $gp_i$ is a sequence of distinct nodes $\{n_1, n_2, ..., n_m\}$ rooted by $n_1$ which describes $(i_1, i_2)$, and respects the following condition: $\nexists\ n_k, n_l \in gp_i$, with $k < l$ and $LC_{n_l}(c) \leq LC_{n_k}(c)$.

Figure 2 presents the identity graphs $IG_1$ and $IG_2$ of the pair of drugs $(drug3, drug4)$.

## 4.2 Algorithm

The goal of the algorithm *DECIDE* (DEtection of Contextual IDEntity) is to determine for each pair of instances $(i_1, i_2) \in I^{tc} \times I^{tc}$ of a target class $tc$ given by the user, the set of the most specific global contexts in which the identity relation $identiConTo$ is true. *DECIDE* requires to have the set of facts $\mathcal{F}$ of the considered knowledge base and the target class $tc$ as inputs. In addition, *DECIDE* may consider different constraint lists $UP$, $NP$, $CP$ given by an expert. In this paper, we restrict the description of this algorithm to its two main functions, nonetheless a more detailed description with different use-cases is available in [22]. The algorithm *DECIDE*, described in Algorithm 1:

**– collects the selected classes** (definition 3.1), in order to indicate the level of abstraction to be considered in building the

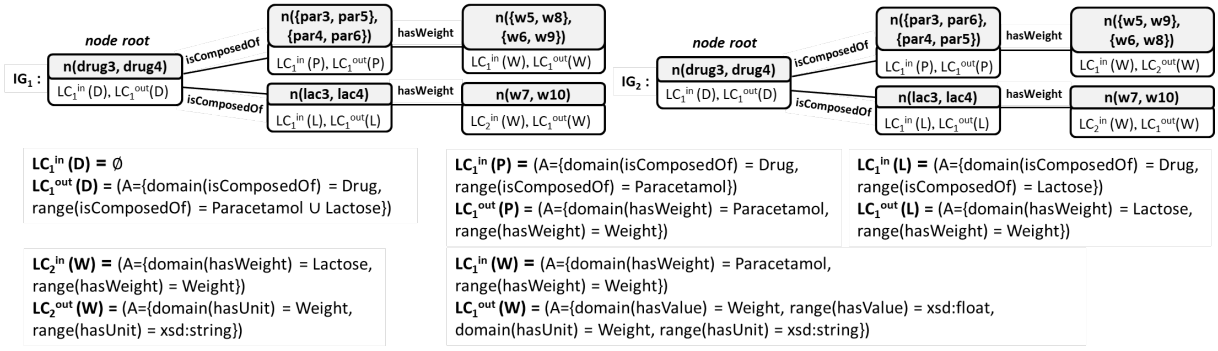


**Figure 2: The two possible Identity Graphs for the pair (drug3, drug4). For simplicity reasons, $C$, $OP$, and $DP$ are not represented in this Figure for all the local contexts.**

**Algorithm 1:** DECIDE

**Input:**
– $tc$: the target class
– $K(NP, UP, CP)$: the expert constraints
– $\mathcal{F}$: the set of RDF triples of the considered knowledge base
**Output:** $MScontexts$: set of most specific global contexts for each pair of instances

$DepC \leftarrow getDepC(\mathcal{F})$ ;
$I^{tc} \leftarrow$ list of instances of $type(tc)$ in $\mathcal{F}$ ;
**foreach** $((i_1, i_2) \in I^{tc} \times I^{tc})$ **do**
  $GCset \leftarrow \emptyset$;
  $IGset \leftarrow constructIdentityGraph(i_1, i_2, DepC, K, \mathcal{F})$ ;
  **foreach** $(IG \in IGset)$ **do**
    $n_0 \leftarrow IG.getNode(i_1, i_2)$ ;
    $N \leftarrow \emptyset$; $a \leftarrow \emptyset$; $GC \leftarrow \emptyset$; $LCset \leftarrow \emptyset$;
    $GC \leftarrow generateGC(n_0, a, GC, LCset, N, IG)$ ;
    $GCset.add(GC)$ ;
    **foreach** $(LC \in LCset)$ **do**
      $GC \leftarrow \emptyset$; $GC.add(LC)$ ;
      $GC \leftarrow generateGC(n_0, a, GC, LCset, N, IG)$;
      **if** $(\nexists\, GC_1 \in GC_{set}$, *such as* $GC_1 \leq GC)$ **then**
        $GCset.add(GC)$ ;
      **if** $(\exists\, GC_2 \in GC_{set}$, *such as* $GC \leq GC_2)$ **then**
        $GCset.remove(GC_2)$ ;
  $MScontexts.add(GCSet, (i_1, i_2))$;
return $MScontexts$;

**Algorithm 2:** Generate GC

**Input:**
– $n$: an identity graph node
– $a_s$: axiom indicating the type of the node source with the property source
– $GC$: the current global context
– $LCset$: set of unused local contexts
– $N$: list of visited nodes
– $IG$: the identity graph
**Output:** $GC$: the current most specific global context

**if** $(n \notin N)$ **then**
  $N.add(n)$ ;
  $LC_n(c) \leftarrow getOutgoingLocalContext(n)$ ;
  $LC_{ex}(c) \leftarrow GC.getExistingLocalContext(c)$ ;
  **if** $(LC_{ex}(c) == null$ *or* $LC_{ex}(c) == LC_n(c))$ **then**
    $GC.add(LC_n(c))$ ; //if it does not exist
    $E^n \leftarrow IG.getOutgoingEdges(n)$ ;
    **foreach** $(e = (op, n, n_d) \in E^n)$ **do**
      $a_d \leftarrow \{domain(op) = c, range(op) = type(n_d)\}$ ;
      $GC \leftarrow generateGC(n_d, a_d, GC, LCset, N, IG)$ ;
  **else**
    **if** $(LC_n(c) \leqslant LC_{ex}(c))$ **then**
      $E^n \leftarrow IG.getOutgoingEdges(n)$ ;
      **foreach** $(e = (op, n, n_d) \in E^n)$ **do**
        $a_d \leftarrow \{domain(op) = c, range(op) = type(n_d)\}$ ;
        **if** $(a_d \in LC_{ex}(c))$ **then**
          $GC \leftarrow$ $generateGC(n_d, a_d, GC, LCset, N, IG)$;
    **else**
      $c_s \leftarrow a_s.getDomain()$ ;
      $LC(c_s) \leftarrow GC.getExistingLocalContext(c_s)$ ;
      $LC(c_s).remove(a_s)$ ;
      $GC.replace(LC(c_s))$; //replace existing $LC(c_s)$
    $LCset.add(LC_n(c))$ ; //if it does not exist
    $LCset.add(intersect(LC_n(c), LC_{ex}(c)))$ ; //if it does not exist
return $GC$;

identity graphs and generating the most specific global contexts. Then for each pair of individuals of the target class $tc$:

– **constructs the identity graph(s)** (definition 4.2), using a depth-first search algorithm. When different mappings between instances of the same class can be considered, a new identity graph identity is constructed.

– **generates the most specific global context(s)** by relying on the constructed identity graphs. A global context $GC$ is constructed using the set of local contexts and insures the presence of no more than one local context per class in the same global context. The most specific global contexts are generated using the function *generateGC*, which traverses the identity graph $IG$ using also a depth-first search algorithm. This function, described in Algorithm 2, aims to add its most specific outgoing local context $LC_n(c)$, which is already calculated in $IG$, to the current global context $GC$ (i.e. the most specific global context). There is three cases:

(1) If $GC$ does not contain a local context $LC_{ex}(c)$ for the class $c$, or if $GC$ contains $LC_{ex}(c)$ with $LC_{ex}(c)$ equal to the local context $LC_n(c)$ of $n$, then $LC_n(c)$ is added to $GC$. This function is then recursively recalled for each node $n_d$ in $IG$, such as there is an edge from $n$ to $n_d$.

(2) If $GC$ contains a local context $LC_{ex}(c)$ for the class $c$, and $LC_n(c)$

is more specific than $LC_{ex}(c)$, then this function is recursively recalled for each destination node $n_d$ in $IG$, such as there is an edge from $n$ to $n_d$ labelled $op$ and we have in the axioms of $LC_{ex}(c)$: domain of $op = c$ and $type(n_d) \sqsubseteq range(op)$.
(3) If $GC$ contains a local context $LC_{ex}(c)$ for the class $c$, and $LC_n(c)$ is not more specific than $LC_{ex}(c)$, then this function is not recalled for this graph node, and the domain representing the type of the node source and the range representing $c$ of the object property $op$ that led to this graph element will be removed from $LC_{ex}(c)$.
In both (2) and (3), $LC_n(c)$ and the most specific local context that generalizes $LC_n(c)$ and $LC_{ex}(n)$ will be added to a list $LCset$, in order to guarantee the presence of these local contexts in other global contexts. Therefore, resulting in several most specific global contexts for the same pair.

The time complexity of this algorithm is $O(n \times I^2)$, with $n$ = the number of pairs of the target class $tc$, and $I$ = the number of instances in $\mathcal{F}$. *DECIDE* is implemented in *Java* using the *Jena TDB* triple store, and is available at http://github.com/raadjoe/DECIDE_v2.

When applied on the pair $(drug1, drug2)$, *DECIDE* results in two global contexts $GC_1$ and $GC_2$, representing the most specific contexts in which these two drugs are identical:

$GC_1$=($C = \{Drug, Paracetamol, Lactose, Weight\}$,
$OP = \{isComposedOf, hasWeight\}$, $DP = \{hasUnit\}$,
$A = \{domain(isComposedOf) = Drug$,
$range(isComposedOf) = Lactose \sqcup Paracetamol$,
$domain(hasWeight) = Lactose \sqcup Paracetamol$,
$range(hasWeight) = Weight$,
$domain(hasUnit) = Weight, range(hasUnit) = xsd{:}string\}$)

$GC_2$=($C = \{Drug, Paracetamol, Lactose, Weight\}$,
$OP = \{isComposedOf, hasWeight\}$, $DP = \{hasValue, hasUnit\}$,
$A = \{domain(isComposedOf) = Drug$,
$range(isComposedOf) = Lactose \sqcup Paracetamol$,
$domain(hasWeight) = Paracetamol$,
$range(hasWeight) = Weight$,
$domain(hasValue) = Weight, range(hasValue) = xsd{:}float$,
$domain(hasUnit) = Weight, range(hasUnit) = xsd{:}string\}$)

# 5 EXPERIMENTS

## 5.1 Datasets description

Our approach has been evaluated on two scientific datasets exploited using the 1.4 version[5] of the ontology $PO^2$ [17], which aims at modelling transformation processes. Each process can be conducted over several itineraries, with each itinerary representing a sequence of transformation steps (drying, heating, etc.). In this ontology, as in most knowledge bases used to model scientific experiments, a distinction is made between the actual experiments that include these steps with their participants, and between the observations conducted at the end of each step. These observations contain a large number of missing information, since not every measure (e.g. temperature, pH) is consistently observed in each experiment's step. The distinction between the experiments and the observations can be seen in the ontology's core model[6].
– The first dataset in which we have tested our approach describes the process of micro-organisms' stabilization, conducted in 20 different itineraries in the context of the INRA[7] CellExtraDry project. This dataset contains 1 721 979 statements, 208 instantiated selected classes, 415 136 individuals and 159 properties (83 object properties).
– The second dataset describes the process of the dairy gels' transformation, conducted in 12 itineraries in the context of the INRA Carredas project. This dataset contains 237 838 statements, 555 instantiated selected classes, 42 269 individuals, and 159 properties (83 object properties).

We have tested the algorithm *DECIDE* separately on each of these datasets, in order to detect the most specific global contexts in which the individuals of the target class *Mixture* are identical. A mixture, similarly to the class *Drug* in Figure 1, is composed of a set of products and is transformed during the different steps of the process. *DECIDE* has been executed on an 8GB RAM Windows 10 machine, with an Intel Core 4 × 2.6 GHz process.

## 5.2 Discovered contextual identity links

Table 1 presents the results of *DECIDE* applied on these two scientific datasets, without considering their observations (i.e. the properties related to the observations have been declared as unwanted properties). In the *CellExtraDry* dataset, the 210 instances of the target class *Mixture* which can form 21945 pairs, have resulted in 31092 contextual identity links valid in 28 global contexts in total, while the 191271 pairs of mixtures in the *Carredas* dataset have resulted in 239410 identity links valid in 231 different global contexts in total. On average in the *CellExtraDry* and *Carredas* datasets, each identity graph of each pair of mixtures is composed of 11 nodes (7 respectively), and each pair is identical in 1.41 most specific global contexts (1.25 respectively).

Each global context is represented as a named graph [4] in the original dataset, with each named graph containing the detected identity statements. A contextual identity statement between two instances $i_1$ and $i_2$ indicates that this context represents the most specific global context in which these two instances are identical (definition 3.5), with each contextual identity statement being symmetric, transitive, and reflexive. Some of the detected contexts contain up to 20 classes and 35 properties, while less specific ones contain only one class and one property.

We have repeated the experiments on each dataset, while taking into account a constraint $cp$ that expresses that a weight value cannot be considered without its unit of measure and vice versa. While the number of distinct most specific global contexts have remained unchanged in both datasets, we have noticed a change in around 40 % of the generated most specific global contexts. More precisely, each global context containing one of the properties without the other has been replaced by another (new) global context where these two properties are not considered for the class $Weight$.

[5]The core ontology of $PO^2$ is available at: http://agroportal.lirmm.fr/ontologies/PO2

[6]http://github.com/raadjoe/DECIDE_v2/blob/master/PO2_model.jpg
[7]The French National Institute for Agriculture

**Table 1: Results of *DECIDE* on the CellExtraDry and Carredas datasets with the target class *Mixture***

| | ***CellExtraDry*** | ***Carredas*** |
|---|---|---|
| *# Individuals of target class* | 210 | 619 |
| *# Possible Pairs* | 21 945 | 191 271 |
| *# Dependant Classes (Total Classes)* | 191 (208) | 488 (555) |
| *# Graph Nodes per pair* | 11 | 7 |
| *# Different Global Contexts* | 28 | 231 |
| *# Identity Links* | 31 092 | 239 410 |
| *# Identity Links per pair* | 1.41 | 1.25 |
| *Execution Time (approx. minutes)* | 2 | 26 |

## 5.3 Use of contextual identity links for prediction

The goal of this experimentation is to test if contextual identity links can be exploited for prediction tasks. More precisely, we want to find out the probability of two experiments, being identical in a certain context, to have similar observations. Therefore, we will be able to predict to a certain degree of certainty, some experiments' unobserved measures. Table 1 indicates that the individuals of the target class *Mixture* are connected to most of the datasets' instantiated classes, 191 out of 208 in *CellExtraDry* and 488 out of 555 classes in *Carredas*, thus showing that an identity between two mixtures can also indicate an identity between the experiments' steps in which these two mixtures exist.

According to Leibniz's "Indiscernibility of Identicals" principle [10], a genuine identity between two objects (e.g. experiments), indicates that every property (e.g. an observed measure) asserted to one is asserted to the other: $x = y \cap p(x,z) \longrightarrow p(y,z)$ with $p \in OP \cup DP$. In this prediction task, we aim to detect for each context $GC_i$, the set $\Psi$ of properties $\{p_1, ..., p_n\}$, where $identiConTo_{<GC_i>}(x,y) \cap p(x,z_1) \longrightarrow p(y,z_2)$ with $z_1 \simeq z_2$ and $\Psi \cap (OP^{GC_i} \cup DP^{GC_i}) = \emptyset$. Such rules can be written as $r$: $identiConTo_{<GC_i>}(x,y) \longrightarrow \text{same}(m)$, with $m$ representing a certain measure (e.g. pH measure). Since the detected contextual identity links are only stated for the most specific contexts of each pair, we have exploited the global contexts' order relation (definition 3.3) to obtain the complete set of contextual identity links for each global context.

In order to evaluate the quality of a rule $r$ we calculate:
– **the rule's average error rate:** for each pair $(x, y)$ identical in $GC_i$, we calculate the error rate for their $m$ measure values, if they exist for both $x$ and $y$. For instance, the error rate for the pair $(x, y)$ for the measure pH: $er_{pH}(x, y) = \frac{|pH(x)-pH(y)|\times 100}{|pH(max)-pH(min)|}$ with $pH(max)$ and $pH(min)$ representing the maximum and the minimum values taken for the measure pH in the dataset. From the sum of all this measure's error rate of all these pairs, we obtain the rule's average error rate.
– **the rule's support:** represents the number of pairs identical in $GC_i$ that have the measure $m$, divided by the total number of pairs in $GC_i$.

We have generated 112 rules in the *CellExtraDry* dataset (averaging 4 rules per context), and 3677 rules in the *Carredas* dataset (averaging 15 rules per context). On average, in *CellExtraDry* a rule's average error rate is 7.3% and the rule's support is 0.4%, while in *Carredas* a rule's average error rate is 20% and a rule's support is 1%. This low support in both datasets shows the large number of observational measures that are missing in each experiment. After testing all the rules in each global context, we have deducted that on average, the error rate of a rule decreases by 22% when a global context is replaced by a more specific global context in the *CellExtraDry* dataset, and decreases by 31.5% in the *Carredas* dataset. This decrease shows that rules discovered in more specific global contexts are more precise than the ones discovered in more general contexts, and that the contextual identity links can for example be exploited to predict missing properties values with different confidence level. We have asked the domain experts to evaluate the plausibility of the 20 best detected rules (in terms of error rate and support combined) on a scale of "Strongly Agree", "Agree", "Disagree", and "Strongly Disagree". The experts have strongly agreed on the plausibility of 9 rules, agreed on 4 rules, and strongly disagreed on the plausibility of 1 rule. The experts were not sure of the plausibility of the 6 remaining rules for various reasons. Table 2 presents some of the rules strongly agreed as plausible in the *Carredas* dataset. For instance, the first rule indicates that there is a high probability that mixtures with the same weight of Rennet, Sardine, and Sodium Chloride, and containing (i.e. not necessarily the same weight) Lipids, Water, and Proteins, to have similar adhesiveness.

**Table 2: Examples of Detected Rules in the Carredas dataset**

| ***Rule*** | ***Error Rate*** | ***Support*** |
|---|---|---|
| $identiConTo_{<GC_{102}>}(x, y)$ $\longrightarrow$ same(Adhesiveness) | 2.2 % | 23 % |
| $identiConTo_{<GC_{74}>}(x, y)$ $\longrightarrow$ same(Sweetness) | 4.5 % | 13 % |
| $identiConTo_{<GC_{202}>}(x, y)$ $\longrightarrow$ same(Bitterness) | 7.1 % | 29 % |
| $identiConTo_{<GC_{124}>}(x, y)$ $\longrightarrow$ same(Acidity) | 8.2 % | 21 % |

## 5.4 Discussion

Our collaboration with the domain experts, and the experiments' results conducted on these scientific datasets have shown us that:
– the use of genuine identity links such as the *owl:sameAs* link is rarely required in scientific datasets, since the experiments' environment tend to change, even slightly from one experiment to another, which could result in a propagation of incorrect observational measures.
– asking domain experts to specify the contexts in which two objects are considered identical is not an intuitive task, as the identity contexts can differ from one expert and task to another. Instead, specifying some constraints on these contexts is a more effective way to benefit from the experts' knowledge.
– thousands of explicit contextual identity links can be detected in a reasonable time, despite the high connectivity between all these graph's instances.
– the contextual identity links can be used to generate rules that can help predict some of the missing observational measures.
– the relevancy of a certain context can vary depending on the

conducted observations. For instance, the identity of the mixtures' composition is required in tasks that study the mixtures' acidity, while the identity of the mixtures' steps is required in tasks studying the experiments' environmental impact.
– rules detected in more specific contexts have better error rates than the ones detected in less specific contexts.

## 6 ACKNOWLEDGMENT

This work is supported by the Center for Data Science, funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

## 7 CONCLUSION

We propose in this paper an approach of Detecting Contextual Identity links (*DECIDE*) in a knowledge base, based on the notion of a global context that represents a sub-ontology. *DECIDE* detects for each pair of individuals of a target class given by the user, the most specific contexts in which this pair is identical. More general contexts can be inferred from the detected most specific ones, thanks to the order relation that hierarchizes all the global contexts. Furthermore, this approach can take into account some experts' constraints, which can be in the form of a list of necessary properties for the identity link, list of unwanted properties, and list of properties that must occur together. A first experiment of this approach has been realized on two scientific datasets, in which these contextual identity links have been used to generate rules that can serve for the prediction of missing experimental observations. These prediction rules' certainty varies depending on the specificity of the context.

As a next step, we would like to exploit these contextual identity links in other tasks. In particular, we would like to discover causality rules, in which these contextual identity links can serve for comparing experiments and selecting the relevant variables, that can explain the cause of some of the experiments results' variations.

## REFERENCES

[1] Mustafa Al-Bakri, Manuel Atencia, Steffen Lalande, and Marie-Christine Rousset. 2015. Inferring Same-As Facts from Linked Data: An Iterative Import-by-Query Approach. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. 9–15.

[2] Colin R. Batchelor, Christian Y. A. Brenninkmeijer, Christine Chichester, Mark Davies, Daniela Digles, Ian Dunlop, Chris T. A. Evelo, Anna Gaulton, Carole A. Goble, Alasdair J. G. Gray, Paul T. Groth, Lee Harland, Karen Karapetyan, Antonis Loizou, John P. Overington, Steve Pettifer, Jon Steele, Robert Stevens, Valery Tkachenko, Andra Waagmeester, Antony J. Williams, and Egon L. Willighagen. 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*. 98–113.

[3] Wouter Beek, Stefan Schlobach, and Frank van Harmelen. 2016. A Contextualised Semantics for owl: sameAs. In *International Semantic Web Conference*. Springer, 405–419.

[4] Jeremy J Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. 2005. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*. ACM, 613–622.

[5] Gerard de Melo. 2013. Not Quite the Same: Identity Constraints for the Web of Linked Data.. In *AAAI*, Marie desJardins and Michael L. Littman (Eds.). AAAI Press.

[6] Mike Dean, Guus Schreiber, Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L McGuinness, Peter F Patel-Schneider, and L Andrea Stein. 2004. OWL web ontology language reference. *W3C Recommendation February* 10 (2004).

[7] Li Ding, Tim Finin, Joshua Shinavier, and Deborah L. McGuinness. 2010. owl:sameAs and Linked Data: An Empirical Study. In *In The Semantic Web - ISWC*. 145–160.

[8] Alfio Ferrara, Andriy Nikolov, and François Scharffe. 2011. Data Linking for the Semantic Web. *Int. J. Semantic Web Inf. Syst.* 7, 3 (2011), 46–76.

[9] Sébastien Ferré and Peggy Cellier. 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*. 107–121.

[10] Peter Forrest. 2016. The Identity of Indiscernibles. In *The Stanford Encyclopedia of Philosophy* (winter 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[11] Matthew Gamble and Carole Goble. 2011. Quality, Trust, and Utility of Scientific Data on the Web: Towards a Joint Model. In *Proceedings of the 3rd International Web Science Conference (WebSci '11)*. ACM, New York, NY, USA, Article 15, 8 pages.

[12] Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. 2012. Assessing linked data mappings using network measures. *The Semantic Web: Research and Applications* (2012), 87–102.

[13] Mohamed Rouane Hacene, Marianne Huchard, Amedeo Napoli, and Petko Valtchev. 2013. Relational concept analysis: mining concept lattices from multi-relational data. *Ann. Math. Artif. Intell.* 67, 1 (2013), 81–108.

[14] Harry Halpin, Patrick J Hayes, James P McCusker, Deborah L McGuinness, and Henry S Thompson. 2010. When owl: sameas isn't the same: An analysis of identity in linked data. In *International Semantic Web Conference*. Springer, 305–320.

[15] Harry Halpin, Patrick J Hayes, and Henry S Thompson. 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*. Springer, 47–60.

[16] Wei Hu, Jianfeng Chen, and Yuzhong Qu. 2011. A self-training approach for resolving object coreference on the semantic web. In *WWW*. 87–96.

[17] Liliana Ibanescu, Juliette Dibie, Stéphane Dervaux, Elisabeth Guichard, and Joe Raad. 2016. POˆ 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*. Springer, 155–165.

[18] Afraz Jaffri, Hugh Glaser, and Ian Millard. 2008. URI Disambiguation in the Context of Linked Data.. In *Linked Data on the Web - LDOW (CEUR Workshop Proceedings)*, Christian Bizer, Tom Heath, Kingsley Idehen, and Tim Berners-Lee (Eds.), Vol. 369. CEUR-WS.org.

[19] Andriy Nikolov, Mathieu d'Aquin, and Enrico Motta. 2012. Unsupervised learning of link discovery configuration. In *9th Extended Semantic Web Conference (ESWC)*. Springer-Verlag, Berlin, Heidelberg, 119–133.

[20] Laura Papaleo, Nathalie Pernelle, Fatiha Saïs, and Cyril Dumont. 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*. 373–384.

[21] Peter F. Patel-Schneider, Patrick Hayes, and Ian Horrocks. 2004. *OWL Web Ontology Language Semantics and Abstract Syntax Section 5. RDF-Compatible Model-Theoretic Semantics*. Technical Report. W3C. http://www.w3.org/TR/owl-semantics/rdfs.html#built_in_vocabulary

[22] Joe Raad, Nathalie Pernelle, and Fatiha Saïs. 2017. *DECIDE - Detecting Contextual Identity*. Technical Report. LRI, Paris-Sud University. http://github.com/raadjoe/DECIDE_v2/blob/master/technical-report.pdf

[23] Fatiha Saïs, Nathalie Pernelle, and Marie-Christine Rousset. 2009. Combining a Logical and a Numerical Method for Data Reconciliation. *Journal on Data Semantics* 12 (2009), 66–94.

[24] Fatiha Saïs and Rallou Thomopoulos. 2014. Ontology-aware prediction from rules: A reconciliation-based approach. *Knowl.-Based Syst.* 67 (2014), 117–130.

[25] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. 2009. Discovering and Maintaining Links on the Web of Data. In *Proceedings of the 8th International Semantic Web Conference(ISWC) (ISWC '09)*. Springer-Verlag, Berlin, Heidelberg, 650–665.