NSFC-61261130588

ANR-12-IS02-002-01

Lindicle

Linked data interlinking in a cross-lingual environment
跨语言环境中语义链接关键技术研究
Liage des données dans un environnement interlingue

D3.1 Context-based ontology matching and data interlinking

Coordinator: Jérôme Euzenat
With contributions from: Jérôme Euzenat, Jérôme David, Angela Locoro, Armen Inants

Quality reviewer:	Jérôme Euzenat
Reference:	Lindicle/D3.1/v6
Project:	Lindicle ANR-NSFC Joint project
Date:	July 8, 2015
Version:	6
State:	final
Destination:	public

Executive Summary

We introduced a distinction in ontology matching between content-based ontology matching and context-based ontology matching. The former compares the content of ontologies to decide which entities are alike others; the latter considers the relations that ontology entities entertain with other resources.

The context-based approach is very well suited to matching multilingual resources since it does not consider the linguistic manifestation of concepts which is part of the content. It however requires relations with other resources.

We first introduce the concept of context-based ontology matching by reviewing early work and providing a general framework for this matching approach.

We then describe more precisely the instantiation of this approach as path-based context matching which relies on algebras of relations for providing precise matching results.

Finally, we discuss the application of context-based techniques to data interlinking. We show how it can be used through finding paths across different data sets. However, the type of relations in data sets is relatively limited so far. It can be extended by designing algebras of relations which encompasses ontology and data relations. These allow for inferring relations across data sets, ontologies, alignments and link sets. We illustrate the use of such techniques for finding inconsistent link sets.

The first part of this deliverable has been published in [Euzenat and Shvaiko 2013] and [Locoro et al. 2014]. Element related to data interlinking are discussed in more depth in [Inants and Euzenat 2015].

Document Information

Project number	ANR-NSFC Joint project	**Acronym**	Lindicle
Full Title	跨语言环境中语义链接关键技术研究 Linked data interlinking in a cross-lingual environment Liage des données dans un environnement interlingue		
Project URL	`http://lindicle.inrialpes.fr/`		
Document URL			

Deliverable	**Number**	3.1	**Title**	Context-based ontology matching and data interlinking
Work Package	**Number**	3	**Title**	Cross lingual ontology matching based on aligned cross lingual human-readable knowledge bases

Date of Delivery	**Contractual**	M24	**Actual**	2015-07-07
Status	final		final ☒	
Nature	prototype ☐ report ☒ dissemination ☐			
Dissemination level	public ☒ consortium ☐			

Authors (Partner)	Jérôme Euzenat, Jérôme David, Angela Locoro, Armen Inants			
Resp. Author	**Name**	Jérôme Euzenat	**E-mail**	Jerome.Euzenat@inria.fr
	Partner	INRIA		

Abstract (for dissemination)	Context-based matching finds correspondences between entities from two ontologies by relating them to other resources. A general view of context-based matching is designed by analysing existing such matchers. This view is instantiated in a path-driven approach that (a) anchors the ontologies to external ontologies, (b) finds sequences of entities (path) that relate entities to match within and across these resources, and (c) uses algebras of relations for combining the relations obtained along these paths. Parameters governing such a system are identified and made explicit. We discuss the extension of this approach to data interlinking and its benefit to cross-lingual data interlinking. First, this extension would require an hybrid algebra of relation that combines relations between individual and classes. However, such an algebra may not be particularly useful in practice as only in a few restricted case it could conclude that two individuals are the same. But it can be used for finding mistakes in link sets.
Keywords	Context-based data interlinking, Multilingual data interlinking, Context-based ontology matching, Algebras of relations, Semantic web

Version Log			
Issue Date	**Rev No.**	**Author**	**Change**
2014-09-20	1	J. Euzenat	Initial outline
2014-09-21	2	J. David	Included context-based ontology matching description
2015-06-19	3	J. Euzenat	Summarized experiments
2015-06-24	4	J. Euzenat	Elements on context-based data interlinking
2015-06-29	5	J. Euzenat	Full revision
2015-07-07	6	J. Euzenat	Fixed examples in Data interlinking + executive summary

Table of Contents

1. Introduction

The Semantic Web relies on the expression of formalized knowledge on the Web. Data is expressed in the framework of ontologies (theories describing the vocabulary used for expressing data). However, due to the decentralisation of the Web, ontologies may be heterogeneous and have to be reconciled. One way to reconcile ontologies is to find correspondences between their entities. This is called ontology matching [Euzenat and Shvaiko 2013] and the resulting set of correspondences is called an alignment. Each correspondence relates entities from each of the ontologies with a particular relation, e.g., equivalence, subsumption.

Context-based ontology matching works by taking advantage of intermediate resources to which the two ontologies to be matched can be connected. This is in contrast with content-based matchers, which compare the content of ontologies for matching them, whereas context-based matching use relationships, called anchors, between the entities of the ontologies to be matched and other ontologies on the web. For instance, in Figure 2.1, Beef from the Agrovoc thesaurus and Food from the NAL thesaurus are anchored to the concepts with the same names in the TAP ontology. Then because Food subsumes Beef in TAP, it is assumed that Food from NAL also subsumes Beef from Agrovoc.

Context-based matchers have already been shown beneficial [Sabou et al. 2008a; Mascardi et al. 2010; Locoro et al. 2014]. However, there is a wide latitude in their design: They depend on the type of resources to be considered (ontologies, encyclopedia, fully informal resources, etc.), how relations are obtained within these resources (asserted, inferred, etc.), how many will be considered (the first one that provides a result or as many as possible), how entities are anchored (simple or complex matchers), and how results are combined when there are several correspondences for the same pair of entities (by vote, by conjunction, etc.). We provide a general framework highlighting these aspects.

The goal of this report is to better explain the influence of some of these parameters on the quality of the resulting alignments. For that purpose, we design a flexible context-based matcher that offers various ways to parameterise its behaviour. This renders explicit the various options and allows us to combine them.

The same type of principles put forward for context-based ontology matching may be used for data interlinking. This may be particularly valuable in a cross-lingual environment as context-based techniques do not have to deal with language-based data. We consider here how this may be further developed.

This deliverable is organised as follows: §2 introduces and synthesises the state-of-the-art in context-based matching. §3 presents the architecture of a system, generalising Scarlet, along with the main operations of context-based matching and how they have been parameterised. §4 discusses the application of context-based approach to data interlinking. §5 concludes and outlines future directions for this work.

2. Context-based matching

Ontology matching must identify relations between ontology entities from two ontologies. These are returned as correspondences of the form $\langle e, r, e' \rangle$ such that e is an entity from the first ontology, e' is an entity from the second ontology and r is the relation assumed to hold between them. Often, matchers associate a measure of their confidence with each correspondence they return. In the following, we consider correspondences between named ontology entities (classes, properties, etc.). Relations may be subsumption ($<$ and $>$ and their reflexive versions $\leq$ and $\geq$), equivalence ($=$) or disjointness ($\perp$) between these entities.

Context-based matching contrasts with content-based matching. Matching ontologies with content-based techniques compares ontology entities (classes, properties) by relying only on its internal content, such as their annotations, structures, and/or semantics. For the same purpose, context-based matching also uses the context of these ontologies, e.g., resources that they annotate, and message exchanges between agents that use them. For instance, Figure 2.1 shows two entities from the Agrovoc (FAO)[1] and NAL (US DoA)[2] thesauri that had to be matched in the *food* test case of OAEI-2007 [Euzenat et al. 2007]. When considering concepts `Beef` and `Food`, the use of ontologies found on the Web, such as the TAP[3] ontology, helps deduce that `Beef` is less general than `Food`. The same result can also be obtained with the help of WordNet since `Beef` is a hyponym (is a kind) of `Food`. Thus, multiple sources of background knowledge can simultaneously help.

2.1 Early work on context-based ontology matching

Context can take different forms, such as web pages or pictures that have been annotated with the concepts of an ontology [Stumme and Mädche 2001]. It can also be some general purpose resource such as a dictionary (WordNet is very often used in ontology matchers).

We concentrate here on systems that use ontological resources as context for matching. By ontological resources, we mean ontologies or knowledge bases, e.g., formalised data sets. Even with this restriction, several context-based ontology matchers have been elaborated over the years:

- using domain specific ontologies, e.g., in the field of anatomy [Zhang and Bodenreider 2007; Aleksovski 2008];
- using upper-level ontologies [Mascardi et al. 2010; Jain et al. 2011];
- using linked data as background knowledge [Jain et al. 2010; Hu et al. 2011];
- using all the ontologies available on the Semantic Web, such as in the work on Scarlet [Sabou et al. 2008a].

By focusing on a specific domain, such as in [Aleksovski 2008] and [Zhang and Bodenreider 2007], authors were able to provide deeper insights on ontology concept similarities, especially based on the analysis of its respective structural relations, i.e., not only hierarchical, but also relational in its broadest sense (for example by means of the *partOf* relation), or by approximating matching measures when different local hierarchies contain the same concept or group of concepts.

In [Mascardi et al. 2010] general purpose upper ontologies are exploited to match ontologies by relating entities if and only if they have the same upper level context. GeRoMeSuite

[1] `http://www.fao.org/aims/ag_intro.htm`.

[2] `http://www.nal.usda.gov/`.

[3] `http://139.91.183.30:9090/RDF/VRP/Examples/tap.rdf`.

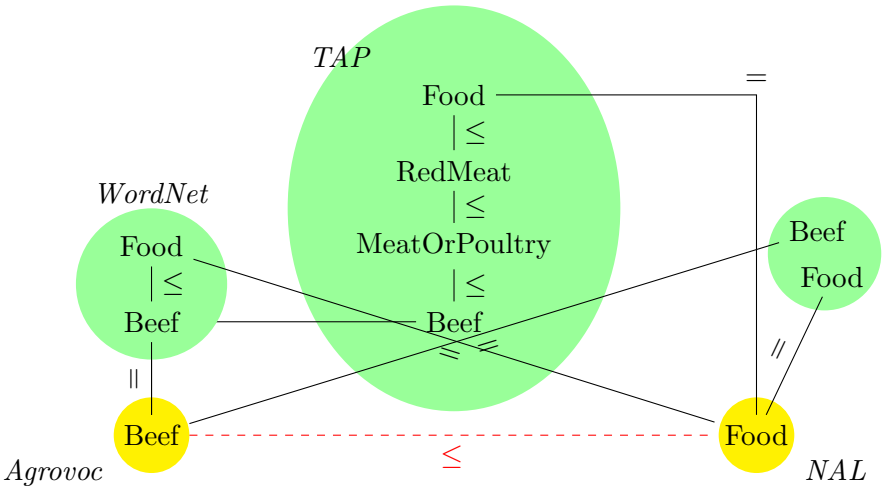

Figure 2.1: Scarlet example: several results are returned and must be aggregated (adapted from [Sabou et al. 2008a]). Two paths are found relating Beef in Agrovoc to Food in NAL and WordNet. The aggregation of their relations indicates that the former is more specific than the latter.

has been extended to select several intermediate ontologies before performing matching [Quix et al. 2011].

The BLOOMS system [Jain et al. 2010] is a first attempt to use Linked Open Data (LOD)[4] for schema-level matching. It tries to connect categories coming from two schemas, transform them in trees of senses for each concept to be matched, and compare such trees of senses for discovering hierarchical relations between such concepts. Its evolution, BLOOMS+ [Jain et al. 2011], exploits the Proton upper-level ontology to enhance the LOD schema-level matching task.

Scarlet [Sabou et al. 2008a] tries to find a relation between two concepts by using all the ontologies on the Web for discovering relational paths that connect them. It is presented in more details in §2.2. In [Hu et al. 2011], a macro scale analysis of thousands of mapped ontologies is carried out in order to detect morphological features as well as power distribution laws in the resulting graphs. In this way, some hints on what exists now and on how to organise and evolve existing knowledge on the Web by means of forthcoming ontologies are provided.

The difficulty of context-based matching is a matter of balance: adding context provides new information, and hence, helps increase recall, but this new information may also generate incorrect correspondences which decrease precision. However, we showed that carefully tuned context-based matching may actually provide more precise results [Locoro et al. 2014].

As can be observed, there are various ways to use ontological resources for context-based ontology matching. Many options can be taken concerning the type of resource to be used or the way it is connected to the ontologies to be matched. Our goal is to explore these options. For that purpose, we decided to extend an existing ontology matcher.

[4] `http://linkeddata.org/`.

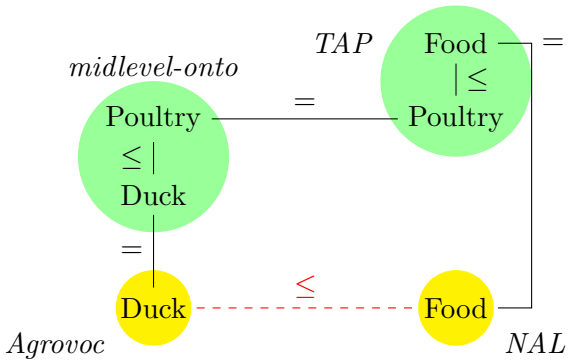

Figure 2.2: Scarlet composition example (adapted from [Sabou et al. 2008a]). There is no intermediate ontology providing a correspondence between Duck and Food. However, two intermediate ontologies (midlevel-onto and NAL) provide a path between these concepts through the Poultry concept. The relations along this path show that Duck is more specific than Food.

2.2 Scarlet

Our starting point was Scarlet[5] [Sabou et al. 2008a] because it already took into account the versatility of context-based matching.

Scarlet [Sabou et al. 2008a; Sabou et al. 2008b] is an ontology matcher that operates by contextualising ontologies with ontologies that can be found on the Web. But Scarlet is more complex than this definition, since it involves selecting ontologies for context, matching entities from the initial ontologies and those of the context, and composing the relations obtained after matching. The rationale behind Scarlet is that using more ontologies improves the results. The problem raised by the heterogeneity of ontologies is solved by taking advantage of these many heterogeneous ontologies, which is based on the following principles:

- using the ontologies on the Web as context;
- composing the relations obtained through these ontologies: this covers reasoning within the ontology for deducing the relations between entities (Figure 2.1) or reasoning across ontologies (Figure 4.2).

In more details, Scarlet processing roughly consists of the following steps:

1. harvest ontologies on the Web with either Swoogle [Ding et al. 2005] or Watson [d'Aquin and Motta 2011];
2. select those which are related to the ontologies to match: usually this is achieved by selecting, for each pair of named entities, the ontologies that contain both names;
3. find anchors between the ontologies to match and those that have been selected: here Scarlet uses simple string equivalence;
4. compose the relations between entities through the intermediate ontologies: this is done by returning the relation found in the ontology (see Figure 4.2);
5. aggregate the obtained results (see Figure 2.1).

When no ontology contains the pair of terms, another implemented variation was to use several ontologies and to bridge them in order to increase the chances to find the pair of terms (see Figure 4.2).

[5] `http://scarlet.open.ac.uk/`.

This can become a very complex procedure so it is restricted to finding, for each pair of ontologies, the intersection between the entities subsuming one term and those subsumed by the other, which helps quickly find subsumption relations (see Figure 3.1).

Three variants of Scarlet have been experimented against Agrovoc (FAO) and NAL (US DoA). The considered variants were:

- $S1$ works with only one intermediate ontology at a time: it retrieves the ontologies covering both candidate terms from both ontologies, and delivers all the correspondences that it finds between matched concepts (Figure 2.1);
- $S1'$ is like S1 but it stops at the first correspondence that it finds;
- $S2$ implements path search in the graph of ontologies (Figure 4.2), but only through direct subsumers (and no subsumees).

A sketch of $S1$ and $S2$ strategies is reported in Figure 2.3.

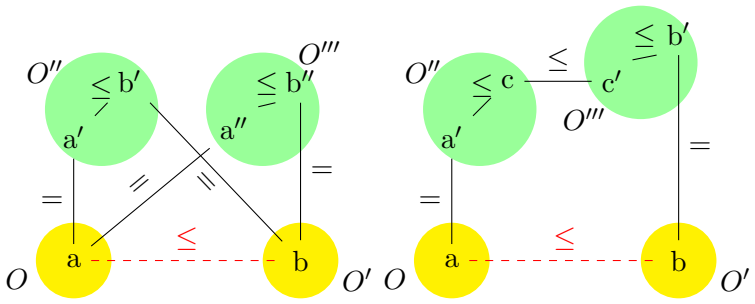

Figure 2.3: Ontology matching (left) within one ontology ($S1$) and (right) across ontologies ($S2$).

In all cases, the search for anchors was provided by strict string matching on terms as bags of words, and candidate ontologies were provided by Swoogle. Because of the lack of a full reference alignment in the data set, results were manually assessed and only reported on precision. They provide an average value of 70% precision. This is expected with the given anchoring strategy, indeed, anchoring with string equivalence usually provides high precision. This result has even been improved by using word-sense disambiguation techniques, which allow for better discriminating similar terms [Gracia et al. 2007]. However, this is rather good given that Scarlet returns subsumption relations.

We went on by further generalising the Scarlet approach [Euzenat and Shvaiko 2013; Locoro et al. 2014].

2.3 A generalised view of context-based matching

Because context-based matching is very versatile, we synthesise its behaviour in a generalised view that aims at covering and extending existing matchers. For that purpose, we decompose the context-based matching process in 7 steps described in Figure 2.4:

Ontology arrangement preselects and ranks the ontologies to be explored as intermediate ontologies. The preselection may retain all the ontologies from the Web or ontologies belonging to a particular type, such as upper ontologies, domain dependent ontologies, e.g., medical or biological ontologies, competencies, popular ontologies, recommended ontologies, or any customised set of ontologies.

The ordering may be based on the likeliness for the ontology to be useful, usually measured by a distance. Such a distance may be based on the proximity of the ontology with the ontology to be matched [David and Euzenat 2008], the existence of alignments between them [David et al. 2010], or the availability of quickly computable anchors.

Contextualisation, or anchoring, finds anchors between the ontologies to be matched and the candidate intermediate ontologies. These anchors are obtained through an ontology matching method or by using existing alignments. They can be correspondences of any types including various relations and confidence measures. In principle, any ontology matching method may be used for anchoring; in practice, this is usually a fast method because anchoring is only a preliminary step.

Ontology selection restricts the candidate ontologies that will actually be used. This selection relies usually on the computed anchors by selecting those ontologies in which anchors are present.

Local inference obtains relations between entities of a single ontology. It may be reduced to logical entailment. It may also use weaker procedures, especially when intermediate resources have no formal semantics, e.g., thesauri. It could then be replaced by the use of asserted relations of the ontologies or relations obtained through composing existing ones.

Global inference finds relations between two concepts of the ontologies to be matched by concatenating relations obtained from local inference and correspondences across intermediate ontologies

Composition determines the relation holding between the source and target entities by composing the relations in the path (sequence of relations) connecting them. The composition method may be functional ($= \cdot = \text{ is } =$), order-based ($< \cdot \leq \text{ is } <$) or relational ($\perp \cdot \geq \text{ is } \perp$).

Aggregation combines relations obtained between the same pair of entities. It can either simply return all correspondences or return only one correspondence with an aggregated relation. Aggregation itself can be based on various methods such as relation aggregation operators (e.g., conjunction), popularity (selecting the relation which is obtained from the most paths) or confidence (selecting the relation with the highest confidence).

These steps extend those provided in the descriptions of Scarlet [Sabou et al. 2008a]: contextualisation was called anchoring, selection was considered, local and global inference as well as composition were gathered in a set of "derivation rules" and aggregation was called combining. GeRoMeSuite has also identified the arrangement (called selection), anchoring, local inference (including composition), and aggregation steps [Quix et al. 2011] to which a consistency check is added. This presentation provides a finer decomposition of context-based matching that can be used for instantiating differently each (optional) step.

We may see context-based matching under a fully logical point of view: local and global inference are replaced by entailment tests and composition and aggregation are replaced by logical deduction. In such a case, beyond anchoring, matching is reduced to reasoning in a network of ontologies. Hence, when the technology of reasoning in networks of ontologies will be fully developed, it will be possible, in principle, to reduce the seven steps to anchoring and reasoning. Matchers such as LogMap [Jiménez-Ruiz and Cuenca Grau 2011] currently apply this, but only between the two ontologies to match.

Such a framework is intellectually very seducing and mostly compatible with the framework proposed above. Indeed, local inference, relation composition and relation aggregation are approximations of their logical counterpart. Only global inference may be too local for fully approximating entailment in a network of ontologies.

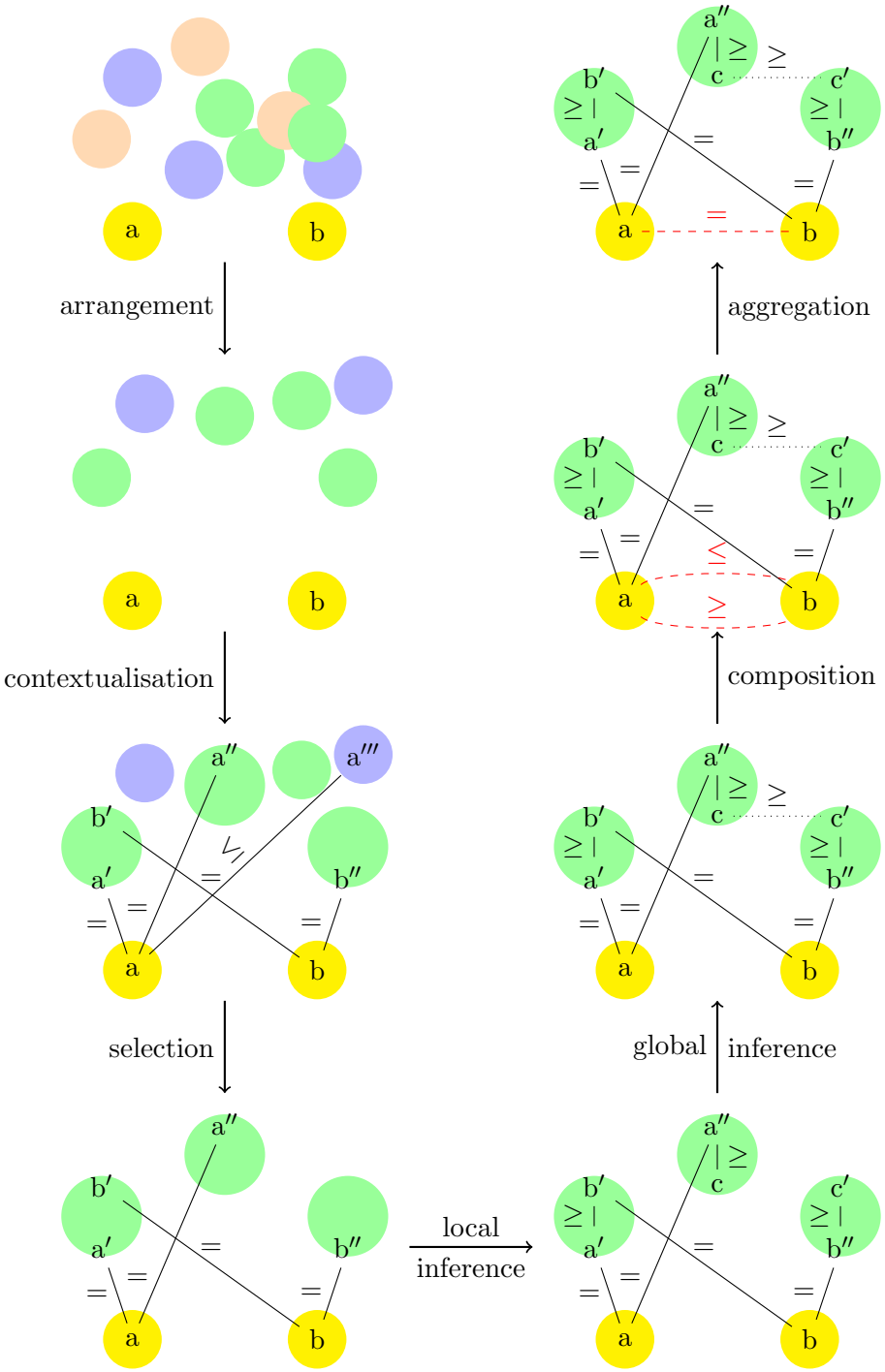

Figure 2.4: The different steps of context-based matching (from [Locoro et al. 2014]).

3. Path-driven context-based matching

A new version of Scarlet, named Scarlet 2.0, has been developed along the framework of the previous section [Locoro et al. 2014]. Its characteristics are as follows:

- it still takes advantage of Watson [d'Aquin and Motta 2011; d'Aquin et al. 2007] giving access to the ontologies of the Web;
- like the initial Scarlet, it uses intensively a path traversal strategy,
- it uses algebras of relations for expressing the relationships between concepts,
- it offers precise parameterisation, so as to study the influence of their values.

We describe this approach as path-driven because the implementation uses the notion of paths, i.e., it considers ontologies and alignments as graphs whose ontology entities are the nodes and the statements and correspondences are the edges. In this setting, matching two concepts consists of (a) finding a path in this graph between them, and (b) computing the relation carried by this path. For instance, in Figure 2.1, there are two paths, one of which is agrovoc:Beef = tap:Beef $\leq$ tap:MeatOrPoultry $\leq$ tap:ReadMeat $\leq$ tap:Food = nal:Food. The composition of the relations in the edges of this path yields $\leq$ as the relation between agrovoc:Beef and nal:Food.

The reason for considering the same restricted framework as Scarlet is that it is possible to control precisely the way the algorithm explores the search space (through ontology selection or limitation of its exploration). Introducing more sophisticated methods, either for anchoring or for inferring, remains mostly possible. We avoided it in order to obtain clear initial observations in the presence of simple methods.

3.1 General overview and parameters

We describe below the techniques implemented in Scarlet 2.0 with respect to the framework of Section 2.3. The parameters governing the behaviour of the system are identified (in *italics*) and their further values are provided in Table 3.1.

Ontology arrangement does not do any preselection and potentially considers all ontologies from the Web as provided by Watson.

Contextualisation, uses a simple matching method. This step is parameterised by the ontology *matching method* used for anchoring. It does not take advantage of confidence measures. Scarlet 2.0 can use any matcher implementing the Alignment API[1]. In this experiment, we will only use a simple token-based string equality (each label is reduced to a set of tokens which are compared with string equality).

Ontology selection is governed by two thresholds on the number of anchors that have to be found between the ontologies to be matched. A first parameter called *minimum local anchors*, is the minimal number of pairs of ontology entities that have anchors in an ontology. A second parameter, *minimum global anchors*, is the minimal total number of anchors found in an intermediate ontology. Obviously, if the first value is greater than or equal to the second one, then the second one is useless. If both values are 0, then all ontologies are selected.

Local inference is implemented by local path exploration: it traverses an intermediate ontology to retrieve paths, i.e., sequences of asserted relations between entities. In this

[1] `http://alignapi.gforge.inria.fr/`.

implementation, it will attempt at finding paths between anchors, or finding subsumption paths of a given length around anchors (for global inference). This exploration process uses three parameters: (1) the *maximum local path length* for restricting the length of the exploration; (2) the *exploration type* for determining which types of relations are followed; (3) the *selection method* for selecting which paths between a pair of entities have to be retained, e.g., the first one, the shortest one, all of them.

Global inference is implemented by global path exploration, i.e., it generates paths between two concepts of the ontologies to be matched by concatenating various local paths from distinct ontologies, such that the concept at the end of each local path is anchored to the concept at the beginning of the next local path. The *maximum global path length* parameter determines the maximal number of ontologies that may be traversed to return a relation between two entities. If this is 0, then the algorithm is in the case of classical (content-based) ontology matching, and matching will be reduced to anchoring. Like before, the *selection method* indicates which paths are selected, e.g., the first one, the shortest one, or all of them. The graphs traversal algorithm is further presented in §3.2.

Composition In this approach, the *composition method* used for composing relations is the standard composition of algebras of relations (see §3.3 for details).

Aggregation relies on an *aggregation method* for aggregating the relations obtained between the same pair of entities. This is either an algebraic operation such as conjunction or disjunction, e.g., the conjunction between $\leq$ and $\geq$ is $=$, though their disjunction is $<, =, >$, or popularity aggregation, which selects the relation obtained from the most paths.

Table 3.1 summarises the parameters identified at each step of this process and the different values that they can take. It also provides approximate values for reproducing the original Scarlet strategies.

We present in more detail three aspects of this procedure: graphs traversal (§3.2), relation composition and aggregation using algebras of relations (§3.3), and minimal path reduction (§3.4).

3.2 Global inference through context traversal

For all pairs of concepts for which a correspondence could not be found in any of the intermediary ontologies used during the context-matching operation, global inference can connect the paths obtained in several context ontologies. We call:

0-context traversal content-based matching;
1-context traversal context-based matching using only one context ontologies;
n-context traversal context-based matching using at most n intermediate ontologies.

We describe the behaviour of 2-context traversal, which traverses two intermediary ontologies. Given two concepts $a \in o$ and $b \in o'$ and their respective set of intermediary ontologies O_a and O_b to which they are anchored, for any pair of ontologies $\langle o_a, o_b \rangle \in O_a \times O_b$, the 2-context traversal algorithm looks if there exists an anchor $\langle c_a, =, c_b \rangle$ between them such that:

1. c_a is either a subsumer or a subsumee of $a \in o_a$, found by exploring o_a until a given path length;

Step	**Parameter**	**Value**	$S1$	$S'1$	$S2$
Arrangement	ontologies	web upper-level ontologies specific domain ontologies specific ontology	√	√	√
Contextualisation	matching method	a matching method (=token based similarity)	string equality		
Selection	minimum local anchors	positive integer (=0)	0	0	0
	minimum global anchors	positive integer (0-10)	0	0	0
Local Inference	maximum local path length	positive integer ([0..4])	∞	∞	1
	exploration type	subsumption disjointness complete	√	√	√
	selection method	all first shortest	√	√	√
Global Inference	maximum global path length	positive integer (0,1,2)	1	1	2
	selection method	all first shortest	√	√	
Composition	composition method	functional order-based relational	√	√	√
Aggregation	aggregation method	none conjunctive disjunctive popularity	√		√

Table 3.1: List of the possible parameters at each step, whose combination generates a new matcher.

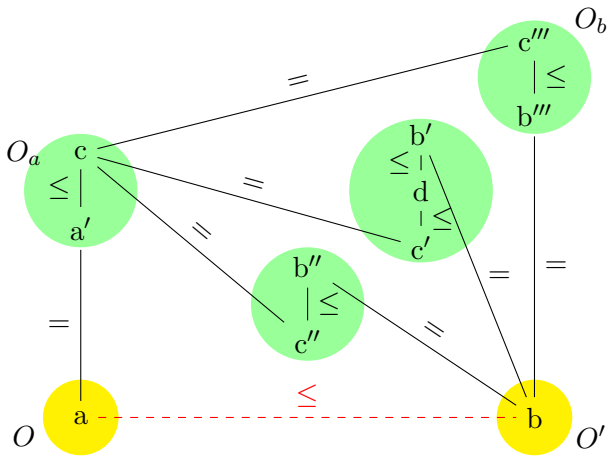

Figure 3.1: 2-context traversal. Of the three paths found between a and b, two return the $\leq$ relation and the third one returns all the possible relations (Γ).

2. c_b is either a subsumer or a subsumee of $b \in o_b$, found by exploring o_b until a given path length.

Once such an anchor is found, the path selection, composition and aggregation for the pair put in correspondence are applied in the usual way. A visual sketch of the algorithm is depicted in Figure 3.1.

The difference between the 2-context traversal strategy and strategy $S2$ of Scarlet lies in that the intermediate concept c is now searched through the whole length of a local path instead of stopping at the direct subsumers or subsumees of the concepts to be matched.

3.3 Composing paths and aggregating correspondences

One of the benefits of an approach exploring different paths for finding relations is that it may return several possible relations between two entities. Such relations may confirm or contradict each other and this has to be considered in the aggregation step. For this reason, we use an algebra of alignment relations [Euzenat 2008; Inants and Euzenat 2015] that structures the set of possible relations.

Such an algebra of relations is based on a set of jointly exhaustive and pairwise disjoint relations Γ. This means that for any pair of concepts, there exists exactly one relation in Γ, which characterises their relative positions. Algebras of relations allow for expressing uncertainty through the use of subsets of Γ, which are interpreted disjunctively. For instance, the often used relation $\leq$, standing for "subsumed by" or "equals to", really stands for $\{<, =\}$, the disjunction of "subsumed by" and "equals to". The complete lattice of the 31 disjunctive relations of such an algebra is reported in [Euzenat 2008]. For our purposes, some of them have been named. Table 3.2 shows the complete list of named relations and some of the disjunctive combinations, along with their symbol, a description of their interpretation, and the short label used to refer to them in the rest of this report.

The benefits of algebras of relations is that they provide well-defined aggregation operators: conjunction or disjunction of such relations correspond to set intersection and union, respectively. Hence, the relations obtained through traversing the graph of ontologies may be aggregated with:

Set of relations	Short Label (symbol)	Description
$=$	equiv ($=$)	equivalence relation
$<$	subClass ($<$)	strict subsumption relation
$>$	superClass ($>$)	strict inverse subsumption relation
$\between$	overlaps ($\between$)	overlaps relation
$\perp$	disjoint ($\perp$)	disjoint relation
$>, =$	subsumesOrEqual ($\geq$)	subsumes or equivalent relations
$<, =$	subsumedOrEqual ($\leq$)	is subsumed or equivalent relations
$>, \between$	subsumesOverlap	subsumes or overlaps relations
$<, \between$	subsumedOverlaps	is subsumed or overlaps relations
$>, \between, \perp$	notSubsumed ($\not\leq$)	is not subsumed relation
$<, \between, \perp$	notSubsumes ($\not\geq$)	does not subsume relation
$>, <, \between, =$	notIncompatible ($\not\perp$)	not disjoint relation
...		other combinations obtained by disjunction or conjunction
$<, >, \between, =, \perp$	all (Γ)	all relations

Table 3.2: Relation symbols that may result from a composition or aggregation operation for the algebra of alignment relations. The first part of the table features the 5 base relations between concepts.

conjunction if we consider that each path provides an exact, but non precise, relation and that several paths contribute precising it. When the conjunction gives the $\emptyset$ relation, the resulting correspondence is inconsistent.

disjunction if we consider that each path provides a possible relation without excluding the others.

An alternative aggregation method, independent from the algebra, is the popularity method, which retains the most frequent relation in the set of correspondences between a pair of entities. If several relations have the same popularity, then they are disjunctively aggregated.

Algebras of relations also provide a composition operation ($\cdot$), usually based on a table. For instance, $\{>, =\} \cdot \{>\}$ is $\{>\}$ and $\{>, =\} \cdot \{<\}$ is $\not\perp$. This operation is important in context traversal. These traversals return paths which carry sequences of relations between concepts. The composition operator reduces this sequence to a relation preserving as much information as possible.

For instance, a real path found by the system is:
BodyOfWater = BodyOfWater $\geq$ FreshWaterLake $\leq$ Lake = Lake
whose composition brings to the notIncompatible relation ($\not\perp$). Intuitively, this means that if BodyOfWater and Lake are two concepts with a sub-concept in common, viz., FreshWaterLake, they should not be disjoint (because in this algebra concepts are assumed non empty). Thanks to composition, the information that the two concepts are not disjoint is preserved.

3.4 Minimal path reduction in path concatenation

During the path exploration procedure, it may happen that paths are extensions of shorter paths. Figure 3.2 shows an example of two such paths for the same pair of concepts, one path (Path 2) being the extension of the other (Path 1). They are:

1. Hotel = Hotel $\leq$ ResidenceBuilding = ResidenceBuilding $\leq$ Residence = Residence, which by composition ($= \cdot \leq \cdot = \cdot \leq \cdot =$) yields $\leq$

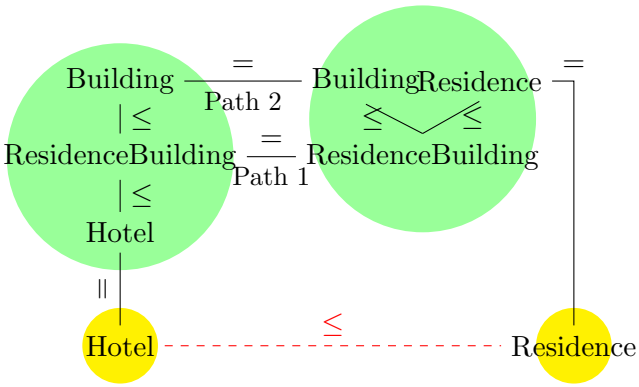

Figure 3.2: Two alternative paths between the same pair or concepts. One path is the extension of the other.

2. Hotel = Hotel ≤ ResidenceBuilding ≤ Building = Building ≥ ResidenceBuilding ≤ Residence = Residence, which by composition ($= \cdot \leq \cdot \leq \cdot = \cdot \geq \cdot \leq \cdot =$) yields Γ.

If both paths are retained, they will be aggregated:

- conjunction gives the final correspondence ⟨Hotel ≤ Residence⟩;
- disjunction gives the final correspondence ⟨Hotel Γ Residence⟩;
- popularity based aggregation would result in both final correspondences ⟨Hotel ≤ Residence⟩, and ⟨Hotel Γ Residence⟩, as they are equally occurring. In this case a disjunction of both the final relations is computed, the final correspondence being ⟨Hotel Γ Residence⟩.

So, there is a risk of having non precise correspondences if all such paths are gathered and it is preferable to select them. There may be several ways to do it:

- select the one that goes across less ontologies: both paths traverse two ontologies, so in this example both paths 1 and 2 would be selected;
- select the shortest one: the former is the shortest, so in this example the path selected would be path 1;
- select the most precise one: the former is the most precise one because $\leq \subseteq \Gamma$.

In our case, a procedure for always selecting the shortest path between the source and the target concepts is applied.

3.5 Summary of experiments

In [Locoro et al. 2014], we conducted a pinpoint analysis on context-based matching by varying some of these parameters.

These experiments establish general observations on the behaviour of such systems, and confirm what was previously observed:

- Not restricting the considered ontologies provides significantly more correspondences than selecting them a priori and this increases F-measure, although precision decreases.
- Increasing global and local path length also provides more correspondences and increases F-measure; the effect of local path length increase is higher than that of global path length.
- Ontology selection is the main parameter impacting time performance.

Algebras of relations allowed for finely characterising the added benefits of these parameter values from the standpoint of the correctness of returned correspondences and the influence of the type of correspondences on this correctness. The observations are as follows:

- As paths get longer, new correct correspondences are still found;
- As paths get longer, correct correspondences may become non precise by additional relations;
- As paths get longer, incorrect correspondences do not become more correct and imprecise correspondences do not become more precise.

In summary, these experiments show once again that context-based ontology matching increases the quality of obtained results through multiplying sources of information. Even if conjunction obtains the best results, it seems that finer strategies could still improve the quality of alignments.

We plan to further develop the implementation and investigate more configurations in more situations. Developing and testing alternative aggregation strategies will also be an outcome of this work.

We disregarded confidence measures returned by matchers. They could be considered at each step of the framework and combined with relations [Euzenat 2008; Atencia et al. 2012] for refining the obtained results. Similarly, logical reasoning may be integrated within context-based matching.

4. Context-based data interlinking

The same principles that have been applied to context-based ontology matching could be considered for data interlinking. This could be especially beneficial in cross-lingual matching/interlinking since there is no need that matched resources use the same natural language. We discuss below how these principles may, or may not, be adapted.

4.1 Path-based data interlinking

The first application simply consists of composing relations between individuals of various data sets. `owl:sameAs` and `owl:differentFrom` are such types of relations between individuals. We will note them as $=$ and $\neq$ respectively. The composition table of such relations may be used in order to generate new ones. Indeed, their well known composition table is displayed in Table 4.1.

$r \backslash r'$	$=$	$\neq$
$=$	$=$	$\neq$
$\neq$	$\neq$	$=, \neq$

Table 4.1: $\mathbb{A}_2$ composition table.

Figure 4.1 provides a simple example of using an authority list for interlinking data. Indeed, Viaf is a sort of meta-authority that aggregates authors of many national libraries. Hence, it is like a huge link set. The DBPedia entry for Emilly Brontë is asserted as the same as VIAF entry 97097302 which is itself assured as the same as the entry for Ellis Bell in the German national library (Deutschen Nationalbibliothek) 1061106306. This allows to deduce the equivalence between these two entities.

Figure 4.2 generalizes this approach by taking advantage of more data sets showing that links across data sets may be used in order to link chinese DBPedia for Emily Brontë to the Ellis Bell entry in the German National Library.

We plan to apply this approach to data interlinking when there is sufficient intermediary data sets.

4.2 Context-based data interlinking through ontologies

In principle, whatever has been considered for context-based ontology matching applies to data interlinking: it should be possible to compose relations between individuals and classes freely. There are relations between classes [Euzenat 2008], relations between individuals (§4.1) as well as relations between classes and individuals such as `rdf:type` ($\in$, $\ni$).

However, there are no immediate composition tables across heterogenous domains. We have investigated this topic both from the perspective of these specific algebras of relations [Inants and Euzenat 2015] and more generally. The resulting composition table can be found in Table 4.2.

For instance, in Figure 4.3, the relation $\neq$ between the instances "Amanda Cross" and "Carolyn Gold Heilbrun" can be inferred by composition. Indeed, composing $\{\in\} \cdot \{<, =\} \cdot \{=\} \cdot \{\perp\} \cdot \{\ni\}$ is equivalent to $\{\in\} \cdot \{\perp\} \cdot \{\ni\}$ which actually yields $\{\neq\}$. More precisely, in the notation of $\mathbb{A}16$, $\{\in\} \diamond \{<, =_n\} \diamond \{=_n\} \diamond \{\|\} \diamond \{\ni\}$ can progressively reduced

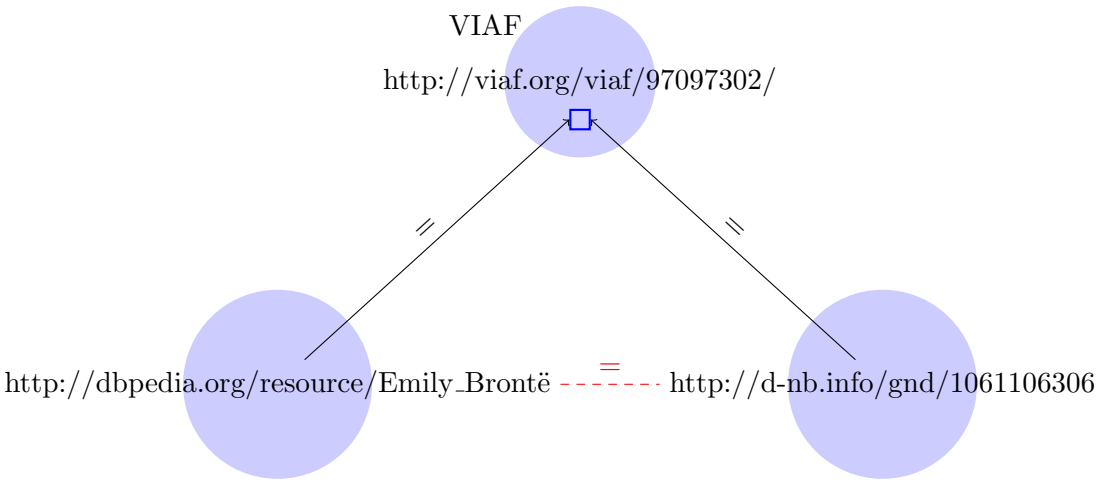

Figure 4.1: Context-based data interlinking using Viaf as a context. Because VIAF resolves both Emilly Brontë and Ellis Bell as the same person, the link can be found (this was already true of the initial sources actually).

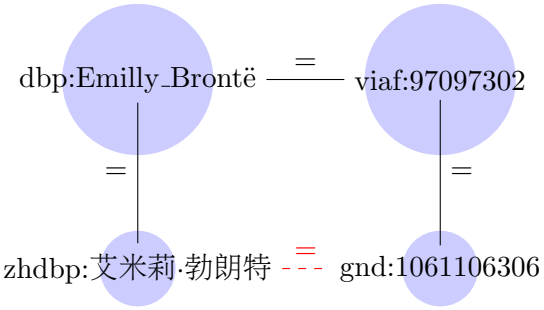

Figure 4.2: Path-based data interlinking using Viaf as a context.

to $\{\in, \in\}\diamond\{=_n\}\diamond\{\|\}\diamond\{\ni\}$, to $\{\in\}\diamond\{\|\}\diamond\{\ni\}$, to $\{\not\in_{in}\}\diamond\{\ni\}$, and finally to $\{\neq_i\}$. This is an instance of path-based data interlinking in which some relation can actually be inferred. This may be useful when such `owl:differentFrom` relations are needed; for instance, for extracting link keys [Atencia et al. 2014].

Unfortunately, this approach is unlikely to provide results as relations between classes are not constraining enough to provide precise relations between individuals (the resulting relation will, in most cases, be $\{=, \neq\}$).

4.3 Link set debugging through context

However, such reasoning may be used for debugging link sets. Indeed, if the composition of relations leads to the empty relation, then it is for sure that there is a problem.

For instance, in Figure 4.3 again, the relation $\{\neq\}$ has been inferred between "Amanda Cross" and "Carolyn Gold Heilbrun". However, they also have the = relation between them and $\{\neq\} \cap \{=\} = \varnothing$ showing that the displayed data is inconsistent.

There may be several ways to deal with this inconsistency: suppressing any of the involved relations, i.e., suppressing relations in any of the ontologies (Mystery novelist$\leq$Paperback

$\diamond$	$=_n$	$>$	$<$	$\between$	$\parallel$	$\ni$	$\not\ni_{ni}$	NE
$=_n$	$=_n$	$>$	$<$	$\between$	$\parallel$	$\ni$	$\not\ni_{ni}$	NE
$<$	$<$	$=_n><\between\parallel$	$<$	$<\between\parallel$	$\parallel$	$\ni\not\ni_{ni}$	$\not\ni_{ni}$	NE
$>$	$>$	$>$	$=_n><\between$	$>\between$	$>\between\parallel$	$\ni$	$\ni\not\ni_{ni}$	NE
$\between$	$\between$	$>\between\parallel$	$<\between$	$=_n><\between\parallel$	$>\between\parallel$	$\ni\not\ni_{ni}$	$\ni\not\ni_{ni}$	NE
$\parallel$	$\parallel$	$\parallel$	$<\between\parallel$	$<\between\parallel$	$=_n><\between\parallel$	$\not\ni_{ni}$	$\ni\not\ni_{ni}$	NE
$\in$	$\in$	$\in\notin_{in}$	$\in$	$\in\notin_{in}$	$\notin_{in}$	$=_i\neq_i$	$\neq_i$	IE
$\notin_{in}$	$\notin_{in}$	$\notin_{in}$	$\in\notin_{in}$	$\in\notin_{in}$	$\in\notin_{in}$	$\neq_i$	$=_i\neq_i$	IE
EN	EN	EN	EN	EN	EN	EI	EI	$=_e$

$\diamond$	$=_i$	$\neq_i$	$\in$	$\notin_{in}$	IE
$=_i$	$=_i$	$\neq_i$	$\in$	$\notin_{in}$	IE
$\neq_i$	$\neq_i$	$=_i\neq_i$	$\in\notin_{in}$	$\in\notin_{in}$	IE
$\ni$	$\ni$	$\ni\not\ni_{ni}$	$=_n><\between$	$>\between\parallel$	NE
$\not\ni_{ni}$	$\not\ni_{ni}$	$\ni\not\ni_{ni}$	$<\between\parallel$	$=_n><\between\parallel$	NE
EI	EI	EI	EN	EN	$=_e$

$\diamond$	$=_e$	EN	EI
$=_e$	$=_e$	EN	EI
NE	NE	$=_n><\between\parallel$	$\ni\not\ni_{ni}$
IE	IE	$\in\notin_{in}$	$=_i\neq_i$

Table 4.2: The class-instance algebra $\mathbb{A}16$ (from [Inants and Euzenat 2015]).

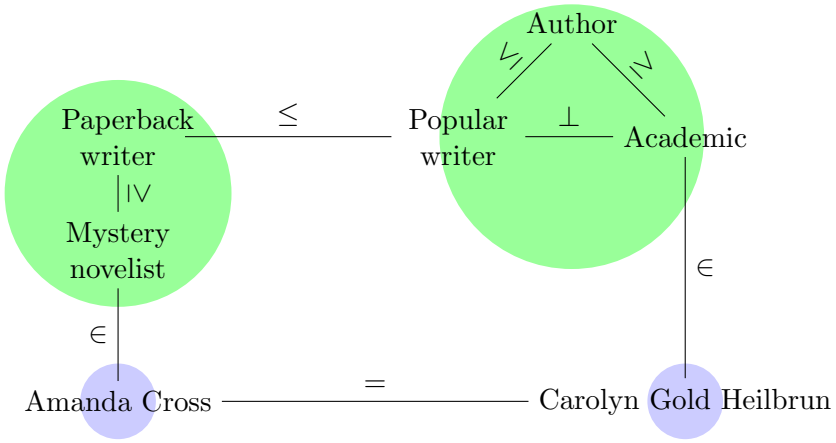

Figure 4.3: An example of inconsistent linked data sets that can be detected through simple composition of relation across ontologies, data, links and correspondences.

writer or Popular writer⊥Academic) in the alignment (Paperback writer=Author), across data sets and ontologies (Amanda Cross∈Mystery novelist or Carolyn Gold Heilbrun∈Academic) or the link (Amanda Cross=Carolyn Gold Heilbrun).

Hence, path-based reasoning may be used in order to find inconsistent sets of links (more generally, inconsistent networks of ontologies).

5. Conclusion

Context-based matching is based on the assumption that putting ontologies in the context of other ontologies may improve matching. In this report, we provided a framework identifying important steps of context-based ontology matching and parameters that may influence its behaviour.

We explained how such a framework may be extended to data interlinking but this requires to compose relations across classes and individuals. Such techniques may also be used for debugging ontologies, data sets, alignments and link sets. We plan to put the path based approach using links transitivity into practice by using either multilingual thesauri or the transitivity from XLore to dpbedia.fr through dbpedia.en.

BIBLIOGRAPHY

Aleksovski, Zharko (2008). "Using background knowledge in ontology matching". PhD thesis. Vrije Universiteit Amsterdam (cit. on p. 6).

Atencia, Manuel, Alexander Borgida, Jérôme Euzenat, Chiara Ghidini, and Luciano Serafini (2012). "A formal semantics for weighted ontology mappings". In: *Proc. 11th International semantic web conference (ISWC), Boston (MA US)*, pp. 17–33 (cit. on p. 19).

Atencia, Manuel, Jérôme David, and Jérôme Euzenat (2014). "Data interlinking through robust linkkey extraction". In: *Proc. 21st european conference on artificial intelligence (ECAI), Praha (CZ)*, pp. 15–20 (cit. on p. 21).

d'Aquin, Mathieu and Enrico Motta (2011). "Watson, more than a Semantic Web search engine". In: *Semantic web journal* 2.1, pp. 55–63 (cit. on pp. 8, 13).

d'Aquin, Mathieu, Claudio Baldassarre, Laurent Gridinoc, Sophia Angeletou, Marta Sabou, and Enrico Motta (2007). "Characterizing Knowledge on the Semantic Web with Watson". In: *EON*, pp. 1–10 (cit. on p. 13).

David, Jérôme, Jérôme Euzenat, and Ondrej Sváb-Zamazal (2010). "Ontology Similarity in the Alignment Space". In: *International Semantic Web Conference (1)*, pp. 129–144 (cit. on p. 10).

David, Jérôme and Jérôme Euzenat (2008). "On fixing semantic alignment evaluation measures". en. In: *Proc. 3rd ISWC workshop on ontology matching (OM), Karlsruhe (DE)*. Ed. by Pavel Shvaiko, Jérôme Euzenat, Fausto Giunchiglia, and Heiner Stuckenschmidt, pp. 25–36 (cit. on p. 10).

Ding, Li, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng, and Pranam Kolari (2005). "Finding and Ranking Knowledge on the Semantic Web". In: *Proc. 4th International Semantic Web Conference (ISWC)*, pp. 156–170 (cit. on p. 8).

Euzenat, Jérôme and Pavel Shvaiko (2013). *Ontology matching*. en. 2nd. Heidelberg (DE): Springer-Verlag. 520 pp. (cit. on pp. 2, 5, 9).

Euzenat, Jérôme, Antoine Isaac, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb, Vojtěch Svátek, Willem Robert van Hage, and Mikalai Yatskevich (2007). "Results of the Ontology Alignment Evaluation Initiative 2007". In: *Proc. 2nd International Workshop on Ontology Matching (OM) at the International Semantic Web Conference (ISWC) and Asian Semantic Web Conference (ASWC)*. Busan (KR), pp. 96–132 (cit. on p. 6).

Euzenat, Jérôme (2008). "Algebras of ontology alignment relations". en. In: *Proc. 7th international semantic web conference (ISWC), Karlsruhe (DE)*. Vol. 5318. Lecture notes in computer science, pp. 387–402 (cit. on pp. 16, 19, 20).

Gracia, Jorge, Vanessa Lopez, Mathieu d'Aquin, Marta Sabou, Enrico Motta, and Eduardo Mena (2007). "Solving Semantic Ambiguity to Improve Semantic Web based Ontology Matching". In: *Proc. 2nd ISWC Ontology Matching workshop (OM 2007), Busan (KR)*, pp. 1–12 (cit. on p. 9).

Hu, Wei, Jianfeng Chen, Hang Zhang, and Yuzhong Qu (2011). "How Matchable Are Four Thousand Ontologies on the Semantic Web". In: *Proc. 8th Extended Semantic Web Conference (ESWC)*, pp. 290–304 (cit. on pp. 6, 7).

Inants, Armen and Jérôme Euzenat (2015). "An Algebra of Qualitative Taxonomical Relations for Ontology Alignments". In: *Proc. 14th international semantic web conference (ISWC), Bethlehem (PA US)*. Lecture notes in computer science (cit. on pp. 2, 16, 20, 22).

Jain, Prateek, Pascal Hitzler, Amit Sheth, Kunal Verma, and Peter Yeh (2010). "Ontology alignment for linked open data". In: *Proc. 9th International Semantic Web Conference*

(ISWC). Vol. 6496. Lecture Notes in Computer Science. Shanghai (CN), pp. 401–416 (cit. on pp. 6, 7).

Jain, Prateek, Peter Yeh, Kunal Verma, Reymonrod Vasquez, Mariana Damova, Pascal Hitzler, and Amit Sheth (2011). "Contextual Ontology Alignment of LOD with an Upper Ontology: A Case Study with Proton". In: *Proc. 8th Extended Semantic Web Conference (ESWC)*, pp. 80–92 (cit. on pp. 6, 7).

Jiménez-Ruiz, Ernesto and Bernardo Cuenca Grau (2011). "LogMap: Logic-Based and Scalable Ontology Matching". In: *Proc. 10th International Semantic Web Conference (ISWC)*. Vol. 7031. Lecture Notes in Computer Science. Springer, pp. 273–288 (cit. on p. 10).

Locoro, Angela, Jérôme David, and Jérôme Euzenat (2014). "Context-based matching: design of a flexible framework and experiment". en. In: *Journal on data semantics* 3.1, pp. 25–46 (cit. on pp. 2, 5, 7, 9, 12, 13, 18).

Mascardi, Viviana, Angela Locoro, and P. Rosso (2010). "Automatic Ontology Matching via Upper Ontologies: A Systematic Evaluation". In: *IEEE Trans. Knowl. Data Eng.* 22.5, pp. 609–623 (cit. on pp. 5, 6).

Quix, Christoph, Pratanu Roy, and David Kensche (2011). "Automatic selection of background knowledge for ontology matching". In: *Proc. International Workshop on Semantic Web Information Management (SWIM), Athens (GR)*. ACM, p. 5 (cit. on pp. 7, 10).

Sabou, Marta, Matthieu d'Aquin, and Enrico Motta (2008a). "Exploring the Semantic Web as Background Knowledge for Ontology Matching". In: *J. Data Semantics* 11, pp. 156–190 (cit. on pp. 5–8, 10).

Sabou, Martha, Mathieu d'Aquin, and Enrico Motta (2008b). "SCARLET: SemantiC RelAtion DiscoveRy by Harvesting OnLinE OnTologies". In: *ESWC*, pp. 854–858 (cit. on p. 8).

Stumme, Gerd and Alexander Mädche (2001). "FCA-Merge: Bottom-Up Merging of Ontologies". In: *Proc. 17th International Joint Conference on Artificial Intelligence (IJCAI)*. Seattle (WA US), pp. 225–234 (cit. on p. 6).

Zhang, Songmao and Olivier Bodenreider (2007). "Experience in aligning anatomical ontologies". In: *International Journal on Semantic Web and Information Systems* 3.2, pp. 1–26 (cit. on p. 6).