UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

# Interactive Ontology Alignment: An Approach Based on the Interactive Modification of the Set of Candidate Correspondences

Jomar da Silva

**Orientadoras**

Fernanda Araujo Baião Amorim

Kate Cerqueira Revoredo

Rio de Janeiro, RJ – Brasil
Fevereiro de 2017

Interactive Ontology Alignment: An Approach Based on the Interactive Modification of the Set of Candidate Correspondences

Jomar da Silva

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovado por:

Fernanda Araujo Baião Amorim, D.Sc.– UNIRIO

Kate Cerqueira Revoredo, D.Sc – UNIRIO

Marcio de Oliveira Barros, D.Sc. – UNIRIO

Aline Marins Paes Carvalho, D.Sc. – UFF

Rio de Janeiro, RJ - Brasil
Fevereiro de 2017

Silva, Jomar da.
S586 Interactive ontology alignment: an approach based on the interactive modification of the set of candidate correspondences / Jomar da Silva, 2017.
94 f. ; 30 cm

Orientadora: Fernanda Araujo Baião Amorim.
Coorientadora: Kate Cerqueira Revoredo.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2017.

1. Tecnologia de informação. 2. Ontologia. 3. Alinhamento de ontologias. I. Amorim, Fernanda Araujo Baião. II. Revoredo, Kate Cerqueira. III. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnológicas. Curso de Mestrado em Informática. IV. Título.

CDD – 004

"A grandeza não consiste em receber
honras, mas em merecê-las".
- Aristóteles

Dedico esta dissertação à minha mãe,
Maria Antônia da Silva (in memoriam),
pelo amor incondicional e apoio que
me concedeu em todos os momentos.

# Agradecimentos

Agradeço às minhas orientadoras, Fernanda e Kate, pela paciência e dedicação que me ajudaram a chegar até aqui e por me deixarem a vontade para seguir o rumo que eu desejava para a pesquisa. Agradeço a elas e ao professor Márcio Barros a forma como me prepararam, seja com ideias, com revisões, nas escritas de texto, nas apresentações. Agradeço à minha mãe Maria Antônia por ser um exemplo de estudo, e por me incentivar desde criança a jamais parar de estudar, pois apesar de não ser uma fórmula exata, é a mais próxima para se alcançar o sucesso pessoal e profissional. À minha noiva, Joyce Faria, pelo apoio, pela torcida, e pela compreensão ao esforço e tempo necessário para concluir o curso. A minha família, principalmente para minha irmã Joelma, pela torcida. A minha gerente Felícia, pelo apoio dado nesta jornada. Finalmente, a todos os professores e funcionários da UNIRIO, pelo trabalho, atenção e tratamento concedidos durante todo o curso.

SILVA, Jomar. **Interactive Ontology Alignment: An Approach Based on the Interactive Modification of the Set of Candidate Correspondences .** UNIRIO, 2017. 21 Páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

## Resumo

O progresso nas tecnologias da informação e de comunicação tornou disponível uma grande quantidade de repositórios de dados, mas com uma grande heterogeneidade semântica, o que dificulta a sua integração. Ontologias têm sido usadas, dentre outras coisas, para a definição e estruturação dos conceitos que definem os dados armazenados em cada repositório. Por isso, um processo que tem sido utilizado para resolver o problema da integração entre repositórios de dados é o alinhamento de ontologias, que tenta descobrir as correspondências existentes entre as entidades de duas ontologias distintas. Existem várias abordagens na literatura para o alinhamento de ontologias, dentre as quais destacam-se as que aplicam uma estratégia interativa, que considera a participação de especialistas para melhorar a qualidade do alinhamento final. Apesar dos avanços nos resultados obtidos na literatura, há ainda erros recorrentes nos alinhamentos obtidos pelas propostas de alinhamento interativo de ontologias, o que pode ser comprovado por uma iniciativa de avaliação conduzida anualmente pela comunidade científica (OAEI). A grande maioria das ferramentas de alinhamento busca construir um conjunto de correspondências candidatas, dentre todas as correspondências possíveis entre duas ontologias, para ser trabalhado pela abordagem. Este trabalho propõe uma abordagem interativa para o alinhamento de ontologias, chamada ALIN, que modifica o conjunto de correspondências candidatas de maneira interativa, ou seja, dependendo da interação com o especialista novas correspondências são escolhidas para a sua apreciação enquanto outras são descartadas. A abordagem ALIN foi avaliada no *interactive track* da OAEI 2016, com a utilização do *conference dataset.* Nos resultados reportados pela iniciativa, ALIN obteve o primeiro lugar em termos de qualidade em cenários interativos e sem erros do especialista, enquanto em cenários não interativos foi destaque em termos de consistência.

**Palavras-chave:** alinhamento de ontologias, anti-padrões de correspondência, alinhamento interativo de ontologias.

# Abstract

The progress in information and communication technologies has made a large number of data repositories available, These repositories, however, are highly heterogeneous, which makes integration difficult. Ontologies have been used, among other things, to define and structure the concepts that define the data stored in each repository. Therefore, a process that has been used to solve the problem of integration among data repositories is ontology alignment process, which tries to discover the correspondences between the entities of two different ontologies. There are several approaches in the literature for the ontology alignment, among which we highlight the ones that apply an interactive strategy. An interactive ontology alignment strategy considers the participation of experts to improve the quality of the final result. Despite the advances in the obtained results in the literature, there are still recurrent errors in the results of the state-of-the-art proposals as stated by the most recent reports of an evaluation initiative conducted annually by the scientific community (OAEI). Most of the ontology alignment tools seeks to construct a set of candidate correspondences to be worked through by the approach. This work proposes an interactive approach for ontology alignment, called ALIN, that modifies the set of candidate correspondences in an interactive way, that is, depending on the interaction with the expert, new correspondences are chosen for his appreciation while others are discarded. The ALIN approach was evaluated in the OAEI 2016 interactive track, using the conference dataset. In the official reports from OAEI, ALIN obtained the first place in terms of quality in the interactive scenario and with no expert mistakes, and was specifically highlighted in terms of the consistency in the non-interactive scenarios.

**Keywords:** ontology matching, correspondence anti-patterns, interactive ontology matching, ontology alignment.

# Summary

# List of Figures

# List of Tables

# List of Algorithms

# List of Formulas

# Glossary

| | |
|---|---|
| API | Application Programming Interface |
| OAEI | Ontology Alignment Evaluation Initiative |
| OWL | Ontology Web Language |
| WS4J | WordNet Similarity for Java |

# 1. Introduction

*This chapter provides an overview of the thesis, presenting the motivation to improve the ontology alignment process, as well as a brief description of the concepts needed to understand the proposed solution. The hypothesis that guides the research and the methodology used to validate it are also presented here.*

## 1.1. Motivation and Characterization of the Problem

An ontology, from a computational perspective, typically provides a vocabulary that describes a domain of interest and a specification of the meaning of terms used in this domain [1], also showing the relationships existing between these terms. In recent years many ontologies have been developed and many of those contain overlapping information. We often need to use multiple ontologies. For instance, companies may want to use community standard ontologies and use them together with company specific ontologies. Applications may need to use ontologies from different areas or from different views on one area. In each of these cases it is important to know the correspondences between the concepts (and properties) in the different ontologies. Further, the data in different data sources in the same domain may have been annotated with different but similar ontologies. Knowledge of the correspondences would in this case lead to improvements in search, integration and analysis of data. It has been realized that this is a major issue and much research has recently been done on ontology alignment [2].

The ontology alignment process seeks to discover correspondences between entities of different ontologies [1]. The ontology alignment can be done in manual, semi-automatic or automatic way [1].

Among the ontology alignment processes carried out in a semi-automatic way, the ones that follow an interactive strategy stand out, considering the participation of experts in the domain that was modeled by the ontologies [3]. The use of a domain expert is not always possible, as it is an expensive, scarce and time-consuming resource. But when it is possible to use it, this strategy has achieved superior results to automatic (non-interactive) strategies, but there is still room for better results [3], as can be seen in the evaluation of interactive tools in the OAEI[1] (Ontology Alignment Evaluation Initiative), where no tool reached the 100%.

When there is participation of an expert an existing concern is the efficient use of this participation, which today is measured by OAEI[1] by the number of interactions with the expert during the interactive ontology alignment process.

The problem focused by ALIN will be to increase the quality of the results of the ontology alignment process, but keeping the number of interactions at a level compatible with other tools. The solution proposed by ALIN is a combination of techniques, some already used by other approaches, and the use of these techniques in different phases, in relation to other approaches, in the ontology alignment process.

According to Meilicke [4], ontology alignment techniques that are based on the analysis of entity names usually have two phases. First there is the creation of a set of candidate correspondences. To not work with all possible pairs of entities between two ontologies, the techniques select a subset of this total set. This subset is commonly

[1] Available at http://oaei.ontologymatching.org/2016/results/interactive/index.html, last accessed on Nov, 19, 2016.

called set of candidate correspondences. In the second phase, each correspondence in this set of candidate correspondences is classified by the ontology alignment approach as true or false. In an interactive strategy, at least part of these correspondences are directly classified by the expert, and the other part is classified by some other technique, as for example, the use of a threshold applied on the similarity metrics or the use of correspondence anti-patterns, which are situations in which two or more supposed correspondences may lie, but in which only one of them may be true. In the ALIN approach, the anti-patterns are used to classify correspondences of the set of candidate correspondences not directly classified by the expert. In ALIN, the classification by anti-patterns is called indirect classification.

In order to select correspondences for the set of candidate correspondences, algorithms, called matchers, are executed, by multiple approaches besides ALIN, to select correspondences based on some criterion of similarity. These similarities may be terminological or structural, as well as others. The ALIN approach uses terminological matchers inserted into an algorithm called the stable marriage algorithm. The ALIN approach also uses structural matchers, not in the creation phase of the set of candidate correspondences like most of others approaches, but in the interactive phase, to include new correspondences into the set. In ALIN, this use of structural matchers in the interactive phase is called retrieval of correspondences.

Since both the use of indirect classification and the retrieval of correspondences modifies the set of candidate correspondences in an interactive manner, the combination of these two techniques will be called, in this work, interactive modification of the set of candidate correspondences.

### 1.2. Objective

This work proposes ALIN, an interactive ontology alignment approach that applies suitable techniques, like stable marriage algorithm, anti-patterns, interactive use of structural matchers to select and classify the set of candidate correspondences.

The goal of this work is to show that the techniques used in it, when used together and in the specified way, generate a high quality alignment without increasing the number of interactions with the expert in an exaggerated way. Some techniques will serve to increase the recall of the generated alignment, others the precision and others to keep under control the number of interactions with the expert.

### 1.3. Hypothesis

The hypothesis guiding this research is stated as follows:

IF expert's feedback (either direct or indirect) is used to classify all the set of candidate correspondences and to modify it through correspondence anti-patterns and retrieval of correspondences, THEN the quality of the final result of the ontology alignment process is increased, keeping a reasonable number of interactions with the expert.

The quality of the final result of the ontology alignment process is traditionally evaluated using the precision and recall measures, and the harmonic mean between them, the f-measure [1], which will be further defined in detail.

### 1.4. Scientific Method

In this work, the quantitative research method is used. This method is focused on the collection of quantitative data with the objective of measuring the state of some variable of a given domain in the real world. In this work, measured variables are the number of interactions, precision, recall and f-measure of the results of the ontology alignment process. The quantitative research method used here follows the process proposed in [5], which comprises the following activities:

**1. Generation of the theory and hypothesis**: Initially, a literature review was carried out regarding the challenges in the area of ontology alignment, which resulted in the research question presented in this thesis. The result of this study was the research hypothesis presented earlier in this chapter.

**2. Development of measurement instruments:** In order to make feasible the evaluation of the proposal presented in this work, an ontology alignment software was developed that simulates the interaction with the expert using the proposed techniques.

**3. Empirical data collection:** In order to verify the cause and effect relationships between the variables present in the proposed approach, an experimental approach was used. Different scenarios were defined with variation of the techniques used, each scenario including a new technique, in addition to the use of all previous techniques. During the execution of these scenarios data was collected for precision, recall, f-measure and number of interactions with the expert, for each pair of ontologies that compose the data set used in this experiment.

**4. Data analysis:** The data collected was analyzed using the descriptive technique. Following the OAEI approach, the data was aggregated by data set and an average value for the quality measures was determined, in addition to the sum of the interactions with the expert, in each scenario.

**5. Evaluation of results:** The results obtained were compared to each new scenario, verifying if the technique used improves the expected variable. In addition, the developed program participated in the track of Interactive Matching of OAEI 2016[2], allowing the comparison of our proposal with other existing ones.

[2] Available at http://oaei.ontologymatching.org/2016/results/interactive/, last accessed on Dec, 19, 2016.

### 1.5. Organization of the Thesis

This thesis is structured as follows: Section 2 describes the theoretical basis needed to understand the work, section 3 describes the ALIN approach, section 4 shows the tool evaluation, section 5 shows the related works and section 6 shows the conclusion reached with the work, also including future works.

# 2. Theoretical foundation

*This chapter presents the concepts needed to understand the research proposal of this work. The chapter presents the concept of an ontology in the computational sense, the ontology alignment as well as ontology alignment process. This chapter also presents existing techniques for execution of the ontology alignment process and metrics for evaluation of the results of the execution. The chapter also presents the concept of correspondence anti-pattern and the concept of matcher.*

## 2.1. Ontology Alignment Process

An ontology typically provides a vocabulary describing a domain of interest and a specification of the meaning of the terms in this vocabulary [1]. An ontology is formed of entities that can be: class (concepts), relationships (object properties) or attributes (data properties). In addition to entities, an ontology may also contain individuals who belong to the concepts and types of attributes.

An ontology alignment is a set of correspondences, where each correspondence is a relation (of equivalence, generalization or disjunction) between two entities of these ontologies. When there is an equivalence correspondence between two concepts this indicates that every real-world object that can be instantiated in one concept will necessarily be instantiated in the other.

The ontology (O1) shown in Figure 1 shows some of the concepts involved in the scope of a cultural product store.

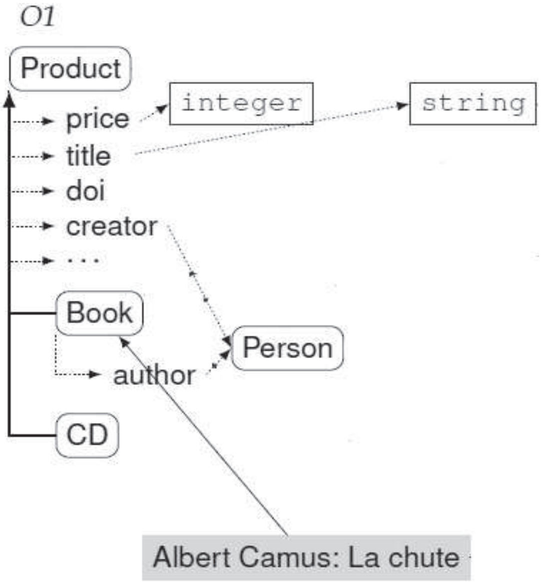


*Figure 1: Ontology of a cultural product store*

The concepts are represented by rounded squares. A specialization-generalization relationship is represented by an arrow which moves from the more specific concept (e.g. Book) to the more general concept (e.g. Product). The attribute of each concept ares represented by names preceded by dashed arrows coming out or from a specialization-generalization relationship (e.g. attribute 'title') or from a concept (e.g. attribute 'author'). Attribute types are represented by rectangles (e.g. integer). Relationships are represented by dashed arrows coming from an attribute and going to a concept (e.g. from creator to Person that represents the relationship 'Person is the creator of the Product'). An instance of a concept is represented by a gray squarer with an arrow (e.g. Albert Camus:La chute is an instance of Book).

The ontology (O2) shown in Figure 2 shows some of the concepts involved in a book publisher. If it is necessary for the cultural product store to buy a book, for example, La chute of Albert Camus, and if it wants that its system that uses the database defined by its ontology to contact the publisher's system, how will they understand each other to do the transaction, since each system uses different concepts to categorize the book? One of the answers to this problem is the ontology alignment process.

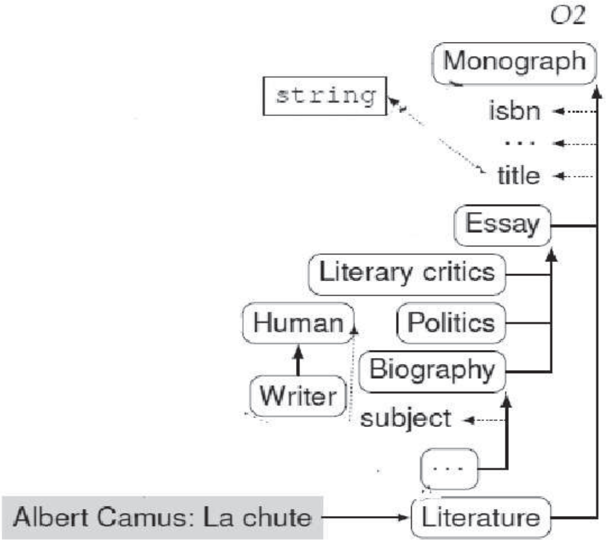


*Figure 2: Ontology of a book publisher*

In Figure 3, one can see an example of an ontology alignment, where the arrows linking entities of the two ontologies are correspondences, and the symbol above them is the semantics of the relation (⊒ being generalization and = equivalence). In this case we could search the book La Chute in the book publisher for the key 'title' (which is equivalent in both concepts containing the book) and search the database tables defined by the concepts Monograph, Essay and Literature, which are Book specializations.

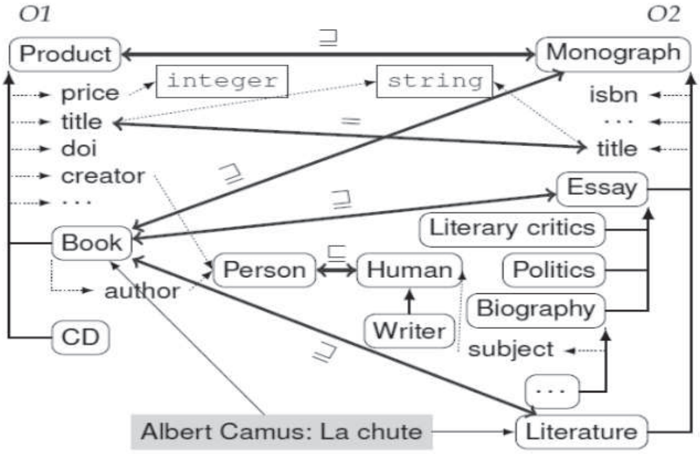


*Figure 3:* **O***ntology alignment between the cultural product store and the book publisher*

Ontology alignment process is a process whose final result is an alignment between the ontologies. Ontology alignment process is also called 'ontology matching',

'matching process' or 'ontology matching process' in the literature and one of these terms can be used in this thesis in order to avoid confusion with the term 'alignment' when one wants to refer to set of correspondences. Figure 4 shows an overview of the interfaces of this process: from the two input ontologies (O and O') that are to be aligned, an alignment A' is generated. Optionally, there may be an initial alignment A (usually generated by another ontology matching tool) to serve as a starting point in the search of A'.

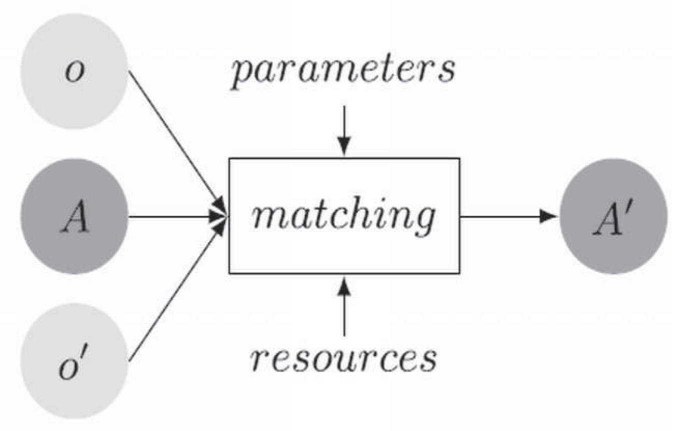


***Figure 4: Ontology matching process [1]***

The behavior of the matching process is tuned by parameters (such as a threshold, that is the minimum value a correspondence should have in order to participate in the matching process). In addition, the matching process can take into account external resources, such as reference ontologies [1].

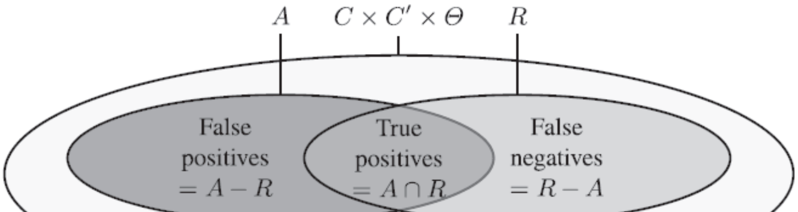


***Figure 5: Sets of correspondences of the generated alignment (A) and reference alignment (R) [1]***

To evaluate the quality of the generated alignment, two measures are commonly used: precision and recall. Such measures are calculated by comparing the generated alignment (from which the quality is to be assessed) with the reference alignment (the set of correspondences that are known to be true). In Figure 5 we have the set CxC'xΘ (C and C' the source ontologies, Θ the set of the possible semantics of relation), with all possible correspondences to be formed by the entities of the two ontologies. Positives, in this context, are the correspondences that the matching process indicated as belonging to the alignment. Negatives are correspondences that are not positive. True positives are the correspondences that the process indicated as belonging to the reference alignment and which actually belong. False positives are the correspondences that the process indicated as belonging to the reference alignment but do not belong. A is the alignment generated by the process and R is the reference alignment. Precision measures the ratio of the total of correspondences found correctly (true positives) to the cardinality of the generated alignment. Therefore, precision can be calculated as showed in Formula 1.

(1) $$\text{precision} = | A \cap R | / | A |$$

Recall measures the ratio of the number of correspondences found correctly (true positives) to the cardinality of the reference alignment. Therefore the recall can be calculated as showed in Formula 2.

(2) $$\text{recall} = | A \cap R | / | R |$$

There is a third measure, called the f-measure, which is the harmonic mean between precision and recall and can be calculated as showed in Formula 3.

(3) $$F = ( P \times R ) / ( ( 1 - \alpha ) \times P + \alpha \times R ),$$

where P is the precision value and R is the recall value, and α is a value between 0 and 1, generally being chosen the value 0.5.

Several techniques are used by multiple ontology matching approaches, including ALIN, to generate their alignment, such as similarity metrics between the ontology entities, the use of logic and ontology characteristics (called anti-patterns in

ALIN), the stable marriage algorithm and matchers.

## 2.2. Terminological Similarities

There are several similarity functions in the literature, grouped according to the perspective of analysis that is considered (string-based, linguistic etc.) as described by Shvaiko and Euzenat [1].

Six similarity metrics are used in ALIN, three string-based (Jaccard, Jaro-Winkler, and n-Gram) and three linguistic (Wu-Palmer, Jiang-Conrath and Lin) metrics. The program made to implement the ALIN approach did not implement these metrics, standard APIs were used. These metrics were used in conjunction with other techniques to form the set of candidate correspondences in the Jarvis approach [6]. The process of selecting metrics for Jarvis was based on two criteria: available implementations and the outcome of these metrics in assessments, such as those carried out in [7] and [8]. ALIN uses the techniques used in the Jarvis approach to form its set of candidate correspondences.

### 2.2.1. String-based Similarity Metric

According to Shvaiko and Euzenat [1], string-based similarity metrics compare names and descriptions of ontology entities, considering them as sequences of letters in an alphabet. They are typically based on the following intuition: the more similar the strings, the greater the likelihood of such entities denoting the same concepts in real life. Usually, a similarity function receives a pair of strings as input and returns a real number between 0 and 1, indicating the similarity between them.

#### 2.2.1.1. Jaccard Similarity Metric

The Jaccard similarity metric [1], for example, serves to show similarity between sets and is calculated between sets A and B, as shown in Formula 4.

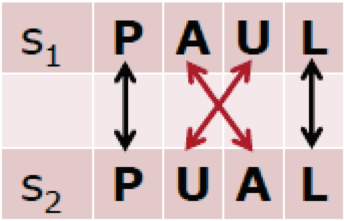
(4) $$J(A,B)=|A\cap B|/|A\cup B|$$

If the two sets are empty it is defined that J (A, B) = 1. In [7], each string word that designates the name of the entities is considered an element of the set. For the Jaccard similarity calculation, the string letters can be used as elements of the sets instead of the words [9]. For the ALIN approach it was assumed that each letter of the string is an element of the set, because it achieved better results than the use of words, when evaluated for the OAEI conference dataset (the conference dataset and the OAEI are explained in section 4.1). Thus, when comparing two strings, are compared two sets whose elements are the letters of these strings.

### 2.2.1.2. Jaro-Winkler Similarity Metric

Before explain the Jaro-Winkler Similarity Metric, the Jaro Similarity Metric will be explained. Jaro Similarity Metric formula can be seen in Formula 5.

(5) $$sim_{jaro} = \frac{1}{3}\left(\frac{m}{|x|} + \frac{m}{|y|} + \frac{m-t}{m}\right)$$

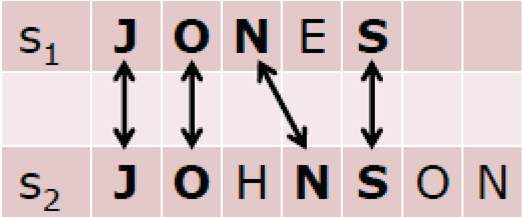


$$m = 4 \qquad t = \frac{2}{2} = 1$$

$$sim_{jaro} = \frac{1}{3} \cdot \left(\frac{4}{4} + \frac{4}{4} + \frac{4-1}{4}\right) \approx 0.92$$

*Figure 6: Jaro similarity metric for strings PAUL and PUAL*

Where |x| is the length of the first character string, |y| is the size of the second character string, m is the number of characters in the two strings that appear in the same order and t the number of character inversions (transpositions), as we can see in the Figures 6 and 7, where the strings PAUL and PUAL, and the strings JONES and JOHNSON are compared.



$$m = 4 \qquad t = \frac{0}{2} = 0$$

$$sim_{jaro} = \frac{1}{3} \cdot \left( \frac{4}{5} + \frac{4}{7} + \frac{4-0}{4} \right) \approx 0.79$$

***Figure 7: Jaro similarity metric for strings JONES and JOHNSON***

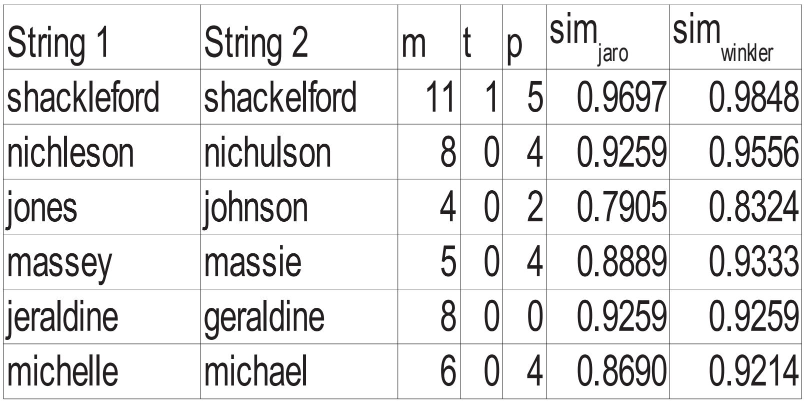
Jaro-Winkler similarity metric assumes that the beginning of the string has a value greater than all its characters. It takes advantage of the Jaro metric and modifies it by giving weight to the first p equal characters, being calculated as shown in Formula 6, where p is the number of first equal characters (common prefix).

(6) $$sim_{winkler}(x, y) = sim_{jaro}(x, y) + (1 - sim_{jaro}(x, y)) \frac{p}{10}$$

□ = 1 if common prefix is ≥10

The calculation of the Jaro-Winkler metric can be seen in Table 1.

*Table 1: Jaro-Winkler similarity metric examples*

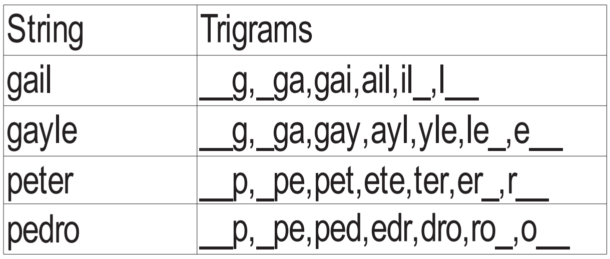


| String 1 | String 2 | m | t | p | $sim_{jaro}$ | $sim_{winkler}$ |
|---|---|---|---|---|---|---|
| shackleford | shackelford | 11 | 1 | 5 | 0.9697 | 0.9848 |
| nichleson | nichulson | 8 | 0 | 4 | 0.9259 | 0.9556 |
| jones | johnson | 4 | 0 | 2 | 0.7905 | 0.8324 |
| massey | massie | 5 | 0 | 4 | 0.8889 | 0.9333 |
| jeraldine | geraldine | 8 | 0 | 0 | 0.9259 | 0.9259 |
| michelle | michael | 6 | 0 | 4 | 0.8690 | 0.9214 |

### 2.2.1.3. n-Gram Similarity Metric

The n-Gram similarity (or k-gram or q-gram) metric divides the strings to be compared into small substrings of size n (in the case of ALIN, n = 3 , called trigrams). As an example of the division of the strings we can see in the Table 2.

*Table 2: Trigrams of strings*



| String | Trigrams |
|---|---|
| gail | __g,_ga,gai,ail,il_,l__ |
| gayle | __g,_ga,gay,ayl,yle,le_,e__ |
| peter | __p,_pe,pet,ete,ter,er_,r__ |
| pedro | __p,_pe,ped,edr,dro,ro_,o__ |

$$\frac{|A| + |B| - (|A \cup B| - |A \cap B|)}{|A| + |B|} \quad (7)$$

The formula of n-Gram similarity is shown in Formula 7, where A is the set of trigrams of the first string and B is the set of trigrams of the second string.

### 2.2.2. Linguistic Similarity Metrics

Linguistic similarity metrics use linguistic resources (such as domain-specific dictionaries or thesauri) to compare words (in this case, the name or description words of the entity of an ontology are considered natural language words). The metrics used in this thesis use the Wordnet [1]. Wordnet consists of synonym synsets [10]. A synset denotes a group of terms with the same meaning. Wordnet provides different semantic relationships between synsets, such as synonym (similarity) / antonym (opposition), hyperonymy / hyponymy (subsumption), meronymy (part of) / holonomy (have one). The relationship employed in the metrics used in this work are those of superconcepts (hyperonymy) and subconcepts (hyponymy), which generate a taxonomy between the synsets.

As with strings, usually the linguistic similarity functions return a value between 0 and 1, 1 when the greatest similarity occurs.

To talk about linguistic similarity metrics, some other related concepts such as Probability of Random Word Being an Instance of a Concept and Lowest Common Subsumer must be talked.

#### 2.2.2.1. Probability of Random Word Being an Instance of a Concept

Probability of Random Word Being an Instance of a Concept is calculated by the number of words belonging to the concept c plus the number of all the words belonging to the concepts that are hyponymies divided by the total of words contained in the linguistic resource (eg. a thesaurus) used, the result is symbolized as P(c).

In Figure 8, part of a generalization hierarchy of a hypothetical thesaurus is shown. A randomly chosen word has 39.5% of being in the 'entity' concept, or in the concepts below it, as inanimate-object.

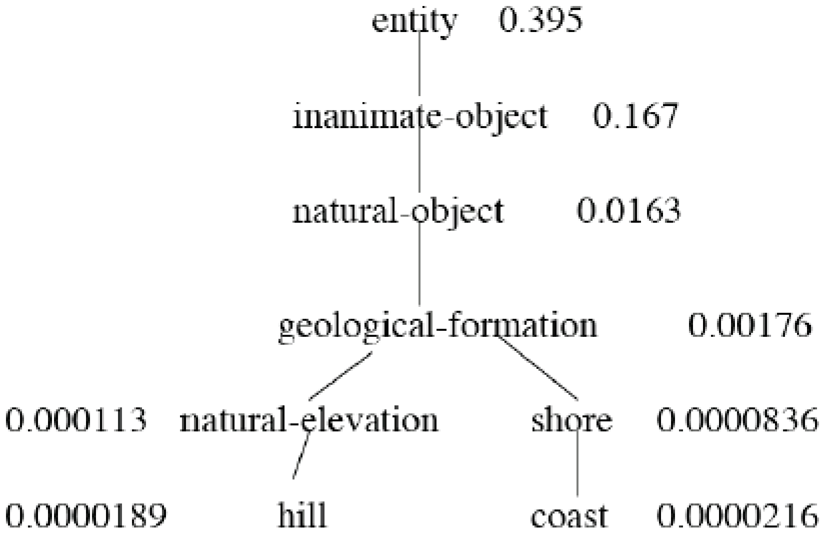


***Figure 8: Generalization hierarchy with probability of random word being an instance of a concept example***

#### 2.2.2.2. Lowest Common Subsumer

LCS (C1, C2) = lowest node in hierarchy that is a hypernym of C1 and C2. In the case of Figure 8, CLS (hill, coast) = geological-formation.

#### 2.2.2.3. Wu-Palmer Similarity Metric

Wu-Palmer linguistic similarity metric [1] is calculated as in Formula 8.

(8) sim(C1,C2) = 2*N3/(N1+N2+2*N3)

where N1 is the number of nodes in the path between C1 and LCS(C1,C2). N2 is the number of nodes in the path of C2 and LCS(C1,C2). N3 is the number of nodes in the path of LCS(C1,C2) and the root.

An example of the Wu-Palmer metric can be seen using the generalization hierarchy shown in Figure 8. We will find the metric for the words hill and coast. In this case, LCS (hill, coast)=geological-formation, which is the lowest common subsumer of both. N3 is 3, (it will be assumed that root is 'entity' to simplify calculations, which is not true, since root would have P(c) = 1), N1 is 2 and N2 is 2. Then the Wu- Palmer of the two words is 2 * 3 / (2 + 2 + 2 * 3) = 6/10 = 0.6.

#### 2.2.2.4. Lin Similarity Metric

The calculation of the Lin similarity metric is shown in Formula 9.

(9) $$sim_{Lin}(c_1, c_2) = \frac{2 \times \log P(LCS(c_1, c_2))}{\log P(c_1) + \log P(c_2)}$$

For example, the Lin similarity between hill and coast in Figure 8 is shown below.

$$sim_{Lin}(hill, coast) = \frac{2 \times \log P(geological\text{-}formation)}{\log P(hill) + \log P(coast)} = 0.59$$

#### 2.2.2.5. Jiang-Conrath Similarity Metric

The calculation of Jiang-Conrath similarity is shown in Formula 10.

(10) $$sim_{JC}(c_1, c_2) = \frac{1}{2 \times \log P(LCS(c_1, c_2)) - (\log P(c_1) + \log P(c_2))}$$

## 2.3. Stable Marriage

Stable Marriage problem (SM) was introduced in the seminal paper of Gale and Shapley [11]. In its classical form, an instance of SM involves n men and n women, each of whom specifies a preference list, which is a total order on the members of the opposite sex. A matching M is a set of (man,woman) pairs such that each person belongs to exactly one pair. If (m,w) ∈ M, we say that w is m's partner in M, and vice versa, and we write M(m) = w, M(w) = m. We say that a person x prefers y to y' if y precedes y' on x’s preference list [12].

A matching M is stable if it admits no blocking pair, namely a pair (m,w) such that m prefers w to M(m) and w prefers m to M(w).

Consider the following example.

Let there be two men m1 and m2 and two women w1 and w2.

Let m1‘s list of preferences be {w1, w2}

Let m2‘s list of preferences be {w1, w2}

Let w1‘s list of preferences be {m1, m2}

Let w2‘s list of preferences be {m1, m2}

The matching { {m1, w2}, {m2, w1} } is not stable because m1 and w1 would prefer each other over their assigned partners. The matching {m1, w1} and {m2, w2} is stable because there are no two people of opposite sex that would prefer each other over their assigned partners.

Gale and Shapley [11] proved that every instance of SM admits at least one stable matching, and described an algorithm (the Gale / Shapley algorithm) that finds such a matching in time that is linear in the input size.

### 2.3.1. Incomplete Lists

A variety of extensions to the basic stable marriage problem were studied. In the problem of stable marriage with incomplete lists (SMI), the amount of men and women need not be the same and the list of preferences of each person consists of a subset of members of the opposite sex in strict order. A pair (m, w) is acceptable if each member of the pair appears in the preference list of the other. An alignment M is now a set of acceptable pairs such that each person belongs to at most one pair [12].

### 2.3.2. Preference Lists with Limited Size

In the context of many alignment schemes, the preference lists of at least one of the sets tends to be short. For example, until recently it was asked to the students to participate in the Scottish medical alignment scheme to indicate, in order of preference, only three hospitals. This type of stable marriage problem with a maximum of options per preference list is called stable marriage with incomplete lists of limited size [12].

### 2.4. Correspondence Anti-Patterns

A correspondence anti-pattern (also called problematic alignment pattern) is, more precisely, a combination of correspondences that generates inconsistency [13].

This inconsistency may be a logical inconsistency, or a broken rule defined for existing ontologies or a broken rule defined for the alignment to be generated.

In this work, anti-patterns are used in the interactive modification of the set of candidate correspondences, ie, making the set be modified in an interactive way, in order to reduce the number of interactions with the expert.

Figure 9, Figure 10 and Figure 11 illustrate some correspondence anti-patterns, empirically defined by Guedes [14], extracted from the results of ontology matching tools evaluated by OAEI [15]. The Ontology Alignment Evaluation Initiative (OAEI) is a coordinated international initiative whose one of the goals is assess strengths and weaknesses of matching systems.

In the Figure 9, Figure 10 and Figure 11 the rounded rectangles are entities of an ontology, the bold double arrow with the equal sign represents a correspondence, the bold double arrow with the difference sign represents a disjunction, and the unidirectional arrow with the '⊑' sign identifies a hierarchy, being the superclass the tip side of the arrow. The name oX:eX indicates that the entity eX belongs to the ontology oX.

Each correspondence forms a (possibly empty) set of other correspondences that are in an anti-pattern with it. If the correspondence is true all correspondences in the set are certainly false. ALIN will take advantage of this feature to decrease the number of interactions with the expert.

### 2.4.1. Anti-pattern of Multiple Entities

To exist this anti-pattern, there must be a restriction that a single entity does not participate in two correspondences in the generated alignment.

In Figure 9, entity o2:e1 participates in correspondences (represented by the double arrow with the sign of =) with o1:e2 and with o1:e1. Since it can only participate in a correspondence, at least one of the correspondences must be false.

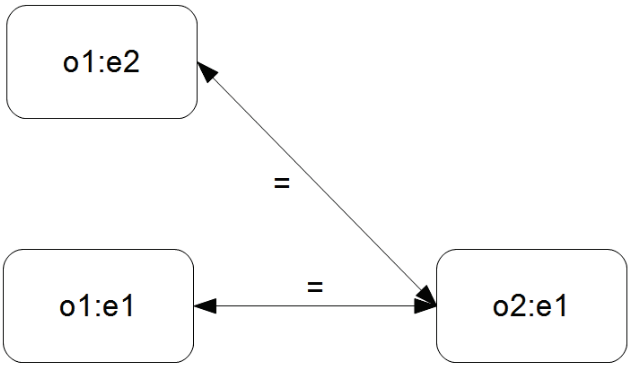


*Figure 9: Anti-pattern of multiple entities [14]*

### 2.4.2. Anti-pattern of Cross Correspondences

In order to exist this anti-pattern, there must be a restriction that no subclass can be equivalent to its superclass, that is, the superclass must be able to instantiate some object that can not be an instance of its subclass.

In Figure 10, entity o2:e1 participates in a correspondence with o1:e2, and o1:e1 participates in a correspondence with o2:e2, which makes the alignment inconsistent as o1:e1 is a subclass of o1:e2, then there may be objects that are instantiated in o1:e2 and can not be instantiated in o1:e1. Since o2:e1 is equivalent to o1:e2 then there are objects that can be instances in o2:e1 and that can not be in o1:e1. Since o2:e1 is the subclass of o2:e2 (represented by the arrow with the sign ⊑) then there may be elements that can be instantiated in o2:e2 and can not be instantiated in o1:e1, which generates a contradiction because, in the alignment, o1:e1 is equivalent to o2:e2.

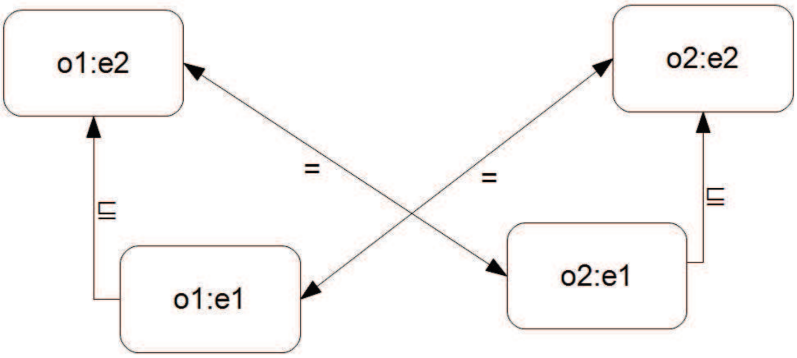


*Figure 10: Anti-pattern of cross correspondences [14]*

### 2.4.3. Anti-pattern of Disjunction and Generalization

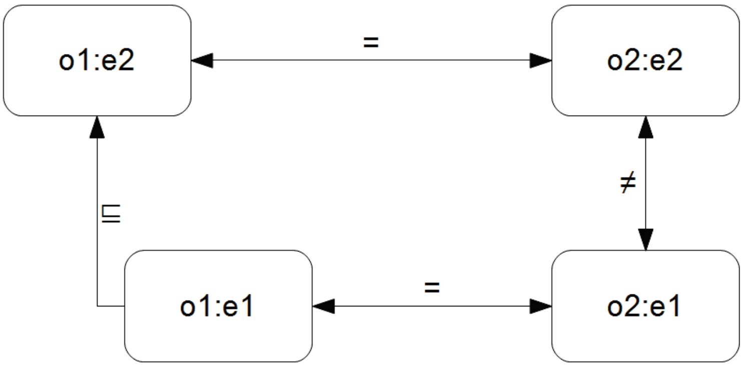


*Figure 11: Anti-pattern of disjunction and generalization [8]*

The alignment of Figure 11 is logically inconsistent because each instance of class o1:e1 is also instance of class o2:e1 (because of equivalence between these classes), and each instance of o1:e1 is also instance of o1:e2 (by fact that o1:e2 is superclass of o1:e1).

Also, each instance of o1:e2 is also instance of o2:e2 (by equivalence). Therefore, one can deduce that the instances of o2:e1 are also instances of o2:e2, which generates a logical contradiction, since o2:e1 and o2:e2 are disjoint, that is, there is no real-world object that is simultaneously an instance of o2:e1 and o2:e2.

## 2.5. Matchers

An ontology matching system can include several matchers, that are algorithms that calculate similarities between the entities from the different source ontologies, returning a set of correspondences based on their similarities[16]. They often implement algorithms based on terminological, structure, constraint, or instance characteristics, or based on auxiliary information or a combination of these. Each matcher utilizes knowledge from one or multiple sources [2].

Structural matchers use, in addition to the source ontologies, a set of correspondences as input, returning a different set of correspondences related to the input correspondences according to their algorithm.

The ALIN approach uses matchers in two points. In the initial selection of correspondences that will be part of the set of candidate correspondences, when using the stable marriage algorithm with six terminological matchers and in the interactive phase, where it uses three structural matchers that return correspondence of concepts, of attributes and of relationships, associated with the input correspondences.

# 3. The ALIN Approach

*This chapter presents the techniques used by the ALIN approach, its sets and processes.*

## 3.1. The ALIN Approach

The ALIN approach, like several other approaches to interactive ontology matching, has two main phases:

- Generation of the set of candidate correspondences, a non-interactive phase;
- Classification of the set of candidate correspondences, an interactive phase.

In the classification phase the correspondences in the set of candidate correspondences are classified as belonging or not belonging to the alignment, or by the expert or by some other technique.

Evaluating a series of ontology matching approaches, three characteristics were perceived that, if modified, could generate a superior quality alignment.

The first characteristic is:

- **Use of certain techniques (threshold, for example) to automatically classify part of the set of candidate correspondences, which can lead to classification errors, even if the expert doesn't make mistakes [6][17][18][19][20][21][22][23][24][25].**

The advantage of using these techniques is to decrease the number of interactions with the expert. An example of such a technique is the use of a threshold to classify the correspondences from the set of candidate correspondences not classified by the expert.

The ALIN, different from these approaches, uses techniques (anti-patterns) that do not make classification errors, if the expert don't make classification errors.

The second characteristic is:

- **Use of structural matchers in the phase of generation of the set of candidate correspondences [26][27][28].**

When structural matchers are used in the non-interactive phase, to create the set of candidate correspondences, entry correspondences to the matchers are automatically found, so there is a certain probability these entry correspondences are false.

The ALIN approach will use structural matchers only in the interactive phase, and only correspondences classified by the expert as true will serve as input for the matchers. If it is assumed that the expert does not make mistakes, only true correspondences will serve as input for the matchers. The use of structural matchers in the interactive phase will cause new correspondences to be inserted into the set of candidate correspondences, this action of inserting new correspondences in the interactive phase is called retrieval of correspondences.

The third characteristic is:

- **Automatic creation of an initial alignment before the interactive phase [6][26][27].**

This is done by several approaches to decrease the number of interactions with the expert, but because it is automatic, it can include false correspondences into the alignment.

ALIN will continue to automatically generate an initial alignment, but with more stringent criteria, to decrease the number of false correspondences automatically included in the alignment.

In the interactive phase, in the classification of the set of candidate correspondences, ALIN will use the following techniques:

- **Retrieval of correspondences (Interactive use of structural matchers);**
- **Indirect classification (Use of anti-patterns in the classification of candidate correspondences not classified directly by the expert).**

Some of these techniques have already been used individually by other approaches. ALIN takes a step forward towards combining all of them.

### 3.1.1. Using Wordnet in ALIN to Calculate Similarity Metrics Between Entity Names

The ALIN approach uses the WS4J[3] API to compute some similarity metrics (Wu-Palmer, Jiang-Conrath and Lin). This API uses Wordnet to calculate these metrics. ALIN also uses Wordnet to remove correspondences with semantically different names from the set of candidate correspondences  (section 3.3.1.1.2, page 38).

In the ALIN approach, entity names are divided into their component words, and linguistic similarity metrics are applied to these words.

The OAEI ontologies have their entities named as terms that contain several words, such as: Contribution_co-author, Review_preference, Conference_document, etc. Wordnet does not contain a lot of terms with multiple words, so is better search for terms with one word, such as: Contribution, Review, Document, etc. In order to have access to Wordnet, each entity name of each ontology is considered as an open compound noun or

[3] “WS4J”. Available at  https://code.google.com/archive/p/ws4j/ Last accessed on Apr, 11, 2016.

a noun followed by a prepositional phrase as post-modifier.

A compound noun contains two or more words which join together to make a single noun. Compound nouns can be words written together, words that are hyphenated (separated with a hyphen), or separated words (open compound noun)  that go together by meaning.

Most compound nouns contain at least one noun. The other word or words may be an adjective, preposition, or verb. In a two word compound name, the second word is almost always the main word, that means that the first word modifies or adds meaning to the second one.

Most English compound nouns that consist of more than two words can be constructed recursively by combining two words at a time. Combining "science" and "fiction", and then combining the resulting compound with "writer", for example, can construct the compound "science fiction writer".

A prepositional phrase can be a post-modifier of a noun. A prepositional phrase is a phrase which begins with a preposition. The examples of prepositions are in, on, at , for, of, with, by, to, above, under, near, and wihtout. The prepositional phrase consists of a preposition plus a noun or a noun phrase. The examples of prepositional phrases are at home, in the house, to campus, with my friend near the post office and for my wife.

A post-modifier of a noun is a modifier which comes after a noun head in a noun phrase. Look at the examples below:

- man in my house
- students of the university
- people at home
- house near the post offic

ALIN selects the main word of the compound noun or the noun which is followed by a post-modifier as main word of the name of entities. This technique is recursive, that is, after removing the main word and the prepositions, the next main word is selected, that is the second most important word, and so it continues until all the words have been placed in order of importance. The words, after that, are placed in their canonical form and searched in the Wordnet. To make the comparison is associated the most frequent meaning of Wordnet for the chosen word, which may occasionaly associate an incorrect meaning to a word. The total similarity is calculated by a weighted average, with the most important with a greater weight of all the words of the term. The weights given to the words were found by testing the ALIN using the OAEI conference dataset. These weights showed the best result in terms of final quality.

As an example we can show how we would organize the words below:

A) ProgramCommitteeChair: It is an open compound noun, the main word is Chair, the rest is ProgramCommittee, which is also an open compound noun whose main word is Committee, so the list with the words in order of importance is: Chair, Committee and Program;

B) Deadline_for_notification_of_acceptance: is a noun (Deadline) followed by a prepositional phrase (for_notification_of_acceptance). Notification_of_acceptance, in turn, is also a noun (notification) followed by a prepositional phrase (of_acceptance), so the list with the words in order of importance is: Deadline, notification and acceptance.

C) Fee_for_extra_trip: It is a noun (Fee) followed by a prepositional phrase (for_extra_trip). The remainder of the term without the preposition (extra_trip) in turn is an open compound noun whose main word is trip, so the list with the words in order of importance is: Fee, trip and extra.

ALIN has an alternative to using Wordnet. This occurs when Cartesian product of the two ontologies has cardinality greater than 100,000. In this case the calculation of the semantic metrics is not done using the cited similarity metrics, but a more simplified algorithm is made, which looks for entities that have the name with more equal words, in the order given by the main words.

## 3.2. Sets Used in ALIN

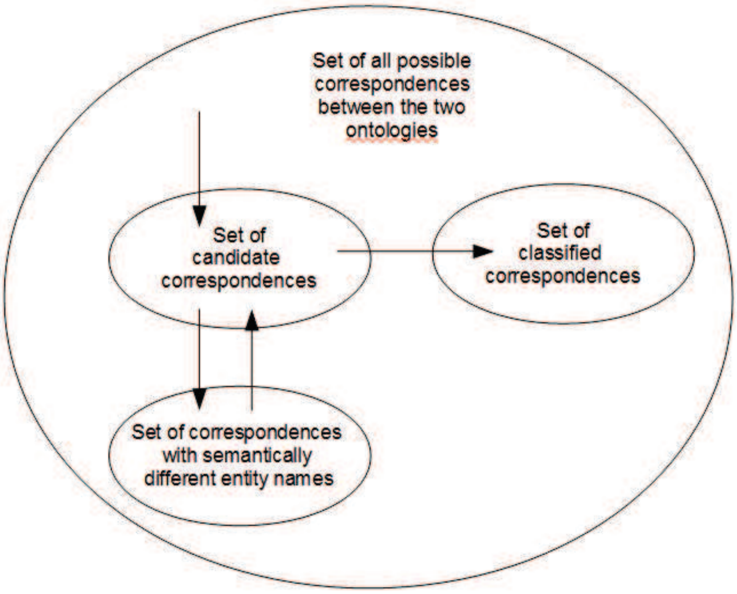

Set of all possible correspondences between the two ontologies
Set of candidate correspondences
Set of classified correspondences
Set of correspondences with semantically different entity names

*Figure 12: Sets through which a correspondence can pass during the ALIN approach to the interactive ontology matching process*

ALIN is a set-based approach for ontology matching, as illustrated in Figure 12. Its three most important sets are the set of candidate correspondences, the set of all possible correspondences between the two ontologies and the set of classified correspondences. The set of classified correspondences is the union of its two subsets, the set of correspondences classified as belonging to the alignment and the set of correspondences classified as not belonging to the alignment as can be seen at Figure 13.

The set of candidate correspondences contains the correspondences that will be classified as true or false by the approach. The set of all possible correspondences between the two ontologies is the Cartesian product between the sets of entities of each ontology. The set of classified correspondences is the set of all the correspondences the approach has already classified as true or false.

Another important set is the set of correspondences with semantically different entity names. This set is formed by all correspondences selected, in principle, to the set of candidate correspondences, but removed from it since they have pairs of entities with low linguistic similarity.

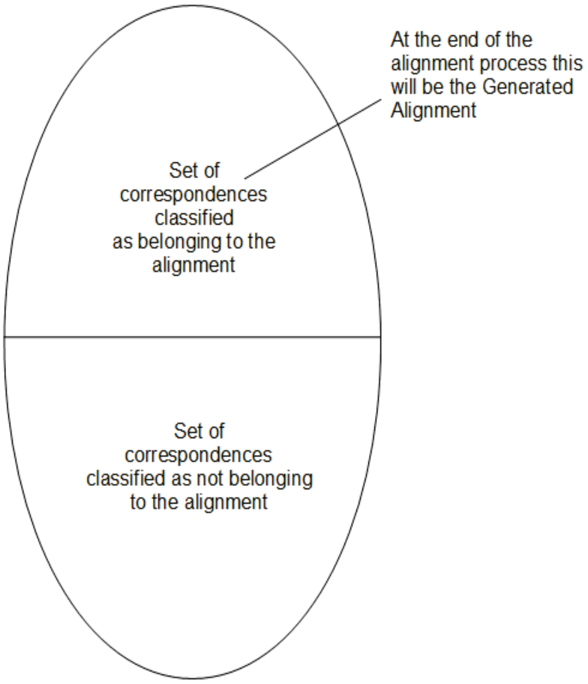

At the end of the alignment process this will be the Generated Alignment
Set of correspondences classified as belonging to the alignment
Set of correspondences classified as not belonging to the alignment

*Figure 13: Subsets of set of classified correspondences*

### 3.2.1. Relations between ALIN sets and sets defined to calculate the quality of the generated alignment

The set defined as CxC'xΘ (Figure 5, page 10) to calculate the quality of the generated alignment is exactly the set of all possible pairs of correspondences between the two ontologies.

In order to measure the quality of the generated alignment at the end of the alignment process, it is sufficient to place as set A (Figure 5, page 10) the set of correspondences classified as belonging to the alignment.

It is possible to measure the quality of any set of the ontology matching process in this way, placing it in the place of A and performing the calculations, even those sets that will still be modified during the process, thus having a view of the quality of the set at the time of measurement. As an example of these sets we can have the set of candidate correspondences, the initial alignment generated with the criterion of maximum similarity, which is the set of correspondences classified as belonging to the alignment before the beginning of the interactive phase, and the set of correspondences classified as belonging to the alignment after each interaction with the expert, thus being able to verify how the quality of the generated alignment to each interaction varies.

### 3.3. ALIN Process

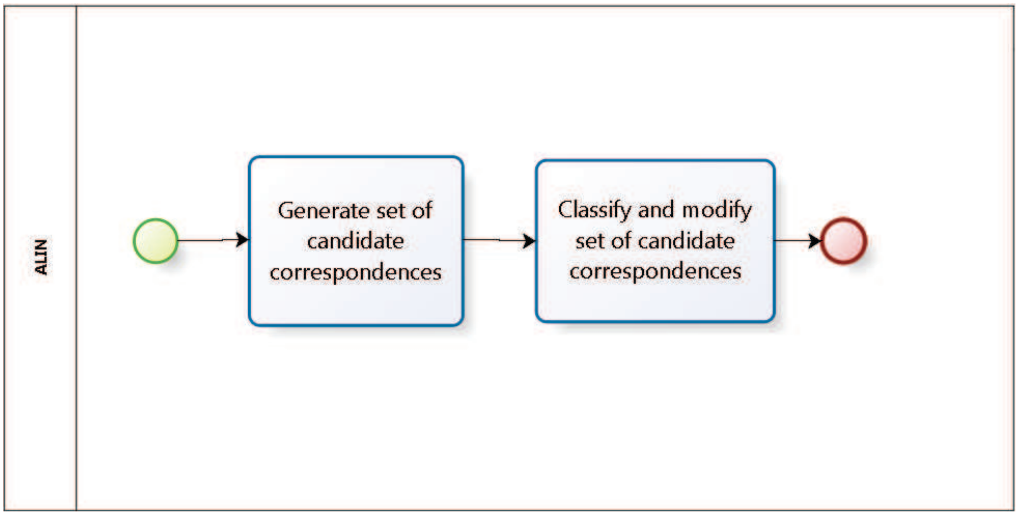

ALIN
Generate set of candidate correspondences
Classify and modify set of candidate correspondences

*Figure 14: Main phases of the ALIN process*

From a procedural perspective, the above mentioned sets evolve during the execution of the processes that are part of the ALIN approach. There are two sequential processes, which are represented in Figure 14, and detailed as follows

A) Generate set of candidate correspondences;

B) Classify and modify set of candidate correspondences.

#### 3.3.1. Generate set of candidate correspondences

As stated earlier, the set of candidate correspondences contains the correspondences that will be classified as true or false by the ALIN approach ( by expert feedback or by use of anti-patterns ). This set exists, in an interactive approach, with the objective of reducing the number of interactions with the expert, since the number of possible correspondences between two ontologies can be very large. Therefore, the objective of this phase is to form the initial set of candidate correspondences that generates the fewest interactions with the expert, yet maintaining a good recall.

It is divided into two subphases, which can be seen in Figure 15:

A) Select correspondences that will be part of the set of candidate correspondences;

B) Generate initial alignment.

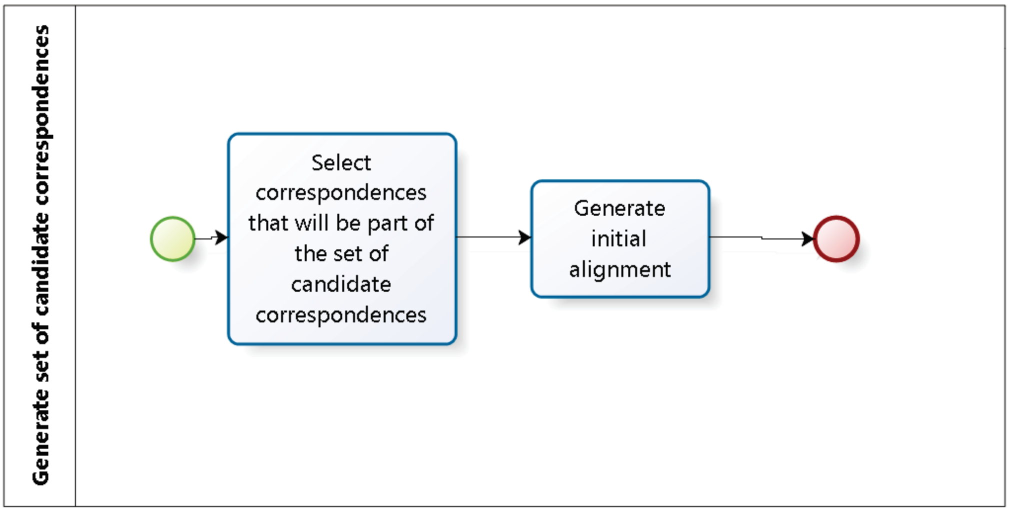

Generate set of candidate correspondences
Select correspondences that will be part of the set of candidate correspondences
Generate initial alignment

*Figure 15: Generate set of candidate correspondences*

The generation of the set of candidate correspondences was based on the Jarvis approach [6]. The differences of ALIN in relation to that work is that ALIN has added the withdrawal of correspondences with semantically different entity names and the review of automatic classification according to the maximum similarity premise.

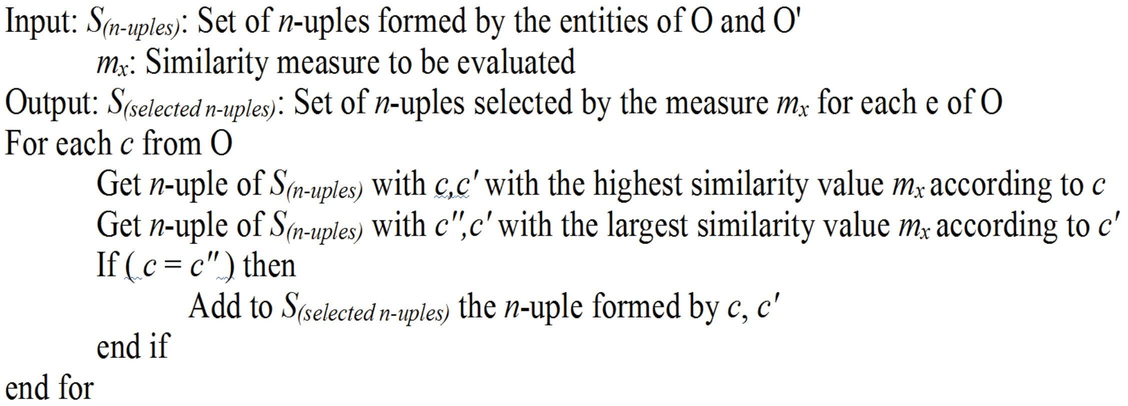
#### 3.3.1.1. First Selection of the Correspondences that Will Be Part of the Set of Candidate Correspondences

In this subphase is made the first selection of the correspondences that will be part of the set of candidate correspondences. The objective of this phase is to form the set of candidate correspondences that balances two goals: generating the smaller number of interactions with the expert, yet maintaining a large recall. It is important to note that these are contradictory goals, since a very small set of candidate correspondences tends to generate low recall, while a set that has high recall tends to be large. In ALIN, two techniques were used to balance these goals: the stable marriage algorithm and the withdrawal of correspondences whose entity names do not belong to the same synset in Wordnet. For the creation of the set of candidate correspondences the stable marriage algorithm will only select correspondences between classes, not taking into account the correspondences between properties. The results shown in [7] indicate that terminological similarity metrics are more efficient when they measure the similarity between class names than when they measure the similarity between property names. Correspondence between properties will be added to the set of candidate correspondences in the classifying phase of the set and not in the generation phase.

#### 3.3.1.1.1. Selection of Candidate Correspondences through the Stable Marriage Algorithm with Incomplete Lists of Limited Size 1

Now, it will be shown how the ALIN generate a set of candidate correspondences using the Stable Marriage Algorithm with Incomplete List of Limited Size 1 (Section 2.3, page 18).

Consider a similarity metric $m_x$: For each class c (for execution of this algorithm only the classes of the ontologies will be taken into account, not all the entities) of the ontology O, is obtained the class c' of the ontology O' with whom c forms the pair of highest similarity value for the metric $m_x$ according to c (first place in the preference list, actually the only one in the preference list of c, since its size = 1). Given the class c' returned, we get the class c'' of the ontology O, with c' forming the pair of greatest similarity value for the metric $m_x$ (c'' is the class in the list of preferences of c'). If c and c'' represent the same class in the ontology O (c and c' is an acceptable pair), then the pair c and c' is selected as candidate correspondence in the perspective of $m_x$. This algorithm can be seen in Algorithm 1.



```
Input: $S_{(n\text{-}uples)}$: Set of $n$-uples formed by the entities of O and O'
        $m_x$: Similarity measure to be evaluated
Output: $S_{(selected\ n\text{-}uples)}$: Set of $n$-uples selected by the measure $m_x$ for each e of O
For each $c$ from O
        Get $n$-uple of $S_{(n\text{-}uples)}$ with $c,c'$ with the highest similarity value $m_x$ according to $c$
        Get $n$-uple of $S_{(n\text{-}uples)}$ with $c'',c'$ with the largest similarity value $m_x$ according to $c'$
        If ( $c = c''$ ) then
                Add to $S_{(selected\ n\text{-}uples)}$ the $n$-uple formed by $c$, $c'$
        end if
end for
```

***Algorithm 1: Selection of candidate correspondences through the stable marriage algorithm with incomplete lists of limited size 1***

#### 3.3.1.1.1. Example of Stable Marriage with Incomplete Lists of Limited Size to 1

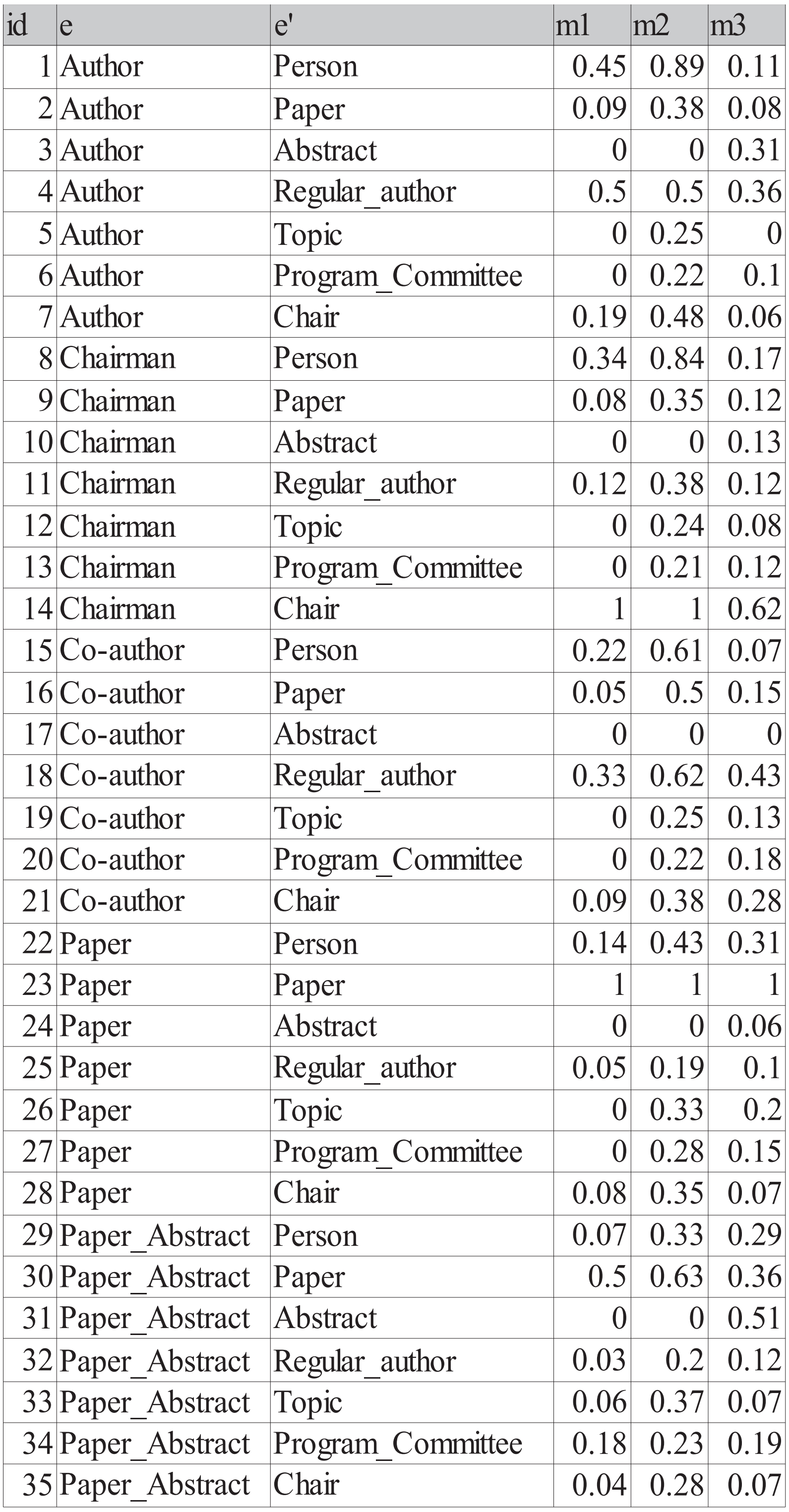
The Algorithm 1 will be executed with the ontologies cmt and conference of the OAEI conference dataset to illustrate the generation of the set of candidate correspondences.

The OAEI ontologies are grouped into sets in which the ontologies belonging have a common domain, which are called datasets by OAEI. One of the datasets is the conference dataset that contains ontologies of academic conferences. Two of the ontologies of this dataset are the conference ontology and the cmt ontology. The conference ontology was developed to develop academic conference management software for SOFSEM (SOFtware SEMinar). The cmt ontology refers to the Microsoft conference management software, the Conference Management Toolkit. Table 3, page 36, shows a subset of the pairs formed from the ontologies cmt and conference and the values obtained by applying three hypothetical similarity metrics (m1, m2 and m3) for each pair of classes. The pairs of classes of this subset will be used throughout this chapter to exemplify the proposed approach and have been selected from a set consisting of 1,800 pairs of classes for space reasons.

The pair (id = 04) (Author,Regular_author) Author in the cmt ontology, and Regular_author in the conference ontology is selected as candidate correspondence because Regular_author is the class with which Author forms the highest value pair of similarity for the metric m1, in the case, 0.50 and the same holds for the inverse, that is, given the class Regular_author, in the ontology conference, the class Author is obtained, in the ontology cmt, because Author is the class with which Regular_author forms the pair of greater value of similarity for the metric m1. In this case, therefore, the n-uple <04, Author, Regular_author, 0.50, 0.50, 0,36> is selected as candidate correspondence according to metric m1. Tables 4, 5 and 6 present the subsets of n-uples selected as candidate correspondences from the Table 3 by the algorithm for the similarity metrics m1, m2 and m3, respectively.

Table 7 shows the resulting set of candidate correspondences (the union of Tables 4, 5 and 6), considering the three similarity metrics used in the example.

***Table 3: Subset of n-uples of the ontologies cmt ( e ) and conference ( e′) [6]***



| id | e | e' | m1 | m2 | m3 |
|---|---|---|---|---|---|
| 1 | Author | Person | 0.45 | 0.89 | 0.11 |
| 2 | Author | Paper | 0.09 | 0.38 | 0.08 |
| 3 | Author | Abstract | 0 | 0 | 0.31 |
| 4 | Author | Regular_author | 0.5 | 0.5 | 0.36 |
| 5 | Author | Topic | 0 | 0.25 | 0 |
| 6 | Author | Program_Committee | 0 | 0.22 | 0.1 |
| 7 | Author | Chair | 0.19 | 0.48 | 0.06 |
| 8 | Chairman | Person | 0.34 | 0.84 | 0.17 |
| 9 | Chairman | Paper | 0.08 | 0.35 | 0.12 |
| 10 | Chairman | Abstract | 0 | 0 | 0.13 |
| 11 | Chairman | Regular_author | 0.12 | 0.38 | 0.12 |
| 12 | Chairman | Topic | 0 | 0.24 | 0.08 |
| 13 | Chairman | Program_Committee | 0 | 0.21 | 0.12 |
| 14 | Chairman | Chair | 1 | 1 | 0.62 |
| 15 | Co-author | Person | 0.22 | 0.61 | 0.07 |
| 16 | Co-author | Paper | 0.05 | 0.5 | 0.15 |
| 17 | Co-author | Abstract | 0 | 0 | 0 |
| 18 | Co-author | Regular_author | 0.33 | 0.62 | 0.43 |
| 19 | Co-author | Topic | 0 | 0.25 | 0.13 |
| 20 | Co-author | Program_Committee | 0 | 0.22 | 0.18 |
| 21 | Co-author | Chair | 0.09 | 0.38 | 0.28 |
| 22 | Paper | Person | 0.14 | 0.43 | 0.31 |
| 23 | Paper | Paper | 1 | 1 | 1 |
| 24 | Paper | Abstract | 0 | 0 | 0.06 |
| 25 | Paper | Regular_author | 0.05 | 0.19 | 0.1 |
| 26 | Paper | Topic | 0 | 0.33 | 0.2 |
| 27 | Paper | Program_Committee | 0 | 0.28 | 0.15 |
| 28 | Paper | Chair | 0.08 | 0.35 | 0.07 |
| 29 | Paper_Abstract | Person | 0.07 | 0.33 | 0.29 |
| 30 | Paper_Abstract | Paper | 0.5 | 0.63 | 0.36 |
| 31 | Paper_Abstract | Abstract | 0 | 0 | 0.51 |
| 32 | Paper_Abstract | Regular_author | 0.03 | 0.2 | 0.12 |
| 33 | Paper_Abstract | Topic | 0.06 | 0.37 | 0.07 |
| 34 | Paper_Abstract | Program_Committee | 0.18 | 0.23 | 0.19 |
| 35 | Paper_Abstract | Chair | 0.04 | 0.28 | 0.07 |

| | | | | | |
|---|---|---|---|---|---|
| 36 | Person | Person | 1 | 1 | 1 |
| 37 | Person | Paper | 0.14 | 0.43 | 0.31 |
| 38 | Person | Abstract | 0 | 0 | 0.13 |
| 39 | Person | Regular_author | 0.22 | 0.44 | 0.13 |
| 40 | Person | Topic | 0 | 0.29 | 0 |
| 41 | Person | Program_Committee | 0 | 0.24 | 0.17 |
| 42 | Person | Chair | 0.28 | 0.53 | 0 |
| 43 | Subject_Area | Person | 0.16 | 0.41 | 0.08 |
| 44 | Subject_Area | Paper | 0.05 | 0.4 | 0.14 |
| 45 | Subject_Area | Abstract | 0 | 0 | 0.22 |
| 46 | Subject_Area | Regular_author | 0.07 | 0.24 | 0.14 |
| 47 | Subject_Area | Topic | 0.5 | 0.4 | 0.08 |
| 48 | Subject_Area | Program_Committee | 0.06 | 0.42 | 0.1 |
| 49 | Subject_Area | Chair | 0.09 | 0.34 | 0.14 |

***Table 4: N-uples selected from $m_1$[6]***

| id | e | e' | m1 | m2 | m3 |
|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.5 | 0.5 | 0.36 |
| 14 | Chairman | Chair | 1 | 1 | 0.62 |
| 23 | Paper | Paper | 1 | 1 | 1 |
| 36 | Person | Person | 1 | 1 | 1 |
| 47 | Subject_Area | Topic | 0.5 | 0.4 | 0.08 |

***Table 5: N-uples selected from $m_2$[6]***

| id | e | e' | m1 | m2 | m3 |
|---|---|---|---|---|---|
| 14 | Chairman | Chair | 1 | 1 | 0.62 |
| 18 | Co-author | Regular_author | 0.33 | 0.62 | 0.43 |
| 23 | Paper | Paper | 1 | 1 | 1 |
| 36 | Person | Person | 1 | 1 | 1 |
| 48 | Subject_Area | Program_Committee | 0.06 | 0.42 | 0.1 |

***Table 6: N-uples selected from $m_3$[6]***

| id | e | e' | m1 | m2 | m3 |
|---|---|---|---|---|---|
| 14 | Chairman | Chair | 1 | 1 | 0.62 |
| 18 | Co-author | Regular_author | 0.33 | 0.62 | 0.43 |
| 23 | Paper | Paper | 1 | 1 | 1 |
| 36 | Person | Person | 1 | 1 | 1 |
| 48 | Subject_Area | Program_Committee | 0.06 | 0.42 | 0.1 |

In the ALIN the stable marriage algorithm with incomplete lists of limited size to 1 is executed six times, each with a different metric (Jaro-Winkler, Jaccard, q-Gram, Jiang-Conrath, Lin and Wu-Palmer). This set of candidate correspondences has as attributes the six similarity metrics used. The approach described was taken from Lopes [6] that also uses this same algorithm to generate the set of candidate correspondences.

***Table 7: Candidate correspondences after the stable marriage algorithm [6]***

| id | e | e' | m1 | m2 | m3 |
|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.5 | 0.5 | 0.36 |
| 14 | Chairman | Chair | 1 | 1 | 0.62 |
| 18 | Co-author | Regular_author | 0.33 | 0.62 | 0.43 |
| 23 | Paper | Paper | 1 | 1 | 1 |
| 31 | Paper_Abstract | Abstract | 0 | 0 | 0.51 |
| 36 | Person | Person | 1 | 1 | 1 |
| 47 | Subject_Area | Topic | 0.5 | 0.4 | 0.08 |
| 48 | Subject_Area | Program_Committee | 0.06 | 0.42 | 0.1 |

The stable marriage algorithm with incomplete lists of limited size to 1, associated with a metric is a matcher, so ALIN uses six terminological matchers to select correspondences for its initial set of candidate correspondences. Each matcher generates a set of correspondences that are then combined through a union operation.

#### 3.3.1.1.2. Withdrawal of Correspondences with Semantically Different Entity Names

Next step, after the stable marriage algorithm, to generate the initial set of candidate correspondences is the removal, from the set created up to now, of all correspondences whose meanings (the most frequent meaning of the word is always chosen in wordnet) of the most important words (see section 3.1.1) of class names are not semantically related. Correspondences are considered non-semantically related if all three linguistic metrics (Wu-Palmer, Lin and Jiang-Conrath) have similarity values <= 0.9. These removed correspondences will form another set called set of correspondences with semantically different entity names, which can go back to the set of candidate correspondences in the interactive phase depending on the interactions with the expert. The set of correspondences with semantically different entity names can be seen in Table 8.

***Table 8: Set of candidate correspondences (above) and set of correspondences with semantically different entity names (below) after the withdrawal of correspondences with semantically different entity names***

| id | e | e' | m1 | m2 | m2 |
|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 |
| 23 | Paper | Paper | 1.00 | 1.00 | 1.00 |
| 31 | Paper_Abstract | Abstract | 0.00 | 0.00 | 0.51 |
| 36 | Person | Person | 1.00 | 1.00 | 1.00 |

| id | e | e' | m1 | m2 | m2 |
|---|---|---|---|---|---|
| 14 | Chairman | Chair | 1.00 | 1.00 | 0.62 |
| 18 | Co-author | Regular_author | 0.33 | 0.62 | 0.43 |
| 47 | Subject_Area | Topic | 0.50 | 0.40 | 0.08 |
| 48 | Subject_Area | Program_Committee | 0.06 | 0.42 | 0.10 |

### 3.3.1.2. Generate Initial Alignment

The goal of this phase is also to decrease the number of interactions with the expert. To do this, correspondences that have a high degree of probability of being correct, because have maximum similarity, are removed from the set of candidate correspondences, not needing to be presented to the expert. This phase forms the initial alignment with the highest cardinality possible, but with high precision. Again, two contradictory goals need to be balanced. The initial alignment is formed by all correspondences that are placed in the set of classified correspondences in the non-interactive phase as true correspondences. All correspondences with maximum similarity are inserted into the initial alignment. Later, the criteria for the selection of the correspondences of initial alignment will be reviewed.

#### 3.3.1.2.1. Automatic Classification According to the Maximum Similarity Premise

This technique, used in [6], is based on the following premise: When a correspondence is analyzed from different perspectives through similarity metrics and all of them return the maximum similarity value for this pair, then it is considered true. Thus, all candidate correspondences that fit the above premise are automatically classified and integrate into the set of classified correspondences, with their class attribute receiving the value 'YES' as a true correspondence and ceasing to be part of the set of candidate correspondences. With similarity metrics used, the maximum similarity occurs only when the names of the classes of the correspondence are completely equals.

Table 9 shows the set of classified correspondences, updated with the candidate correspondences of Table 8 that were automatically classified. Table 9 also presents the updated set of candidate correspondences, that is, by removing the correspondences already automatically classified.

***Table 9: Set of candidate correspondences (above) and set of classified correspondences (below), after the generation of initial alignment***

| id | e | e' | m1 | m2 | m2 |
|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 |
| 31 | Paper_Abstract | Abstract | 0.00 | 0.00 | 0.51 |

| id | e | e' | Belongs to alignment |
|---|---|---|---|
| 23 | Paper | Paper | yes |
| 36 | Person | Person | yes |

#### 3.3.1.2.2. Review of Automatic Classification According to the Maximum Similarity Premise

After the withdrawal of the correspondences with maximum similarity, it was verified that the precision could be improved if some additional criteria were employed. Additional criteria were then created to restricting the flow of correspondences to the set of classified correspondences, in order to increase precision. This is due to the sending, from the set of candidate correspondences, of correspondences with maximum similarity, which are supposed to be true, but which are not true. Not all correspondence with maximum similarity are in the reference alignment, which decreases precision. In this section will be shown new criteria, that the correspondences with maximum similarity must obey, and if they do not obey they are left in the set of candidate correspondences. It is expected that with these new criteria the precision will be increased, and thus the f-measure, although it should also increase the number of interactions. The criteria are shown in Table 10.

***Table 10: Additional criteria for automatic classification according to the maximum similarity premise***

| | |
|---|---|
| 1 | The correspondence is not inconsistent, according to some anti-pattern, with any other correspondence classified as belonging to the alignment. |
| 2 | Size of the names of the two classes of correspondence >= 6. |
| 3 | If the correspondence between classes has its classes as subclasses of classes of another automatically classified correspondence between classes, then both are immediate subclasses. |
| 4 | If the two classes of a correspondence between classes have subclasses, then there is some correspondence between these subclasses. |
| 5 | The correspondence between classes has its classes with equal number of subclasses. |

Criterion 1 refers to two or more correspondences, selected to the set of classified correspondences, that are in some anti-pattern, that is, if a group of correspondences enters one of the anti-patterns described in the sections 2.4.1 (page 21), 2.4.2 or 2.4.3, illustrated by the Figures 9 (page 21), 10 or 11. If it occurs then all the correspondences involved are not sent to the set of classified correspondences, remaining in the set of candidate correspondences, even though the class names are exactly the same.

As an example of criterion 2 there is the correspondence, from the set of candidate correspondences, that associates two classes Paper and Paper, as both are less than six characters so this correspondence would remain in the set of candidate correspondences. The size of the class names in the number 2 criterion was found by testing the ALIN using the OAEI conference dataset. Size 6 was what generated the greatest increase in precision.

The criterion number 3 may be better illustrated in the following example: in Figure 16 the correspondences (Student, Student) and (Woman, Woman), in the set of candidate correspondences, would be chosen, because they have maximum similarity. With this new criterion the correspondence (Woman, Woman) would be left out because there is the Secondary Student class interposing between the two classes. In this case, specifically, one class would indicate "secondary student woman" and the other "student woman". Of course, the fact of having an interposed class does not always indicate that the correspondence is not true, but it puts a suspicion on it, so it is best to place it in the set of candidate correspondences.

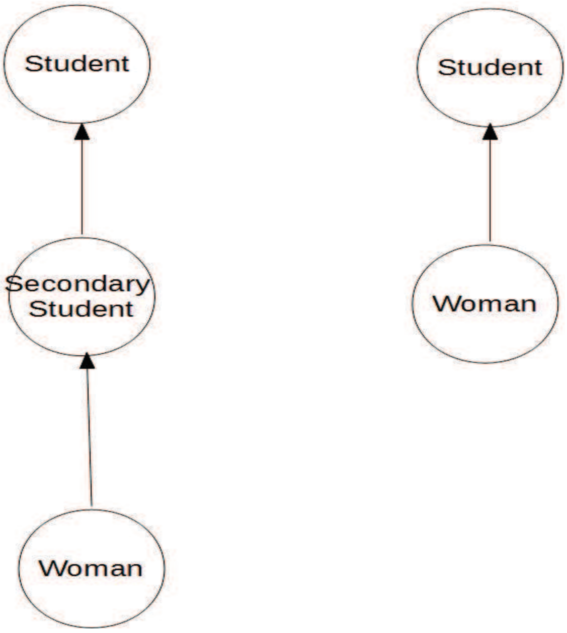


*Figure 16: Hierarchies of classes with classes of identical names with an interposed class*

As an example of criterion number 4 one can see the Figure 17. If there is correspondence A, in the set of candidate correspondences, but there is no correspondence B, the relationship between Student and Student remains in the set of candidate correspondences, not going to the set of classified correspondences.

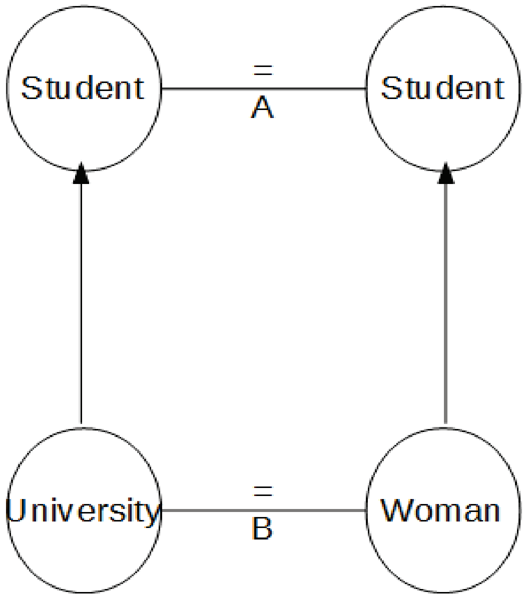


*Figure 17: Criterion number 4*

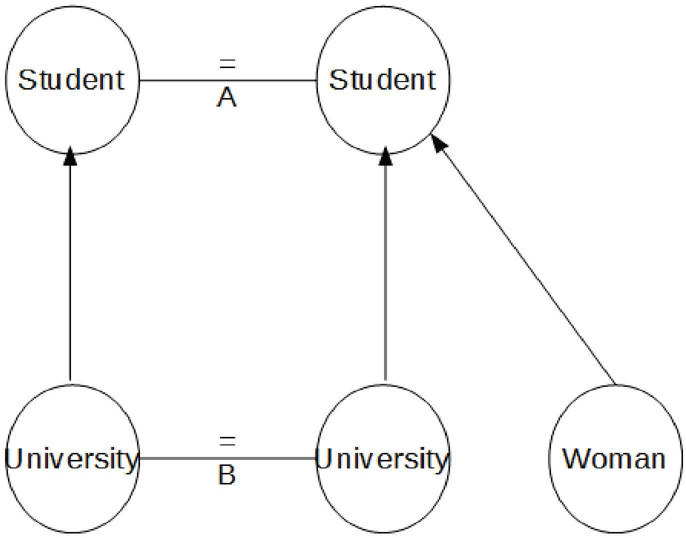


*Figure 18: Criterion number 5*

As an example of criterion number 5 one can see the Figure 18, if there are correspondences A and B, in the set of candidate correspondences, the correspondence A does not go to the set of classified correspondences, since the two Student classes have different numbers of subclasses, one has one subclass and the other has two subclasses.

*Table 11: Set of candidate correspondences (above) and set of classified correspondences (below) after withdrawal of correspondence 23 from the set of classified correspondences*

| id | e | e' | m1 | m2 | m2 |
|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 |
| 23 | Paper | Paper | 1.00 | 1.00 | 1.00 |
| 31 | Paper_Abstract | Abstract | 0.00 | 0.00 | 0.51 |

| id | e | e' | Belongs to alignment |
|---|---|---|---|
| 36 | Person | Person | yes |

Updating Table 9, page 40, we would have the Table 11, with the correspondence 23 remaining in the set of candidate correspondences because the class names have less than 6 characters in the entity names, thus disobeying criterion 2 of Table 10, page 41.

### 3.3.2. Classify and Modify Set of Candidate Correspondences

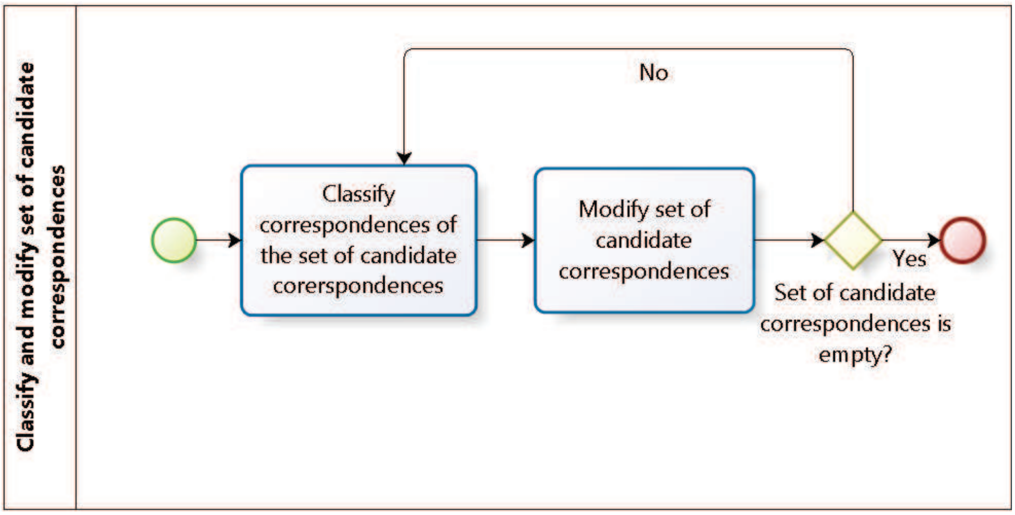

Classify and modify set of candidate correspondences
No
Classify correspondences of the set of candidate corerspondences
Modify set of candidate correspondences
Yes
Set of candidate correspondences is empty?

*Figure 19: Classify and modify set of candidate correspondences*

The objective of this phase is to achieve the highest possible quality for the generated alignment, based on the set of candidate correspondences developed in the previous phase.

This phase is divided into two subphases, which can be seen in Figure 19:

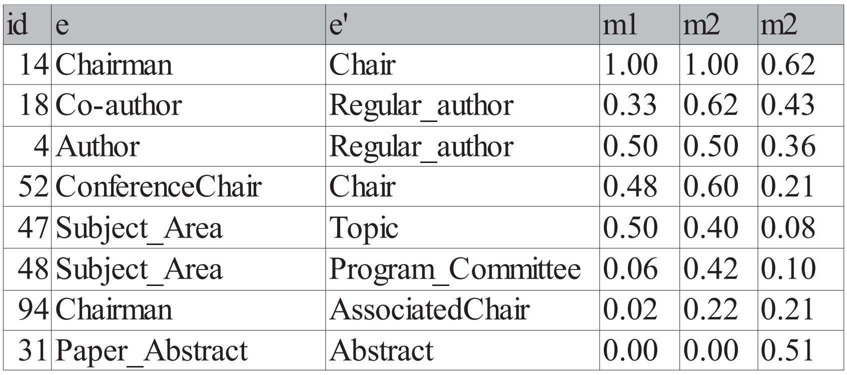
A) Classify correspondences of the set of candidate correspondences;

B) Modify set of candidate correspondences.

#### 3.3.2.1. Classify Correspondences of the Set of Candidate Correspondences

In this subphase, at each iteration, correspondences from the set of candidate correspondences are selected and presented to the expert to receive his feedback. An interaction with the expert corresponds to a question asked about at most three correspondences, as long as they have at least one of the entities in common. This is compliant with the OAEI definition [29]. For example, if the following correspondences are shown to the expert at the same time (ConferenceChair,Chair), (Chairman,Chair) e (Chairman,AssociatedChair), they will be counted as only one interaction since each correspondence has at least one entity of another correspondence.

*Table 12: Selecting correspondences for an interaction*



| id | e | e' | m1 | m2 | m2 |
|---|---|---|---|---|---|
| 14 | Chairman | Chair | 1.00 | 1.00 | 0.62 |
| 18 | Co-author | Regular_author | 0.33 | 0.62 | 0.43 |
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 |
| 52 | ConferenceChair | Chair | 0.48 | 0.60 | 0.21 |
| 47 | Subject_Area | Topic | 0.50 | 0.40 | 0.08 |
| 48 | Subject_Area | Program_Committee | 0.06 | 0.42 | 0.10 |
| 94 | Chairman | AssociatedChair | 0.02 | 0.22 | 0.21 |
| 31 | Paper_Abstract | Abstract | 0.00 | 0.00 | 0.51 |

ALIN chooses the candidate correspondences to be presented to the expert as follows:

A.1) Choose the correspondence with the highest confidence value (sum of similarity metrics) between all correspondences in the set of candidate correspondences;

A.2) Choose, following the order of the confidence value, up to two correspondences that have at least one entity equal to one entity of one correspondence previously chosen.

If the Table 12 is the set of candidate correspondences, the first correspondence selected to be part of the interaction will be that of id 14, since it is the one with the highest confidence value (the sum of the similarity metrics). The next to be selected will be the id 52, since it is next in order which has an equal entity (Chair) to an entity of the first correspondence. The third and last will be that of id 94, since it is next in order which has an equal entity (Chairman) to an entity of one of the two previous correspondences.

#### 3.3.2.2. Modify set of candidate correspondences

According to the expert feedback, other correspondences that didn't receive feedback may be removed from, and new correspondences may be inserted into, the set of candidate correspondences. These modifications are made using techniques (correspondence anti-patterns, interactively used structural matchers) placing or removing correspondences from the set of candidate correspondences. The correspondences inserted into the set of candidate correspondences comes from the set of all possible correspondences between the two ontologies or from the set of correspondences with semantically different entity names.

This process only ends when the set of candidate correspondences is empty.

##### 3.3.2.2.1. Interactive Ontology Matching with Use of Anti-Patterns and Retrieval of Correspondences

The set of candidate correspondences is modified at each interaction with the expert. Whenever the expert gives his feedback about a correspondence, this correspondence is taken from the set of candidate correspondences and placed into the set of classified correspondences. This process of modifying the set of candidate correspondences is called, in this work, a trivial modification of the set of candidate correspondences. The proposal under this approach is that the set of candidate correspondences be modified in a deeper way than simply withdrawing the correspondence that received feedback from the expert. There will be shown a series of techniques that will remove other correspondences besides those that received feedback, as also will add new correspondences, in both situations the correspondences having some relationship with the correspondence that received positive feedback.

### 3.3.2.2.2. Interactive Modification of the Set of Candidate Correspondences Using Anti-Patterns

The first technique used for the interactive modification of the set of candidate correspondences is the use of correspondence anti-patterns for the removal of correspondences from the set of candidate correspondences. This technique does not directly serve to increase the quality of the generated alignment by increasing the f-measure, but it helps to decrease the number of interactions with the expert. The technique acts on the set of candidate correspondences as follows: if the expert answers "yes", indicating that a correspondence is a true correspondence, all the correspondences of the set of candidate correspondences inconsistent with it, according to the anti-patterns, are classified as "no", indicating that these correspondences are false and are thus taken from the set of candidate correspondences and placed in the set of classified correspondences.

#### 3.3.2.2.2.1. Example of Modification of the Set of Candidate Correspondence Using Anti-Patterns

For the use of anti-patterns, one more field must be added to the Table 11, page 44. This field will indicate correspondences inconsistent with the candidate correspondences, as shown in Table 13. The correspondence of id = 4, for example, is an inconsistent correspondence to correspondence of id = 31, that is, the correspondences of id = 4 and id = 31 are in an anti-pattern, that is, if one of them is true, the other is sure to be false.

***Table 13: Set of candidate correspondences (above) and set of classified correspondences (below) before the classification of correspondence 23***

| id | e | e' | m1 | m2 | m2 | Incompatible correspondences |
|---|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 | 31 |
| 23 | Paper | Paper | 1.00 | 1.00 | 1.00 | 31 |
| 31 | Paper_Abstract | Abstract | 0.00 | 0.00 | 0.51 | 4,23 |

| id | e | e' | Belongs to alignment |
|---|---|---|---|
| 36 | Person | Person | yes |

If the correspondence of id = 23 is selected to be shown to the expert and the expert's answer is "yes" to the question if this is a true correspondence, this correspondence will be taken from the set of candidate correspondences and placed in the set of classified correspondences with the classification = "yes", and the correspondence of id = 31 will be placed in the set of classified correspondences with the classification = "no" because it is in an anti-pattern with that of id = 23 and so only one of them can be classified as "yes", which was done by the expert. Then the resulting sets are shown in Table 14.

***Table 14: Set of candidate correspondences (above) and set of classified correspondences (below) after the classification of correspondence 23***

| id | e | e' | m1 | m2 | m2 | Incompatible correspondences |
|---|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 | |

| id | e | e' | Belongs to alignment |
|---|---|---|---|
| 23 | Paper | Paper | yes |
| 31 | Paper_Abstract | Abstract | no |
| 36 | Person | Person | yes |

### 3.3.2.2.3. Interactive Modification of Set of Candidate Correspondences Through Retrieval of Correspondences

From this section will be described more techniques that aim to modify the set of candidate correspondences in an interactive manner. These techniques will be composed of three structural matchers that will be used interactively to include new correspondences in the set of candidate correspondences, which is called, in this work, retrieval of correspondences. These techniques aim to increase the recall and thus increase the quality of the generated alignment.

The techniques are:

A) Retrieval of correspondences between relationships;

B) Retrieval of correspondences between attributes;

C) Retrieval of correspondences between subclasses of the set of correspondences with semantically different entity names.

#### 3.3.2.2.3.1. Retrieval of Correspondences Between Relationships

In the technique defined in this section will be added, to the set of candidate correspondences, correspondences between relationships which will be retrieved from the set of all possible correspondences allowing to increase the recall to be reached by the generated alignment.

As mentioned before, the approach so far has only taken into account the correspondences between classes. With this new criterion will be brought the correspondences between relationships between the classes of correspondences already identified as true by the expert. The new criterion is best illustrated in Figure 20. This is expected to increase the recall of the generated alignment, although it should increase the number of interactions with the expert.

In Figure 20 it is assumed that the correspondences between o1:e1 and o2:e1 and between o1:e2 and o2:e2 have already been confirmed by the expert. Once this happens the ALIN approach includes in the set of candidate correspondences the correspondence between the relationships o1:r1 and o2:r1 with the similarity metrics being the average of the metrics of the two class correspondences. Once in the set of correspondences candidates this correspondence of relationships enters the process, and if they are chosen they will be evaluated by the expert.

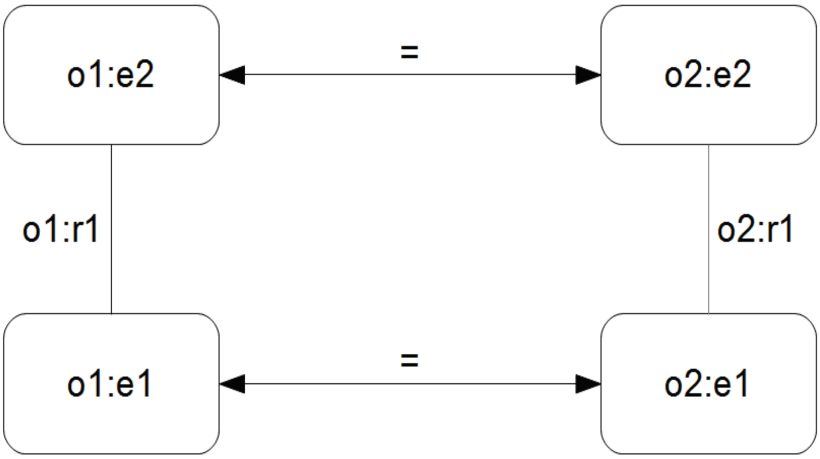


*Figure 20: Relationships between classes of class correspondences*

In addition to the relationships directly connected to the classes, a set of correspondences between relationship formed by all the super-relationships of these, in addition to the relationships among all subclasses of the classes involved, is formed. A maximum depth of 3 levels was established in the generalization hierarchy. Without this limit the number of correspondences between relationships is very large. The maximum depth of 3 levels was found by testing the ALIN using the OAEI conference dataset. This level showed the best balance between increased recall and increased number of interactions. The found set of correspondences between relationship is placed in the set of candidate correspondences. Once a correspondence between relationships has been chosen by the expert as true, all other correspondences between relationships with a common relationship are marked as false (Figure 9 - Anti-pattern of multiple entities, page 21) and are placed in the set of classified correspondences.

#### 3.3.2.2.3.1.1. Example of Modification of the Set of Candidate Correspondences through Retrieval of Correspondences Between Relationships

As shown in Table 14, page 48, the expert indicated as true the correspondence of id = 23. If there were the relationship *authorIs* between the classes Paper and Person, both of the ontology cmt, and the relationship *hasAutor* between the classes Paper and Person of the ontology conference then the correspondence authorIs and hasAuthor would enter into the set of candidate correspondences as shown in Table 15.

***Table 15: Set of candidate correspondences (above) and set of classified correspondences (below) after the classification of correspondence 23***

| id | e | e' | m1 | m2 | m2 | Incompatible correspondences |
|---|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 | |
| 51 | authorIs | hasAuthor | 1.00 | 1.00 | 1.00 | |

| id | e | e' | Belongs to alignment |
|---|---|---|---|
| 23 | Paper | Paper | yes |
| 31 | Paper_Abstract | Abstract | no |
| 36 | Person | Person | yes |

#### 3.3.2.2.3.2. Retrieval of Correspondences between Attributes

Here we have another technique that will cause, depending on the interactions with the expert, some correspondences between entities, more specifically correspondences between attributes, be rescued from the original set of all possible correspondences.

In this technique, once the expert has indicated that a correspondence between classes is true, correspondences between the attributes of these classes are selected to be placed in the set of candidate correspondences. Again, recall is expected to increase.

In order to choose which correspondences between attributes of two classes are going to be inserted in the set of candidate correspondences, the same criteria that are used for the initial selection of correspondences between classes are applied, that is, the stable marriage algorithm shown in Algorithm 1, page 34, will be applied with the classes taking the place of the ontologies and the attributes taking the place of the classes.

At each iteration with the expert, these correspondences between attributes are evaluated to see if they do not fall into the anti-pattern of multiple entities (Figure 9, page 21), this anti-pattern being checked only in relation to other correspondences of attributes.

Once in the set of candidate correspondences this correspondence between attributes enters the process, and if they are chosen they will be evaluated by the expert.

#### 3.3.2.2.3.2.1. Example of Modification of the Set of Candidate Correspondences through Retrieval of Correspondences between Attributes

*Table 16: Set of candidate correspondences (above) and set of classified correspondences (below) after classifying the correspondence 14*

| id | e | e' | m1 | m2 | m2 | Incompatible correspondences |
|---|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 | |
| 51 | authorIs | hasAuthor | 1.00 | 1.00 | 1.00 | |
| 71 | name | hasName | 1.00 | 1.00 | 1.00 | |

| id | e | e' | Belongs to alignment |
|---|---|---|---|
| 23 | Paper | Paper | yes |
| 31 | Paper_Abstract | Abstract | no |
| 36 | Person | Person | yes |

Consider the correspondence of id = 36, the class Person of the ontology cmt has the attribute "name" and the class Person of the ontology conference has the attribute "hasName". As the expert said that the correspondence of id = 36 is true then the correspondence formed by the attributes "name" and "hasName" will enter the set of candidate correspondences, as shown in Table 16.

#### 3.3.2.2.3.3. Retrieval of Correspondences between Subclasses of the Set of Correspondences with Semantically Different Entity Names

This technique aims to increase the recall of the generated alignment, thus increasing its quality. In this technique, all correspondences of classes that are subclasses of the classes of a correspondence indicated as true by the process and that are in the set of correspondences with semantically different entity names are placed in the set of candidate correspondences.

##### 3.3.2.2.3.3.1. Example of Modification of the Set of Candidate Correspondences through Retrieval of Correspondences between Subclasses of the Set of Correspondences with Semantically Different Entity Names

*Table 17: Set of candidate correspondences (above), set of classified correspondences, and set of correspondences with semantically different entity names (below) after classifying correspondence 36*

| id | e | e' | m1 | m2 | m2 | Incompatible correspondences |
|---|---|---|---|---|---|---|
| 4 | Author | Regular_author | 0.50 | 0.50 | 0.36 | |
| 18 | Co-author | Regular_author | 0.33 | 0.62 | 0.43 | |
| 51 | authorIs | hasAuthor | 1.00 | 1.00 | 1.00 | |
| 71 | name | hasName | 1.00 | 1.00 | 1.00 | |

| id | e | e' | Belongs to alignment |
|---|---|---|---|
| 23 | Paper | Paper | yes |
| 31 | Paper Abstract | Abstract | no |
| 36 | Person | Person | yes |

| id | e | e' | m1 | m2 | m2 |
|---|---|---|---|---|---|
| 14 | Chairman | Chair | 1.00 | 1.00 | 0.62 |
| 47 | Subject_Area | Topic | 0.50 | 0.40 | 0.08 |
| 48 | Subject_Area | Program_Committee | 0.06 | 0.42 | 0.10 |

Assume that Co-author is a subclass of Person in the ontology cmt and Regular_Author is a subclass of Person in the ontology conference. So the correspondence of id = 18 would enter into the set of candidate correspondences, as can be seen in Table 17. Table 8, page 39, shows the previous set of correspondences with semantically different entity names.

### 3.3.3. Summary of the Techniques used by the ALIN Approach

As part of the ALIN approach, the following techniques are used to generate and modify the set of candidate correspondences and to classify the correspondences:

1 - Stable marriage with incomplete list with limited size to 1;

2 - Withdraw of correspondences with semantically different entity names;

3 - Automatic classification according to the maximum similarity premise;

4 - Review of automatic classification according to the maximum similarity premise;

5 – Direct classification (Interaction with the expert);

6 – Indirect classification (Use of correspondence anti-patterns);

7 - Retrieval of correspondences between relationships (Interactively used structural matcher);

8 - Retrieval of correspondences between attributes (Interactively used structural matcher);

9 - Retrieval of correspondences between subclasses of the set of correspondences with semantically different entity names (Interactively used structural matcher).

# 4. Evaluation

*This chapter evaluates the techniques used in the ALIN approach. The techniques will be added one by one to observe the impact of each of them. Then a comparison with other OAEI participant approaches will be made.*

## 4.1. Ontology Alignment Evaluation Initiative (OAEI)

Ontology Alignment Evaluation Initiative (OAEI) is a coordinated international initiative, which organizes the evaluation of an increasing number of ontology matching systems. Its main goal is to compare systems and algorithms openly and on the same basis, in order to allow anyone to draw conclusions about the best alignment strategies [30]. OAEI provides reference alignments for a number of evaluation domains (datasets).

One of the datasets available at OAEI is the domain of academic conferences. The conference dataset consists of 7 ontologies. There is a reference alignment between each pair of ontologies, totaling 21 reference alignments. The ontologies of datasets have three types of entities: concepts (classes), data properties (attributes) and object properties (relationships).

## 4.2. ALIN Architecture

The proposed approach was implemented using the following Java APIs:

Stanford-corenlp[4], with a routine to put a word in canonical form; Simmetrics[5], with string-based similarity metrics; WS4J[6], with Wordnet base-based linguistic metrics; And Alignment[7], which contains routines for handling ontologies written in OWL.

### 4.3. Evaluation Overview and Designed Analysis

In order to verify the cause and effect relationships between the variables present in the proposed approach, an experimental approach was used. Different scenarios were defined with variation of the techniques used, each scenario including a new technique, in addition to the use of all previous techniques. During the execution of these scenarios, data was collected for precision, recall, f-measure, number of interactions with the expert, true positives found automatically and true positives answered by the expert for each pair of ontologies that compose the dataset used in this experiment.

The data collected was analyzed using the descriptive technique. Following the OAEI approach, the data was aggregated by dataset and an weighted average value for the quality measures was determined, in addition to the sum of the interactions with the expert, in each scenario.

The results obtained were compared to each new scenario, verifying if the technique used improves the expected variable. In addition the developed program participated in the track of Interactive Matching of OAEI 2016[8], allowing the comparison of our proposal with other existing ones.

[4] "Stanford CoreNLP". Available at http://stanfordnlp.github.io/CoreNLP/ Last accessd on Sept, 15, 2016.

[5] "String Similarity Metrics for Information Integration". Available on http://www.coli.uni-saarland.de/courses/LT1/2011/slides/stringmetrics.pdf. Last accessed on Apr, 19, 2016.

[6] "WS4J". Available at https://code.google.com/archive/p/ws4j/ Last accessed on Apr, 11, 2016.

[7] "Alignment API". Available at http://alignapi.gforge.inria.fr/ Last accessed on Apr, 11, 2016.

[8] Available at http://oaei.ontologymatching.org/2016/results/interactive/, last accessed on Dec, 19, 2016.

## 4.4. Analysis of the Results of the Stable Marriage With Incomplete List with Limited Size to 1 Algorithm and Withdraw of Correspondences with Semantically Different Entity Names

First, ALIN was executed using the “stable marriage with incomplete list with limited size to 1”, “withdraw of correspondences with semantically different entity names” and direct classification techniques, described in section 3.3.1.1.2, page 38, (this was called T1). The result was compared to an execution of ALIN without the use of any technique, that is, only with the interaction with the expert as described in the section 3.3.2.1, page 45, (this was called T0).

That is, the execution T0 was done only using the direct classification technique, item 5 of section 3.3.3, page 54. The T1 implementation used the techniques stable marriage with incomplete list with limited size to 1, withdraw of correspondences with semantically different entity names and direct classification, respectively the items 1, 2 and 5 of section 3.3.3.

The execution of the 21 alignments of the OAEI conference dataset was performed in each ALIN run.

***Table 18: Comparison between matching executions T0 and T1***

| | NI | Recall |
|---|---|---|
| T0 | 91864 | 1,0 |
| T1 | 316 | 0,61 |

The result is shown in Table 18 (NI is the sum of all interactions with the expert of the 21 alignments, the recall shown is the weighted average, where the weight is proportional to the number of correspondences of the reference alignment of each alignment, of the recall of the 21 alignments), which shows that the techniques used greatly decrease the number of interactions with the expert (- 99%) proportionally decreasing much less recall (about 40%). That is the objective of the subphase 'selection of the set of candidate correspondences': decrease the number of interaction with the expert maintaining a good recall. The precision in both executions is 1 because the only way to classify the candidate correspondences, up to now, is the interation with the expert, and it is assumed he not make mistakes.

The use of the weighted average for the calculation of the precision and recall of the 21 conference dataset alignments, the weight being the cardinality of the reference alignment, is the standard to OAEI. This standard is a good one, because if a simple mean were used, a technique that would make a false positive correspondence become true positive one would have much more weight on the final result in an alignment with few correspondences than another technique that would do the same on a alignment with more correspondences.

During the evaluation the generated alignments with the use of each presented technique were compared. The comparison was made through a table where each line represents the execution of the approach with the inclusion of a certain technique plus all the techniques of the previous line, using the OAEI conference dataset.

## 4.5. Analysis of the Results of the Automatic Classification According to the Maximum Similarity Premise

Table 19 shows the results obtained using the generation of the initial set of classified correspondences (called T2), as described in section 3.3.1.2.1, page 40.

*Table 19: Comparison between different matching executions*

| | NI | True positives found automatically | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| T0 | 91864 | 0 | 305 | 1.0 | 1.0 | 1.0 |
| T1 | 316 | 0 | 187 | 1.0 | 0.61 | 0.75 |
| T2 | 203 | 118 | 69 | 0.93 | 0.61 | 0.73 |

The execution T2 was done using the “stable marriage with incomplete list with limited size to 1”, “withdraw of correspondences with semantically different entity names", "automatic classification according to the maximum similarity premise" and direct classification techniques, respectively described in the items 1, 2, 3 and 5 of section 3.3.3. It can be seen that the number of interactions with the expert has decreased, but with a precision of less than 1, which we will try to mitigate with the next technique used.

## 4.6. Analysis of the Results of the Review of Automatic Classification According to the Maximum Similarity Premise

***Table 20: Comparison between different matching executions***

| | NI | True positives found automatically | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| T0 | 91864 | 0 | 305 | 1.0 | 1.0 | 1.0 |
| T1 | 316 | 0 | 187 | 1.0 | 0.61 | 0.75 |
| T2 | 203 | 118 | 69 | 0.93 | 0.61 | 0.73 |
| T3 | 233 | 81 | 106 | 0.95 | 0.61 | 0.74 |

Table 20 shows the result for the new execution with the use of additional criteria (T3), described in the section 3.3.1.2.1, page 40. The execution T3 was done using the "stable marriage with incomplete list with limited size to 1", "withdraw of correspondences with semantically different entity names", "automatic classification according to the maximum similarity premise", "review of automatic classification according to the maximum similarity premise" and direct classification techniques, respectively described in the items 1, 2, 3, 4 and 5 of section 3.3.3. The result shows an increase in precision, which was expected.

## 4.7. Analysis of Results of Anti-Pattern Usage

The use of anti-patterns in the ALIN approach aims to reduce the number of interactions with the expert, by eliminating candidate correspondences that are inconsistent with those classified as true.

The number of correspondences that was included in one of the three anti-patterns, in all 21 alignments of the conference dataset, was counted and the result can be seen in Table 21. The graph in Figure 21 shows the percentage of correspondences that was included in some correspondence anti-pattern.

*Table 21 - Number of correspondences in some anti-pattern*

| | Number of correspondences |
|---|---|
| anti-pattern of multiple entities | 2615 |
| anti-pattern of cross correspondences | 39 |
| anti-pattern of disjunction and generalization | 1603 |

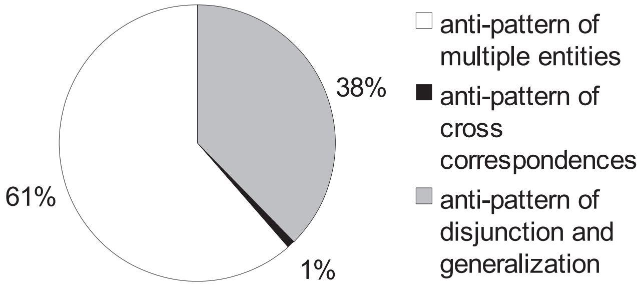


*Figure 21: Percentage of correspondences in some anti-pattern*

*Table 22: Comparison between different matching executions*

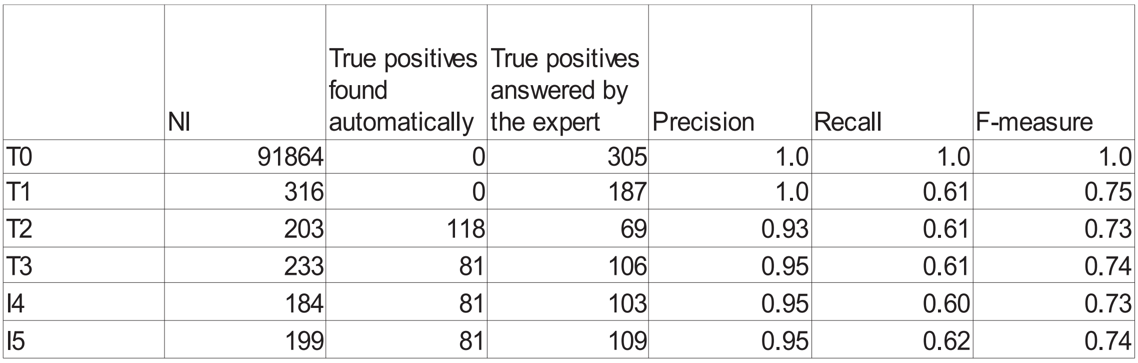
| | NI | True positives found automatically | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| T0 | 91864 | 0 | 305 | 1.0 | 1.0 | 1.0 |
| T1 | 316 | 0 | 187 | 1.0 | 0.61 | 0.75 |
| T2 | 203 | 118 | 69 | 0.93 | 0.61 | 0.73 |
| T3 | 233 | 81 | 106 | 0.95 | 0.61 | 0.74 |
| I4 | 184 | 81 | 103 | 0.95 | 0.60 | 0.73 |

By performing the approach with the use of anti-pattern (I4) described in the section 3.3.2.2.2, page 47, the result shown in Table 22 was found. The execution I4 was done using the “stable marriage with incomplete list with limited size to 1”, “withdraw of correspondences with semantically different entity names”, “automatic classification according to the maximum similarity premise”, “review of automatic classification according to the maximum similarity premise”, direct classification and indirect classification techniques, respectively described in the items 1, 2, 3, 4, 5 and 6 of section 3.3.3.

The result shows a decrease in the number of interactions with the expert, which is to be expected, since several of the correspondences are excluded from the candidate correspondences besides those that the expert indicates as true. There is a small decrease in quality that is explained by the fact that there are wrong automatic classification (with maximum similarity) of correspondences as true and so some true correspondences that were inconsistent with them were discarded and not presented to the expert.

## 4.8. Analysis of the Results of the Retrieval of Correspondence between Relationships

***Table 23: Comparison between different matching executions***



| | NI | True positives found automatically | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| T0 | 91864 | 0 | 305 | 1.0 | 1.0 | 1.0 |
| T1 | 316 | 0 | 187 | 1.0 | 0.61 | 0.75 |
| T2 | 203 | 118 | 69 | 0.93 | 0.61 | 0.73 |
| T3 | 233 | 81 | 106 | 0.95 | 0.61 | 0.74 |
| I4 | 184 | 81 | 103 | 0.95 | 0.60 | 0.73 |
| I5 | 199 | 81 | 109 | 0.95 | 0.62 | 0.74 |

In this section, an increase in recall was attempted by retrieving correspondences between relationships as described in section 3.3.2.2.3.1, page 49. The execution I5 was done using the “stable marriage with incomplete list with limited size to 1”, “withdraw of correspondences with semantically different entity names”, “automatic classification according to the maximum similarity premise", "review of automatic classification according to the maximum similarity premise", direct classification, indirect classification techniques and "retrieval of correspondences between relationships", respectively described in the items 1, 2, 3, 4, 5, 6 and 7 of section 3.3.3. With the addition of the new technique we have the result shown in Table 23, which shows an increase in the number of interactions with the expert and in the recall, generating an improvement in the quality of the alignment.

## 4.9. Analysis of the Results of the Retrieval of Correspondence between Attributes

***Table 24: Comparison between different matching executions***

| | NI | True positives found automatically | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| T0 | 91864 | 0 | 305 | 1.0 | 1.0 | 1.0 |
| T1 | 316 | 0 | 187 | 1.0 | 0.61 | 0.75 |
| T2 | 203 | 118 | 69 | 0.93 | 0.61 | 0.73 |
| T3 | 233 | 81 | 106 | 0.95 | 0.61 | 0.74 |
| I4 | 184 | 81 | 103 | 0.95 | 0.60 | 0.73 |
| I5 | 199 | 81 | 109 | 0.95 | 0.62 | 0.74 |
| I6 | 236 | 81 | 123 | 0.95 | 0.67 | 0.78 |

In this section an increase in recall was attempted by retrieving correspondences between attributes as described in section 3.3.2.2.3.2, page 51. The execution I6 was done using the "stable marriage with incomplete list with limited size to 1", "withdraw of correspondences with semantically different entity names", "automatic classification according to the maximum similarity premise", "review of automatic classification according to the maximum similarity premise", direct classification, indirect classification techniques, "retrieval of correspondences between relationships" and "retrieval of correspondences between attributes", respectively described in the items 1, 2, 3, 4, 5, 6, 7 and 8 of section 3.3.3. With the addition of new technique we have the result shown in Table 24, which shows an increase in the number of interactions with the expert and in the recall, generating an improvement in the quality of the alignment.

## 4.10.Analysis of the Results of the Retrieval of Correspondences between Subclasses of the Set of Correspondences with Semantically Different Entity Names

*Table 25: Comparison between different matching executions*

| | NI | True positives found automatically | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|
| T0 | 91864 | 0 | 305 | 1.00 | 1.00 | 1.00 |
| T1 | 316 | 0 | 187 | 1.00 | 0.61 | 0.75 |
| T2 | 203 | 118 | 69 | 0.93 | 0.61 | 0.73 |
| T3 | 233 | 81 | 106 | 0.95 | 0.61 | 0.74 |
| I4 | 184 | 81 | 103 | 0.95 | 0.60 | 0.73 |
| I5 | 199 | 81 | 109 | 0.95 | 0.62 | 0.74 |
| I6 | 236 | 81 | 123 | 0.95 | 0.67 | 0.78 |
| I7 | 326 | 81 | 144 | 0.96 | 0.74 | 0.83 |

In this section an increase in recall was attempted by retrieving correspondences between subclasses of the set of correspondences with semantically different entity names as described in section 3.3.2.2.3.3, page 53. The execution I7 was done using the "stable marriage with incomplete list with limited size to 1", "withdraw of correspondences with semantically different entity names", "automatic classification according to the maximum similarity premise", "review of automatic classification according to the maximum similarity premise", direct classification, indirect classification techniques, "retrieval of correspondences between relationships", "retrieval of correspondences between attributes" and "retrieval of correspondences between subclasses of the set of correspondences with semantically different entity names", respectively described in the items 1, 2, 3, 4, 5, 6, 7, 8 and 9 of section 3.3.3. After the inclusion of this new technique in the interactive modification of the set of candidate correspondences, the results shown in Table 25 were reached.

## 4.11.Inconsistencies in the Generated Alignment of ALIN Approach

*Table 26: Statistics of consistency*

| Matcher | #Align. | #Incoh.Align. | #TotConsist.Viol. | #AvgConsist.Viol. |
|---|---:|---:|---:|---:|
| Alin | 21 | 0 | 0 | 0 |
| AML | 21 | 0 | 0 | 0 |
| CroMatch | 21 | 8 | 25 | 1.25 |
| DKPAOM | 21 | 0 | 0 | 0 |
| FCAMap | 21 | 12 | 150 | 7.14 |
| Lily | 21 | 13 | 167 | 8.79 |
| LogMap | 21 | 0 | 0 | 0 |
| LogMapBio | 21 | 0 | 0 | 0 |
| LogMapLt | 21 | 6 | 81 | 3.86 |
| LPHOM | 21 | 0 | 0 | 0 |
| LYAM | 21 | 1 | 3 | 0.14 |
| NAISC | 21 | 20 | 701 | 50.07 |
| XMap | 21 | 0 | 0 | 0 |

The OAEI evaluation on the conference data track has two modes: there is a non-interactive execution evaluation of the tools, where participate the automatic ontology matching tools, and there is the evaluation of interactive execution. Although the main focus of ALIN was participation in interactive execution, it also participated in non-interactive evaluation in OAEI 2016 and in this evaluation was computed the number of  logical inconsistencies generated by the tools and ALIN did very well, as can be seen in Table 26. The OAEI did not evaluate the logical inconsistencies in the interactive execution, but because the ALIN send for evaluation by the expert (or with the use of anti-patterns) all the correspondences of the set of candidate correspondences and if you assume that the expert does not make mistakes, even in the interactive phase the ALIN should not have generated many logical inconsistencies.

## 4.12.Comparison among Tools that Participated in the OAEI Interactive Conference Track

The OAEI provides a comparison between tool performance in the ontology matching process each year, and one of the ontology groups used is the conference dataset used in this thesis.

***Table 27: Comparison of OAEI interactive conference track participant tools***

| | Year of participation in OAEI | Number of distinct questions | Number of questions | NI | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|---|---|---|
| Alin | 2016 | | 574 | 326 | 144 | 0.957 | 0.735 | 0.831 |
| AML | 2016 | | 270 | 271 | 47 | 0.912 | 0.711 | 0.799 |
| LogMap | 2016 | | 142 | 142 | 49 | 0.886 | 0.610 | 0.723 |
| Xmap | 2016 | | 4 | 4 | 0 | 0.837 | 0.574 | 0.681 |
| Jarvis | 2015 | 154 | 154 | | 53 | 0.810 | 0.550 | 0.650 |
| ServOMBI | 2015 | 295 | 535 | | 156 | 1.000 | 0.650 | 0.788 |
| WeSeE | 2014 | | | | | 0.734 | 0.404 | 0.473 |
| Hertuda | 2014 | | | | | 0.790 | 0.497 | 0.582 |

Table 27 and Figure 22 shows a comparison of the tools that participated in the OAEI interactive conference track. The tools did not all participate in the same year. NI means number of interactions. In each interaction there can be up to three questions, the number of questions can contain repeated questions. In 2015 the number of interactions was not calculated, but the number of different questions asked to the expert. Before 2015 there was no calculation of the number of questions. ALIN with the results achieved with the use of all the techniques, the same that participated in the OAEI interactive conference track, which we call I7 at this thesis [31], can be seen at Table 27.

Table 27 shows the performance of the interactive tools in OAEI, with the expert hitting 100% of the answers in relation to the conference dataset, in which ALIN ranked first in quality, with very reasonable values in the number of interactions with the expert. Therefore, the use of expert feedback to modify the set of candidate correspondences, using the techniques shown in this work, generates a high quality alignment in case the expert does not miss the answers. In addition, the alignment generated by ALIN in its non-interactive phase did not generate inconsistencies, as can be seen in Table 26, page 64, and by the characteristics of the ALIN of classification of the correspondences of the set of candidate correspondences, probably the final generated alignment has few inconsistencies.

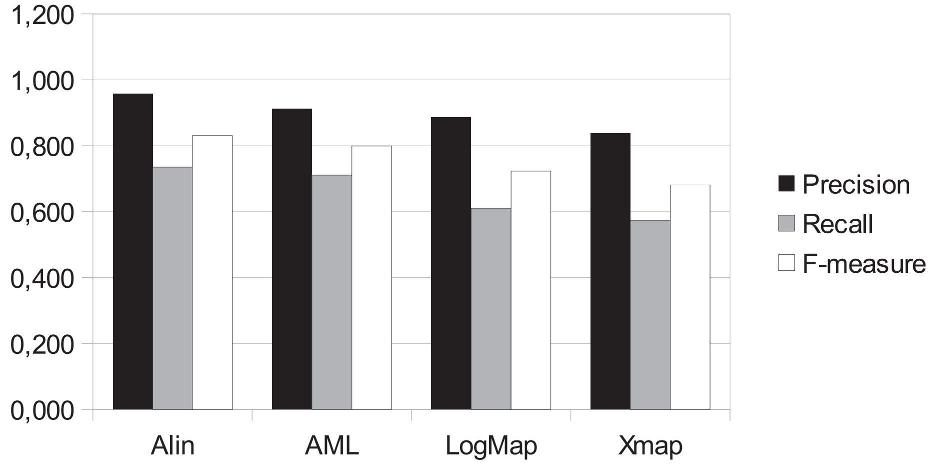


*Figure 22: Graphic comparing the performance of different tools*

## 4.13.Comparison among Tools that Participated in the OAEI Interactive Conference Track with no 100% Hit Rate

*Table 28: Comparison of ALIN with OAEI participating tools, interactive matching of the conference dataset with 90% hit rate*

| | NI | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| Alin | 315 | 124 | 0,794 | 0,67 | 0,727 |
| AML | 285 | 51 | 0,847 | 0,703 | 0,768 |
| LogMap | 140 | 45 | 0,847 | 0,6 | 0,702 |
| Xmap | 4 | 0 | 0,837 | 0,574 | 0,681 |
| Jarvis | | 34,3 | 0,73 | 0,53 | 0,61 |
| ServOMBI | | 137,7 | 0,89 | 0,57 | 0,70 |

In the OAEI evaluation, executions are performed simulating a percentage of error by the expert, specifically each execution of the tool to make an alignment is executed four times: one with 100% of hits, already shown in section 4.12 in relation to the conference dataset. The results shown in the Tables 28, 29 and 30 are related to the executions with 90%, 80% and 70% hit rate. One can see the comparison between the variables measured in these executions in the Figures 23, 24, 25, 26 and 27. That shows that the number of interactions falls with the decrease of the hit rate. This is due to ALIN's characteristic of placing new correspondences in the set of candidate correspondences. When the expert misses a correspondence saying that it is false when it is true, all the properties that would be inserted into the set of candidate correspondences are not, which decreases the number of interactions. This could be compensated for by the number of properties brought by the false correspondences and indicated as true, but the probability of two concepts of a false correspondence having object properties is smaller than that of a true correspondence, which explains the smaller number of interactions.

***Table 29: Comparison of ALIN with OAEI participating tools, interactive matching of the conference dataset with 80% hit rate***

| | NI | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| Alin | 303 | 108 | 0,672 | 0,615 | 0,642 |
| AML | 290 | 53 | 0,767 | 0,681 | 0,721 |
| LogMap | 143 | 38 | 0,822 | 0,588 | 0,686 |
| Xmap | 4 | 0 | 0,837 | 0,574 | 0,681 |
| Jarvis | | 28,7 | 0,67 | 0,52 | 0,58 |
| ServOMBI | | 122,0 | 0,80 | 0,50 | 0,61 |

The number of true positives that goes down can also be explained by the wrong marking of correspondences, because the wrong marking of correspondences makes the approach doesn't bring properties between the correspondences, properties that could be right. And it is also explained by the use of anti-patterns, because an false correspondence that is indicated as true can remove a true one that is in some anti-pattern with it.

The drop in precision is mainly due to the expert's error, because an expert's error indicating a correspondence as true causes a drop in precision. The fall in recall is mainly

*Table 30: Comparison of ALIN with OAEI participating tools, interactive matching of the conference dataset with 70% hit rate*

| | NI | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| Alin | 303 | 93 | 0,57 | 0,568 | 0,569 |
| AML | 284 | 47 | 0,718 | 0,651 | 0,683 |
| LogMap | 144 | 33 | 0,803 | 0,585 | 0,677 |
| Xmap | 4 | 0 | 0,837 | 0,574 | 0,681 |
| Jarvis | | 24,3 | 0,62 | 0,51 | 0,56 |
| ServOMBI | | 105,0 | 0,66 | 0,43 | 0,52 |

due to the expert's error, indicating a true correspondence as false, and the use of anti-patterns, which causes true correspondences to be taken from the set of candidate correspondences, and the non-inclusion of true properties.

It is also possible to notice that the ALIN has a sharper drop than the other tools, due to the fact that ALIN relies more on the expert's feedback than the other tools, besides the fact that ALIN uses techniques of adding new correspondences that are impaired when the expert misses.

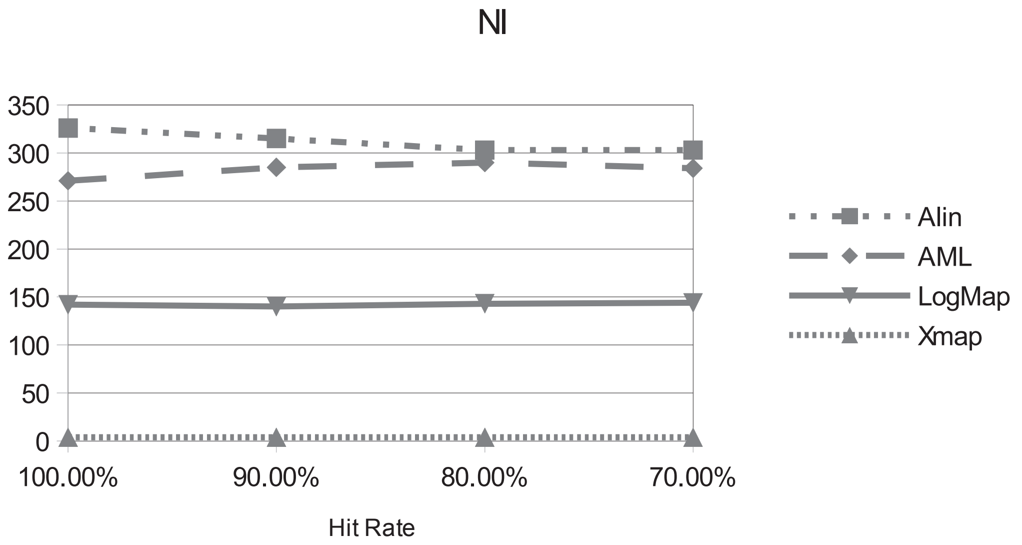


*Figure 23: NI of the evaluation of the tools*

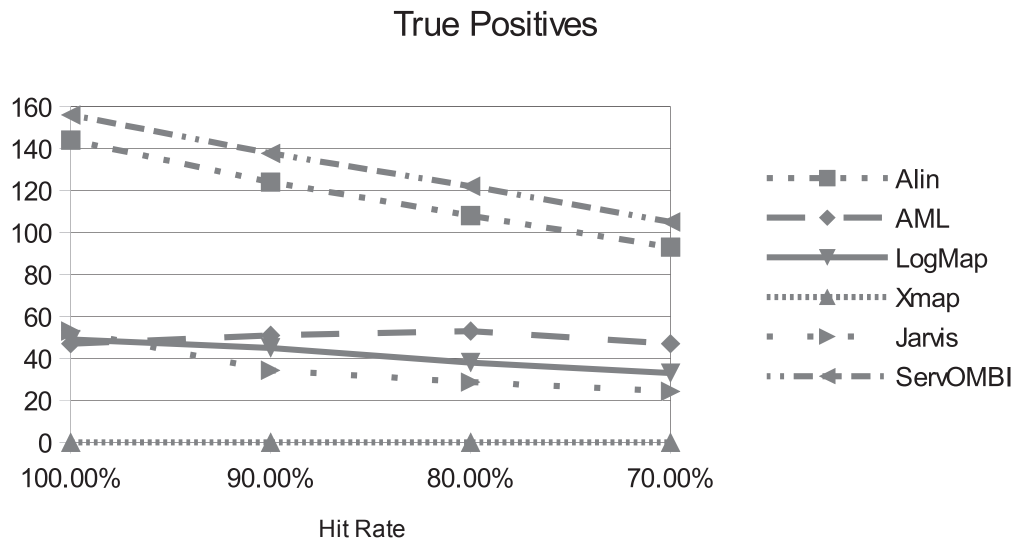


*Figure 24: True positives of the evaluation of the tools*

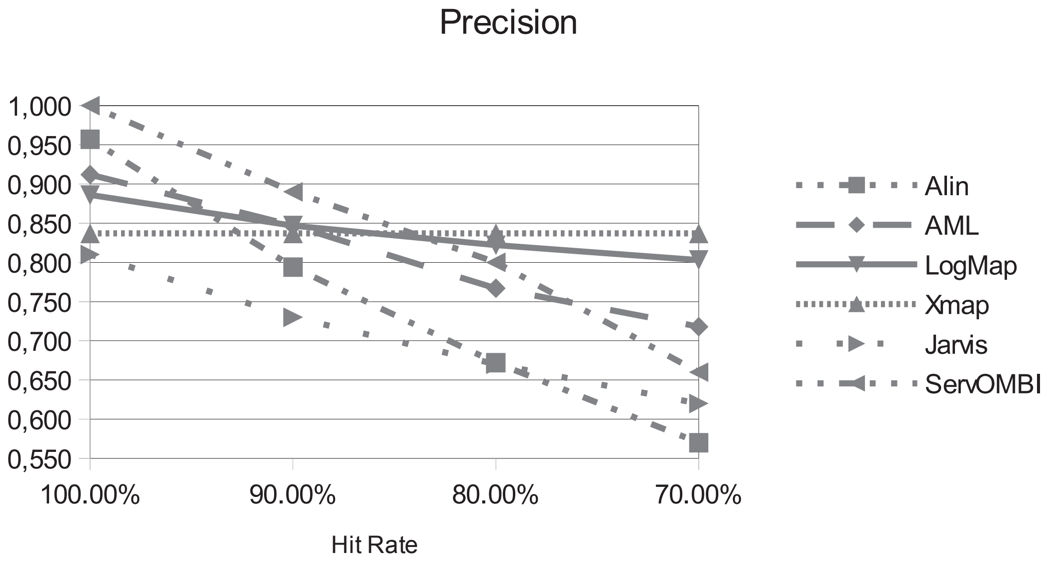


*Figure 25: Precision of the evaluation of the tools*

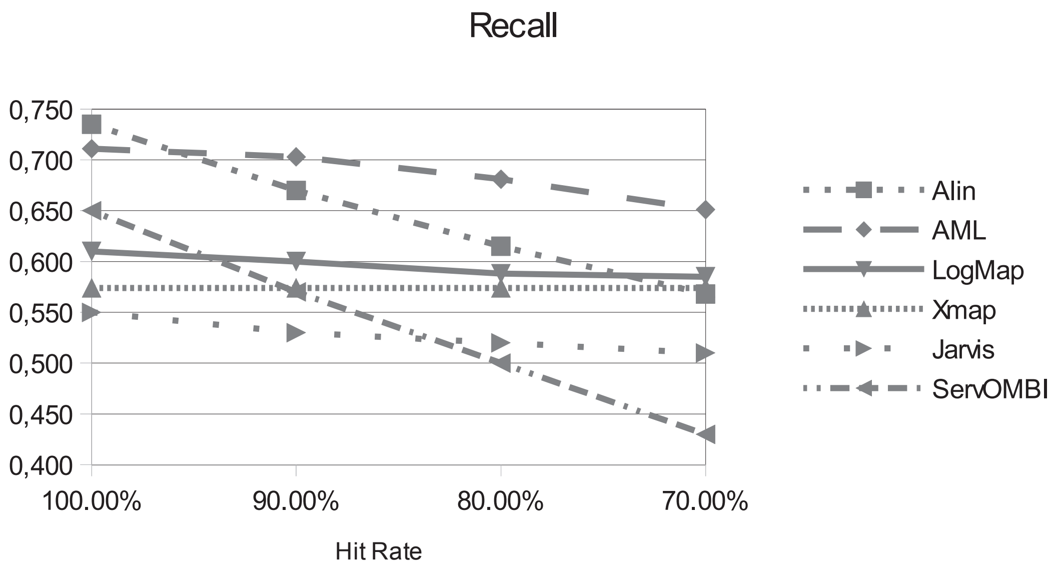


*Figure 26: Recall of the evaluation of the tools*

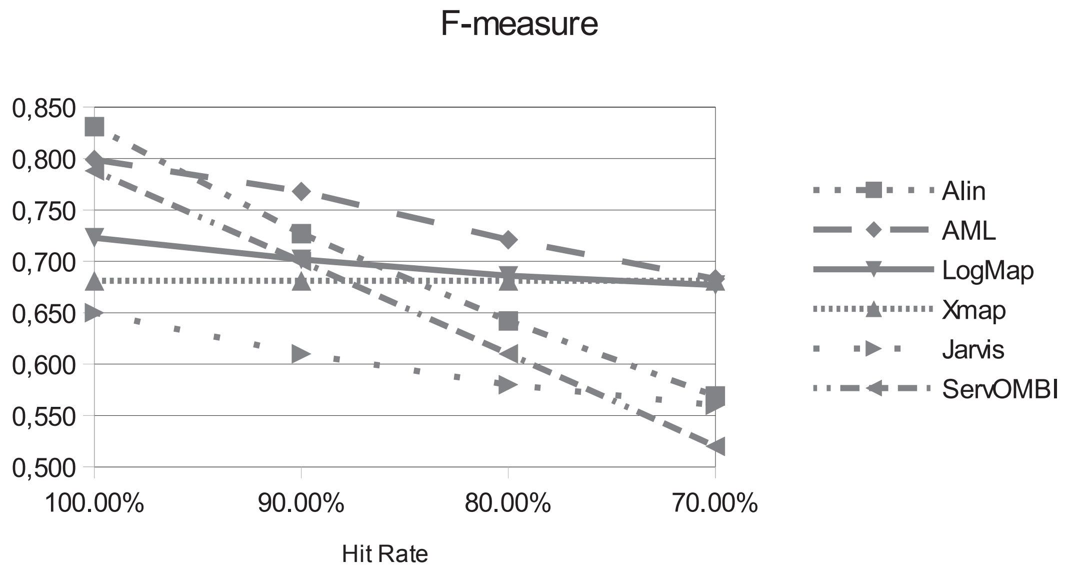


*Figure 27: F-measure of the evaluation of the tools*

## 4.14. Comparison among Tools that Participated in the OAEI Interactive Anatomy Track

*Table 31: Comparison of 2016 OAEI interactive anatomy track participant tools*

| | NI | True positives answered by the expert | Precision | Recall | F-measure |
|---|---|---|---|---|---|
| Alin | 803 | 626 | 0,993 | 0,749 | 0,854 |
| AML | 241 | 51 | 0,968 | 0,948 | 0,958 |
| LogMap | 590 | 287 | 0,982 | 0,846 | 0,909 |
| Xmap | 35 | 5 | 0,929 | 0,867 | 0,897 |

ALIN also participated in the Interactive Anatomy Track of OAEI 2016. In this track the alignment between two ontologies is made, one being the anatomy of a human being and the other the anatomy of a mouse. The total number of possible pairs between the two ontologies is more than 9 million, which has meant that ALIN does not use Wordnet, harming the stable marriage algorithm, which uses three metrics using Wordnet, and making it impossible to use retrieval of subclasses. In addition there are virtually no properties (attributes and relationships), so which ALIN almost can not use the retrieval of correspondences between attributes and the retrieval of correspondences between relationships. The results of this track are shown in Table 31.

# 5. Related work

*This chapter will present the state-of-the-art research-related works. Here we present a comparative study of the works related to the proposal of this research.*

## 5.1. Description of the Related Approaches

After reviewing the state of the art of ontology matching approaches, a set of proposals that consider the participation of the domain expert in this process as a way to improve the quality of the alignment was selected. The approaches described below have as a characteristic the request of expert feedback on correspondences, presenting different strategies for selecting these correspondences and for propagating the effect of expert feedback to other correspondences.

### 5.1.1. AML

AML has been participating in the OAEI evaluation since 2009.

AML [27][16] is an ontology matching system that can be used both for automatic ontology matching and for interactive ontology matching. The AML initially selects a set of candidate correspondences based on lexical, semantic and structural similarities. In automatic mode, correspondences that are above a given threshold are placed in the final alignment, which then goes through a repair process. In the interactive mode two thresholds are used, the correspondences that are above the greater threshold and generate inconsistencies are presented to the expert and the others are automatically approved. The correspondences that remain between the two thresholds and do not generate inconsistencies are presented to the expert and the others are automatically discarded. Those that are below the lower threshold are discarded automatically. If a correspondence receives positive feedback all correspondences that have an entity in common with it are classified as false, as well as all correspondences that generate inconsistencies with it.

### 5.1.2. LogMap

LogMap has participated in OAEI evaluations since 2011.

LogMap [26][32] is a highly scalable interactive ontology matching system with built-in features of logical reasoning and diagnosis and repair of inconsistencies. LogMap initially selects a set of candidate correspondences with a high degree of similarity, then correspondences that are semantically and structurally related to these correspondences are added. After this step there is a new one in which correspondences are drawn from the set of candidate correspondences if they are considered unreliable according to lexical and semantic characteristics. The remaining correspondences of the set of candidate correspondences are presented to the expert in order of highest similarity value. Correspondences that have an entity equal to or conflict with correspondences classified by the expert as true are classified as false. The expert can stop the interaction at any time, with the remaining correspondences in the set of candidate correspondences automatically decided.

### 5.1.3. XMAP

XMAP [19][33] has participated in OAEI evaluations since 2013.

XMAP is an interactive ontology matching system that uses three types of similarity metrics to select correspondences: structural, string and semantic. From this choice and from two previously selected thresholds, the correspondence that goes to the final alignment is chosen by the higher threshold. Correspondences between the two thresholds are presented to the expert. There is no propagation of feedback.

### 5.1.4. Jarvis

Jarvis participated in the OAEI evaluation in 2015.

Lopes [6] proposed Jarvis, an interactive approach to the ontology matching process that applies the query-by-committee technique. Jarvis uses a classifier committee that is composed of 3 classifiers (Perceptron, Naive Bayes and Random Forest). The classifiers are used in two moments in the process: when to choose which correspondences will be presented to the expert for feedback and at the time of automatically classify correspondences that have not been classified by the expert.

The set of candidate correspondences is formed using the stable marriage algorithm, using terminological metrics. In a next step, all the correspondences that have reached the maximum value in all the metrics of similarity are withdraw from the set of candidate correspondences and automatically classified as true.

The correspondences to be presented to the expert are ordered at each iteration in decreasing order of disagreement between classifiers, there is disagreement among classifiers when one of them does not agree with the others.

But not all the correspondences in the set of candidate correspondences are presented to the expert, the number of interactions is limited. The rest of the correspondences in the set of candidate correspondences not classified by the expert are classified by the classifiers.

Correspondences that have maximum similarity, those that received positive feedback from the expert, and those that have been classified by the classifiers as true will be part of the generated alignment.

### 5.1.5. ALIN with query-by-committee

This version of ALIN has never participated in OAEI.

This earlier version of ALIN [34] did not receive feedback from all correspondences in the set of candidate correspondences nor did it retrieve correspondences. They were classified as part of the set of candidate correspondences by the expert and the remaining part was classified by a set of classifiers, the same as Jarvis. ALIN already used anti-patterns. The results obtained were higher than those of Jarvis, since part of the correspondence classified by the classifiers in Jarvis was classified by the anti-patterns in ALIN, which reduced the error rate.

### 5.1.6. WeSeE

The WeSeE system participated in the OAEI evaluation between 2012 and 2014.

The WeSeE System [17] does the ontology matching process by calculating the similarity between the entities of the ontologies involved. For each entity there are associated documents, which are formed by titles and summaries of web pages searched on the internet. The similarity between the entities of a correspondence is calculated based on these documents. A similarity matrix is formed with all the entities between the two ontologies. For each column and each row of the matrix the most similar correspondences are selected for the set of candidate correspondences if the similarity value is greater than a threshold, after which the correspondences of data properties with different ranges are removed.

In order to determine the true correspondences, a new threshold is used, and the correspondences are those that have their similarity value above this threshold. The interactive part of the system is used to calculate this threshold, and some correspondences of the set of candidate correspondences are presented for this purpose.

The final alignment consists of all the correspondences above the threshold plus the correspondences that received positive feedback minus the correspondences that received negative feedback.

### 5.1.7. Hertuda

The Hertuda system participated in the OAEI evaluation between 2012 and 2014.

Hertuda [18] compares the ontology entity names using the Damerau-Levenstein metric. Only pairs of homogeneous entities, such as classes with classes, data properties with data properties, etc. are selected for the set of candidate correspondences. Correspondences that have a value greater than the threshold of the specific type are chosen.

After this, a filter is applied in these correspondences to separate those that are in the alignment of those that are not. This filter is made from a new threshold value found with the help of an expert.

Working in a similar way to WeSeE, the interactive part of the system is used to calculate this new threshold, and correspondences are presented to the expert for this purpose.

The final alignment consists of all correspondences above the threshold plus the correspondences that received positive feedback minus the correspondences that received negative feedback.

### 5.1.8. ServOMBI

ServOMBI [28] participated in the OAEI evaluation in 2015.

In this approach there is a lexical selection of correspondences, followed by a semantic selection, from these two are added new correspondences by structural selection. Then the stable marriage algorithm is used. After that, correspondences that generate logical inconsistencies with other correspondences of the set of candidate correspondences and have less similarity are removed.  All the remaining correspondences in the set of candidate correspondences receive feedback directly from the expert.

The final alignment is formed of all correspondence that received the expert's positive feedback.

### 5.1.9. MAPSOM

The MAPSOM [20] has never participated in the OAEI evaluation.

The tool consists of two main parts. The first part is aggregation of similarity metrics with the help of self-organizing map (a type of neural network). The second part incorporates user feedback for refining self-organizing map outcomes.

The process can be summarized as follows: first a clustering of the set of candidate correspondences is done using neural networks. After this the expert can explore the set of candidate correspondences by visual means and can change the classification of the correspondences. After that, the iterative phase begins, where it is presented, to the expert, correspondences of the set of candidate correspondences. The classification made by the expert are propagated to the other correspondences through the neural networks.

### 5.1.10. Approach proposed in Shi et al. [22]

The solution proposed in Shi et al. [22] has two steps.

In the first stage, the participation of the expert is used to choose a threshold, which will be used to choose, among the correspondences from the set of candidate correspondences, those that will be in the generated alignment. In this phase, the correspondences to be shown to the expert with the similarity metrics closest to a given initial threshold are selected. Depending on the expert's response the presumed threshold value will be increased or decreased until reaching stability.

A similarity propagation graph is constructed, which is derived from the structures of the ontologies being aligned, which shows how the fact of a certain correspondence, depending on the response of the expert, can influence the possibility of others being true or false. There is a measure called an error rate that indicates the probability that a correspondence has been automatically classified in the wrong way.

It is chosen, to be shown to the expert, the correspondence, probably misclassified, that can propagate more information using the similarity graph. After feedback from the expert, his information is propagated by modifying the similarity metrics through the similarity propagation graph. This is repeated until there are no further changes by propagation or the maximum number of interactions is reached.

The final alignment consists of all correspondences above the threshold plus the correspondences that received positive feedback minus the correspondences that received negative feedback.

### 5.1.11. Approach proposed in To et al. [35]

The system from To et al. [35] can be executed in two ways: supervised and semi-supervised. If there is a pre-alignment with many data, the supervised approach is chosen; otherwise, the semi-supervised approach is chosen. The learned model is used to predict whether the existing correspondences between the concepts of the two ontologies are true or false.

If the semi-supervised approach is chosen, some correspondences are chosen to be presented to the expert to receive feedback. Correspondences that received positive feedback are placed in the training set and the correspondences that did not received feedback are classified. The correspondences with the lowest confidence of classification are chosen among those classified to be shown to the expert. Again, the classified correspondences are placed in the training set and the correspondences that did not receive feedback are classified. The process is repeated for a number of iterations.

Those that received positive feedback from the expert and those that were classified as true will be part of the generated alignment.

### 5.1.12. Approach proposed in Wagner et al. [36]

The general idea of the approach of Wagner et al. [36] consists of the following properties:

- Incremental: in each interaction, the process considers only ontology partitions instead of entire ontologies.

- Interactive: The process takes into account expert feedback to improve precision and recall of alignment results.

- Iterative: The expert will work on iterations, where, in each iteration, a different matching method can be used for the generation and refinement of the correspondences of each partition. In addition, correspondences obtained from each iteration are stored and can be reused in subsequent iterations.

In the approach, there is the generation of ontology partitions, depending on the expert, through the choice of a central concept for the partitioning of the ontology and a maximum distance that the other concepts of the partition must be of the central concept, after there is the selection of an algorithm to generate an alignment between the entities of the two partitions of the ontologies. After, there is the feedback from the expert, where he can point false positives between the chosen correspondences, in addition to being able to add new correspondences. The approach continue until the expert decides to terminate the process.

### 5.1.13. Approach proposed in Cruz et al. [37]

In Cruz et al.'s [37] approach, similarity metrics are calculated for all concept pairs between two ontologies.

To determine the final alignment, a threshold is chosen, and only correspondences whose similarity metrics are above this threshold are considered true correspondences.

Only correspondences in which there is no concept repeated in relation to another true correspondence chosen is considered true.

Correspondences are presented to the expert whose similarity metrics are in greater disagreement, that is, there are a similar number of metrics that indicate that the correspondence is true in relation to those that indicate that it is false, and that have not been presented before.

Similar clustering is done in relation to similarity metrics, where two thresholds indicate a group of correspondences whose one of the metrics are between these two thresholds.

When a correspondence is chosen by the expert as true, the metrics of the other correspondences in the cluster are increased, and when a correspondence is indicated by the expert as false, the metrics of the other correspondences in the cluster are decreased. After classified by the expert the correspondences have all their metrics set to 1 (or 0) and are no longer modified.

### 5.1.14. Approach proposed in Duan et al. [21]

In Duan et al.'s [21] approach, lexical, aggregate and structural similarity metrics are used.

Initially, the lexical similarity metrics for each possible ontology correspondence are calculated. There are two types of lexical similarity, one based on the entity name and another on the entity documentation.

A selection of candidate correspondences is made, selecting the k most similar entities in the other ontology for each entity of the first ontology, based on the lexical similarities.

A set of true correspondences are inserted into the system by the expert. All correspondences that have at least one common entity with other correspondence considered true are labeled as false correspondence.

The system aggregates the similarity metrics of each correspondence into a single value and chooses a threshold. Aggregation is done by weighing the values of the lexical metrics. The correspondences indicated by the expert as true serve to indicate the weight of each lexical similarity metric. Only correspondences that have an aggregate value greater than the threshold are labeled true.

Structural metrics are created for each correspondence, which are metrics that depend on the aggregate similarity values, and on the lexical similarity values of neighboring correspondences.

The correspondences indicated by the expert gain added value of similarity equal to 1, those that are considered false gain value of added similarity equal to 0.

From that point on, iteration begins, with the modification of the structural metrics of each correspondence at each iteration, modification based on the lexical and aggregate metric values of neighboring correspondences, this allows inclusion of new correspondences in the set of candidate correspondences based also on its metrics of structural similarity. A stop criterion is used to finish the iterations.

In the end, the correspondences of the set of candidate correspondences with the aggregated metric above a given threshold are labeled true.

The final alignment consists of all correspondences above the threshold plus the correspondences that received positive feedback minus the correspondences that received negative feedback minus all correspondences that have at least one common entity with other correspondence considered true.

### 5.1.15. Approach proposed in Cruz, Loprete et al. [24]

The approach of Cruz, Loprete et al. [24] is based on a multiuser model, where several experts interact with the tool to validate the supposed true correspondences found by automatic means.

Initially, a set of k matchers is run. The results of the individual matchers are combined into a global similarity. An optimization algorithm is run to choose the final alignment in order to maximize the overall similarity and satisfy the alignment cardinality. After that the interaction with the user begins. The expert asks for a correspondence to validate. To the expert is shown the correspondence with the lowest quality according to a quality calculation algorithm. The same correspondence can receive feedback from multiple experts.

After, there is the feedback propagation. This method updates the global similarity by changing the similarity score for the validated correspondence and for the correspondences are close to the correspondence that was just validated, according to a distance measure.

### 5.1.16. Approach proposed in Li et al. [25]

The final alignment in the approach of Li et al. [25] is done through a threshold and the threshold is chosen through interactions with the expert. The approach determine the threshold by presenting correspondences to the expert and collecting the feedback. The algorithm for selecting thresholds is designed as follow: several thresholds are chosen for testing, each sampled threshold is applied to run on the dataset. From the results, we collect all correspondences that are found by at least one, but not by all thresholds. Those correspondences are put in a list ordered by the disagreement, and presented to the expert for validation. Based on the expert response, a score for each threshold is computed. For a true positive, all thresholds that have found the correspondence increase their score by 3. For a false positive, all thresholds that have not found the correspondence increase their score by 1. The threshold with highest score is returned.

This approach uses propagation formulas that can add new correspondences that were not, in principle, in the set of candidate correspondences.

At first, a set of candidate correspondences is chosen, associating a similarity metric to all the correspondences and selecting the set of candidate correspondences by the threshold.

To the set of candidate correspondences are applied the following rules:

Rule 1: All correspondences with a common entity with other correspondences are removed from the set of candidate correspondences.

Rule 2: A consistency constraint is used, it is similar to the generalization-disjunction anti-pattern.

Rule 3: Stability constraints are used, these constraints do not remove the correspondences from the set of candidate correspondences, but decrease their probability of being true, decreasing their measure of similarity.

Rule 4: Correlation propagation: it increases the probability of structurally associated correspondences with supposedly true correspondences, increasing its measure of similarity. This rule can add new correspondences that were not, in principle, in the set of candidate correspondences.

### 5.1.17. Approach proposed in Balasubramani et al. [23]

In Balasubramani et al. [23] approach, to choosing the set of candidate correspondence, five matching methods are performed and the resulting correspondences of these alignments are aggregated and the correspondences receive five metrics of similarity, of the five methods performed. Then a new method is executed that generates a new metric which is the weighted sum of the five previous metrics, with the weight being calculated according to a confidence calculation made by the method. The interaction with the expert is made to change the value of the weights to be given to the metrics associated with the correspondences, for that, a classifier is used.

Correspondences to be shown to the expert are selected using three added criteria, one being the similarity between the first five metrics, the least similar ones shown first. Multiple correspondences can be shown to the expert on each iteration.

A logistic regression classifier is used to modify the weights associated to each metric of the weighted value, this is done at each iteration until the values converge.

Then all correspondences with a common entity with another selected correspondence are removed from the final alignment.

### 5.2. Comparison of approaches with ALIN

Main characteristics of ALIN are:

The ALIN approach seeks to achieve an improvement in the quality of the alignment, but maintains the number of interactions with the expert in a level compatible with the other existing approaches. For this will be used the following strategies:

1 - All the correspondences of the set of candidate correspondences will be classified directly by the expert or else, indirectly, by logic or application of rules present in the involved ontologies. With this it is expected to achieve a high precision alignment.

2 - A set of classified correspondences is produced, as an initial alignment, with correspondences with high probability of being correct. This set tends to be small, generating an alignment with a high precision but a not so high recall.

3 – The approach will try to increase the recall of the final alignment by the interactive modification of the set of candidate correspondences by including new correspondence related to the correspondences that received positive feedback from the

expert. As there is the assumption that the expert does not make mistakes, it is expected that the new correspondences associated with those identified as true are also more likely to be true, increasing the recall without increasing too much the size of the set of candidate correspondences, so without increasing too much the number of interactions with the expert. All the new correspondences inserted will also be classified by the expert, or by the logic and rules of the ontologies.

The LogMap [26][32] approach resembles ALIN in the first feature, and it is able to classify all correspondences in the set of candidate correspondences with expert feedback or by logic and application of rules. But there is no interactive modification of the set of candidate correspondences with the insertion of new correspondences.

The approach described in Li et al [25] uses the concept of 'stability constraints' and uses these constraints to decrease the similarity metrics of correspondences that fit the constraints, thereby decreasing their likelihood of being in the final alignment. One such stability constraint is similar to criterion 4 of the additional criteria for automatic classification according to the maximum similarity premise.

In the approach presented by Duan et al. [21], expert feedback can influence the choice of new correspondences for the set of candidate correspondences, but in a different way from ALIN. In the approach presented by Duan et al., the new correspondences are not necessarily associated with those that received positive feedback from the expert, but are those that are in the vicinity of those that are in the set of candidate correspondences, some of them being marked by the expert as true. These new correspondences are not evaluated by the expert, different from the ALIN approach.

Several approaches remove correspondences, using the logic or characteristics of the ontologies or of the alignment to be generated, from the set of candidate correspondences, related to a correspondence that received positive feedback from the expert, but do not include the inclusion of new correspondences nor allow the classification of all the candidate correspondences,  directly or indirectly, by the expert. They are:

- AML [27][16];
- LogMap [26][32];
- Duan et al. [21];
- Li et al [25];
- Cruz et al. [37];
- Cruz, Loprete et al [24];
- Balasubramani et al [23].

Some approaches considered anti-patterns as resources during the ontology matching process. Guedes [14] constituted a repository of correspondence anti-patterns. In addition, Guedes [14] proposed a matching approach that submitted the generated alignment for verification, after the ontology matching process has already been fully executed. The correspondences of this alignment were evaluated to see if they instantiated any anti-pattern of the repository. If so, wrong correspondences was removed from the generated alignment and submitted again for verification. The result of the work indicated that the use of the anti-patterns improved the quality of the alignment. However, the defined anti-patterns are considered by Guedes [14] only at the end of the process, and not in an interactive way. The tool ASMOV [38][39], a non-interactive tool, used correspondence anti-patterns, which were called types of inconsistencies, to remove from the generated alignment all correspondences of the generated alignment that were in an anti-pattern and had a lower confidence value. The objective of the use of anti-patterns was reduce the inconsistency of the generated alignment.

Jarvis chooses the set of candidate correspondences in a similar way to ALIN, with the difference of the use, by ALIN, of additional criteria for the selection by maximum similarity and the withdrawal of correspondences with semantically different entity names. With this, it is expected that ALIN generates a set of candidate correspondences, up to this stage, with greater precision. In the interactive part of the process the two approaches work in a very different way, with Jarvis not inserting any new correspondence in the set of candidate correspondences, generating a training set, being classified by the expert part of the set of candidate correspondences, and the other part being automatically classified by the classifiers. As, in the ALIN, all the correspondences are classified by the expert, directly or indirectly, it is expected that the high precision of the first phase is maintained.

There is a selection of new correspondences, based on structural analysis of the ontologies (such as relationship correspondences between classes of class relationships already selected for the set of candidate correspondences), for inclusion in the set of candidate correspondences in the tools LogMap [26][32], AML [27][16], but in them the inclusion of new correspondences occurs in the generation of the set of candidate correspondences before the interaction with the expert, that is, the interaction with the expert does not influence the choice of new correspondences by structural matchers, different from the ALIN approach.

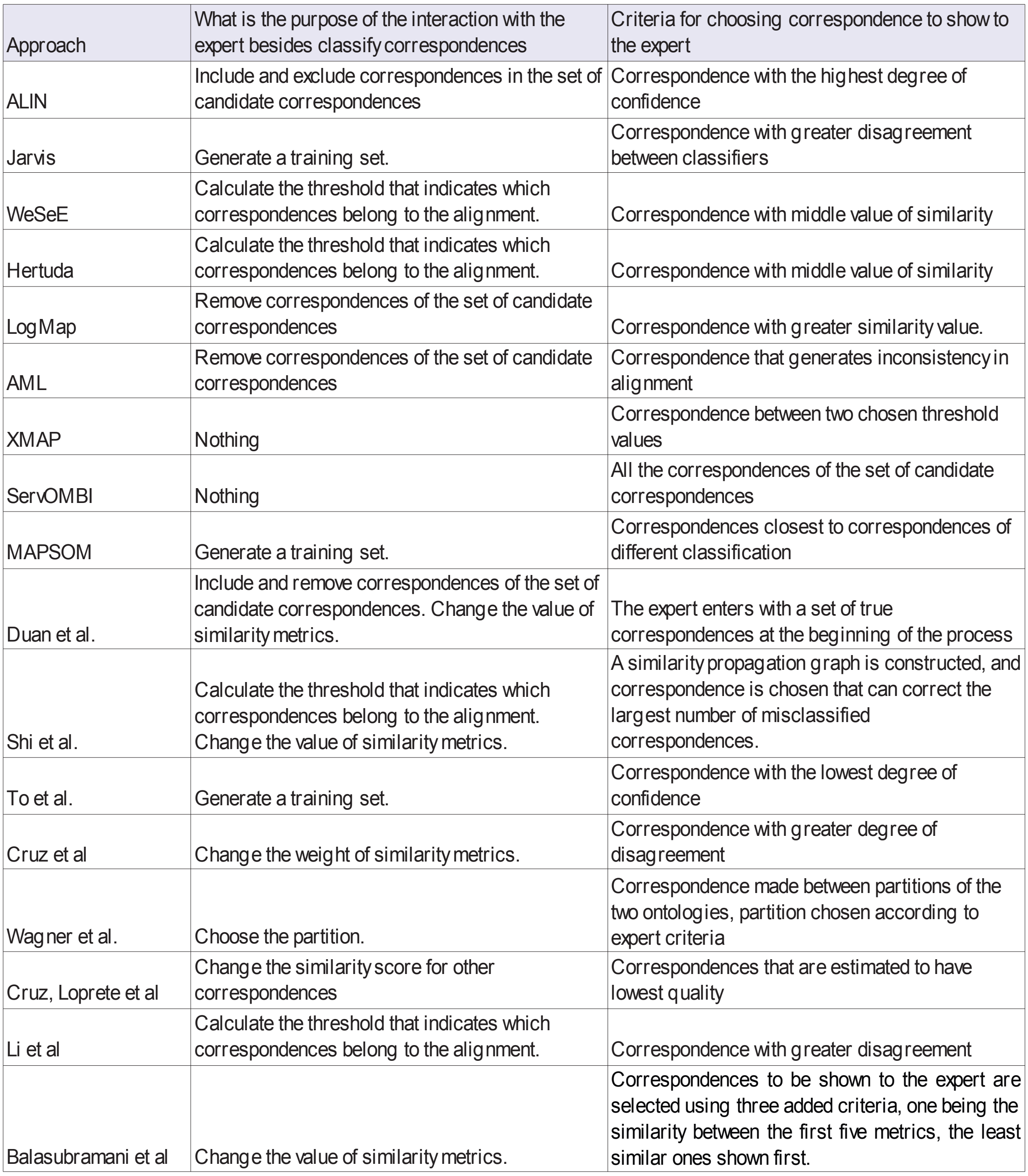
The approach described in Li et al [25] allows the entry of new correspondences, which were not in the original set of candidate correspondences, in the final alignment. It does this through an increase in the similarity metrics of the correspondences structurally associated with those that are considered true, but not all correspondences that increase similarity measures come from the expert's feedback, some may come from correspondences believed true because they are above the threshold.

In the Tables 32, 33 and 34 are shown the interactivity characteristics of the studied approaches.

*Table 32: Interactivity characteristics of studied approaches*

| Approach | Criteria for creating the set of candidate correspondences before the interactive phase | Techniques used to classify candidate correspondences not classified by the expert |
|---|---|---|
| ALIN | Generation of a set, with concepts of the ontologies, using the stable marriage algorithm. Removal, from the generated set, of the correspondences whose entities are not in the same synset and some of those that have maximum similarity across all their metrics | Logic and characteristics of ontologies and of the alignment to be generated |
| Jarvis | Generation of a set, with concepts of the ontologies, using the stable marriage algorithm. Removal, from the generated set, those that have maximum similarity across all their metrics | Use of classifiers |
| WeSeE | The highest metric entity pairs if the metric value is greater than a constant value, and then the data properties with different ranges are removed. | Use of a threshold |
| Hertuda | Only pairs of homogeneous entities are selected, such as classes with classes, data properties with data properties, and so on. Correspondences that have a value greater than the threshold of the specific type are chosen. | Use of a threshold |
| LogMap | Choice of all pairs of entities with equal names. Selection of new pairs semantically and structurally linked to those chosen at the beginning. Withdrawal of inconsistent correspondences. | Use of a threshold. Logic and characteristics of alignment to be generated. |
| AML | Use of lexical measures to form a initial set of candidate correspondences of classes with equal names.<br>Use of semantic measures to add correspondences to the initial set of candidate correspondences.<br>Use of string and word comparators to add new correspondences to the initial set of candidate correspondence.<br>Inclusion of correspondences structurally close to those already chosen.<br>Inclusion of properties related to the correspondences already chosen, using string and semantic metrics to choose these correspondences.<br>And an algorithm is executed to reduce the cardinality of the generated set, removing the correspondences with repeated entities and lower level of confidence. | Use of a threshold. Logic and characteristics of alignment to be generated. |
| XMAP | Use of a threshold | Use of a threshold |
| ServOMBI | There is a lexical selection of correspondences, followed by a semantic selection, from these two are added new correspondences by evaluation structure of the already chosen correspondences. Then the stable marriage algorithm is used.<br>Then, correspondences of less similarity are removed that generate logical inconsistencies with other correspondences of the set of candidate correspondences. | All the correspondences are classified by the expert |
| MAPSOM | Correspondences whose similarity metrics are above a given threshold are selected, one threshold for each metric. | Use of a classifier |
| Duan et al. | Selection by lexical metrics of a constant number of correspondences for each entity choosing those of greater similarity. | Use of a threshold. Characteristics of alignment to be generated. |
| Shi et al. | All possible correspondences | Use of a threshold |
| To et al. | All pairs of concepts are chosen. | Use of a classifier |
| Cruz et al | All pairs of concepts are chosen. | Use of a threshold. Characteristics of alignment to be generated. |
| Wagner et al. | The approach chooses a partition of each ontology and e is run an algorithm of choosing correspondences between the entities of the partitions | Correspondences not classified by the expert are considered true |
| Cruz, Loprete et al | Several matchers are used, its results aggregated and an optimization algorithm is run to select the final alignment so as to maximize the overall similarity and satisfy the mapping cardinality. | By an optimization algorithm that is run to select the final alignment so as to maximize the overall similarity. Characteristics of alignment to be generated. |
| Li et al | Use of a threshold. | Use of a threshold. Logic and characteristics of alignment to be generated. |
| Balasubramani et al | Five linguistic matchers are used and its results aggregated by another matcher | Characteristics of alignment to be generated. |

***Table 33: Interactivity characteristics of studied approaches***



| Approach | What is the purpose of the interaction with the expert besides classify correspondences | Criteria for choosing correspondence to show to the expert |
|---|---|---|
| ALIN | Include and exclude correspondences in the set of candidate correspondences | Correspondence with the highest degree of confidence |
| Jarvis | Generate a training set. | Correspondence with greater disagreement between classifiers |
| WeSeE | Calculate the threshold that indicates which correspondences belong to the alignment. | Correspondence with middle value of similarity |
| Hertuda | Calculate the threshold that indicates which correspondences belong to the alignment. | Correspondence with middle value of similarity |
| LogMap | Remove correspondences of the set of candidate correspondences | Correspondence with greater similarity value. |
| AML | Remove correspondences of the set of candidate correspondences | Correspondence that generates inconsistency in alignment |
| XMAP | Nothing | Correspondence between two chosen threshold values |
| ServOMBI | Nothing | All the correspondences of the set of candidate correspondences |
| MAPSOM | Generate a training set. | Correspondences closest to correspondences of different classification |
| Duan et al. | Include and remove correspondences of the set of candidate correspondences. Change the value of similarity metrics. | The expert enters with a set of true correspondences at the beginning of the process |
| Shi et al. | Calculate the threshold that indicates which correspondences belong to the alignment. Change the value of similarity metrics. | A similarity propagation graph is constructed, and correspondence is chosen that can correct the largest number of misclassified correspondences. |
| To et al. | Generate a training set. | Correspondence with the lowest degree of confidence |
| Cruz et al | Change the weight of similarity metrics. | Correspondence with greater degree of disagreement |
| Wagner et al. | Choose the partition. | Correspondence made between partitions of the two ontologies, partition chosen according to expert criteria |
| Cruz, Loprete et al | Change the similarity score for other correspondences | Correspondences that are estimated to have lowest quality |
| Li et al | Calculate the threshold that indicates which correspondences belong to the alignment. | Correspondence with greater disagreement |
| Balasubramani et al | Change the value of similarity metrics. | Correspondences to be shown to the expert are selected using three added criteria, one being the similarity between the first five metrics, the least similar ones shown first. |

***Table 34: Interactivity characteristics of studied approaches***

| Approach | How the expert's feedback influences other correspondences |
|---|---|
| ALIN | New correspondences are included in the set of candidate correspondences for expert evaluation. Correspondences are indicated as not belonging to the alignment if they are in an anti-pattern with the correspondence chosen as belonging to the alignment. |
| Jarvis | Correspondences that received feedback from the expert form a training set that serves to classify correspondences from the set of candidate correspondences that did not receive feedback |
| WeSeE | Correspondences that received expert feedback help you calculate a threshold that stipulates which of the other correspondences belong to the alignment or not |
| Hertuda | Correspondences that received expert feedback help you calculate a threshold that stipulates which of the other correspondences belong to the alignment or not |
| LogMap | Correspondences are indicated as not belonging to the alignment, if they have entities in common or when they conflict with the one indicated by the expert as belonging to the alignment. |
| AML | Correspondences are indicated as not belonging to the alignment, if they have entities in common or when they conflict with the one indicated by the expert as belonging to the alignment. |
| XMAP | It does not influence other correspondences. |
| ServOMBI | It does not influence other correspondences. |
| MAPSOM | Correspondences that received feedback from the expert form a training set that serves to classify correspondences from the set of candidate correspondences that did not receive feedback |
| Duan et al. | Association of correspondences as not belonging to the alignment if it has an entity equal to another of a correspondence indicated by the expert as belonging to the alignment. A correspondence chosen as belonging to alignment by the expert modifies the weight of similarity metrics. Correspondences referenced as belonging to the alignment by the expert gain aggregate similarity value equal to 1, those that were indicated as not belonging to the alignment gain 0. These values are propagated to the adjacent correspondences. New correspondences are added to the set of candidate correspondences depending on this spread. In the end, the correspondences with weighted aggregate of the metrics above a given predefined threshold are chosen as belonging to the alignment. |
| Shi et al. | The similarity of the correspondences is modified according to the similarity propagation graph. Depending on a threshold, the correspondences that belong to the alignment are chosen. The threshold is also chosen through interaction with the expert. |
| To et al. | Correspondences that received feedback from the expert generate a training set that serves to classify correspondences from the set of candidate correspondences that did not receive feedback |
| Cruz et al | Increase or decrease of correspondence similarity values according to expert feedback. A predefined threshold indicates what correspondences belongs to the alignment or not. |
| Wagner et al. | It does not influence other correspondences. |
| Cruz, Loprete et al | it updates the similarity score for the validated correspondence ( to 1 or 0 depending on the answer of the expert ) and for the correspondences whose signature vector is close to the signature vector of the correspondence that was just validated, according to a distance measure. |
| Li et al | Correspondences that received expert feedback help you calculate a threshold that stipulates which of the other correspondences belong to the alignment or not |
| Balasubramani et al | Increase or decrease of correspondence similarity values according to expert feedback. A correspondence with modified similarity value modifies your chance of being considered true. |

# 6. Conclusion

*This chapter presents the final considerations of the work, highlighting its main contributions, limitations identified in conducting the research as well as opportunities for future works.*

Progress in information and communication technologies has made a large number of data repositories available, but with a great deal of semantic heterogeneity, which makes it difficult to integrate. A process that has been used to solve this problem is the ontology matching, which tries to discover the existing correspondences between the entities of two distinct ontologies, which in turn structure the concepts that define the data stored in each repository. There are several approaches in the literature for ontology matching, among which the interactive strategy, which considers the participation of experts to improve the quality of the final alignment, stands out.

This work presented an interactive approach for ontology matching, based on interactive modification of the set of candidate correspondences. This approach seeks to include or exclude correspondences from the set of candidate correspondences at each interaction with the expert, in order to increase the alignment quality.

In order to evaluate if the generated alignment is of good quality, a comparison was made with other tools that have participated in the track of interactive ontology matching in OAEI 2016.

The results obtained show that ALIN generates an alignment with a superior quality to other tools evaluated by the OAEI in interactive alignment, considering the measures of precision, recall and f-measure, when the expert never make mistakes and the ontologies involved are not very large.

## 6.1. Main Contributions

The main contribution of this work is to show that the use of expert feedback as input of structural matchers and the application of correspondence anti-patterns as the only technique to classify the correspondences of the set of candidate correspondences not classified by the expert can generate a high quality alignment and with low level of inconsistencies. This is due to the fact that, in the other approaches these tasks are done in a way that is at least partially automatic, which increases the possibility of error or are implemented in such a way that only one of the tasks depends on the expert feedbacks. The innovation of ALIN was the use of expert feedback in both cases and the use of both in the same approach.

## 6.2. Limitations of the Proposal

As shown in the OAEI 2016 evaluation, ALIN performs well when the ontology matching has some characteristics, such as:

A) Proportionately large number of correspondences between attributes and between relationships in the correct alignment, because of the phases of retrieval of correspondences of attributes and of retrieval of correspondences of relationships.

B) Ontologies are small in size (with the Cartesian product of the two ontologies being less than 100,000). ALIN does not scale very well, taking a very long time to align large ontologies. After this size the algorithm is modified to not use the Wordnet, generating loss of quality.

C) The expert does not make classification mistakes. When an expert misclassify a correspondence, his error it is propagated in the set of candidate correspondences, leading to a rapid decrease in the quality of the alignment.

If the listed characteristics occur, the generated alignment is of high quality, as shown in the evaluation results.

## 6.3. Future works

Each technique used in ALIN has the potential to be improved, and therefore can generate research work. Examples include:

A) In the application of the stable marriage algorithm with incomplete lists of size limited to 1, one may wonder if the use of another limit will increase the recall, since each interaction can have up to 3 questions.

B) In the evaluation by semantic similarity one can investigate if the use of disambiguation algorithms can improve the set of candidate correspondences. Today is used the most frequent meaning of words, which is not necessarily correct.

C) Verify that the constant placed in the revision of the maximum similarity, the 6-character size for entity names is applicable to other ontologies in addition to the conference dataset. Check whether maximum depth of 3 levels for the retrieval of correspondences between relationships also serves for other ontologies besides the conference dataset. Verify that the constant 0.9, placed in Withdraw of Correspondences with Semantically Different Entity Names, is applicable to other ontologies in addition to the conference dataset.

D) Check that the use of compound nouns and nouns with post modifiers is better than other approaches to using wordnet.

E) One can search algorithms that scale more easily. ALIN in this regard, too, can be improved. The stable marriage algorithm has complexity O $(n^2)$, but allows the use of parallelism. This is a topic that could be studied. The use of Wordnet is slow, which could be paralleled as well.

F) By using correspondence anti-patterns, a very large number of correspondences that are known to be false are generated. There is still no way to use this knowledge to improve the set of candidate correspondences, only when a correspondence is indicated as true,  it is used for this purpose. One way to use this knowledge would probably increase the quality of the generated alignment.

G) One way to find out, between the answers given by the expert, which ones are false would cause the quality of the generated alignment to continue high even if an expert error rate is greater than zero.

H) The order in which the correspondences are shown to the expert influences the efficiency of the use of the anti-patterns, because if correct correspondences are shown in advance, more correspondences will be removed using anti-patterns. If some criterion is found to be better than the confidence value to show the correct correspondences before then we will have an ontology matching process with less interactions with the expert.

# 7. References

[1] J. Euzenat and P. Shvaiko, *Ontology Matching - Second Edition*, 2°. Springer-Verlag, 2013.

[2] P. Lambrix and R. Kaliyaperumal, “A Session-based Ontology Alignment Approach enabling User Involvement,” *Semant. Web*, vol. 1, pp. 1–28, 2016.

[3] H. Paulheim, S. Hertling, and D. Ritze, “Towards Evaluating Interactive Ontology Matching Tools,” *Lect. Notes Comput. Sci.*, vol. 7882, pp. 31–45, 2013.

[4] C. Meilicke and H. Stuckenschmidt, “A New Paradigm for Alignment Extraction,” *CEUR Workshop Proc.*, vol. 1545, pp. 1–12, 2015.

[5] J. Recker, “Scientific Research in Information Systems,” in *Scientific Research in Information Systems*, Intergovernmental Panel on Climate Change, Ed. Cambridge University Press, 2013.

[6] V. Lopes, F. Baião, and K. Revoredo, “Alinhamento Interativo de Ontologias Uma Abordagem Baseada em Query-by-Committee, Dissertação de Mestrado,” UNIRIO, 2015.

[7] M. Cheatham and P. Hitzler, “String similarity metrics for ontology alignment,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8219 LNCS, no. PART 2, pp. 294–309, 2013.

[8] E. G. M. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou, “Design and Evaluation of Semantic Similarity Measures for Concepts Stemming from the Same or Different Ontologies object instrumentality,” *Proc. 4th Work. Multimed. Semant.*, vol. 4, pp. 233–237, 2006.

[9] M. Cheatham and P. Hitzler, “The Role of String Similarity Metrics in Ontology Alignment,” *Semant. Web–ISWC 2013.*, pp. 1–54, 2013.

[10] F. Lin and K. Sandkuhl, “A survey of exploiting WordNet in ontology matching,” *IFIP Int. Fed. Inf. Process.*, vol. 276, pp. 341–350, 2008.

[11] D. . Gale and L. . S. . Shapley, “College Admissions and the Stability of Marriage,” *Am. Math. Mon.*, vol. 69, no. 1, pp. 9–15, 2014.

[12] R. W. Irving, D. F. Manlove, and G. O’Malley, “Stable marriage with ties and

bounded length preference lists," *J. Discret. Algorithms*, vol. 7, no. 2, pp. 213–219, 2009.

[13] C. Meilicke, "Alignment Incoherence in Ontology Matching - Ph.D. dissertation," University of Mannheim, Germany, 2011.

[14] A. Guedes, F. Baião, and K. Revoredo, "On the Identification and Representation of Ontology Correspondence Antipatterns," *Proc. 5th Int. Conf. Ontol. Semant. Web Patterns (WOP'14), CEUR Work. Proc.*, vol. 1302, pp. 38–48, 2014.

[15] A. Guedes, F. Baião, and K. Revoredo, "Digging Ontology Correspondence Antipatterns," *Proceeding WOP'14 Proc. 5th Int. Conf. Ontol. Semant. Web Patterns*, vol. 1302, pp. 38–48, 2014.

[16] D. Faria, C. Pesquita, E. Santos, M. Palmonari, I. F. Cruz, and F. M. Couto, "The AgreementMakerLight Ontology Matching System," in *OTM 2013: On the Move to Meaningful Internet Systems: OTM 2013 Conferences*, 2013, pp. 527–541.

[17] H. Paulheim and S. Hertling, "WeSeE-match results for OAEI 2013," *CEUR Workshop Proc.*, vol. 1111, pp. 197–202, 2013.

[18] S. Hertling, "Hertuda Results for OEAI 2012," *OM'12 Proc. 7th Int. Conf. Ontol. Matching*, vol. 946, pp. 141–144, 2012.

[19] W. E. Djeddi and M. T. Khadir, "XMap: Results for OAEI 2016," *CEUR Workshop Proc.*, vol. 1766, 2016.

[20] R. Jirkovsky, Václav and Ichise, "MAPSOM: User Involvement in Ontology Matching," *Semant. Technol.*, vol. 8388, pp. 348–363, 2014.

[21] S. Duan, A. Fokoue, and K. Srinivas, "One Size Does Not Fit All: Customizing Ontology Alignment Using User Feedback," in *Lecture Notes in Computer Science (LNCS)*, 2010, pp. 177–192.

[22] F. Shi, J. Li, J. Tang, G. Xie, and H. Li, "Actively learning ontology matching via user interaction," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5823 LNCS, no. 60703059, pp. 585–600, 2009.

[23] B. S. Balasubramani, A. Taheri, and I. F. Cruz, "User Involvement in Ontology Matching Using an Online Active Learning Approach," *CEUR Workshop Proc.*, vol. 1545, pp. 45–49, 2015.

[24] I. Cruz, F. Loprete, and M. Palmonari, "Pay-As-You-Go Multi-user Feedback Model for Ontology Matching," *Knowl. Eng. …*, pp. 80–96, 2014.

[25] C. Li, Z. Cui, P. Zhao, J. Wu, J. Xin, and T. He, "Improving ontology matching with propagation strategy and user feedback," *Seventh Int. Conf. Digit. Image Process.*, vol. 9631, p. 6, 2015.

[26] E. Jim and B. C. Grau, "LogMap : Logic-based and Scalable Ontology

Matching," *Lect. Notes Comput. Sci.*, vol. 7031, pp. 273–288, 2011.

[27] D. Faria, C. Martins, A. Nanavaty, D. Oliveira, B. S. Balasubramani, A. Taheri, C. Pesquita, F. M. Couto, and I. F. Cruz, "AML results for OAEI 2015," *CEUR Workshop Proc.*, vol. 1545, pp. 116–123, 2015.

[28] N. Kheder and G. Diallo, "ServOMBI at OAEI 2015," *CEUR Workshop Proc.*, vol. 1545, no. Ml, pp. 200–207, 2015.

[29] D. Faria, "Using the SEALS Client ' s Oracle in Interactive Matching," 2016. [Online]. Available: https://github.com/DanFaria/OAEI_SealsClient/blob/master/OracleTutorial.pdf.

[30] M. Achichi and M. Cheatham, "Results of the Ontology Alignment Evaluation Initiative 2016," *Proc. 11th Int. Work. Ontol. Matching co-located with 15th Int. Semant. Web Conf. (ISWC 2016) Kobe, Japan, Oct. 18, 2016.*, 2016.

[31] J. Silva, F. A. Baião, and K. Revoredo, "ALIN Results for OAEI 2016," *CEUR Workshop Proc.*, vol. 1766, 2016.

[32] E. Jiménez-Ruiz, B. C. Grau, Y. Zhou, and I. Horrocks, "Large-scale interactive ontology matching: Algorithms and implementation," *Front. Artif. Intell. Appl.*, vol. 242, no. ii, pp. 444–449, 2012.

[33] W. E. Djeddi and M. T. Khadir, "A Dynamic Multistrategy Ontology Alignment Framework Based on Semantic Relationships using WordNet," *Proc 3rd Int. Conf. Comput. Sci. its Appl. (CIIA 11)*, 2011.

[34] J. Da Silva, F. A. Baião, and K. Revoredo, "Alinhamento Interativo de Ontologias usando Anti- Padrões de Alinhamento: Um Primeiro Experimento Alternative Title: Interactive Ontology Alignment using Alignment Antipatterns: A First Experiment," *Proc. XII Brazilian Symp. Inf. Syst.*, pp. 208–215, 2016.

[35] I. R. To H. and H. Le, "An Adaptive Machine Learning Framework with User Interaction for Ontology Matching," *Twenty-first Int. Jt. Conf. Artif. Intell.*, 2009.

[36] F. Wagner, J. A. F. Macedo, and B. Lóscio, "An incremental and user feedback-based ontology matching approach," *13th Int. Conf. Inf. Integr. Web-based Appl. Serv. - iiWAS '11*, p. 4, 2011.

[37] I. F. Cruz, C. Stroe, and M. Palmonari, "Interactive user feedback in ontology matching using signature vectors," *Proc. - Int. Conf. Data Eng.*, pp. 1321–1324, 2012.

[38] Y. R. Jean-mary, E. P. Shironoshita, and M. R. Kabuka, "ASMOV : Results for OAEI 2010," *CEUR Workshop Proc.*, 2010.

[39] Y. R. Jean-mary, E. P. Shironoshita, and M. R. Kabuka, "Ontology matching with semantic verification," *"Web Semant. Sci. Serv. Agents World Wide Web,"* 2009.