
# D2.2.9: Description of alignment evaluation and benchmarking results

**Coordinator: Pavel Shvaiko (University of Trento)**
***with contributions from***
**Jérôme Euzenat (INRIA Rhône-Alpes & LIG), Heiner Stuckenschmidt (University of Mannheim), Malgorzata Mochol (Free University of Berlin), Fausto Giunchiglia, Mikalai Yatskevich, Paolo Avesani (University of Trento), Willem Robert van Hage (Vrije Universiteit Amsterdam), Ondřej Šváb, Vojtěch Svátek (University of Economics, Prague)**

**Abstract.**
This document discusses its latest advances in ontology matching evaluation: $(i)$ by providing new measures for ontology matching evaluation, such as semantic precision and recall and $(ii)$ by presenting the results of the Ontology Alignment Evaluation Initiative: OAEI-2006 campaign.

Keyword list: ontology matching, ontology alignment, ontology mapping, evaluation, benchmarking, contest, performance measures.



| | |
|---|---|
| Document Identifier | KWEB/2004/D2.2.9/v1.2 |
| Project | KWEB EU-IST-2004-507482 |
| Version | v1.2 |
| Date | August 8, 2007 |
| State | Final |
| Distribution | Public |

## Knowledge Web Consortium

This document is part of a research project funded by the IST Programme of the Commission of the European Communities as project number IST-2004-507482.

**University of Innsbruck (UIBK) - Coordinator**
Institute of Computer Science
Technikerstrasse 13
A-6020 Innsbruck
Austria
Contact person: Dieter Fensel
E-mail address: dieter.fensel@uibk.ac.at

**École Polytechnique Fédérale de Lausanne (EPFL)**
Computer Science Department
Swiss Federal Institute of Technology
IN (Ecublens), CH-1015 Lausanne
Switzerland
Contact person: Boi Faltings
E-mail address: boi.faltings@epfl.ch

**France Telecom (FT)**
4 Rue du Clos Courtel
35512 Cesson Sévigné
France. PO Box 91226
Contact person : Alain Leger
E-mail address: alain.leger@rd.francetelecom.com

**Freie Universität Berlin (FU Berlin)**
Takustrasse 9
14195 Berlin
Germany
Contact person: Robert Tolksdorf
E-mail address: tolk@inf.fu-berlin.de

**Free University of Bozen-Bolzano (FUB)**
Piazza Domenicani 3
39100 Bolzano
Italy
Contact person: Enrico Franconi
E-mail address: franconi@inf.unibz.it

**Institut National de Recherche en Informatique et en Automatique (INRIA)**
ZIRST - 655 avenue de l'Europe -
Montbonnot Saint Martin
38334 Saint-Ismier
France
Contact person: Jérôme Euzenat
E-mail address: Jerome.Euzenat@inrialpes.fr

**Centre for Research and Technology Hellas / Informatics and Telematics Institute (ITI-CERTH)**
1st km Thermi - Panorama road
57001 Thermi-Thessaloniki
Greece. Po Box 361
Contact person: Michael G. Strintzis
E-mail address: strintzi@iti.gr

**Learning Lab Lower Saxony (L3S)**
Expo Plaza 1
30539 Hannover
Germany
Contact person: Wolfgang Nejdl
E-mail address: nejdl@learninglab.de

**National University of Ireland Galway (NUIG)**
National University of Ireland
Science and Technology Building
University Road
Galway
Ireland
Contact person: Christoph Bussler
E-mail address: chris.bussler@deri.ie

**The Open University (OU)**
Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA
United Kingdom
Contact person: Enrico Motta
E-mail address: e.motta@open.ac.uk

**Universidad Politécnica de Madrid (UPM)**
Campus de Montegancedo sn
28660 Boadilla del Monte
Spain
Contact person: Asunción Gómez Pérez
E-mail address: asun@fi.upm.es

**University of Karlsruhe (UKARL)**
Institut für Angewandte Informatik und Formale
Beschreibungsverfahren - AIFB
Universität Karlsruhe
D-76128 Karlsruhe
Germany
Contact person: Rudi Studer
E-mail address: studer@aifb.uni-karlsruhe.de

**University of Liverpool (UniLiv)**
Chadwick Building, Peach Street
L697ZF Liverpool
United Kingdom
Contact person: Michael Wooldridge
E-mail address: M.J.Wooldridge@csc.liv.ac.uk

**University of Manchester (UoM)**
Room 2.32. Kilburn Building, Department of Computer
Science, University of Manchester, Oxford Road
Manchester, M13 9PL
United Kingdom
Contact person: Carole Goble
E-mail address: carole@cs.man.ac.uk

**University of Sheffield (USFD)**
Regent Court, 211 Portobello street
S14DP Sheffield
United Kingdom
Contact person: Hamish Cunningham
E-mail address: hamish@dcs.shef.ac.uk

**University of Trento (UniTn)**
Via Sommarive 14
38050 Trento
Italy
Contact person: Fausto Giunchiglia
E-mail address: fausto@dit.unitn.it

**Vrije Universiteit Amsterdam (VUA)**
De Boelelaan 1081a
1081HV. Amsterdam
The Netherlands
Contact person: Frank van Harmelen
E-mail address: Frank.van.Harmelen@cs.vu.nl

**Vrije Universiteit Brussel (VUB)**
Pleinlaan 2, Building G10
1050 Brussels
Belgium
Contact person: Robert Meersman
E-mail address: robert.meersman@vub.ac.be

# Work package participants

The following partners have taken an active part in the work leading to the elaboration of this document, even if they might not have directly contributed to writing parts of this document:

Centre for Research and Technology Hellas
École Polytechnique Fédérale de Lausanne
Free University of Bozen-Bolzano
Institut National de Recherche en Informatique et en Automatique
National University of Ireland Galway
Universidad Politécnica de Madrid
University of Innsbruck
University of Karlsruhe
University of Manchester
University of Sheffield
University of Trento
Vrije Universiteit Amsterdam
Vrije Universiteit Brussel

# Changes

| Version | Date | Author | Changes |
|---|---|---|---|
| 0.1 | 13.02.2006 | Jérôme Euzenat | Creation |
| 0.2 | 11.05.2007 | Pavel Shvaiko | Structure of the deliverable |
| 0.3 | 14.05.2007 | Pavel Shvaiko | Structure of the deliverable refined and a first input from various partners consolidated |
| 0.4 | 17.05.2007 | Pavel Shvaiko | Introduction, abstract, and executive summary added |
| 0.5 | 22.05.2007 | Pavel Shvaiko | The directory test case expanded |
| 0.6 | 25.05.2007 | Pavel Shvaiko | Incorporating the feedback based on the Kweb GA WP2.2 meeting in Innsbruck |
| 0.7 | 29.05.2007 | Pavel Shvaiko | The conference track contents updated |
| 0.8 | 04.06.2007 | Pavel Shvaiko | The anatomy test case updated |
| 0.9 | 06.06.2007 | Pavel Shvaiko | The food test case updated |
| 1.0 | 15.06.2007 | Pavel Shvaiko | Polishing pass over the deliverable done |
| 1.1 | 05.07.2007 | Jérôme Euzenat | Edited for typos |
| 1.2 | 06.08.2007 | Pavel Shvaiko | Comments of the reviewer addressed |

# Executive summary

This document is concerned with the latest advances in ontology matching evaluation. It describes new measures for ontology matching evaluation, new methodology for evaluating matching systems and results of the last OAEI campaign.

On measures, it presents semantic precision and recall which draw on previous syntactic generalizations of precision and recall, by introducing semantically justified measures that satisfy maximal precision and maximal recall for correct and complete alignments. These new measures are compatible with classical precision and recall.

It also presents the Ontology Alignment Evaluation Initiative 2006 campaign as well as its results. The OAEI campaign aims at comparing ontology matching systems on precisely defined test cases. OAEI-2006 is built over previous campaigns by having four tracks followed by ten participants. These tracks include: $(i)$ benchmarks, $(ii)$ expressive ontologies, $(iii)$ directories and thesauri, and $(iv)$ conference and consensus building workshop. In turn, the systems that participated are: Falcon, HMatch, DSSim, COMA++, AUTOMS, JHU/APL, PRIOR, RiMOM, OCM, and NIH. In OAEI-2006 we had more participants than in previous campaigns, namely: 4 in 2004, 7 in 2005 and 10 in 2006.

OAEI-2006 showed clear improvements over the results of the previous campaigns. Specifically, for example, on the benchmarks of OAEI-2005, Falcon dominated the other systems, while on the same test series of OAEI-2006, COMA++ and RiMOM (together with Falcon) managed to arrive at the same level of quality, e.g., RiMOM showed 96% precision and 88% recall. For what concerns the expressive ontologies track, in OAEI-2005 none of the participants was in a position to submit a result for the anatomy dataset. Almost all participants reported major difficulties in processing the ontologies due to their size. In OAEI-2006, five out of ten participants submitted results for the anatomy data set. This clearly shows the advance of matching systems on the technical level and also suggests that matching technology is potentially ready for large scale applications. In turn, on the directories test case of OAEI-2006, the participants in general demonstrated better results than in OAEI-2005. The thesauri (food) test case as well as the conference track and consensus building workshop (of reference alignments) were the new tracks with respect to OAEI-2005. These introduced particular modalities of evaluation and contributed to better insights on strengths and weaknesses of the matching systems. Finally, we also noted the increase in tools compliance and robustness: there were less problems to carry the tests on the side of participants as well as there were less problems to evaluate the results on the side of organizers.

This evaluation effort has proved very successful and will be pursued beyond knowledge web. The OAEI is an independent initiative which will continue to operate thanks to its dedicated members and the low expenses that are required for running these evaluation campaigns.

The material of this deliverable has been published in various publications: [Euzenat, 2007, Shvaiko *et al.*, 2006, Šváb *et al.*, 2007, Zhang and Bodenreider, 2007b].

# Contents

# Chapter 1

# Introduction

Ontology matching is an important problem for which many matching algorithms have been provided [Euzenat and Shvaiko, 2007]. Given a pair of ontologies, these algorithms compute a set of correspondences (called an alignment) between entities of these ontologies. This deliverable is focused on one of the hardly explored topics of ontology matching: ontology matching evaluation. In particular, it presents its several latest advances:

**Semantic evaluation measures.** In order to evaluate matching algorithms, they are confronted with ontologies to be matched and the resulting alignment ($A$) is compared with a reference alignment ($R$) based on some criterion. The usual approach for evaluating the returned alignments is to consider them as sets of correspondences and to compute precision and recall measures originating from information retrieval [van Rijsbergen, 1975] and adapted to the matching task. Precision and recall are the ratio of the number of true positives ($|R \cap A|$) on that of the retrieved correspondences ($|A|$) and those expected ($|R|$), respectively. These criteria are well understood and widely accepted.

However, when the objects to compare are semantically defined, like ontologies and alignments, it can happen that a fully correct alignment has low precision. This is due to the restricted set-theoretic foundation of these measures. Drawing on previous syntactic generalizations of precision and recall, semantically justified measures that satisfy maximal precision and maximal recall for correct and complete alignments are proposed. These new measures are compatible with classical precision and recall and can be computed.

**OAEI-2006 campaign.** The Ontology Alignment Evaluation Initiative[1] (OAEI) is a coordinated international initiative that organizes the evaluation of the increasing number of ontology matching systems. The main goal of the Ontology Alignment Evaluation Initiative is to be able to compare systems and algorithms on the same basis and to allow anyone for drawing conclusions about the best matching strategies. Our ambition is that from such evaluations, tool developers can learn and improve their systems. The OAEI campaigns are the evaluation of matching systems on consensus test cases.

Two first events have been organized in 2004: $(i)$ the Information Interpretation and Integration Conference (I3CON) held at the NIST Performance Metrics for Intelligent Systems (PerMIS) workshop and $(ii)$ the Ontology Alignment Contest held at the Evaluation of

[1] `http://oaei.ontologymatching.org`

Ontology-based Tools (EON) workshop of the annual International Semantic Web Conference (ISWC) [Sure *et al.*, 2004]. The first unique OAEI evaluation campaign has been presented at the workshop on Integrating Ontologies held in conjunction with the International Conference on Knowledge Capture (K-Cap) 2005 [Ashpole *et al.*, 2005]. The campaign of 2006 was presented at the Ontology Matching (OM) workshop at ISWC, in Athens, Georgia, USA [Shvaiko *et al.*, 2006]. In reaction over last year's remarks, within the OAEI-2006 campaign we had a variety of test cases that emphasized different aspects of the matching needs. From three test cases of 2005, in 2006 we had six very different test cases. Some of these tests introduced particular modalities of evaluation, such as a consensus building workshop (of reference alignments) and application-oriented evaluation.

The rest of the deliverable is organized as follows. Chapter 2 presents new evaluation measures: semantic precision and recall. Chapter 3 provides introduction to the OAEI-2006 campaign, while its constituent tracks are discussed in sequel. Specifically, Chapters 4, 5, 6, 7 discuss the benchmark, the expressive ontologies, the directories and thesauri, the conference tracks, respectively. Finally, Chapter 8 summarizes the lessons learned and outlines future plans for the ontology matching evaluation activities.

# Chapter 2

# A semantic measure for precision and recall

As reported in [Ehrig and Euzenat, 2005], the measures of precision and recall have the drawback to be of the all-or-nothing kind. An alignment may be very close to the expected result and another quite remote from it and both sharing the same precision and recall values. The reason for this is that the criteria only compare two sets of correspondences without considering if these correspondences are close or remote to each other: if they are not the same exact correspondences, they score zero. They both score identically low, despite their different quality.

Moreover, there can be semantically equivalent alignments which are not identical. A fair measure of alignment quality should rank these alignments with the same values (or at least, closer values than non equivalent alignments). It is thus necessary to design semantically grounded alignment measures instead of measures based on their syntactic equality.

In this chapter we investigate some measures that generalize precision and recall in order to overcome the problems mentioned above. We first provide the basic definitions of alignments, semantics, precision and recall as well as a motivating example (§2.1). We then present the framework for generalizing precision and recall (§2.2). From that point we investigate and propose new semantically justified evaluation versions of precision and recall. We discuss their properties and define complementary measures (§2.3). We show on the motivating example that the proposed measures improve on previous ones (§2.4). Finally, we overview the related work and summarize the major findings of this chapter (§2.5).

The work presented in this chapter has been published in [Euzenat, 2007].

## 2.1 Foundations

Let us first precisely define what the alignments are through their syntax (§2.1.1) and semantics (§2.1.2). Then we introduce precision and recall adapted to alignments (§2.1.3).

We will consider ontologies as logics. The languages used in the semantics web such as RDF or OWL are indeed logics. The semantics of the ontologies are given through their sets of models.

### 2.1.1 Alignments

The result of matching, called an alignment, is a set of pairs of entities $\langle e, e' \rangle$ from two ontologies $o$ and $o'$ that are supposed to satisfy a certain relation $r$ with a certain confidence $n$.

**Definition 1** (Alignment, correspondence)**.** *Given two ontologies $o$ and $o'$, an alignment between $o$ and $o'$ is a set of correspondences (i.e., 4-uples): $\langle e, e', r, n \rangle$ with $e \in o$ and $e' \in o'$ being the two matched entities, $r$ being a relationship holding between $e$ and $e'$, and $n$ expressing the level of confidence in this correspondence.*

For the sake of simplicity, we will here only consider correspondences as triples $\langle e, e', r \rangle$. The best way to compare results with confidence is to plot their precision/recall functions. The examples are only provided for simple ontologies, which are class hierarchies but do not depend on this simple language.

Figure 2.1 presents two ontologies together with five alignments $R$, $A_1$, $A_2$, $A_3$, and $A_4$.

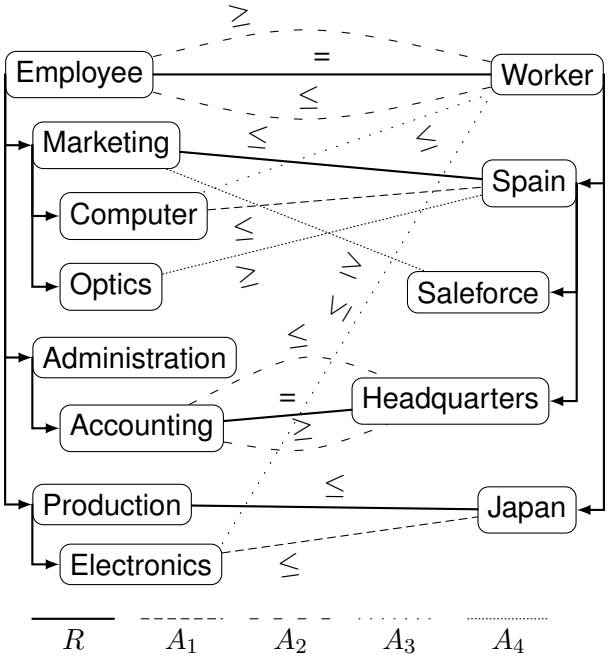



Figure 2.1: Five class alignments between two ontologies (only the classes involved in correspondences are displayed).

$R$ is the reference alignment and can be expressed by the following equations:

$$R = \begin{cases} \text{Employee} = \text{Worker} & \text{Accounting} = \text{Headquarters} \\ \text{Production} \leq \text{Japan} & \text{Marketing} \leq \text{Spain} \end{cases}$$

The other alignments share the first two correspondences with the reference alignment and have additional ones:

$$A_1 = \left\{ \begin{array}{ll} \mathsf{Employee} = \mathsf{Worker} & \mathsf{Accounting} = \mathsf{Headquarters} \\ \mathsf{Electronics} \leq \mathsf{Japan} & \mathsf{Computer} \leq \mathsf{Spain} \end{array} \right.$$

$$A_3 = \left\{ \begin{array}{ll} \mathsf{Employee} = \mathsf{Worker} & \mathsf{Accounting} = \mathsf{Headquarters} \\ \mathsf{Electronics} \leq \mathsf{Worker} & \mathsf{Computer} \leq \mathsf{Worker} \end{array} \right.$$

$$A_4 = \left\{ \begin{array}{ll} \mathsf{Employee} = \mathsf{Worker} & \mathsf{Accounting} = \mathsf{Headquarters} \\ \mathsf{Optics} \geq \mathsf{Spain} & \mathsf{Marketing} \geq \mathsf{Saleforce} \end{array} \right.$$

Alignment $A_2$ contains more correspondences than the others; it is made up of the following correspondences:

$$A_2 = \left\{ \begin{array}{ll} \mathsf{Employee} \leq \mathsf{Worker} & \mathsf{Accounting} \leq \mathsf{Headquarters} \\ \mathsf{Employee} \geq \mathsf{Worker} & \mathsf{Accounting} \geq \mathsf{Headquarters} \\ \mathsf{Production} \leq \mathsf{Japan} & \mathsf{Marketing} \leq \mathsf{Spain} \end{array} \right.$$

### 2.1.2 Semantics of alignments

In line of the work on data integration [Ghidini and Serafini, 1998], we provide a first-order model theoretic semantics. It depends on the semantics of ontologies but does not interfere with it. In fact, given a set of ontologies and a set of alignments between them, we can evaluate the semantics of the whole system in function of the semantics of each individual ontology. The semantics of an ontology is given by its set of models.

**Definition 2** (Models of an ontology)**.** *Given an ontology $o$, a model of $o$ is a function $m$ from elements of $o$ to elements of a domain of interpretation $\Delta$. The set of models of an ontology is denoted as $\mathcal{M}(o)$.*

Because the models of various ontologies can have different interpretation domains, we use the notion of an equalising function, which helps making these domains commensurate.

**Definition 3** (Equilising function)**.** *Given a family of interpretations $\langle I_o, \Delta_o \rangle_{o \in \Omega}$ of a set of ontologies $\Omega$, an equalising function for $\langle I_o, \Delta_o \rangle_{o \in \Omega}$ is a family of functions $\gamma = (\gamma_o : \Delta_o \longrightarrow U)_{o \in \Omega}$ from the domains of interpretation to a global domain of interpretation $U$. The set of all equalising functions is called $\Gamma$.*

When it is unambiguous, we will use $\gamma$ as a function. The goal of this $\gamma$ function is only to be able to (theoretically) compare elements of the domain of interpretation. It is simpler than the use of domain relations in distributed first order logics [Ghidini and Serafini, 1998] in the sense that there is one function per domain instead of relations for each pair of domains.

The relations used in correspondences do not necessarily belong to the ontology languages. As such, they do not have to be interpreted by the ontology semantics. Therefore, we have to provide semantics for them.

**Definition 4** (Interpretation of alignment relations)**.** *Given $r$ an alignment relation and $U$ a global domain of interpretation, $r$ is interpreted as a binary relation over $U$, i.e., $r^U \subseteq U \times U$.*

The definition of correspondence satisfiability relies on $\gamma$ and the interpretation of relations. It requires that in the equalised models, the correspondences are satisfied.

**Definition 5** (Satisfied correspondence)**.** *A correspondence $c = \langle e, e', r\rangle$ is satisfied for an equalising function $\gamma$ by two models $m$, $m'$ of $o$, $o'$ if and only if $\gamma_o \cdot m \in \mathcal{M}(o)$, $\gamma_{o'} \cdot m' \in \mathcal{M}(o')$ and*

$$\langle \gamma_o(m(e)), \gamma_{o'}(m'(e'))\rangle \in r^U$$

*This is denoted as $m, m' \models_\gamma c$.*

For instance, in the language used as example, if $m$ and $m'$ are respective models of $o$ and $o'$:

$$\begin{aligned}
&m, m' \models_\gamma \langle c, c', =\rangle \text{ if and only if } \gamma_o(m(c)) = \gamma_{o'}(m'(c'))\\
&m, m' \models_\gamma \langle c, c', \leq\rangle \text{ if and only if } \gamma_o(m(c)) \subseteq \gamma_{o'}(m'(c'))\\
&m, m' \models_\gamma \langle c, c', \geq\rangle \text{ if and only if } \gamma_o(m(c)) \supseteq \gamma_{o'}(m'(c'))\\
&m, m' \models_\gamma \langle c, c', \perp\rangle \text{ if and only if } \gamma_o(m(c)) \cap \gamma_{o'}(m'(c')) = \emptyset
\end{aligned}$$

**Definition 6** (Satisfiable alignment)**.** *An alignment $A$ of two ontologies $o$ and $o'$ is said satisfiable if and only if*

$$\exists m \in \mathcal{M}(o), \exists m' \in \mathcal{M}(o'), \exists \gamma \in \Gamma; \forall c \in A, m, m' \models_\gamma c$$

Thus, an alignment is satisfiable if there are models of the ontologies that can be combined in such a way that this alignment makes sense.

**Definition 7** (Models of aligned ontologies)**.** *Given two ontologies $o$ and $o'$ and an alignment $A$ between these ontologies, a model of these aligned ontologies is a triple $\langle m, m', \gamma\rangle \in \mathcal{M}(o) \times \mathcal{M}(o') \times \Gamma$, such that $A$ is satisfied by $\langle m, m', \gamma\rangle$.*

In that respect, the alignment acts as a model filter for the ontologies. It selects the interpretation of ontologies which are coherent with the alignments. Note, this allows to transfer information from one ontology to another since reducing the set of models will entail more consequences in each aligned ontology.

In this chapter we consider those consequences of aligned ontologies that are correspondences.

**Definition 8** ($\alpha$-Consequence of aligned ontologies)**.** *Given two ontologies $o$ and $o'$ and an alignment $A$ between these ontologies, a correspondence $\delta$ is a $\alpha$-consequence of $o$, $o'$ and $A$ (denoted as $A \models \delta$) if and only if for all models $\langle m, m', \gamma\rangle$ of $o$, $o'$ and $A$, $m, m' \models_\gamma \delta$ (the set of $\alpha$-consequences is denoted as $Cn(A)$).*

For $\alpha$-consequences, $A_2$ is strictly equivalent to $R$ (i.e., $A_2 \models R$ and $R \models A_2$). In fact, the aligned ontologies with $A_2$ and $R$ have exactly the same models.

It is noteworthy that, given an alignment, the $\alpha$-consequences of this alignment can be larger than it. If the alignment is not satisfiable, then any correspondence is a $\alpha$-consequence of it.

Such a formalism helps defining the meaning of alignments: it tells what are the consequences of ontologies with alignments. It is particularly useful for deciding if delivered alignments are consistent and for specifying what is expected from matching algorithms and how they should be designed or evaluated. It can be naturally extended to distributed systems in the sense of [Ghidini and Serafini, 1998], i.e., sets of ontologies and alignments.

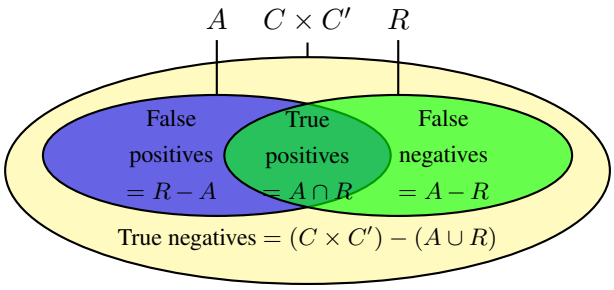



Figure 2.2: Two alignments as sets of correspondences and their relations.

### 2.1.3 Precision and recall

Precision and recall are commonplace measures in information retrieval. They are based on the comparison of an expected result and the effective result of the evaluated system. These results are considered as a set of items, e.g., the documents to be retrieved.

Since these measures are commonly used and well understood, they have been adapted for ontology matching evaluation [Do *et al.*, 2002]. In this case, sets of documents are replaced by sets of correspondences, i.e., alignments. The alignment ($A$) returned by the system to evaluate is compared to a reference alignment ($R$).

Like in information retrieval, precision measures the ratio of correctly found correspondences (true positives) over the total number of returned correspondences (true positives and false positives). In logical terms this is supposed to measure the correctness of the method. Recall measures the ratio of correctly found correspondences (true positives) over the total number of expected correspondences (true positives and false negatives). In logical terms, this is a completeness measure. This is displayed in Figure 2.2.

**Definition 9** (Precision and recall)**.** *Given a reference alignment $R$, the precision of some alignment $A$ is defined as follows:*

$$P(A, R) = \frac{|R \cap A|}{|A|}.$$

*Recall, in turn, is defined as follows:*

$$R(A, R) = \frac{|R \cap A|}{|R|}.$$

Notice that in the definition above, the letter $R$ stands for both the recall function and the reference alignment. Since one is a function and the other is a set, these are easy to be distinguished by their use even if referred by the same letter.

Other measures, such as fallout, F-measure, noise and silence can be derived from precision and recall [Do *et al.*, 2002].

For the examples of Figure 2.1, the two first columns of Table 2.1 (see p.14) display precision and recall. As expected, evaluating the reference against itself yields a maximal precision and recall. All the other alignments share the two first correspondences with $R$ and miss the two other ones so they have a precision and recall of 50% (but $A_2$ which is handicapped by having more than 4 correspondences).

The semantics of alignments has not been taken into account since an alignment, like $A_2$ equivalent to $R$ scores worse than an incorrect and incomplete alignment, like $A_4$.

## 2.2 Generalizing precision and recall

Even being well understood and widely accepted measures, precision and recall have the drawback that whatever correspondence has not been found is definitely not considered. As a result, they do not discriminate between a bad and a better alignment.

Indeed, when considering the example of Figure 2.1, alignment $A_4$ is arguably worse than the others, because its additional correspondences are measurably more different from the reference ones, but it scores the same as the other alignments.

As precision and recall are easily explained measures, it is useful to maintain the precision and recall structure when looking for new measures. This also ensures that measures derived from precision and recall (e.g., F-measure) still can be computed easily. [Ehrig and Euzenat, 2005] proposed an abstract generalization of precision and recall that has been instantiated with syntactic measures. This allows to take into account "near misses", i.e., incorrect correspondences that are close to the target (and that, for instance, can be more easily repaired).

Instead of comparing alignments set-theoretically, [Ehrig and Euzenat, 2005] proposes to measure the proximity of correspondence sets rather than the strict size of their overlap. Instead of taking the cardinality of the intersection of the two sets ($|R \cap A|$), the natural generalizations of precision and recall measure their proximity ($\omega(A, R)$).

**Definition 10** (Generalized precision and recall)**.** *Given a reference alignment $R$ and an overlap function $\omega$ between alignments, the precision of an alignment $A$ is given by*

$$P_\omega(A, R) = \frac{\omega(A, R)}{|A|}$$

*and recall is given by*

$$R_\omega(A, R) = \frac{\omega(A, R)}{|R|}.$$

In order, for these new measures to be true generalizations, $\omega$ has to share some properties with $|R \cap A|$. In particular, the measure should be positive:

$$\forall A, B, \omega(A, B) \geq 0 \qquad \text{(positiveness)}$$

and should not exceed the minimal size of both sets:

$$\forall A, B, \omega(A, B) \leq min(|A|, |B|) \qquad \text{(maximality)}$$

This guarantees that the given values are within the unit interval $[0\ 1]$. Further, this measure should only add more flexibility to the usual precision and recall so their values cannot be worse than the initial evaluation:

$$\forall A, B, \omega(A, B) \geq |A \cap B| \qquad \text{(boundedness)}$$

Hence, the main constraint faced by the proximity is:

$$|A \cap R| \leq \omega(A, R) \leq min(|A|, |R|)$$

This is indeed a true generalization because, $\omega(A, R) = |A \cap R|$ satisfies all these properties.

## 2.3 Semantic precision and recall and other measures

Our main goal is to design a generalization of precision and recall that is semantically grounded. As a result, those correspondences that are consequences of the evaluated alignments have to be considered as recalled and those that are consequences of the reference alignments as correct.

For that purpose we will attempt to follow the guidelines introduced in [Ehrig and Euzenat, 2005] as far as possible. We add some more constraints to a semantic precision and recall which consider correctness and completeness of an alignment as their limit:

$$\begin{aligned} R \models A \Rightarrow & P_{sem}(A,R) = 1 && \text{(max-correctness)} \\ A \models R \Rightarrow & R_{sem}(A,R) = 1 && \text{(max-completeness)} \\ Cn(A) = Cn(R) \text{ if and only if } & P_{sem}(A,R) = 1 && \\ \text{and } & R_{sem}(A,R) = 1 && \text{(definiteness)} \end{aligned}$$

The classical positiveness depends on $A = R$, it is replaced here by its semantic counterpart. In addition, we rephrase the previous properties by applying them to precision and recall instead of $\omega$ ($M$ is any of precision and recall and $M'$ is its generalized counterpart):

$$\begin{aligned} & M'(A,R) \geq 0 && \text{(positiveness)} \\ & M'(A,R) \leq 1 && \text{(maximality)} \\ & M'(A,R) \geq M(A,R) && \text{(boundedness)} \end{aligned}$$

### 2.3.1 Ideal model

The natural semantic extension of these measures consists of using the set of $\alpha$-consequences (or deductive closure on the prover side) instead of $|A \cap R|$. This corresponds to taking as $\omega$ the size of the set identified by $d$ instead of that identified by $a$ in Figure 2.3.

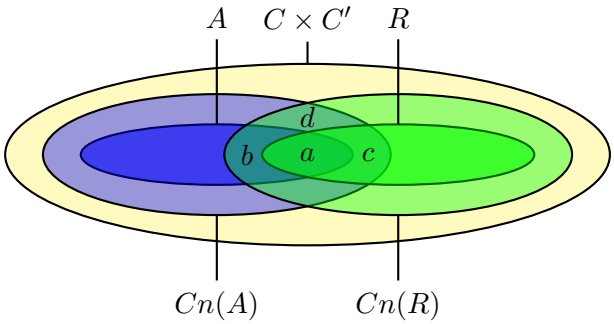



Figure 2.3: Two alignments and their relations through the set of their consequences.

In this case, the true positives become the correspondences that are consequences of both alignments and the usual definitions of true and false positives and negatives are only extended to alignment consequences.

**Definition 11** (Ideal semantic precision and recall)**.** *Given a reference alignment $R$, the precision of some alignment $A$ is given by*

$$P_{ideal}(A,R) = \frac{|Cn(R) \cap Cn(A)|}{|Cn(A)|} = P(Cn(A), Cn(R))$$

*and recall is given by*

$$R_{ideal}(A,R) = \frac{|Cn(R) \cap Cn(A)|}{|Cn(R)|} = R(Cn(A), Cn(R)).$$

This ideal way of dealing with semantic precision and recall can be applied to any language with a semantics. It is not restricted to alignments and as soon as the notion of consequence is defined from an alignment, it can be applied.

These measures are different from the extensions of [Ehrig and Euzenat, 2005] (also reported in Deliverable 2.2.4) because the divisor has changed. However, for a language with a well-defined semantics this measure is a natural extension of precision and recall. Most of the required properties are satisfied by these ideal measures:

**Property 1.** *$P_{ideal}$ and $R_{ideal}$ satisfy:*

- *positiveness;*
- *maximality;*
- *completeness-maximality;*
- *correctness-maximality;*
- *definiteness;*

*$P_{ideal}$ and $R_{ideal}$ do not necessarily satisfy boundedness.*

This extension has two drawbacks: (1) both numerator and divisor could be infinite, yielding an undefined result, and (2) contrary to the objective of [Ehrig and Euzenat, 2005], these measures do not guarantee to provide better results than precision and recall in general, i.e., we do not have neither $P(A,R) \leq P_{ideal}(A,R)$ nor $R(A,R) \leq R_{ideal}(A,R)$. This is because there is no direct relation between the size of an alignment (or a set of axioms) and the size of its $\alpha$-consequences.

### 2.3.2 Semantic precision and recall

In order to deal with the problems raised by the infinite character of the set of $\alpha$-consequences, a natural way would be to compare the deductive reductions instead of the deductive closures. Unfortunately, the deductive reduction is usually not unique. We thus propose to use the deductive closure bounded by a finite set so that the result is finite. It is based on different sets of true positives:

$$TP_P(A,R) = \{\delta \in A; R \models \delta\} = A \cap Cn(R)$$

and

$$TP_R(A,R) = \{\delta \in R; A \models \delta\} = Cn(A) \cap R.$$

These two sets correspond respectively to the sets $b$ and $c$ in Figure 2.3. They are obviously finite since they are the intersection of a set of $\alpha$-consequences with a finite alignment. They are not, however, a real count of the true positive by any means. The semantic precision and recall are based on these sets.

**Definition 12** (Semantic precision and recall)**.** *Given a reference alignment $R$, the precision of some alignment $A$ is given by*

$$P_{sem}(A,R) = \frac{|A \cap Cn(R)|}{|A|}$$

*and recall is given by*

$$R_{sem}(A,R) = \frac{|Cn(A) \cap R|}{|R|}.$$

Both values are defined when the alignments are finite. Moreover, the considered values can be computed if there exists a complete and correct prover for the languages because there is always a finite set of assertions to check (i.e., $Cn(A) \cap R = \{\delta \in R; A \models \delta\}$).

**Property 2.** *$P_{sem}$ and $R_{sem}$ satisfy:*

- *positiveness;*
- *maximality;*
- *boundedness;*
- *completeness-maximality;*
- *correctness-maximality;*
- *definiteness;*

These measures satisfy positiveness and boundedness (since it is clear that $Cn(X) \supseteq X$). They have the classically expected values: $P_{sem}(R,R) = 1$ and $R_{sem}(R,R) = 1$. They do not satisfy anymore, if $A \cap R = \emptyset$, then $P_{sem}(A,R) = 0$ which is replaced by if $Cn(A) \cap Cn(R) = \emptyset$, then $P_{sem}(A,R) = 0$.

### 2.3.3 Compactness and independence

Now we can find that a particular alignment is semantically equivalent to some reference alignment. However, what makes that the reference alignment has been chosen otherwise? Are there criteria that enable to measure the quality of these alignments so that some of them are better than others?

Indeed, more compact alignments turn out to be preferable. Compactness can be measured as the number of correspondences. So it is possible to measure either, this absolute number or the ratio of correspondences in the reference alignments and the found alignments:

$$Compactness(A,R) = \frac{|R|}{|A|}.$$

There is no reason that this measure cannot be higher than 1 (if the found alignment is more compact than the reference alignment). Compactness depends on the raw set of correspondences is, however, primitive and non semantically grounded (it is especially useful for complete and correct alignments). A more adapted measure is that of independence, which checks that alignments are not redundant:

$$Ind(A) = \frac{|\{c \in A; A - \{c\} \not\models c\}|}{|A|}.$$

It measures the ratio of independent correspondences in an alignment independently of a reference. If we want to measure independence with regard to the reference alignment, it is possible to measure the independence of correspondences that do not appear in the reference alignment:

$$Ind(A, R) = \frac{|\{c \in A - R; A - \{c\} \not\models c\}|}{|A - R|}.$$

In the examples considered here independence measures return 1. for all alignments.

## 2.4 Examples

In order to compare the behavior of the proposed measures, we compare it with previously provided measures on the examples of Figure 2.1. Additionally to "standard" precision and recall we compare them with two measures introduced in [Ehrig and Euzenat, 2005]: "symmetry" considers as close a correspondence in which a class is replaced by its direct sub- or super-class; "effort" takes as proximity the measure of the effort to produce for correcting the alignment.

| $\omega$ | standard | | symmetry | | effort | | semantic | | |
|---|---|---|---|---|---|---|---|---|---|
| $A$ | P | R | P | R | P | R | P | R | compactness |
| $R$ | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. | 1. |
| $A_1$ | .5 | .5 | .75 | .75 | .8 | .8 | 1. | .5 | 1. |
| $A_2$ | .33 | .5 | .5 | .75 | .5 | .75 | 1. | 1. | .66 |
| $A_3$ | .5 | .5 | .5 | .5 | .5 | .5 | 1. | .5 | 1. |
| $A_4$ | .5 | .5 | .5 | .5 | .65 | .65 | .5 | .5 | 1. |

Table 2.1: Precision and recall results on the alignments of Figure 2.1.

The results are provided in Table 2.1. As can be expected, the results of the semantic measures match exactly the correctness and completeness of the alignment. These results are far more discriminating than the other ones as far as $A_4$ is concerned. The equivalent alignment $A_2$ which did not stand out with the other measures is now clearly identified as equivalent. An interesting aspect for $A_1$ which is correct but incomplete, is that the other measures fail to recognize this asymmetry. $A_1$ and $A_3$ are both correct and incomplete, they thus have the same semantic values, while $A_1$ is arguably more complete than $A_3$ (in fact $A_1 \models A_3$): this is accounted better by the syntactic measures.

Finally, no redundancy was present in the selected alignments, so they all score 100% in independence. $A_2$ is less compact than $R$ (and the others) as expected.

## 2.5 Discussion

Let us first briefly overview the related work. In particular, relevant work is that of [Langlais *et al.*, 1998] in computational linguistics, [Sun and Lin, 2001] in information retrieval and [Ehrig and Euzenat, 2005] in artificial intelligence. All rely on a syntactic distance between entities of the ontologies (or words or documents). They have been compared in [Ehrig and Euzenat, 2005]. However, these works tackle the problem of measuring how far is a result from the solution. Here, the semantic foundations only consider valid solutions: the semantic definitions account for the semantically equivalent but syntactically different results.

This explains why, with regard to [Ehrig and Euzenat, 2005], we used different but compatible properties not fully relying on a proximity measure $\omega$. Hence, it should be possible to combine the semantic precision and recall with the proposed syntactic relaxed measures.

The results of Table 2.1 show how the semantic approach compares with [Ehrig and Euzenat, 2005] on a small example. One important difference is that the syntactic measures can be easily computed because they only work on the syntax of the alignments while the semantic measures require a prover for the ontology languages (and even for the alignment semantics).

Let us summarize major findings of this chapter. We have provided a semantics for alignments based on the semantics of ontologies and we designed semantic precision and recall measures that take advantage of this semantics. The definition of these measures is thus independent from the semantics of ontologies. We showed that these measures are compliant with (an adapted version of) constraints put on precision and recall generalization of [Ehrig and Euzenat, 2005] and that they behave as expected in the motivating example:

– they are still maximal for the reference alignment;
– they correspond to correctness and completeness;
– they help discriminating between irrelevant alignments and not far from target ones.

Examples have been given based on the alignment of simple taxonomies with very limited languages. However, all the definitions are independent from this language.

# Chapter 3

# Introduction to the OAEI-2006 campaign

This chapter serves as an introduction to the evaluation campaign of 2006 and to the results provided in the following chapters. We present the general methodology for the 2006 campaign as it was defined and report its execution. Specifically, first we introduce OAEI-2006 tracks and test cases (§3.1). Then, we discuss the preparatory, execution and evaluation phases of the campaign, respectively (§3.2, §3.3, and §3.4). Finally, we comment on the OAEI-2006 campaign execution (§3.5).

The results of the OAEI-2006 campaign have been published in [Shvaiko *et al.*, 2006].

## 3.1 Tracks and test cases

The OAEI-2006 campaign has consisted of four tracks gathering six data sets and different evaluation modalities.

**The benchmark track:** Like in previous campaigns, systematic benchmark series have been produced. The goal of this benchmark series is to identify the areas in which each matching algorithm is strong or weak. The test is based on one particular ontology dedicated to the very narrow domain of bibliography and a number of alternative ontologies of the same domain for which alignments are provided.

**The expressive ontologies track:**

**Anatomy:** The anatomy real world case covers the domain of body anatomy and consists of two ontologies with an approximate size of several ten thousands classes and several dozens of relations.

**Jobs:** The jobs test case is an industry evaluated real world business case. A company has a need to improve job portal functionality with semantic technologies. To enable higher precision in retrieval of relevant job offers or applicant profiles, OWL ontologies from the employment sector are used to describe jobs and job seekers and matching with regard to these ontologies provides the improved results. For confidentiality reasons, the test is run by the company team (WorldWideJobs - WWJ GmbH) with software provided by the participants.

It turned out that due to changes in management of the company (WorldWideJobs - WWJ GmbH) this test case has not been completed, and therefore, we do not provide its results.

**The directories and thesauri track:**

**Directory:** The directory real world case consists of matching web directories, such as open directory, Google and Yahoo. It has more than four thousands of elementary tests.

**Food:** Two SKOS thesauri about food have to be matched using relations from the SKOS mapping vocabulary. Samples of the results are evaluated by domain experts.

**The conference track and consensus workshop:** Participants have been asked to freely explore a collection of conference organization ontologies (the domain being well understandable for every researcher). This effort was expected to materialize in usual alignments as well as in interesting individual correspondences ("nuggets"), aggregated statistical observations and/or implicit design patterns. There is no a priori reference alignment. For a selected sample of correspondences, consensus was sought at the workshop and the process of reaching consensus was recorded.

Table 3.1 summarizes the variation in the results expected from these tests.

| test | language | relations | confidence | modalities |
|---|---|---|---|---|
| benchmarks | OWL | = | [0 1] | open |
| anatomy | OWL | = | 1 | blind |
| jobs | OWL | = | [0 1] | external |
| directory | OWL | = | 1 | blind |
| food | SKOS | narrowMatch, exactMatch, broadMatch | 1 | blind+consensual |
| conference | OWL-DL | =, $\leq$ | 1 | blind+consensual |

Table 3.1: Characteristics of test cases (open evaluation is done with already published expected results, blind evaluation is done by organizers from reference alignments unknown to the participants, consensual evaluation is obtained by reaching consensus over the found results and external evaluation is performed independently of the organizers by running the actual systems).

## 3.2 Preparatory phase

The ontologies and alignments of the evaluation have been provided in advance during the period between June 1st and June 28th. This gave potential participants the occasion to send observations, bug reports, remarks and other test cases to the organizers. The goal of this preparatory period is to ensure that the delivered tests make sense to the participants. The tests still evolved after this period, but only for ensuring a better participation to the tests. The final test base was released on August 23rd.

## 3.3 Execution phase

During the execution phase the participants used their systems to automatically match the ontologies from the test cases. Participants have been asked to use one algorithm and the same set of parameters for all tests in all tracks. It is fair to select the set of parameters that provide the best results (for the tests where results are known). Beside parameters, the input of the algorithms must be the two ontologies to be matched and any general purpose resource available to everyone, i.e., no resource especially designed for the test. In particular, the participants should not use the data (ontologies and reference alignments) from other test sets to help their algorithms.

In most cases ontologies are described in OWL-DL and serialized in the RDF/XML format. The expected alignments are provided in the Alignment format expressed in RDF/XML. All the participants also provided the papers that are published in [Shvaiko *et al.*, 2006] and a link to their systems and their configuration parameters.

## 3.4 Evaluation phase

The organizers have evaluated the results of the algorithms used by the participants and provided comparisons on the basis of the provided alignments.

In order to ensure that it is possible to process automatically the provided results, the participants were requested to provide preliminary results by September 4th. In the case of blind tests only the organizers did the evaluation with regard to the withheld reference alignments. In the case of double blind tests, the participants provided a version of their system and the values of the parameters if any.

The standard evaluation measures are precision and recall computed against the reference alignments. For the matter of aggregation of the measures we used weighted harmonic means (weights being the size of the true positives). This clearly helps in case of empty alignments. Another technique that was used is the computation of precision/recall graphs so it was advised that participants provide their results with a weight to each correspondence they found.

New measures addressing some limitations of precision and recall have also been used for testing purposes. These were presented at the OM-2006 workshop[1] discussion in order for the participants to provide feedback on the opportunity to use them in further evaluation.

## 3.5 Comments on the execution

In OAEI-2006 we had more participants than in previous years: 4 in 2004, 7 in 2005 and 10 in 2006. We also noted the increase in tools compliance and robustness: they had less problems to carry the tests and we had less problems to evaluate the results.

We have had not enough time so far to validate the results which have been provided by the participants. In 2005, validating these results has proved feasible so we plan to do it again in future (at least for those participants who provided their systems).

We summarize the list of participants in Table 3.2. Similar to OAEI-2005 not all participants provided results for all tests. They usually did those which are easier to run, such as benchmark, directory and conference. The jobs line corresponds to the participants who have provided an

[1] `http://www.om2006.ontologymatching.org`

| Test \ System | Falcon | HMatch | DSSim | COMA++ | AUTOMS | JHU/APL | PRIOR | RiMOM | OCM | NIH | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Benchmark | √ | √ | √ | √ | √ | √ | √ | √ | √ | | 9 |
| Anatomy | √ | √ | | √ | | | √ | | | √ | 5 |
| Jobs | √ | √ | | √ | √ | | √ | | √ | | 6 |
| Directory | √ | √ | | √ | √ | | √ | √ | √ | | 7 |
| Food | √ | √ | | √ | | | √ | √ | | | 5 |
| Conference | √ | √ | | √ | √ | | | √ | √ | | 6 |
| Certified Confidence | √ | √ | | √ | | | √ | √ | | | 5 |
| Time | | | | | √ | | √ | √ | √ | √ | 5 |

Table 3.2: Participants and the state of their submissions. Confidence is ticked when given as non boolean value. Time indicates when participants included execution time with their tests.

executable version of their systems. The variety of tests and the short time given to provide results have certainly prevented participants from considering more tests.

Like in 2005, the time devoted for performing these tests (three months) and the period allocated for that (summer) is relatively short and does not really allow the participants to analyze their results and improve their algorithms. On the one hand, this prevents having algorithms to be particularly tuned for the tests. On the other hand, this can be frustrating for the participants.

# Chapter 4

# The benchmark track

The goal of the benchmark tests is to provide a stable and detailed picture of each algorithm. For that purpose, the algorithms are run on systematically generated test cases. In particular, first we discuss the test cases (§4.1), and then we provide their results for the participated systems (§4.2).

## 4.1 Test set

The domain of this first test is Bibliographic references. It is, of course, based on a subjective view of what must be a bibliographic ontology. There can be many different classifications of publications, for example, based on area and quality. The one chosen here is common among scholars and is based on publication categories; as many ontologies (tests #301-304), it is reminiscent to BibTeX.

The systematic benchmark test set is built around one reference ontology and many variations of it. The reference ontology is that of test #101. The participants have to match this reference ontology with the variations. These variations are focusing the characterization of the behavior of the tools rather than having them compete on real-life problems. The ontologies are described in OWL-DL and serialized in the RDF/XML format. This reference ontology contains 33 named classes, 24 object properties, 40 data properties, 56 named individuals and 20 anonymous individuals.

Since the goal of these tests is to offer some kind of permanent benchmarks to be used by many, the data set is an extension of the 2004 EON Ontology Alignment Contest. The reference ontology has been improved in 2005 by including circular relations that were missing from the first test. In 2006, we put the UTF-8 version of the tests as standard, the ISO-8859-1 being optional. Test numbering (almost) fully preserves the numbering of the first EON contest of 2004.

The kind of expected alignments is still limited: they only match named classes and properties, they mostly use the "=" relation with confidence of 1. There are three groups of tests in this benchmark:

**Simple tests (1xx)** such as comparing the reference ontology with itself, with another irrelevant ontology (the wine ontology used in the OWL primer) or the same ontology in its restriction to OWL-Lite;

**Systematic tests (2xx)** that were obtained by discarding features from some reference ontology. It aims at evaluating how an algorithm behaves when a particular type of information is lacking. The considered features were:

- *Name of entities* that can be replaced by random strings, synonyms, name with different conventions, strings in another language than English;
- *Comments* that can be suppressed or translated in another language;
- *Specialization hierarchy* that can be suppressed, expanded or flattened;
- *Instances* that can be suppressed;
- *Properties* that can be suppressed or having the restrictions on classes discarded;
- *Classes* that can be expanded, i.e., replaced by several classes or flattened.

**Four real-life ontologies of bibliographic references (3xx)** that were found on the web and left mostly untouched (there were added xmlns and xml:base attributes).

Full description of these tests can be found on the OAEI web site[1], see also Table 4.3 (p.25).

## 4.2 Results

Table 4.1 provides the consolidated results, by groups of tests. We display the results of participants as well as those given by some very simple edit distance algorithm on labels (edna). Like in 2005, the computed values are real precision and recall and not a simple average of precision and recall. The full results are on the OAEI web site.

These results show already that three systems are relatively close (COMA++, Falcon and RiMOM). The RiMOM system is slightly ahead of the others on these raw results. The DSSim system obviously favored precision over recall but its precision degrades with "real world" 3xx series. No system had strictly lower performance than edna.

The results have also been compared with the three measures proposed in [Ehrig *et al.*, 2005] (also reported in deliverable 2.2.4) in 2005 (symmetric, effort-based and oriented). These are generalization of precision and recall in order to better discriminate systems that slightly miss the target from those which are grossly wrong. The three measures provide the same results. This is not really surprising given the proximity of these measures. As expected, they can only improve over traditional precision and recall. The improvement affects all the algorithms, but this is not always strong enough for being reflected in the aggregated results. Moreover, the new measures do not dramatically change the evaluation of the participating systems.

Each algorithm has its best score with the 1xx test series. There is no particular order between the two other series. Again, it is more interesting to look at the 2xx series structure to distinguish the strengths of algorithms.

In 2006 the apparently best algorithms provided their results with confidence measures. It is thus possible to draw precision/recall curves in order to compare them. We provide in Figure 4.1 the precision and recall graphs of 2006. They involve only the results of participants who provided confidence measures different from 1 or 0 (see Table 3.2). They also feature the results for edit distances on class names (edna) and the results of Falcon in 2005 (denoted as Falcon-2005) (presentation of these results have also been provided with another presentation in deliverable 1.2.2.2.1). The graph for Falcon-2005 is not really accurate since it provided 1/0 alignments in 2005. This graph has been drawn with only technical adaptation of the technique used in TREC. Moreover, due to lack of time, these graphs have been computed by averaging the graphs of each of the tests (instead of pure precision and recall).

[1] `http://oaei.ontologymatching.org/2006/`

| Algo | refalign | | edna | | AUTOMS | | COMA++ | | DSSim | | Falcon | | HMatch | | JHU/APL | | OCM | | PRIOR | | RiMOM | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Test | P | R | P | R | P | R | P | R | P | R | P | R | P | R | P | R | P | R | P | R | P | R |
| 1xx | 1.00 | 1.00 | 0.96 | 1.00 | 0.94 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 1.00 | 1.00 | 0.91 | 1.00 | 1.00 | 1.00 | 0.95 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2xx | 1.00 | 1.00 | 0.90 | 0.49 | 0.94 | 0.64 | 0.96 | 0.82 | 0.99 | 0.49 | 0.91 | 0.85 | 0.83 | 0.51 | 0.20 | 0.86 | 0.93 | 0.51 | 0.95 | 0.58 | 0.97 | 0.87 |
| 3xx | 1.00 | 1.00 | 0.94 | 0.61 | 0.91 | 0.70 | 0.84 | 0.69 | 0.90 | 0.78 | 0.89 | 0.78 | 0.78 | 0.57 | 0.18 | 0.50 | 0.89 | 0.51 | 0.85 | 0.80 | 0.83 | 0.82 |
| Total | 1.00 | 1.00 | 0.91 | 0.54 | 0.94 | 0.67 | 0.96 | 0.83 | 0.98 | 0.55 | 0.92 | 0.86 | 0.84 | 0.55 | 0.22 | 0.85 | 0.93 | 0.55 | 0.95 | 0.63 | 0.96 | 0.88 |
| Symmetric | 1.00 | 1.00 | 0.91 | 0.54 | 0.94 | 0.68 | 0.96 | 0.83 | 0.99 | 0.55 | 0.94 | 0.89 | 0.85 | 0.56 | 0.22 | 0.87 | 0.93 | 0.55 | 0.96 | 0.64 | 0.97 | 0.89 |
| Effort-based | 1.00 | 1.00 | 0.91 | 0.54 | 0.94 | 0.68 | 0.96 | 0.83 | 0.99 | 0.55 | 0.94 | 0.89 | 0.85 | 0.56 | 0.22 | 0.87 | 0.93 | 0.55 | 0.96 | 0.64 | 0.97 | 0.89 |
| Oriented | 1.00 | 1.00 | 0.91 | 0.54 | 0.94 | 0.68 | 0.96 | 0.83 | 0.99 | 0.55 | 0.94 | 0.89 | 0.85 | 0.56 | 0.22 | 0.87 | 0.93 | 0.55 | 0.96 | 0.64 | 0.97 | 0.89 |

Table 4.1: Means of results obtained by participants on the benchmark test case (corresponding to harmonic means).

Contrary to 2005, we have three systems competing at the highest level in 2006 (Falcon, COMA++ and RiMOM) and there is a gap between these and the other systems. We have compared the results of OAEI-2006 with the previous years on the basis of 2004 tests, see Table 4.2. The three best systems (Falcon, COMA++ and RiMOM) arrive at the level of 2005 best system (Falcon). However, no system outperforms it. Unfortunately no representant of the group of systems that followed Falcon in OAEI-2005 participated in OAEI-2006.

The best systems are at the level of 2005 best system (Falcon).

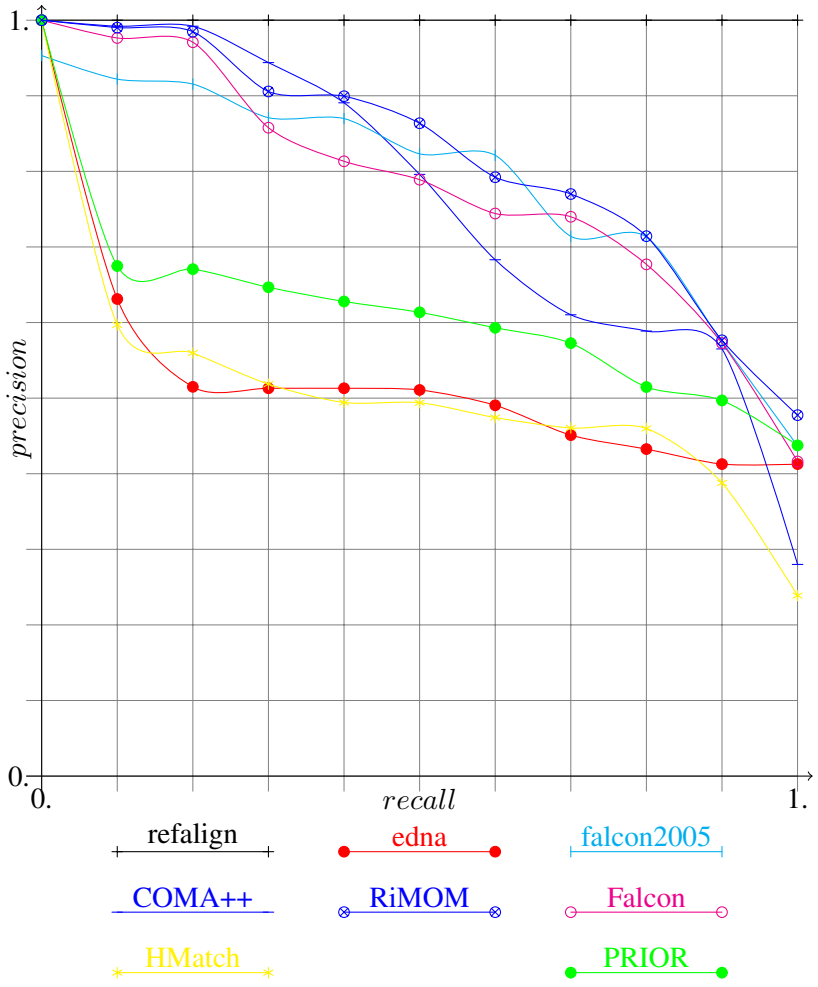



Figure 4.1: Precision/recall graphs for the systems which provided confidence values in their results.

| Year | 2004 | | | | 2005 | | 2006 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| System | Fujitsu | | Stanford | | Falcon | | RiMOM | | Falcon | | COMA++ | |
| test | P | R | P | R | P | R | P | R | P | R | P | R |
| 1xx | 0.99 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2xx | 0.93 | 0.84 | 0.98 | 0.72 | 0.98 | 0.97 | 1.00 | 0.98 | 0.97 | 0.97 | 0.99 | 0.97 |
| 3xx | 0.60 | 0.72 | 0.93 | 0.74 | 0.93 | 0.83 | 0.83 | 0.82 | 0.89 | 0.78 | 0.84 | 0.69 |
| H-means | 0.88 | 0.85 | 0.98 | 0.77 | 0.97 | 0.96 | 0.97 | 0.96 | 0.97 | 0.95 | 0.98 | 0.94 |

Table 4.2: Evolution of the best scores over the years (on the basis of 2004 tests).

| # | Name | Com | Hier | Inst | Prop | Class | Comment |
|---|---|---|---|---|---|---|---|
| 101 | | | | | | | Reference alignment |
| 102 | | | | | | | Irrelevant ontology |
| 103 | | | | | | | Language generalization |
| 104 | | | | | | | Language restriction |
| 201 | R | | | | | | No names |
| 202 | R | N | | | | | No names, no comments |
| 203 | | N | | | | | No comments (was missspelling) |
| 204 | C | | | | | | Naming conventions |
| 205 | S | | | | | | Synonyms |
| 206 | F | F | | | | | Translation |
| 207 | F | | | | | | |
| 208 | C | N | | | | | |
| 209 | S | N | | | | | |
| 210 | F | N | | | | | |
| 221 | | | N | | | | No specialisation |
| 222 | | | F | | | | Flatenned hierarchy |
| 223 | | | E | | | | Expanded hierarchy |
| 224 | | | | N | | | No instance |
| 225 | | | | | R | | No restrictions |
| 226 | | | | | | | No datatypes |
| 227 | | | | | | | Unit difference |
| 228 | | | | | N | | No properties |
| 229 | | | | | | | Class vs instances |
| 230 | | | | | | F | Flattened classes |
| 231* | | | | | | E | Expanded classes |
| 232 | | | N | N | | | |
| 233 | | | N | | N | | |
| 236 | | | | N | N | | |
| 237 | | | F | N | | | |
| 238 | | | E | N | | | |
| 239 | | | F | | N | | |
| 240 | | | E | | N | | |
| 241 | | | N | N | N | | |
| 246 | | | F | N | N | | |
| 247 | | | E | N | N | | |
| 248 | N | N | N | | | | |
| 249 | N | N | | N | | | |
| 250 | N | N | | | N | | |
| 251 | N | N | F | | | | |
| 252 | N | N | E | | | | |
| 253 | N | N | N | N | | | |
| 254 | N | N | N | | N | | |
| 257 | N | N | | N | N | | |
| 258 | N | N | F | N | | | |
| 259 | N | N | E | N | | | |
| 260 | N | N | F | | N | | |
| 261 | N | N | E | | N | | |
| 262 | N | N | N | N | N | | |
| 265 | N | N | F | N | N | | |
| 266 | N | N | E | N | N | | |
| 301 | | | | | | | Real: BibTeX/MIT |
| 302 | | | | | | | Real: BibTeX/UMBC |
| 303 | | | | | | | Real: Karlsruhe |
| 304 | | | | | | | Real: INRIA |

Table 4.3: Structure of the systematic benchmark test case.

# Chapter 5

# The expressive ontologies track

The focus of the anatomy test case used in the expressive ontology track is to confront existing matching technology with real world ontologies. Our aim here is to obtain a better impression of where the matching technology is positioned with respect to real challenges that normally require an enormous manual effort and an in-depth knowledge of the domain.

In particular, first we discuss the test case (§5.1). Then, we provide its early results (§5.2), and a followed up alignment cross-validation analysis for the participated systems (§5.3). Finally, we outline the major findings of this chapter (§5.4).

Some parts of the material presented in this chapter have been published in [Zhang and Bodenreider, 2007b].

## 5.1 Test set

The task is placed in the medical domain as this is the domain where we find large, carefully designed ontologies. The specific characteristics of the ontologies under consideration include:

- Very large models: OWL models of more than 50MBs.
- Extensive class hierarchies: ten thousands of classes organized according to different views on the domain.
- Complex relationships: classes are connected by a number of different relations.
- Stable terminology: the basic terminology is rather stable and should not differ substantially in the different models.
- Clear modeling principles: the modeling principles are well-defined and documented in publications about the ontologies.

As a consequence, the anatomy test case actually assesses existing matching systems with respect to two questions.

- Do existing approaches scale to very large models?
- Are existing approaches able to take advantage of well-documented modeling principles and knowledge about the domain?

The above two questions also mean that the goal of this test case is not to compare the performance of matching systems on a quantitative basis (though we provide some comparisons). We

are rather interested in how many systems are actually able to create an alignment at all and in specific heuristics used for computing it.

The ontologies to be matched are different representations of human anatomy developed independently by teams of medical experts. Both ontologies are available in OWL format and mostly contain classes and relations between them. The use of axioms is limited.

### 5.1.1 The Foundational Model of Anatomy

The Foundational Model of Anatomy (FMA) has been developed by the University of Washington. It is an ontology describing the human anatomy including a taxonomy of body parts, information about anatomical structures and structure transformations. According to the developers the Foundational Model of Anatomy ontology contains approximately 75.000 classes and over 120.000 terms; over 2.1 million relationship instances from 168 relationship types link the FMA's classes into a coherent symbolic model.

We extracted an OWL version of the ontology from a Protégé database. The resulting model is in OWL-full as relations are defined between classes rather than instances.

### 5.1.2 Galen

The second ontology is the anatomy model developed in the OpenGalen Project by the University of Manchester. According to the creators, the ontology contains around 10.000 concepts covering somewhat more than standard textbook anatomy in terms of body parts, anatomical structures and relations between different parts and structures.

The ontology is freely available as a Protégé project file on the OpenGalen web page. We created an OWL version of the ontology using the export functionality of Protégé. The resulting ontology is in OWL-DL thus supporting logical reasoning about inconsistencies.

## 5.2 Initial results

The anatomy use case is part of the ontology alignment evaluation challenge for the second time now. While in 2005, none of the participants was in a position to submit a result for this data set. Almost all participants reported major difficulties in processing the ontologies due to their size and the fact that one of the models is in OWL full. At least these scalability problems appear somewhat to be solved this year. In the OAEI-2006 campaign, five out of ten participants submitted results for the anatomy data set. This clearly shows the advance of matching systems on the technical level and also suggests that matching technology is potentially ready for large scale applications.

On the content level the results are much harder to judge. Due to lack of a reference alignment, we were not able to provide a quantitative judgment and comparison of the different systems for the time of the OAEI-2006 results presentation at the Ontology Matching workshop[1]. Thus, we rather concentrated on the coverage of the ontologies, the degree of agreement amongst the matching systems and on the specific techniques the matching systems used in order to address this task.

A first observation is that none of the systems managed to reach a good coverage of the ontologies. Although both models contain several ten thousand concepts and the fact that we can assume

[1] http://www.om2006.ontologymatching.org/

a high degree of overlap in the two models, the systems were only able to produce correspondences for 2000 to 3000 concepts, which is less than 5% of the concepts of FMA.

We also found out that systems have severe difficulties with irregular concept names. The GALEN ontology contains a subset of concepts with highly irregular concept names. It turned out that only one system (COMA++) was able to determine correspondences for these concepts – at the expense of not being able to match any of the concept names with regular names.

A common pattern can be observed by looking at the actual methods used by the systems. Almost all systems use the linguistic similarity between class names and other features of the class description as a basis for determining match candidates. Normally, the systems combine different similarity measures. On top of this purely linguistic comparison, some systems also apply structural techniques. In particular, they translate the models into a graph structure and propagate the individual similarity in the graph structure. Only one of the systems (AOAS) actually used reasoning techniques to validate hypothesis and to determine matches based on the semantics of the models.

## 5.3 Cross-validation of the alignments

After the OAEI-2006 campaign, the results discussed so far have been further analyzed in [Zhang and Bodenreider, 2007b]. In particular, some of the results were reviewed based on cross-validation in order to obtain insights into the strengths and weaknesses of the various approaches. The key hypothesis of the cross-validation is that correspondences identified by several systems have a better chance to be valid.

### 5.3.1 Systems under consideration

The three matching systems analyzed in this study are the Anatomical Ontology Alignment System (AOAS), which participated in OAEI-2006 as the NIH system, PRIOR and Falcon. Two other systems participating in the OAEI-2006 campaign are not included in this review for the following reasons. Almost all correspondences identified by COMA++ were specific to this system and could not therefore contribute to cross-validation. The result files contributed by IsLab were not available when this study was performed. As most matching systems, the three systems under investigation rely on a combination of lexical and structural methods, based on the assumption that equivalent concepts across ontologies have similar names and similar relations to other concepts.

Notice that both PRIOR and Falcon allow partial matches between concept names (e.g., Adductor magnus of thigh matches Adductor magnus), while only minor term variations are allowed between matches by AOAS. Unlike AOAS or Falcon, PRIOR can exploit the anonymous concepts in GALEN. Finally, while AOAS only identifies correspondences between concepts, PRIOR and Falcon also find correspondences between relationships.

### 5.3.2 Methods and results

**Overlap.** First, the intersection among the lists of correspondences obtained by the three systems has been computed. This partitioned the set of all correspondences into subsets with respect to the systems that identified them (e.g., correspondences identified by AOAS and Falcon, but not by PRIOR). In particular, 1.429 matches were identified by all of the three systems, accounting for approximately 46%, 57% and 55% of concept matches in AOAS, Falcon and PRIOR, respectively.

The proportion of correspondences specific to one system varies largely, from 14% for Falcon to 39% for PRIOR, with 27% for AOAS.

**Manual validation.** Olivier Bodenreider manually reviewed for accuracy all correspondences not identified by AOAS. There are several reasons for explaining the bias towards this system. Unlike the other two systems, AOAS was developed specifically for matching anatomical ontologies. In a recent work [Zhang and Bodenreider, 2007a], those results were evaluated against a reference alignment established manually and against other systems. Recall was about .9 and most correspondences identified specifically by AOAS were deemed valid. Then, the correspondences were classified into the following categories: certain, possible (requires additional domain knowledge) and wrong. The objective of this cursory evaluation is primarily to quantify the false positives for Falcon and PRIOR and the false negatives for AOAS. As a result, 1.183 of the 1.383 (86%) correspondences not identified by AOAS were deemed invalid. More knowledge is required to establish the validity of half of the remaining 14%.

## 5.4 Discussion

For the first time since the anatomy data set has been used in the ontology alignment evaluation challenge, we are in a position, where we can actually compare the results of different matching systems. The results show that there is still a lot of work to do to make matching systems ready for real life applications. The problems above showed that differences in the naming scheme of classes can already cause matchers to fail on a significant subset of the vocabulary. It seems that existing matchers suffer from the need to balance precision and recall in determining correspondences. This conclusion is backed by results from other experiments, where it turned out that matching systems that produce highly precise correspondences miss many correspondences found by other systems. We conclude that using fixed thresholds to determine match candidates is not a good way for trading-off precision and recall.

Lexical matching constitutes an important step in ontology matching. Systems such as PRIOR focusing on bag-of-word matching rather than term matching miss many correspondences identified by the other two systems on the basis of exact matches of concept names. Compared to AOAS, Falcon uses a relaxed model of lexical similarity, based on edit distance. AOAS missed some correspondences due to improper segmentation of the original GALEN strings. For example, the string SupraHyoidMuscle was segmented at points where case changes, leading to the term supra hyoid muscle. However, the proper spelling for this term is suprahyoid muscle and the normalization algorithm used by AOAS could not match the two terms. In contrast, the relaxed approach to string matching employed by Falcon identified the two strings as a match. The analysis of the correspondences identified by Falcon and not AOAS revealed about 10 segmentation issues and 15 misspellings in GALEN (e.g., Mensicus for Meniscus). Conversely, the relaxed model of lexical resemblance can lead to "egregious" correspondences and therefore be extremely detrimental to the alignment. For example, Falcon identified a correspondence between Axillary artery (in the armpit) and Maxillary artery (near the mandible).

We were disappointed to see that only one system actually used some form of reasoning in order to take the meaning of the ontologies into account. As one of the major advantages of OWL is the ability to specify and reason about the semantics of concepts, it is at least surprising that this feature is not exploited by existing matchers. In fact, logical reasoning could be a way of

becoming less dependent on the quality of certain similarity measures that obviously have some limitations when it comes to complex ontologies.

AOAS is the only system to fully take advantage of synonymy for the alignment. Some synonyms are provided by the FMA, but others come from the UMLS metathesaurus. In fact, it has been verified that most of the 856 correspondences identified by AOAS are indeed valid and involve such synonyms. This is the case, for example, of the correspondence between Aortic orifice and Ostium of aorta, and between Shoulder joint and Glenohumeral joint. The more conservative and linguistically-motivated approach to lexical similarity adopted by AOAS [McCray *et al.*, 1994] prevents a large number of false positives. However, it is also more sensitive to misspellings and segmentation issues, as well as missing synonyms. Overall, we believe that the benefit of preventing many false positives largely outweighs the few false negatives. It is clear why generic and domain-independent systems such as Falcon and PRIOR have adopted relaxed lexical models. The resources available for biomedicine (UMLS synonyms, domain-specific model of lexical resemblance) are not available for most domains. However, pairs of long terms encountered in anatomy often differing by one qualifier (e.g., for laterality) have an artificially high similarity value when compared with edit distance or in a vector space model. Calibrating the models for a particular domain is an issue that remains to be addressed.

In summary, the results of the anatomy test case have shown that there is some significant progress in terms of the maturity of matching technology. However, the results also show that there are still a lot of open problems with respect to producing good alignments on real life cases.

# Chapter 6

# The directories and thesauri track

This track includes two test cases: directory (§6.1) and food (§6.2).

## 6.1 The directory test case

The directory test case aims at providing a challenging task for ontology matchers in the domain of large web directories. The manual construction of the reference correspondences for the large applications is usually too demanding to the point of being infeasible, since the number of possible correspondences grows quadratically with the number of entities to be compared. This suggests the need for the development of semi-automatic approaches for acquiring the reference correspondences. First, we introduce the methodology used to build semi-automatically the datasets for evaluating recall (§6.1.1) and precision (§6.1.2). Then, we briefly summarize the web directories test set as used in OAEI-2006 (§6.1.3) and present its results for the participating systems (§6.1.4). Finally, we discuss the major findings of the web directories test case (§6.1.5).

### 6.1.1 A dataset for evaluating recall

We follow the semi-automatic method for an approximation of reference alignment proposed in [Avesani *et al.*, 2005] and applied it to the Google, Yahoo and Looksmart web directories. The key idea is to rely on a reference interpretation for entities (nodes), constructed by analyzing which documents have been classified in which nodes. The assumption is that the semantics of nodes can be derived from their pragmatics, namely from analyzing the documents that are classified under the given nodes. In particular, following on the work described in [Avesani *et al.*, 2005] we argue that the meaning of two nodes is equivalent if the sets of documents classified under those nodes have a meaningful overlap. The basic idea is therefore to compute the relationship hypotheses based on the co-occurence of documents.

Let us consider the example presented in Figure 6.1. Let $N_1$ be a node in the first taxonomy and $N_2$ be a node in the second taxonomy. $D_1$ and $D_2$ stand for the sets of documents classified under the nodes $N_1$ and $N_2$ respectively. $A_2$ denotes the documents classified in the ancestor node of $N_2$; $C_1$ denotes the documents classified in the children nodes of $N_1$. A simple *equivalence* measure is defined as follows:

$$Eq(N_1, N_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2| - |D_1 \cap D_2|} \tag{6.1}$$

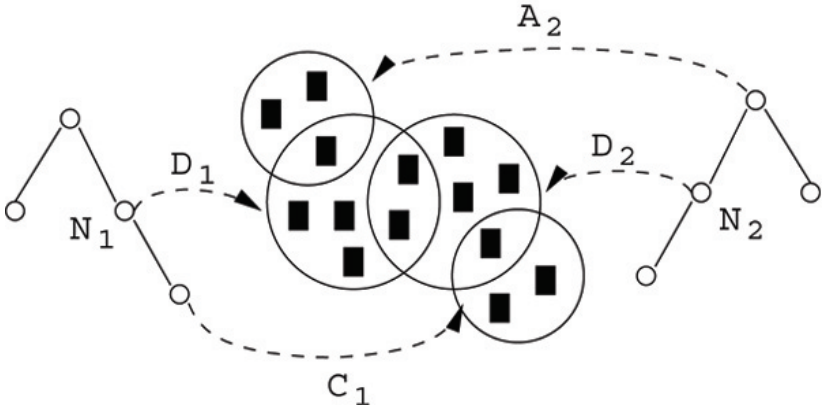



Figure 6.1: *TaxME*. Illustration of a document-driven similarity assessment.

Notice that the range of $Eq(N_1, N_2)$ is [0,$\infty$]. The intuition is that the more $D_1$ and $D_2$ overlap, the bigger $Eq(N_1, N_2)$; with $Eq(N_1, N_2)$ becoming infinite with $D_1 \equiv D_2$. $Eq(N_1, N_2)$ is normalized within [0,1]. The special case of $D_1 \equiv D_2$ is approximated to 1.

Let us first focus on the generalization relation. Given two nodes $N_1$ and $N_2$ and the related document sets $D_1$ and $D_2$, we introduce two additional sets: $(i)$ the set of documents classified in the ancestor node of $N_2$, namely $A_2$, and $(ii)$ the set of documents classified in the children nodes of $N_1$, namely $C_1$.

The generalization relationship holds when the first node has to be considered more general of the second node. Intuitively, this happens when the documents classified under the first node occur in the ancestor of the second node, or the documents classified under the second node occur in the subtree of the first node. Following this intuition we can formalize the generalization hypothesis as follows:

$$Mg(N_1, N_2) = \frac{|(A_2 \cap D_1) \cup (C_1 \cap D_2)|}{|D_1 \cup D_2|} \tag{6.2}$$

The specialization relationship hypothesis $Lg(N_1, N_2)$ can be easily formulated exploiting the symmetry of the problem.

The reference alignment for *TaxME* dataset is computed starting from Google, Yahoo! and Looksmart. The web directories hold many interesting properties: they are widely known, they cover overlapping topics, they are heterogeneous, they are large, and they address the same space of contents. All of this makes the working hypothesis of documents co-occurrence sustainable. The nodes are considered as categories denoted by lexical labels, the tree structures are considered as hierarchical relations, and the URLs classified under a given node are taken to denote documents. Table 6.1 summarizes the total amount of processed data.

Table 6.1: Number of nodes and documents processed in the *TaxME* construction process.

| **Web Directories** | **Google** | **Looksmart** | **Yahoo!** |
|---|---|---|---|
| number of nodes | 335.902 | 884.406 | 321.585 |
| number of urls | 2.425.215 | 8.498.157 | 872.410 |

Let us briefly summarize the five steps process used in the *TaxME* reference alignment construction.

**Step 1** All three web directories are crawled, both the hierarchical structure and the web content;

**Step 2** The URLs that do not exist in at least one web directory are discarded;

**Step 3** The nodes with a number of URLs under a given threshold (10 in the experiment) are pruned;

**Step 4** A manual selection is performed with the goal to restrict the assessment of the similarity metric to the subtrees concerning the same topic; 50 pairs of subtrees are selected;

**Step 5** For each of the subtree pairs selected, an exhaustive assessment of correspondences holding between nodes is performed. This is done by exploiting the equivalence metric defined by Eq. 6.1 and the corresponding generalization (Eq. 6.2) and specialization metrics. The *TaxME* similarity metric is computed as the biggest out of the three metrics, namely as follows:

$$Sim_{TaxME} = max(Eq(N_1, N_2), Lg(N_1, N_2), Mg(N_1, N_2)) \quad (6.3)$$

The distribution of correspondences constructed using $Sim_{TaxME}$ is depicted in Figure 6.2.

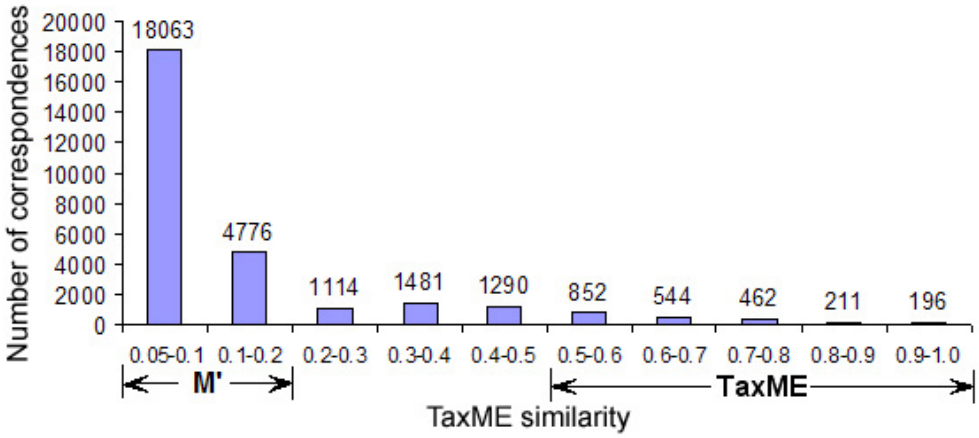



Figure 6.2: Distribution of correspondences according to the *TaxME* similarity metric.

Notice that $Sim_{TaxME}$ is robust. The number of correspondences is in fact very stable and grows substantially, of two orders of magnitude, only with a value of the metric less than 0.1. As a pragmatic decision, the correspondences with $Sim_{TaxME}$ above 0.5 are taken to constitute the reference alignment $TaxME$. As a result, $TaxME$ is composed from 2265 correspondences. Half of them are equivalence relationships and half are generalization relationships.

As depicted in Figure 6.3, $TaxME$ is an incomplete reference alignment since it contains only part of the correspondences in $H$. The key difference between Figure 6.3 and Figure 2.2 is the fact that a complete reference alignment (the area inside the dotted circle in Figure 6.3) is simulated by exploiting an incomplete one (the area inside the dashed circle in Figure 6.3).

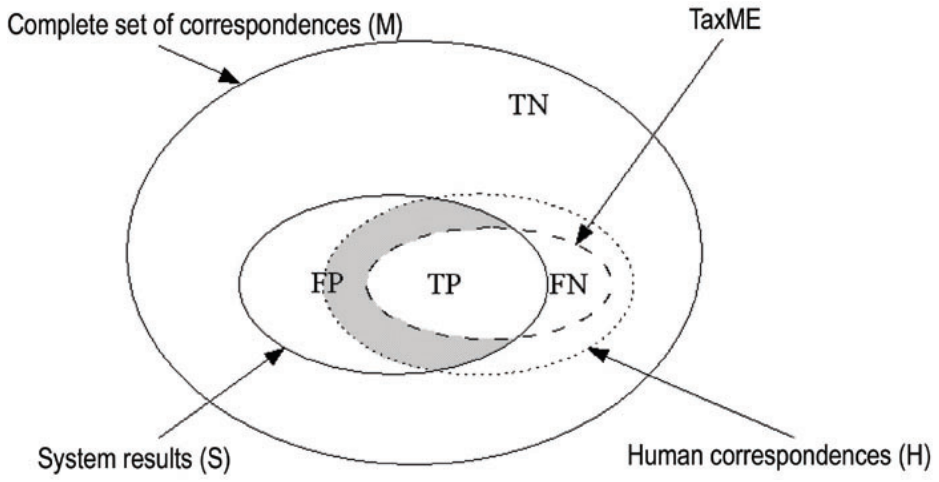



Figure 6.3: Alignment comparison using *TaxME*. $TP$, $FN$ and $FP$ stand for true positives, false negatives and false positives, respectively.

However, if we assume that *TaxME* is a good representative of $H$ we can use definition of recall as defined in §2.1.3 (p.9) for its estimation. In order to ensure that this assumption holds a set of requirements have to be satisfied:

1. *Correctness*: $TaxME \subset H$ (modulo annotation errors).
2. *Complexity*: state of the art matching systems experience difficulties when run on *TaxME*.
3. *Discrimination capability*: different sets of correspondences taken from *TaxME* are hard for the different systems.
4. *Incrementality*: *TaxME* allows for the incremental discovery of the weaknesses of the tested systems[1].

As discussed in [Avesani *et al.*, 2005] *TaxME* satisfies these requirements. We have also evaluated the robustness of $Sim_{TaxME}$. We have randomly selected 100 correspondences in each of the intervals of $Sim_{TaxME}$ values depicted in Figure 6.4 and manually evaluated their correctness. This resulted in a relatively small amount of manual work as we have analyzed around one thousand correspondences. The results are presented in Figure 6.4 and show that $Sim_{TaxME}$ is robust, namely:

– It is very stable with a small percentage of incorrect correspondences for a very large range [0.3,1];

– The number of incorrect correspondences becomes substantial for very small values of $Sim_{TaxME}$, namely with threshold less than 0.1.

[1] We do not consider this property here as insignificant to our goals.

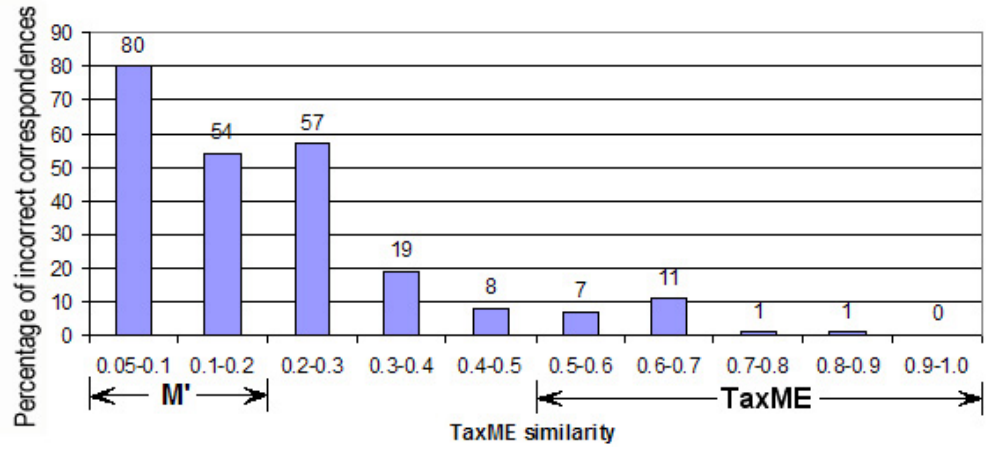



Figure 6.4: Distribution of incorrect correspondences. Each column is calculated evaluating 100 randomly selected correspondences.

### 6.1.2 A dataset for evaluating precision

As from §2.1.3 (p.9) in order to evaluate precision, we need to know $FP$, which in turn, as from Figure 2.2 (and Figure 6.3) requires to know $H$ (reference alignment). However, computing $H$ in the case of a large scale matching task requires an implausible human effort. We cannot either use an incomplete reference alignment composed from positive correspondences, i.e., $TaxME$. In this case, as shown in Figure 6.3, $FP$ can not be computed. This is the case because $FP_{unknown} = S \cap (H - TaxME)$, marked as a gray area in Figure 6.3, is not known.

Our proposal here is to construct a reference alignment for the evaluation of both recall and precision, let us call it $TaxME_2$, defined as follows:

$$TaxME_2 = TaxME \cup N_{T2} \tag{6.4}$$

$N_{T2}$ is an incomplete reference alignment containing *only* negative correspondences (i.e., $N_{T2} \subset M - H$ in Figure 6.3). Of course $TaxME_2$ must be a good representative of $M$ and therefore satisfy the requirements described in the previous section and satisfied by *TaxME*. Notice that the request of correctness significantly limits the size of $N_{T2}$ since each correspondence has to be evaluated by a human annotator (i.e., $|N_{T2}| \ll |M - H|$). At the same time, $N_{T2}$ must be big enough in order to be the source of meaningful results. Therefore, we require $N_{T2}$ to be at least of the same size as $TaxME$, namely $|N_{T2}| \geq |TaxME|$.

$N_{T2}$ is computed from the complete alignment set $M$ in two macro steps. Let us first introduce them briefly and then discuss them in detail.

– *Step 1: Candidate correspondence selection.* The goal of this step is to select a set $M'$ where $M' \subseteq M$ which contains a big number of "hard" negative correspondences.

– *Step 2: Negative correspondence selection.* The goal of this step is to filter all positive correspondences from $M'$. In order to achieve this goal $M'$ is first pruned to the size that allows manual evaluation of the correspondences. Finally, the negative correspondences are manually selected from the remaining set of correspondences.

**Candidate correspondence selection**

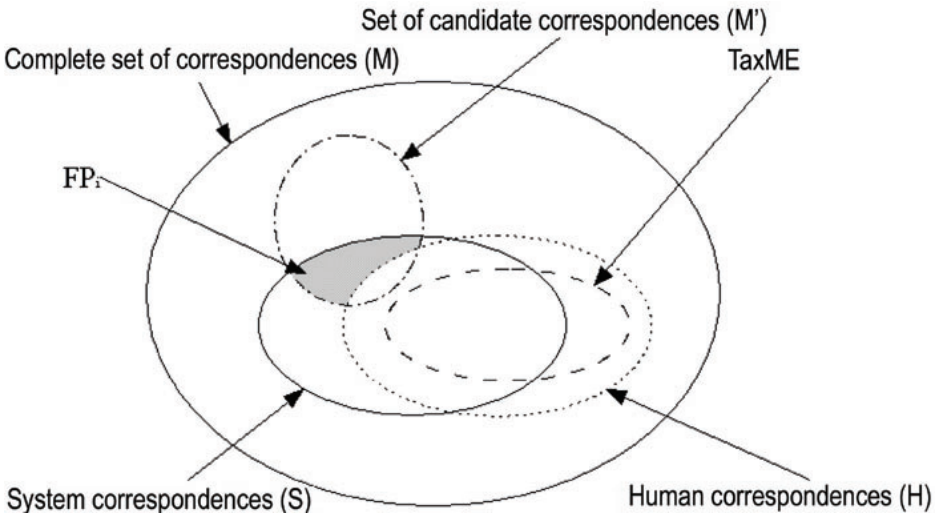



Figure 6.5: Alignment sets in *TaxME 2*. Gray area stands for $FP_i$ a set of FP produced by a matching system on $M'$.

The candidate set of correspondences $M'$ is selected from $M$, as depicted in Figure 6.5. The goal of this step is to ensure that $M'$ contains a big number of "hard" negative correspondences. Intuitively a "hard" negative correspondence is the correspondence with high value of similarity measure which is incorrect according to manual annotation. Taking into account the robustness of $Sim_{TaxME}$ and also complexity and scalability requirements we have selected $M'$ as the correspondences having $Sim_{TaxME}$ values in the 0.05-0.2 range. As from Figure 6.2, this allowed us to obtain 18063+4776=22836 candidate correspondences.

**Negative correspondence selection**

The negative correspondence selection step is devoted to the computation of $N_{T2}$. The process is organized as follows:

– *Step 1: Matching system selection.* The goal of this step is to select a set of matching systems whose results are exploited for constructing $N_{T2}$. The set of the selected systems should be heterogeneous, i.e., the selected systems should make mistakes on different sets of correspondences. Thus, the selected systems have to be the representatives of the different classes of the existing matching techniques. This also prevents $N_{T2}$ from being biased towards a particular class of matching solutions.

  As the result of this step, we have selected three different matching systems, namely COMA [Do and Rahm, 2002], Similarity Flooding (SF) [Melnik *et al.*, 2002] and S-Match (SM) [Giunchiglia *et al.*, 2004], see [Avesani *et al.*, 2005] for reasons of this choice.

– *Step 2: Computation of negative correspondences.* The goal of this step is to compute $N_{T2}$ exploiting the results obtained by running the selected matching systems on $M'$. In particular, $N_{T2}$ is computed from FP as $N_{T2} = \bigcup_i FP_i$, where $FP_i$ stands for the FP produced

by running the $i-th$ matching system on $M'$. The result of this exercise is depicted in Figure 6.5, where the gray area stands for $FP_i$. This construction schema ensures that $N_{T2}$ will be hard for all existing systems and discriminative given that the set of matching systems evaluated on $M'$ is representative and heterogeneous. An implicit constraint is that the number of FPs produced by each of the systems should be comparable. This prevents the existence of a bias towards a particular class of matching solutions. Notice that the computation of FP requires the human annotation of the systems results.

As the result of this step, we have executed COMA, SF and S-Match on $M'$. We also have manually evaluated the correspondences found by the systems and selected the FP from them. Notice that we have not distinguished among different semantic relations while evaluating the matching quality. Therefore, for example, the correspondence $A \sqsubseteq B$ produced by S-Match and $A_1 \equiv B_1$ produced by COMA have been considered as TP if $A \equiv B$ and $A_1 \sqsubseteq B_1$ are TP according to the human judgment. Finally, we have computed $N_{T2}$ as the union of the FPs produced by the matching systems.

Table 6.2 provides a quantitative description of the content of $N_{T2}$ and of the effort needed to build it. As from the first row of Table 6.2 the total number of annotated correspondences was 2553+2163+2151=6867. Notice that this is 6 orders of magnitude lower than the number of correspondences to be considered in the case of complete reference alignment. Notice also that the number of correspondences per system is very balanced, as required.

Table 6.2: Total number of correspondences and number of FP computed by COMA, SF and S-Match on $M'$.

| | COMA | SF | SM |
|---|---|---|---|
| Found (S) | 2553 | 2163 | 2151 |
| Incorrect (FP) | 870 | 776 | 781 |

Figure 6.6 shows how the FPs produced by the systems are partitioned. In particular, there are no FPs found by SM, COMA and SF, or even by SM and COMA together. There are the small intersections between the FPs produced by SM and by SF (0.1%) or by COMA and by SF (2.3 %). These results justify our assumption that all 3 systems belong to different classes.

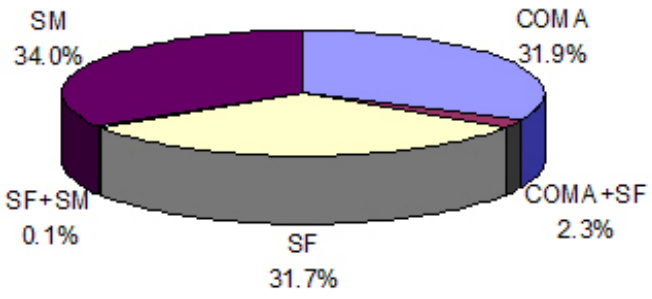



Figure 6.6: Partitioning of the FPs computed by COMA, SF and S-Match on $M'$.

The final result is that $N_{T2}$ consists of 2374 correspondences. Notice that the size of $N_{T2}$ is not equal to the sum of the FPs reported in the second row of Table 6.2 since, as from Figure 6.6,

there is some intersection among these sets. The union of $N_{T2}$ with $TaxME$ has allowed us to compute a reference alignment $TaxMe_2$, in turn, allowing for the evaluation of both recall and precision, of 2265+2374=4639 correspondences.

### 6.1.3 Test set summary

The data set exploited in the directory matching task was constructed from Google, Yahoo and Looksmart web directories following the methodology described in the previous subsections. The dataset is presented as taxonomies where the nodes of the web directories are modeled as classes and classification relation connecting the nodes is modeled as rdfs:subClassOf relation.

There were proposed three representations of this test case:

- one matching task between two taxonomies of $10^3$ categories (full test set),
- one matching task between two taxonomies of $10^2$ categories (10% test set), and
- 4639 matching tasks between two paths of around 10 categories (unit test sets).

The first data set incorporates the matching tasks involved in the unit tests which also correspond to the reference set. The second data set guarantees to contain 10% of these unit tests.

The reference alignment is composed of two parts:

- Representative subset of complete reference alignment ($P \subseteq R$). It contains the positive correspondences, i.e., the correspondences that hold for the matching task.
- Representative subset of negative correspondences ($N \subseteq \bar{R}$), i.e., the correspondences that do not hold for the matching task.

The reference alignment is composed of 2265 positive and 2374 negative correspondences. Therefore, the matching unit test set corresponds to 2265+2374=4639 tasks of finding a relation holding between paths in the web directories modeled as subclass hierarchies.

### 6.1.4 Results

Approximate precision, recall and F-measure of the systems on the directory dataset are presented in Figure 6.7, 6.8 and 6.9, respectively. Given an alignment $A$ and the set $P$ and $N$ of positive and negative correspondences, approximate precision and recall are computed as follows:

$$AP(A,P,N) = \frac{|P \cap A|}{|P \cap A| + |N \cap A|} \qquad AR(A,P) = \frac{|P \cap A|}{|P|}$$

These formula, especially that of $AP$, generalize precision and recall by not taking the whole set of valid correspondences as reference alignment. They are called approximate precision and recall because when $P = R$ and $N = \bar{R}$ ($R$ is the reference alignment), they correspond to precision and recall. How this is an accurate approximation of precision and recall heavily depends (as from the previous sections) on the choice of $P$ and $N$.

Similarly to OAEI-2005, seven matching systems were evaluated on the dataset. However, only one of them (Falcon) participated in both evaluations. In OAEI-2006, the systems in general demonstrated better results than in OAEI-2005. The average approximate recall of the systems

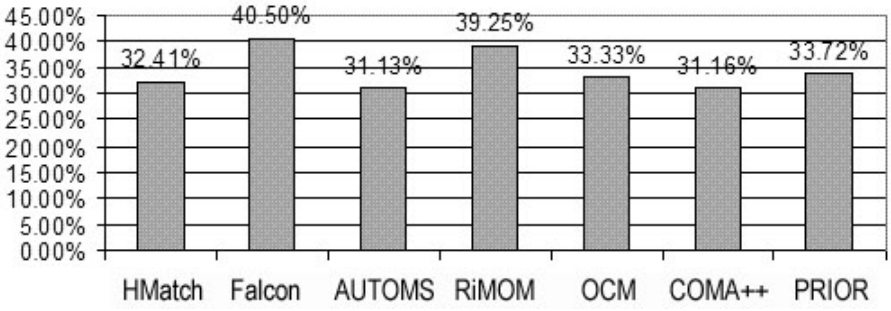



Figure 6.7: Approximate precision for web directories matching task.

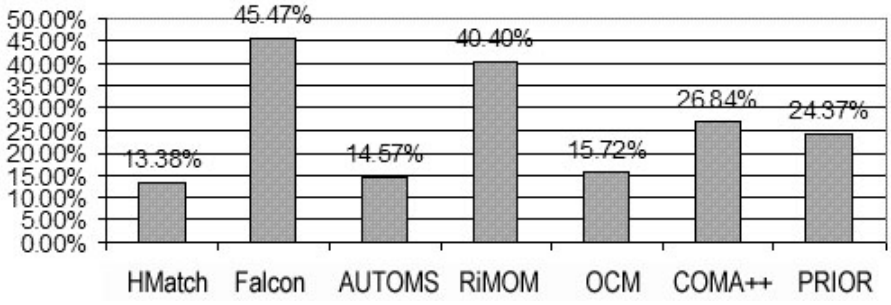



Figure 6.8: Approximate recall for web directories matching task.

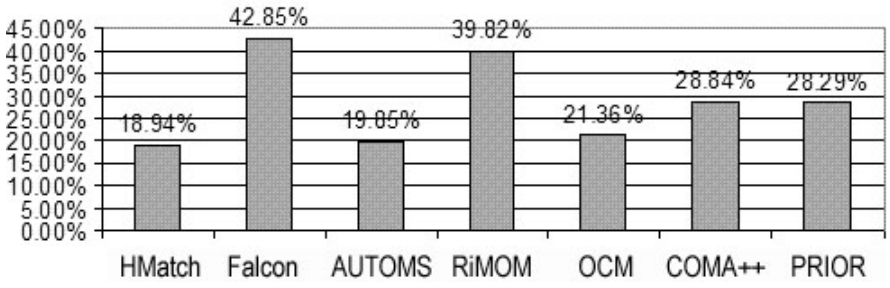



Figure 6.9: Approximate F-measure for web directories matching task.

increased from 22.23% to 25.82%. The highest approximate recall (45.47%) was demonstrated by the Falcon system what is almost a 50% increase with respect to its result of 2005 (31.17%).

Despite this progress the dataset remains difficult for the matching systems. The maximum and average values for approximate precision (40.5% and 34.5%), approximate recall (45.47% and 25.82%) and approximate F-measure (42.85% and 28,56%) are significantly lower than corresponding real values in benchmark tests for example.

Partition of positive and negative correspondences according to the systems results are presented in Figure 6.10 and Figure 6.11.

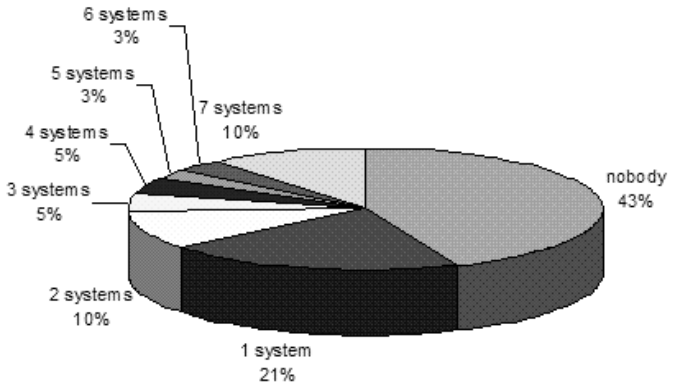



Figure 6.10: Partition of the systems results on positive correspondences.

Figure 6.10 and Figure 6.11 show that 43% of positive correspondences have not been found by any of the systems. At the same time 22% of negative correspondences were found by all the matching systems, i.e., all the matching systems mistakenly returned them as positive ones. Moreover, only 10% of positive correspondences were found by all the matching systems.

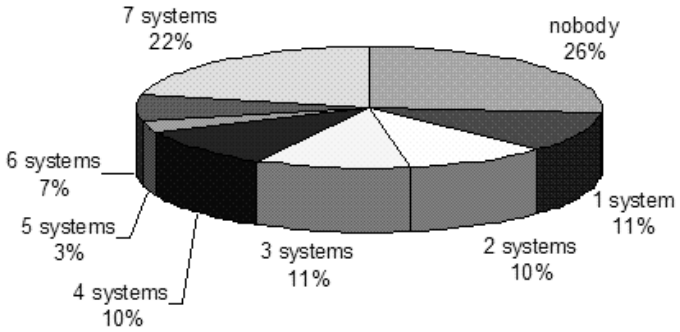



Figure 6.11: Partition of the systems results on negative correspondences.

### 6.1.5 Discussion

Six out of seven systems that participated in the evaluation presented their results only for one of the dataset representations, namely for the representation composed of 4639 node matching tasks. Only one system (HMatch) presented the results also for the other representations. Since, the other tasks were proposed in order to test the scalability of the approaches, this can be interpreted as a sign of poor scalability of the systems participating in the evaluation.

Blind evaluation offered for some of the systems a possibility to improve their final results after preliminary result disclosure. For example, the final results of the COMA++ and PRIOR matching systems were slightly lower than their preliminary results. The final F-measure of COMA++ dropped from 32.56% to 28.84% while F-measure of PRIOR dropped from 28.32% to 28.29%.

## 6.2 The food test case

The food test case aims at providing a realistic task for ontology matchers in the domain of thesauri. First, we introduce the test case (§6.2.1). Then, we outline the evaluation procedure (§6.2.2) and present test case results for the participated systems (§6.2.3). Finally, we discuss general issues that limit matching systems to produce better results on this test case (§6.2.4).

### 6.2.1 Test set

The thesauri used for this task are the Food and Agriculture Organization of the United Nations[2] (FAO) AGROVOC thesaurus[3] and the United States National Agricultural Library (NAL) Agricultural Thesaurus (NALT)[4]. Both thesauri were supplied unaltered to the participants in their native SKOS format[5]. An OWL-Lite translation made by Wei Hu was also supplied to the participants.

**AGROVOC** This thesaurus (version: May 2006) consists of 28.174 descriptor terms (i.e., preferred terms) and 10.028 non-descriptor terms (i.e., alternative terms). This version of AGROVOC is multilingual. It is available in ten languages, including English, French, Spanish, Arabic, Chinese, Portugese, Czech, Japanese, Thai, and Slovak. The AGROVOC thesaurus is used to index a multitude of data sources all over the world, one of which is the AGRIS/CARIS[6] literature reference database. A fragment of AGROVOC is shown in Figure 6.12 on the left side.

**NALT** This thesaurus (version: 2006) consists of 41.577 descriptor terms and 24.525 non-descriptor terms. It is monolingual and is available in English. The NALT thesaurus is used to index, amongst others, the AGRICOLA[7] literature reference database of the USDA[8], various data sources of the Agriculture Network Information Center[9] (AgNIC) and the Food Safety Re-

[2] http://www.fao.org
[3] http://www.fao.org/agrovoc
[4] http://agclass.nal.usda.gov/agt/agt.shtml
[5] http://www.few.vu.nl/~wrvhage/oaei2006
[6] http://www.fao.org/agris
[7] http://agricola.nal.usda.gov
[8] http://www.usda.gov/wps/portal/usdahome
[9] http://www.agnic.org

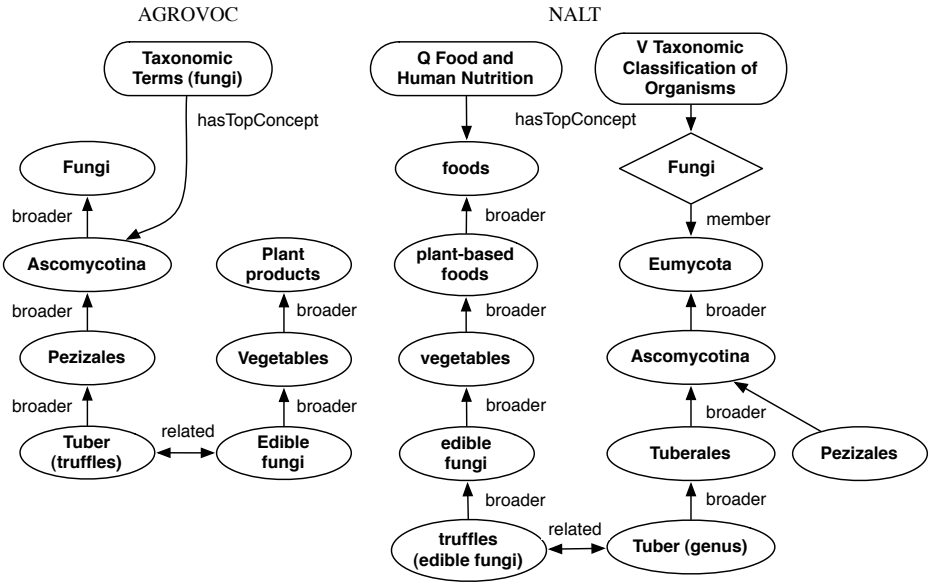



Figure 6.12: The concept of truffles in AGROVOC and NALT.

search Information Office[10] (FSRIO). A fragment of NALT is shown in Figure 6.12 on the right side.

**SKOS** We chose to use SKOS[11] because it closely follows "traditional" thesauri formats, such as the iso2709 format. Let us introduce some notation. We depict skos:Concepts as an oval filled with the skos:prefLabel text. In cases where we explicitly want to show skos:altLabel and skos:prefLabel we depict the skos:Concept as an oval filled with its URI, connected to boxes that represent its various labels. skos:Collections are depicted by diamond shapes and skos:ConceptShemes are shown as boxes with round sides.

**SKOS mapping vocabulary** For the correspondences we use the SKOS mapping vocabulary[12]. The participants were allowed to use the following relations: skosmap:narrowMatch, skosmap:exactMatch, and skosmap:broadMatch. The other relations and boolean combinators (skosmap:minorMatch, skosmap:majorMatch, skosmap:AND, skosmap:OR, skosmap:NOT) of the SKOS mapping vocabulary were not used in the evaluation.

### 6.2.2 Evaluation procedure

**Precision** In order to assess precision of the results within the time span of OAEI-2006 and given a limited number of assessors and man-hours we performed a sample evaluation. The samples were chosen to be representative of the type of topics covered by the thesauri and to be impartial to each participant and impartial to how much consensus amongst the participants there

[10] http://fsrio.nal.usda.gov
[11] http://www.w3.org/TR/swbp-skos-core-spec
[12] http://www.w3.org/2004/02/skos/mapping/spec

was about each correspondence (i.e., the "hardness" or "complexity" of the correspondence). This was accomplished with stratification.

*Stratification of the matching results.* We distinguished three categories of topics in the thesauri. Each of these required a different level of domain knowledge of the assessors: *taxonomical* concepts (plants, animals, bacteria, etc.), *biological and chemical* terms (structure formulas, terms from generics, etc.), and *miscellaneous*, the remaining concepts (geography, agricultural processes, etc.). The correspondences were partitioned into strata corresponding to these topics. Under the authority of taxonomists at the USDA the taxonomical stratum was assessed completely using the strict rules that apply to the naming scheme of a taxonomy. These rules state that two taxonomical concepts are considered exact matches if and only if the skos:prefLabel of one concept is literally the same as either the skos:prefLabel or the skos:altLabel of the other concept. From the latter two strata we took samples that were assessed by two groups: $(i)$ a group of domain experts from the USDA and the FAO, and $(ii)$ a group of computer scientists at the EKAW conference[13]. The size of the strata and the size of the assessed samples are shown in Table 6.3. The samples were selected in such a way that each participant has an equal share in the sample. Also, there is an equal number of correspondences in the sample for each "agreement level", where "agreement level" is defined as the set of correspondences returned by $n$ systems. This means that from each of the columns in Table 6.4 (p.47) marked 1–5, and from each row that represents a system, we took a sample of equal size[14].

Table 6.3: Sizes of the strata and of the samples from those strata that were assessed to evaluate precision.

| Stratum topic | Stratum size ($N_h$) | Sample size ($n_h$) |
|---|---|---|
| Taxonomical | 18.399 | 18.399 |
| Bio/chem | 2.403 | 250 |
| Miscellaneous | 10.310 | 650 |
| All topics | 31.112 | |

*Assessment tool for precision.* For the assessment of precision we used a tool developed by TNO[15], see Figure 6.13 for a screen shot. This tool reads a set of correspondences in the common format for alignments and outputs a web form that is used by judges to assess the correspondences. The results of the form are submitted to the organizer of the food task. The assessment process of a correspondence follows three steps (see also Figure 6.13).

1. The judge decides if the relation specified above the arrow between the two (green) boxes holds between the two concepts displayed in bold face. If the relation holds (s)he skips step 2 and proceeds to step 3 directly, otherwise step 2 is performed.

2. The judge tries to specify an alternative relation, either by changing the relation type, or the concepts. If possible (s)he selects "exactMatch" and specifies the proper concepts be-

[13] `http://ekaw.vse.cz/`

[14] The samples can be downloaded from `http://www.few.vu.nl/~wrvhage/oaei2006/gold_standard`.

[15] `http://www.tno.nl/index.cfm?Taal=2`

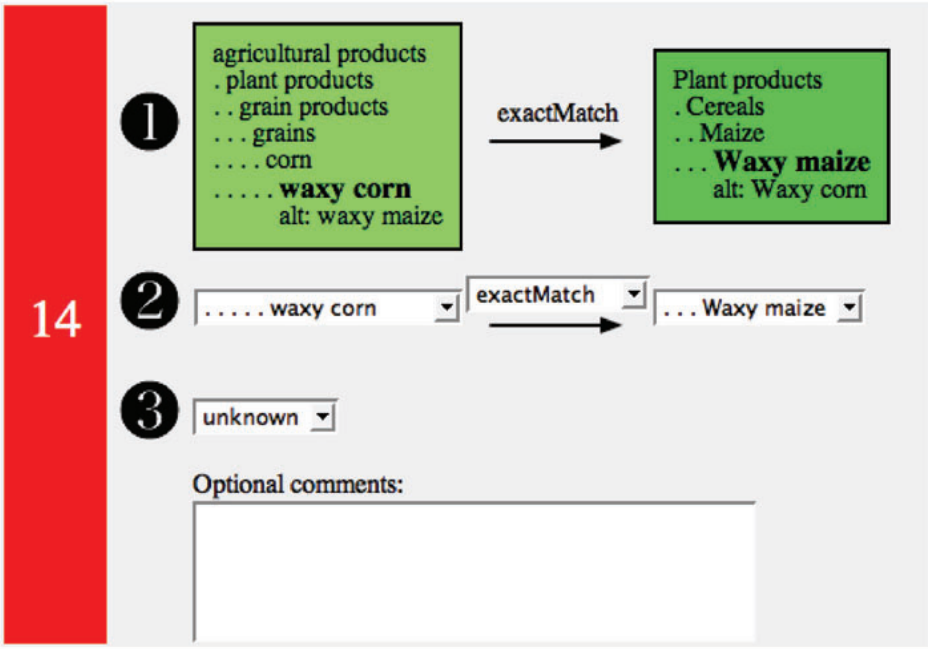



Figure 6.13: Screen shot of the assessment tool used to evaluate precision. It shows the 14th correspondence relation from a sample set of correspondences: nalt:'waxy corn' skosmap:exactMatch agrovoc:'Waxy maize'.

tween which the "exactMatch" relation holds. Otherwise, (s)he selects "broadMatch" or "narrowMatch" and specifies the proper concepts between which that relation holds.

3. The judge changes the default value ("unknown") of the assessment into either "true" or "false". If the relation holds and step 2 was skipped, then (s)he selects "true". If the relation does not hold, but if (s)he successfully selected an alternative relation (at step 2) that does hold, (s)he also selects "true". If the relation does not hold and no correct alternative could be found at step 2, (s)he selects "false".

Finally, if the judge wishes to document the decision taken (s)he may fill in the box called "Optional comments" at the bottom of the assessment form.

*Inter-judge agreement.* The agreement between the group of domain experts and the group of computer scientists was 72%. The computer scientists were less likely to judge a correspondence to be correct than the domain experts. They judged 78% of the sample correspondences to be "true", while the domain experts judged 85% to be "true". The effect of this is that the estimated precision based on the computer scientists' work, as compared to that based on the domain experts' work, will be relatively low. We have no data on the inter-judge agreement within these groups.

*Significance testing.* As a significance test on precision scores of the systems we used the Bernoulli distribution. Precision of system $A$ (denoted by $P_A$) can be considered to be significantly greater than precision of system $B$ (denoted by $P_B$), if their estimated values, $\hat{P}_A$ and $\hat{P}_B$

are far enough apart. In general, based on the Bernoulli distribution, this is the case when the following formula holds:

$$|\hat{P}_A - \hat{P}_B| > \frac{2}{n}\sqrt{\left(\hat{P}_A(1-\hat{P}_A)\right)^2 + \left(\hat{P}_B(1-\hat{P}_B)\right)^2}$$

This significance test was used to determine which of the systems performs best for each of the topic strata.

We calculated these estimations based on three strata. This allows us to distinguish smaller differences in the results than by simple random sampling by combining the results of the strata. We denote the estimated precision of system $A$ on stratum $h$ as $\hat{P}_{A,h}$, the size of stratum $h$ as $N_h$, and the size of the sample from stratum $h$ as $n_h$ (see also Table 6.3). We can conclude that system $A$ performs significantly better than system $B$ when the following formula holds:

$$|\hat{P}_A - \hat{P}_B| > \frac{2}{N}\sqrt{\left(\sum_{h=1}^{L}\hat{P}_{A,h}(1-\hat{P}_{A,h})(\frac{N_h}{n_h}-1)\right)^2 + \left(\sum_{h=1}^{L}\hat{P}_{B,h}(1-\hat{P}_{B,h})(\frac{N_h}{n_h}-1)\right)^2}$$

This significance test was used to determine which of the systems performs best for all the strata.

**Recall** Assessing the recall within the time span of OAEI-2006 was not feasible, so we estimated it based on four sample sub-hierarchies of the thesauri: $(i)$ all oak trees (everything under the concept representing the Quercus genus), $(ii)$ all rodents (everything under Rodentia), $(iii)$ Geographical concepts of Europe, $(iv)$ everything under the NALT concept animal health and all AGROVOC concepts that have correspondences to these concepts and their sub-concepts. These four samples respectively have sizes 41, 42, 74, and 34. Around 30% of the correspondences were broadMatch and narrowMatch, the rest was exactMatch. Currently, the sample for recall is extended for OAEI-2007[16].

*Assessment tool for recall.* To create the samples we used the AIDA Thesaurus Browser. That is a SKOS browser that supports parallel browsing of two thesauri, concept search, correspondence traversal, and the addition, change and removal of correspondences of the SKOS mapping vocabulary, see Figure 6.14 for a screen shot. This tool was developed by TNO[17] in the context of the VL-e project[18].

A preliminary version of the recall samples were made at the Vrije Universiteit Amsterdam and was verified and extended by domain experts at the the FAO and USDA to produce the final recall samples[19]. The guidelines used to make the correspondence were the following:

1. Starting from AGROVOC, identify a skosmap:exactMatch for every concept in the sample. If this is impossible, try to find a skosmap:narrowMatch or skosmap:broadMatch. Always choose the broader concept of these correspondences as narrow as possible and the narrower concept as broad as possible.

[16] http://oaei.ontologymatching.org/2007

[17] http://www.tno.nl/index.cfm?Taal=2

[18] http://www.vl-e.nl/frame_home.htm

[19] The samples can be downloaded from http://www.few.vu.nl/~wrvhage/oaei2006/gold_standard.


Figure 6.14: Screen shot of the AIDA Thesaurus Browser, which was used to create correspondence samples for the evaluation of recall.

2. Investigate the surrounding concepts of the target concept in NALT. If the surrounding concepts are still on the topic for the sample, try to match this concept "back" to AGROVOC using skosmap:exactMatch. If this is impossible, try to find a skosmap:narrowMatch or skosmap:broadMatch.

*Significance tests.* We exploited the same significance tests as used for precision.

### 6.2.3 Results

Five participants took part in the OAEI-2006 food test case: South East University (Falcon) [Hu *et al.*, 2006], University of Pittsburgh (PRIOR) [Mao and Peng, 2006], Tsinghua University (RiMOM) [Li *et al.*, 2006], University of Leipzig (COMA++) [Maßmann *et al.*, 2006], and Universitá degli Studi di Milano (HMatch) [Castano *et al.*, 2006]. Each team provided between 10.000 and 20.000 correspondences. This amounted to 31.112 unique correspondences in total. All of these correspondences were of type skosmap:exactMatch. None of the systems was able to detect skosmap:broadMatch or skosmap:narrowMatch correspondences. There was a "high agreement" between the best three systems, RiMOM, Falcon, and HMatch, see Table 6.4. This table also indicates that there is a relatively large set of "easy" correspondences that are recognized by all systems.

Table 6.4: Distribution of the systems' results. It shows the number of correspondences returned by each system and how many correspondences are returned by $n$ out of 5 systems.

| | # Correspondences | # Correspondences shared amongst $n$ systems | | | | |
|---|---|---|---|---|---|---|
| System | returned | 1 | 2 | 3 | 4 | 5 |
| RiMOM | 13,975 | 868 | 1,042 | 2,121 | 4,389 | 5,555 |
| Falcon | 13,009 | 642 | 419 | 1,939 | 4,400 | 5,555 |
| PRIOR | 11,511 | 1,543 | 1,106 | 676 | 2,631 | 5,555 |
| COMA++ | 15,496 | 11,610 | 1,636 | 629 | 2,028 | 5,555 |
| HMatch | 20,001 | 7,000 | 981 | 2,045 | 4,420 | 5,555 |
| **all systems** | 31,112 | 21,663 | 2,592 | 2,470 | 4,467 | 5,555 |

The taxonomical parts of the thesauri accounted for the largest part of the correspondences, 59% of all submitted correspondences. The more difficult correspondences that required lexical normalization, such as structure formulas, and relations that required background knowledge, such as many of the relations in the miscellaneous domain, accounted for a smaller part of the correspondences. This caused systems that did well at the taxonomical correspondences to have a great advantage over the other systems. The Falcon system performed consistently best at the largest two strata, taxonomical and miscellaneous, and thus achieved high precision, see Table 6.5.

Table 6.5: Precision results based on sample evaluation. "⋆" indicates the best results.

| Precision for | RiMOM | Falcon | PRIOR | COMA++ | HMatch |
|---|---|---|---|---|---|
| taxonomical | 82% | 83%⋆ | 68% | 43% | 48% |
| bio/chem | 85%⋆ | 80% | 81% | 76% | 83% |
| miscellaneous | 78% | 83%⋆ | 74% | 70% | 80% |
| **all topics** | **81%** | **83%⋆** | **71%** | **54%** | **61%** |

All systems only returned skosmap:exactMatch correspondences. This means that the recall of all systems was limited to 71%. For example, RiMOM achieved 50%, while it could achieve 71%, see Table 6.6. The RiMOM system managed to discover more good results than the Falcon system on the four small sample recall bases (at the expense of precision). Notice that the recall was assessed on a small set of examples, therefore we can only draw certain conclusions based on the precision results. Table 6.7 summarizes tentative F-measure results.

Table 6.6: Tentative estimation of recall based on sample evaluation.

| Recall for | RiMOM | Falcon | PRIOR | COMA++ | HMatch |
|---|---|---|---|---|---|
| **all relations** | **50%** | **46%** | **45%** | **23%** | **46%** |
| only exactMatch | 71% | 65% | 64% | 33% | 65% |

Table 6.7: Tentative estimation of F-measure based on sample evaluation.

| F-measure for | RiMOM | Falcon | PRIOR | COMA++ | HMatch |
|---|---|---|---|---|---|
| **all rel. & top.** | **62%** | **59%** | **55%** | **33%** | **53%** |

A potential user of ontology matching systems does not necessarily have to limit himself/herself to only one matching system. Simple ensemble methods such as majority voting can

improve precision. To give an impression of this we list the average precision of the different "agreement levels" in Table 6.8. For agreement level 4 and 5 (i.e., the correspondences that were returned by 4 out of 5 systems or all of the systems) precision is significantly higher than for the best system by itself, Falcon in this case. Nearly all of the 5,555 correspondences found in this way are correct. Obviously, these are the "easy" correspondences. Whether they are useful or not useful depends on the application of the correspondences and remains a topic for future research.

Table 6.8: Consensus: average precision of the correspondences returned by a number of systems.

| correspondence found by # systems | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| average precision | 6% | 35% | 67% | 86% | 99% |
| # correspondences | 21.663 | 2.592 | 2.470 | 4.467 | 5.555 |

### 6.2.4 Discussion

Let us discuss a number of issues that limit the performance of matching systems. Some of these issues are technical and easy to solve, while others are more fundamental and represent challenges to be tackled in a longer term.

**Inappropriate "spelling correction".** Incorrect matches such as nalt:patients skosmap:exactMatch agrovoc:Patents and nalt:aesthetics skosmap:exactMatch agrovoc:anaesthetics are caused by inappropriate spelling correction. In general, spelling tolerance in thesauri is not very effective, but if it is applied nonetheless it should only be applied when there is no exact literal match. For example, there is a concept representing "patients" in both thesauri. Recognizing this should trigger a matching system to refrain from suggesting a correspondence to "patents".

**Labels following naming schemes.** Labels often follow naming schemes. Real-life ontologies often use more than one naming scheme. Both AGROVOC and NALT have a large section on biology. The labels of these concepts follow the Linnaeic system of species names. Concepts in other sections of the thesauri (e.g., the sections on geography) follow different schemes. It is vital that lexical matchers recognize that different naming schemes require different matching rules. One of the common matching rules is "suffix" matching. For instance, "lime stone" and "sand stone" can be found similar using the suffix matcher. In fact, they are both kinds of "stone". However, two terms from the Linnaeic system that end in the same word, such as "Quercus pubescens" (a tree) and "Ibacus pubescens" (a crustacean) are completely dissimilar. Failing to recognize that the Linnaeic system needs also prefix matching and not only suffix matching can lead to many wrong correspondences. The bold arrow in Figure 6.15 indicates this wrong correspondence.

**"Use" and "use for" modeled with skos:altLabel.** When use is modeled using skos:altLabel the difference between synonyms, obsolete terms, and acknowledgment of lack of detail disappears. For example, in Figure 6.16 AGROVOC does not include detailed descriptors for the concept nalt:Sigmodon. In fact, a few levels of taxonomical distinctions are left out. The skos:altLabel "Sigmodon" is added to indicate this omission. It indicates that users that desire to refer to sigmodons should use the agrovoc:c_6633 concept, that symbolizes all rodents. A matcher without

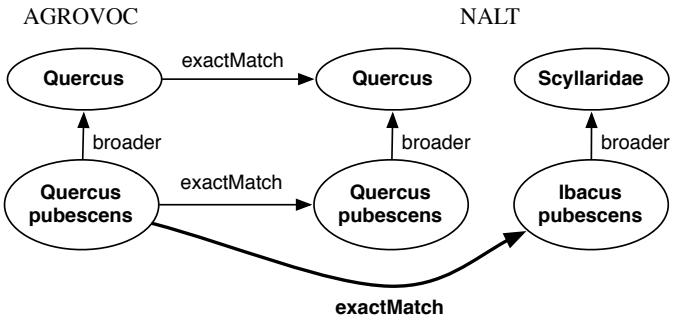



Figure 6.15: Failing to recognize the naming scheme can lead to wrong correspondences.

prior knowledge about this modeling decision cannot distinguish this from synonymy represented with skos:altLabel. This will cause most systems to conclude that there is a skosmap:exactMatch between agrovoc:c_6633 and nalt:Sigmodon, while the proper relation between these concepts is a skosmap:narrowMatch.

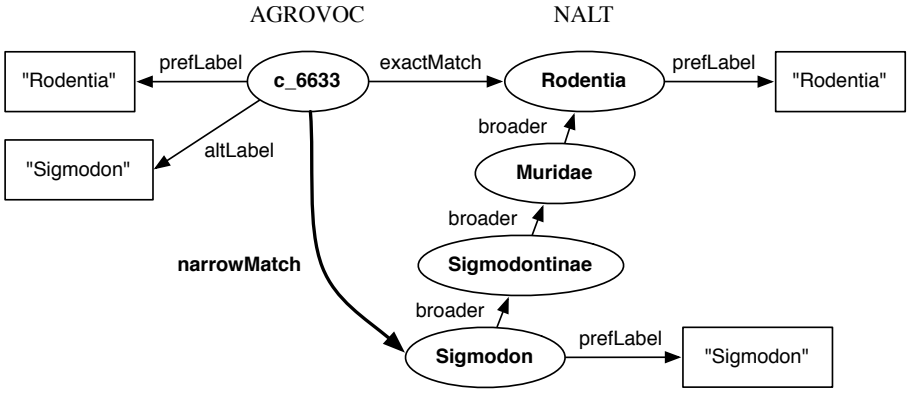



Figure 6.16: use modeled with skos:altLabel in AGROVOC.

**Colloquial names and scientific names.** A delicate problem is that of colloquial versus scientific names for the same species. Take the example illustrated in Figure 6.17 of gerbils with the scientific name "Gerbilinae". In NALT, the two types of names each have their own hierarchy. In AGROVOC these two are combined, because they both refer to the same actual species. This leads us to believe that in this case there should be skosmap:exactMatch correspondences to both hierarchies in NALT. We created the evaluation samples for recall based on this assumption. Whether it is the proper treatment depends on the application of the correspondences. The different names can symbolize different views of the concepts or refer to the same extension. A problem similar to this example occurs with colloquial names, scientific names, and structure formula's of chemicals.

**Clashing senses.** Let us consider "Ireland" and the "British Isles". The British Isles can be partitioned in two ways: the Irish Republic and the United Kingdom; or Ireland and the other islands of the British Isles, which all belong to the United Kingdom. If the distinction is made between Ireland and the United Kingdom, the most obvious interpretation is the former partition.

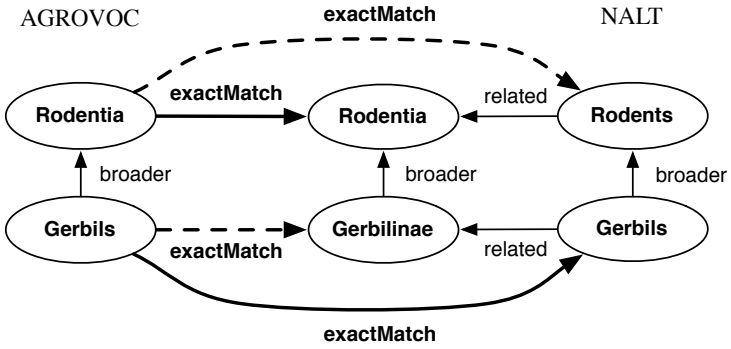



Figure 6.17: Separate hierarchies for colloquial names and scientific names.

This is due to the fact that most people assume sibling concepts to be disjoint. The lack of a broader relation between agrovoc:Northern Ireland and agrovoc:Ireland supports this. Another common assumption is that narrower concepts are strictly narrower than (i.e., not equivalent to) their parents. This means that the existence of the concept nalt:Irish Republic makes people assume that nalt:Ireland refers to the entire island. The narrower concept Northern Ireland confirms this. This means that agrovoc:Ireland should be equivalent to nalt:Irish Republic. In this case, this problem could be solved by adding OWL statements that proclaim siblings to be disjoint and broader concept to be not equivalent to narrower concepts. This kind of approach, however, is likely to cause more harm than good in the entire thesaurus. Thesaurus concepts are inherently vague and such a strict interpretation often causes unintentional inconsistencies. A technique that uses the added axioms as heuristics might be more suitable.

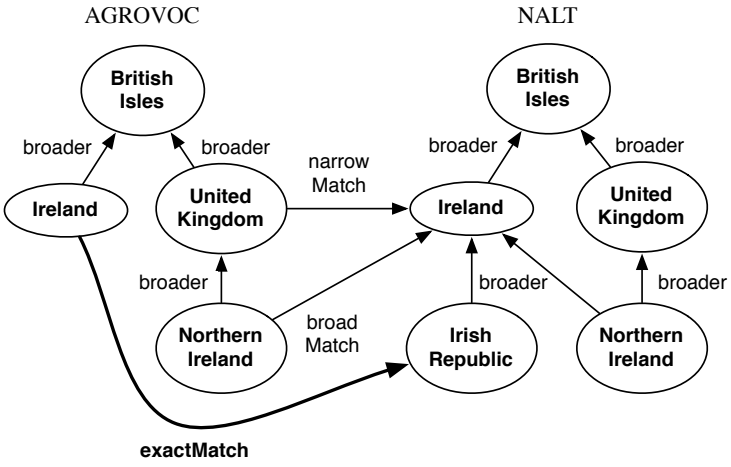



Figure 6.18: Concepts representing different senses of a term.

**No direct evidence in the thesauri for a correct correspondence.** In many cases it is simply impossible to find certain correspondences without resorting to external knowledge sources, such as a third ontology, concrete domain reasoning, text mining, or traditional knowledge acquisition. For example, in Figure 6.19 Western Europe is a named geographical region, but the skos:broader

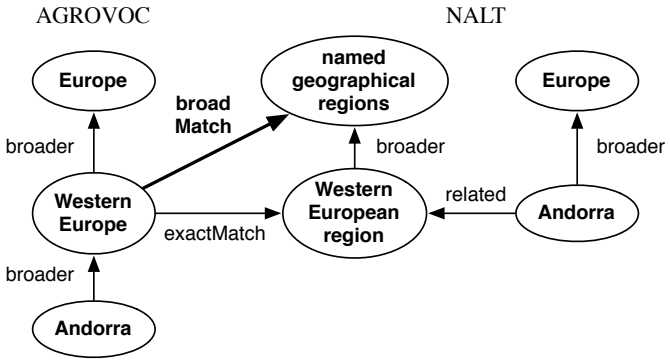



Figure 6.19: The is no evidence in the thesauri for this skosmap:broadMatch.

relation between nalt:Western European region and nalt:named geographical regions alone is not enough evidence to suggest this. For example, AGROVOC contains no concepts that are lexically similar to the latter NALT concept.

# Chapter 7

# The conference track and consensus workshop

The conference test set introduces matching several ontologies together as well as a consensus workshop aiming at studying the elaboration of consensus when establishing reference alignments. First we present the underlying test set (§7.1). Then, we provide analysis of the evaluation results via manual labeling (§7.2), logical reasoning (§7.3), the consensus building workshop (§7.4), and, ultimately, by means of pattern-aware data mining (§7.5). Finally, we outline some related works and summarize the major findings of this chapter (§7.6).

The material of this chapter has been published in [Šváb *et al.*, 2007].

## 7.1 The *OntoFarm* collection

The motivation for initiating the creation of the *OntoFarm*[1] collection (in Spring 2005) was the lack of 'manageable' material for testing ontology engineering (especially, matching) techniques. As underlying domain, we chose that of *conference organization*—among other, for the following reasons:

- Most ontology engineers are academics who themselves submit and review papers and organize conferences: there is zero overhead of acquiring the domain expertise.
- Organization of a conference shares some aspects with (heavier-weighted) business activities: access restrictions, hard vs. soft constraints, temporal dependencies among events, evolution of the meaning of concepts in time, etc. There is also a wide range of supporting software tools covering various aspects of conference organization. Their domain assumptions can also be captured using ontologies (specific for each system). The process of matching the requirements of conference organizers with the capacities of such tools is analogous with that of matching the requirements of a business with the capacities of an off-the-shelf enterprise information system.
- In many cases, even the underlying instance data could be obtained, since legal restrictions are typically not as strong as, e.g., in business or medicine.

[1] `http://nb.vse.cz/~svabo/oaei2006`

In OAEI-2006 the conference test case consisted of ten OWL-DL ontologies, typically of the size of 30–80 concepts and 30–60 properties, see Table 7.1. Most ontologies were equipped with DL axioms of various kinds. Six among the ontologies were derived from different conference organization support tools (for the review process, registration, etc.), using their documentation and experiments with installed tools ('tool' ontologies); two of them are based on the experience of people with personal participation in conference organization ('insider' ontologies); finally, two of them are merely based on the content of web pages of concrete conferences ('web' ontologies). The ontology designers (partly students of a course on Knowledge Modelling and partly experienced knowledge engineers) did not interact among themselves. This should guarantee that, although the ontologies themselves are to some degree artificial (their development not being driven by an application need), their heterogeneity was introduced in a natural way, that possibly simulating the heterogeneity of ontologies developed by different communities in the real world.

Table 7.1: Characteristics of the *OntoFarm* ontologies.

| Name | Type | Number of Classes | Number of Properties | DL expressivity |
|---|---|---|---|---|
| EKAW | Insider | 77 | 33 | $\mathcal{SHIN}(\mathcal{D})$ |
| SOFSEM | Insider | 60 | 64 | $\mathcal{ALCHIF}(\mathcal{D})$ |
| SIGKDD | Web | 49 | 28 | $\mathcal{ELI}(\mathcal{D})$ |
| IASTED | Web | 140 | 41 | $\mathcal{ALCIF}(\mathcal{D})$ |
| Confious | Tool | 57 | 57 | $\mathcal{SHIN}(\mathcal{D})$ |
| PCS | Tool | 23 | 38 | $\mathcal{ELUIF}(\mathcal{D})$ |
| OpenConf | Tool | 62 | 45 | $\mathcal{ALCIO}(\mathcal{D})$ |
| ConfTool | Tool | 38 | 36 | $\mathcal{SIF}(\mathcal{D})$ |
| CRS | Tool | 14 | 17 | $\mathcal{ALCIF}(\mathcal{D})$ |
| CMT | Tool | 36 | 59 | $\mathcal{ALCIF}(\mathcal{D})$ |

## 7.2 Initial manual empirical evaluation

There were six participant groups to the OAEI-2006 conference track, with the following matching systems: AUTOMS, COMA++, OWL-CtxMatch, Falcon, HMatch and RiMOM. The alignments obtained were examined by the organizers, and each individual correspondence was assigned a label. Results from the initial evaluation phase consist in global statistics about the participants results, which more-or-less reflect their quality. Additional, finer-grained results were obtained at the consensus building workshop (§7.4).

The global statistics for each system amount to (among other):

- The distinction whether the correspondence is true/false or ranges between 0 and 1;
- Number of alignments;
- Number of individual correspondences labeled as correct vs. incorrect;
- Number of interesting correct correspondences, namely, those that were subjectively not so easy to identify at first sight (e.g., due to lack of string similarity);

– Number of correspondences that seemed to exhibit an interesting type of error (or problematic feature), specifically for: subsumption mistaken for equivalence, sibling concepts mistaken for equivalent ones, mutually inverse properties matched on each other, relation matched onto class;
– Precision and recall measures.

Additionally, some of the correspondences that were retained as worth discussing by both independent evaluators were then submitted to the consensus building workshop.

## 7.3 Empirical evaluation via logical reasoning

In addition to manual evaluation, we conducted an automatic analysis on a subset of the correspondences. Correspondences between class names in different ontologies were formalized in C-OWL [Bouquet *et al.*, 2003] and the DRAGO system [Serafini and Tamilin, 2005] was used to determine whether the correspondences created by a particular system cause logical inconsistencies in one of the matched ontologies. C-OWL was chosen as basis for the evaluation, as its semantics is tuned towards describing correspondences between ontologies of the same domain; it solves some problems that occur when standard OWL is used for this purpose. A more detailed description of the approach can be found in [Meilicke *et al.*, 2006]. The analysis was performed on six ontologies with four matching systems, namely Falcon, OWL-CTXmatch, COMA++ and HMatch.

Table 7.2 shows the results of the reasoning-based analysis. The first column lists the systems under consideration. The second column lists the number of correspondences produced by the given system that make the target ontology inconsistent, the third column lists the average number of inconsistent concepts per correspondence, and the fourth column shows the overall precision of the alignment. Note that the precision only refers to correspondences between class names and therefore naturally differs from the numbers at the result report page. The precision has been determined by a manual investigation of the correspondences by three independent people (different from those doing almost the same task for the sake of the result report page). In cases of a disagreement the correctness of a correspondence was decided by a majority vote. It however turned out that there was little disagreement with respect to the correctness of correspondence. For only about 3% of the correspondences the result had to be determined by vote.

The results of this evaluation are useful in two ways. First of all, we can see from the numbers that a low number of inconsistent alignments is an indicator for the quality of correspondences (we also see that the actual number of concepts that become unsatisfiable is less relevant). The

Table 7.2: Results of reasoning-based evaluation.

| System | Inconsistent correspondences | Avg. number of inconsistent concepts | Overall precision |
|---|---|---|---|
| Falcon | 4 | 1,5 | 89,7 % |
| OWL-CTXmatch | 6 | 9,6 | 85,67 % |
| COMA++ | 12 | 2,2 | 67,7 % |
| HMatch | 9 | 5,5 | 63,7 % |

second benefit of this evaluation is the fact that the information about inconsistent concepts and correspondences that caused these inconsistencies reveal obvious and also non-obvious errors in correspondences. Some examples of obviously incorrect correspondences produced by matching systems in the experiments are the following:

$$\begin{aligned} Document &= Topic \\ Decision &= Location \\ Reception &= Rejection \end{aligned}$$

The real benefit of this evaluation is its ability to find non-obvious errors in correspondences that can only be detected taking the position of the matched concepts in the concept hierarchy into account. In our experiments, we found a number of such errors. Examples include the following correspondences:

$$\begin{aligned} Regular\_Paper &= Regular \\ Reviewing\_event &= review \\ Main\_office &= Location \end{aligned}$$

In the case of the first correspondence, *Regular* actually denotes the regular participation fee as opposed to the early registration. The error in the second correspondence is caused by the fact that *Reviewing_event* represents the process of reviewing whereas review denotes the review document as such. The last correspondence is not correct, because the concept *Main_office* actually represents the main office as an organizational unit rather than a location. Such correspondences are candidates for a closer inspection in terms of a committee of experts that analyze the reason for the inconsistency and decide whether the problem is in the correspondence or in the ontologies.

## 7.4 Consensus building workshop

The idea of consensus building workshop was to discuss some interesting correspondences in detail. Such interesting correspondences are determined as a result of the manual and the automatic evaluation of the matching results, as shown above. In the case of the manual evaluation correspondences where the evaluators where in doubt or where they disagreed on the correctness of a correspondence are candidates for a consensus workshop. In the automatic evaluation, correspondences that have been shown to cause concepts in the matched ontologies to become inconsistent are such candidates, especially if the correspondences have been annotated as being correct in the manual evaluation. Often, a decision whether a correspondence is correct or not can be made quite easily in a committee of experts. In some cases, however, it turns out that deciding whether a correspondence is correct or not is far from being trivial. In particular, it turns out that sometimes a detailed analysis of the matched ontologies is necessary to come to a decision.

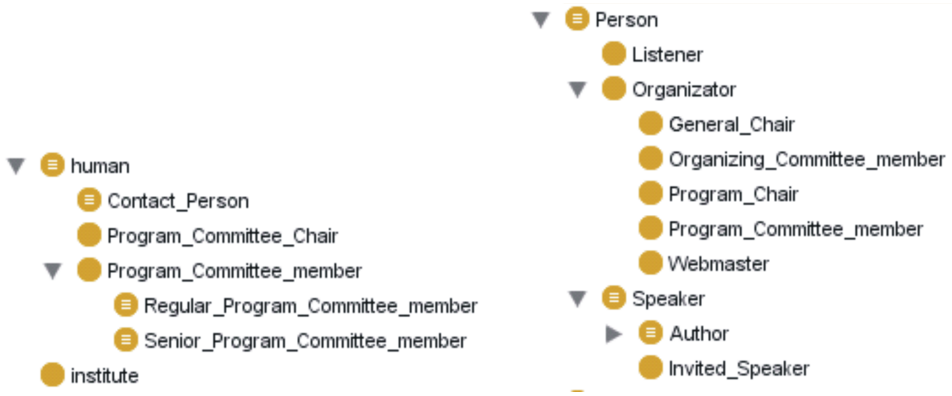
As far as arguments against and for individual correspondences are concerned, we experienced that *lexical* reasons of correspondence were first considered by the workshop participants. Then followed arguments with regard to the *context* of elements in question. This means consideration of certain neighborhood, subclasses and superclasses (in the case of properties, we can consider subproperties and superproperties). This can disclose different extensions of classes (especially

through their subclasses). Also, properties related to classes were considered. As a last resort, *axioms* (more complex restrictions) were taken into account if they were present.

### 7.4.1 Examples of correspondences discussed

In the following, we focus on examples that illustrate the kinds of arguments used in the discussion and the insights gained.

**Person vs. human:** At first sight the equivalence between the concepts person and human looks rather intuitive, it is however not obvious that the two concepts have the same intended meaning in different ontologies. First of all, the concept person can be interpreted in a legal context in which it also refers to organizations. Further, when we look at the hierarchies of the different ontologies, we see that the concepts have completely different sets of subconcepts depending on the scope of the ontology, see Figure 7.1.




(a) IASTED (b) SIGKDD

Figure 7.1: Subtrees rooted at the concepts *human* and *person*.

As we can see, the notion of a person in SIGKDD also contains subclasses not subsumed under human in IASTED (e.g., speakers). As it is clear, however that both ontologies cover the same domain, it was decided that in this case the two concepts actually have the same intended meaning even though they do not share all subclasses.

**PC_Member vs. member_PC:** The concepts *PC_member* and *member_PC* are another example of correspondences that seem to be trivially correct at first sight. In this case the question is whether the ontologies assume the same set of people to belong to the program committee. A look at the hierarchies reveals that the two ontologies use a different interpretation of the set of people belonging to the PC. In particular in one case the *PC_chair* is assumed to be a member of the committee, in the other case not, see Figure 7.2. This seems to imply that the notion of *PC_member* in EKAW is more general than that in *ConfTool*. However, this is only the case if we assume that the concepts *Chair_PC* und *PC_Chair* are equivalent. Another possible interpretation is that the concepts *PC_member* and *member_PC* are equivalent but *Chair_PC* and *PC_Chair* are different concepts, namely one denoting PC chairs that are members of the PC and the other denoting PC chairs that are not members of the PC. While both interpretations are possible, the majority of workshop participants favored the first interpretation where PC chairs are the same concepts.


(a) EKAW (b) ConfTool

Figure 7.2: Subtrees containing the concepts *PC_Member* and *Member_PC*.

### 7.4.2 Lessons learned

The discussions at the consensus workshop revealed a number of insights about the nature of ontology matching and limitations of existing systems that provide valuable input for the design of matching tools. In the following we summarize the three most important insights gained.

**Relevance of context.** Probably the most important insight of the consensus workshop was that in many cases it is not enough to look at the concept names to decide whether a correspondence is correct or not. In all of the examples above, the position of the concept in the hierarchy and in some cases also the scope of the complete ontology had to be taken into account. In some cases, a decision actually requires deep ontological arguments, for instance, to distinguish between a recommendation and the actual decision made on the basis of this recommendation. For existing matching tools this means that the use of lexical matching techniques and often even of local structure matching is not sufficient. Matchers rather have to take the complete ontology and its semantics or even background knowledge about basic ontological distinctions into account. This observation is also supported by the results of the reasoning-based evaluation where automatically created correspondences often turned out to cause inconsistencies in the ontologies.

**Semantic relations.** All of the systems participating in the evaluation were restricted to detecting equivalences between concepts or relations respectively. It turned out that this restriction is a frequent source of errors. Often ontologies contain concepts that are closely related but not exactly the same. In many cases one concept is actually a subclass of the other. Heuristics-based matching tools will often claim these concepts to be equivalent, because they have similar features and similar positions in the hierarchy. As a result, such correspondences often become inconsistent.

We believe that matching tools that are capable of computing subsumption rather than equivalence relations are able to produce more correct and suitable correspondences.

**Alternative interpretations.** The example of *PC_member* illustrates the fundamental dilemma of ontology matching, which tries to determine the intended meaning of concepts based on a necessarily incomplete specification. As a result, it is actually not always possible to really decide whether a correspondence is correct or not. All we can do is to argue that a correspondence is consistent with specifications in the ontologies and with the other correspondences. In the example this leads to a situation where we actually have two possible interpretations each of which makes a different set of correspondences correct. It is not completely clear how this dilemma can be handled by matching tools. The only recommendation we can give is in favor of using methods for checking the consistency of correspondences as an indicator whether the correspondence encodes a coherent view on the system.

## 7.5 Evaluation via pattern-aware data mining

### 7.5.1 Matching patterns

Before discussing the matching patterns, it is useful to briefly consider the notion of patterns as typically treated in ontological engineering research. We will consider three categories of patterns: content patterns, logical patterns and frequent errors. *Content patterns* [Gangemi, 2005] use specific non-logical vocabulary and describe a recurring, often domain-independent state of affairs. An example is the "Descriptions&Situations" pattern, which reflects the typical way a situation (with various entities and events involved) is described using some representation. *Logical patterns*, in turn, capture the typical ways certain modeling problems can be tackled in a specific ontological language. An example is the "Classes as Property Values" pattern[2], which defines multiple ways to satisfy the need for using a class in place of a value of an OWL property. Finally, *frequent errors* (though not usually denoted as patterns, they are clearly so) describe inadequate constructions that are often used by inexperienced modelers [Rector *et al.*, 2004]. All three mentioned types of patterns are used to describe modeling behaviors that are considered as either desirable (content and logical patterns) or undesirable (frequent errors). They can be qualified as *design* patterns; indeed, ontology building is essentially an activity carried out by human intellect (at least at the level of defining logical axioms, which are hard to obtain via automated ontology learning). In contrast, *matching patterns* that will be discussed further are by themselves neither desirable nor undesirable; their desirability depends on the correctness of the correspondences. They don't result from a deliberate activity by humans but can be detected in data output by automated matching systems.

As opposed to ontology design patterns, which concern one ontology, matching patterns deal with (at least) two ontologies. These patterns reflect the *structure of ontologies* on the one side, and on the other side they include *correspondences* between elements of ontologies. A matching pattern is a graph structure, where nodes are classes, properties or instances. Edges represent correspondences, relations between elements (e.g., domain and range of properties) or structural relations between classes (e.g., subclasses or siblings).

[2] `http://www.w3.org/TR/swbp-classes-as-values/`

The simplest (trivial) matching pattern we do not consider here only contains one element from each of the two ontologies (let us call them O1 and O2), and a correspondence between them. In our data mining experiments (described later) we employed three slightly more complex matching patterns.

The first one is depicted in Figure 7.3. The left-hand side (class A) is from O1 and the right-hand side (class B and its subclass C) is from O2. There is a correspondence between A and B and at the same time between A and C.

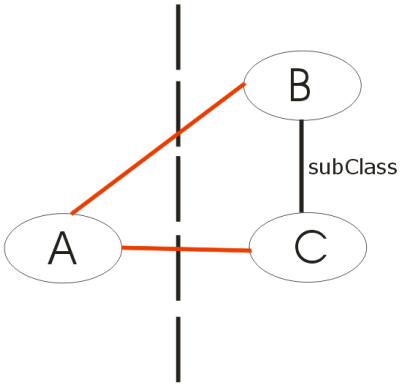



Figure 7.3: Pattern 1 – 'Parent-child triangle'.

The second pattern is depicted in Figure 7.4. It is quite similar to the previous one, but now we consider a child and a parent from each ontology and simultaneous correspondences between parents and between children.

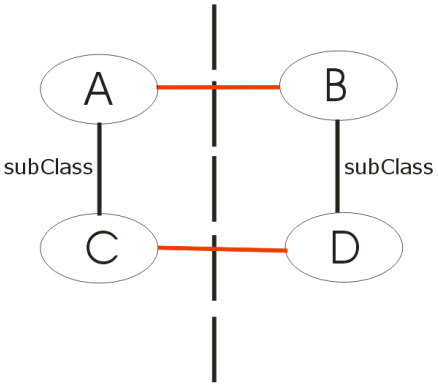



Figure 7.4: Pattern 2 – 'Matching along taxonomy'.

The third matching pattern we consider is depicted in Figure 7.5. It consists of simultaneous correspondences between class A from ontology O1 and two sibling classes C and D from ontology O2.

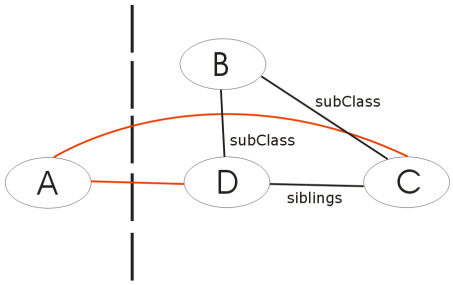



Figure 7.5: Pattern 3 – 'Sibling-sibling triangle'.

### 7.5.2 *4ft-Miner* overview

The *4ft-Miner* procedure is the most frequently used procedure of the *LISp-Miner* data mining system [Rauch and Šimunek, 2005]. *4ft-Miner* mines for association rules of the form $\varphi \approx \psi/\xi$, where $\varphi$, $\psi$ and $\xi$ are called *antecedent*, *succedent* and *condition*, respectively. Antecedent and succedent are conjunctions of *literals*. Literals are derived from attributes, i.e., fields of the underlying data matrix; unlike most propositional mining system, they can be (at run time) equipped with complex *coefficients*, i.e., value ranges. The association rule $\varphi \approx \psi/\xi$ means that on the subset of data defined by $\xi$, $\varphi$ and $\psi$ are associated in the way defined by the symbol $\approx$. The symbol $\approx$, called *4ft-quantifier*, corresponds to some statistical or heuristic test over the four-fold contingency table of $\varphi$ and $\psi$.

The task definition language of *4ft-Miner* is quite rich, and its description goes beyond the scope of this chapter. Let us only declare its two important features for our mining task: it is possible to formulate a wide range of so-called *analytic questions*, from very specific to very generic ones, and the underlying data mining algorithm is very fast due to highly optimized bit-string processing [Rauch and Šimunek, 2005].

### 7.5.3 Using *4ft-Miner* for mining over matching results

For the purpose of data mining, a data matrix with each record capturing all information about one (occurrence of) correspondence was built[3]. This elementary information amounted to: name of matching *system* that detected this (occurrence of) correspondence; *validity* assigned to the correspondence by the system; types of *ontologies* ('tool', 'insider', 'web') on both sides of the correspondence; correctness *label* manually assigned to the correspondence (§7.2). In addition, there is information about *patterns* (those from §7.5) in which the given correspondence participates. There are two data fields for each of the three patterns; the first one contains the correctness *label* of the *other* correspondences within the pattern (note that there are exactly two correspondences in each of these simple patterns), and the second one contains the *validity* assigned to this *other* correspondence by the system.

The analytic questions (i.e., task settings) we formulated for *4FT-Miner* were, for example, as follows:

1. Which systems give higher/lower validity than others to the correspondences that are deemed in/correct?

[3] In total, there are 5238 records.

2. Which systems produce certain matching patterns more often than others?
3. Which systems are more successful on certain types of ontologies?

Due to limited space we do not list complete nor detailed results of the data mining process. We only present some interesting association hypotheses discovered.

For the first question, we found, for example, the following hypotheses:

– Correspondences output by Falcon with medium validity (between 0.5 and 0.8) are almost twice more often incorrect than such correspondences output by all systems (on average).
– Correspondences output by RiMOM and by HMatch with high validity (between 0.8 and 1.0) are more correct than such correspondences output by all systems (on average).

For the second question, we found, for example, the following hypotheses:

– Correspondences output by HMatch with medium validity (between 0.5 and 0.8) are more likely to connect a child with a class that is also connected (with high validity) with a parent (Pattern 1) than such correspondences with all validity values (on average).

– Correspondences output by RiMOM with high validity (between 0.8 and 1.0) are more likely to connect class C with class D whose parent B is connected (with high validity) with A, which is parent of C (Pattern 2), than such correspondences with all validity values (on average).

These two hypotheses seem to have a natural interpretation (at the level of patterns, perhaps it is not so at the level of matching systems). Pattern 1 represents a potential matching conflict, i.e., increasing the validity of one may lead to decreasing the validity of the other. On the other hand, Pattern 2 seems to evoke positive feedback between the two correspondences.

A feature of the *OntoFarm* collection that was clearly beneficial for the data mining approach to matching evaluation was the fact that it contains (far) more than two ontologies that can be matched. Thanks to that, matching patterns frequently arising because of the specific nature of some ontologies could be separated from matching patterns that are frequent in general.

## 7.6 Discussion

To our knowledge, there has been no systematic effort in posterior analysis of ontology alignments without reference alignment involving multiple methods like discussed in this chapter. There are only projects with which we share some isolated aspects. For example, matching patterns are implicitly considered in [Ghidini and Serafini, 2006]. However, that work focuses on heterogeneous correspondences (e.g., class to property) as a special kind of pattern. We also considered this, but it appeared too infrequently (essentially, it was only output by the COMA++ system) to allow for meaningful data mining.

The purpose of the current study was to examine multiple methods of posterior evaluation of ontology alignments, focusing on the situation when there is no reference alignment available and/or we want to obtain deeper insight into the nature of correspondences. Our results could have at least two potential uses: to give the authors of individual matching systems feedback on strong and weak points of the systems (going far beyond the usual precision/recall statistics), and to contribute to better insight of the whole research community into possible argumentation used in the ontology matching process.

Although the methods are largely different, they have certain dependencies. In particular, initial manual empirical evaluation is pre-requisite for selecting representative cases for the consensus building workshop (this role was also played by automated reasoning) as well as for subsequent data mining. Consensus workshop, in turn, helped refine the nature of matching patterns. An outline of general methodology could easily be worked out from these dependencies.

# Chapter 8

# Conclusions

We summarize the major findings of this deliverable as well as outline directions for future activities along its two themes, namely: $(i)$ semantic precision and recall discussed in Chapter 2, and $(ii)$ the results of the OAEI-2006 campaign presented in the rest of the deliverable.

For what concerns the first theme, we plan to implement these measures (semantic precision and recall) and apply them to larger sets of data (results of the OAEI[1] evaluation campaigns for instance). This requires the use of a correct and complete prover for the considered ontology languages.

For what concerns the OAEI-2006 campaign, we have several observations. Firstly, the tests that have been run in 2006 were even more complete than those of the previous years. However, more teams participated and the results tend to be better. This shows that, as expected, the field of ontology matching is becoming stronger (and we conjecture that evaluation has been contributing to this progress). Finally, the Ontology Alignment Evaluation Initiative will continue these tests by improving both test cases and testing methodology for being more accurate.

## 8.1 Lesson learned

From OAEI-2005 lesson learned, we have applied those concerning character encoding, new evaluation measures and having a progressive test suite in the directory case. However, we must admit that not all of them have been applied, partly due to lack of time. So we reiterate those lessons that still apply with new ones, including:

A) It is now a general trend that tools for the semantic web are more robust and compliant. As a consequence, we had comments on the tests this year that concerned problems not discovered in previous years. Obviously the tools can now better handle the ontologies proposed in the tests and they return results that are more easy to handle for the evaluation. Moreover, we had more participants able to handle large scale test cases.

B) Not all the systems from the last year campaign participated in the campaign of this year. Fortunately, the best system participated this year as well. It will be useful to investigate if this is a definitive trend, whether we are evaluating research prototypes or "serious" systems.

[1] `http://oaei.ontologymatching.org`

C) The benchmark test case is not discriminant enough between systems. It is still useful for evaluating the strengths and weaknesses of algorithms but does not appear to be sufficient anymore for comparing algorithms (that have already participated in the campaigns). We will have to look into better alternatives.

D) We have had more proposals for test cases this year (we had actively looked for them). However, the difficult lesson is that proposing a test case is not enough, there is a lot of remaining work in preparing the evaluation, e.g., reference alignments acquisition. For example, as showed in [Zhang and Bodenreider, 2007b] the absence of a reference alignment cannot be adequately compensated by the use of cross-validation. A cursory review also leaves many open questions, while establishing manually a reference alignment would require the collaboration of domain experts and adequate funding. We expect that with tool improvements, it will be easier to perform the evaluation.

E) It would be interesting and certainly more realistic, to provide some random gradual degradation of the benchmark tests (5% 10% 20% 40% 60% 100% random change) instead of a general discarding of a feature. This has not been done so far due to lack of time.

F) Last but not least, similarly to the last year the time line for this evaluation was far from ideal both from the participants and the evaluators points of view. More time must be allocated to the next campaigns.

## 8.2 Future plans

In the short term, future plans for the Ontology Alignment Evaluation Initiative are certainly to go ahead and to improve it. In particular, OAEI-2007 campaign[2] will be held in conjunction with the Ontology Matching workshop[3] at the annual International Semantic Web Conference in Busan, Korea. Further improvements along the technical lines include:

- Finding new real world test cases;
- Improving the tests along the lesson learned;
- Accepting continuous submissions (through validation of the results);
- Improving the measures to go beyond precision and recall (we have done this for generalized precision and recall as well as for using precision/recall graphs, and will continue with other measures);
- Drawing lessons from the new test cases and establishing general rules for consensus reference alignment acquisition and application-oriented evaluation. In particular, for the next campaign, tests were delivered earlier and we will have more time for evaluating the results.

These are only indicative lines of improvement, which are to be further refined during the forthcoming campaigns.

In the medium to long term, it is planned to continue running the Ontology Alignment Evaluation Initiative campaign as long as it is useful. At the moment the effort is tremendously useful and this is why we have increasing numbers of test cases and participants. Moreover, organisers are committed to continue running this event. While the task could be heavy, it is split among

[2] http://oaei.ontologymatching.org/2007/
[3] http://om2007.ontologymatching.org/

several organisers: this make it more acceptable. The OAEI has been created three years ago as a sustainable instrument from scratch. It is an independent initiative supported by solid institutions like INRIA or the university of Trento. It has a visible web site, steering committees and dedicated leaders. It does not require much resource to be run so there is no reason to make more elaborate plans for sustainability: as soon as this effort is useful, it will exist in one form or another, and if it becomes useless, then it would certainly have met its goal.

## Acknowledgements

We appreciate technical feedback from Mark Ehrig, Stefano Zanobini and Antoine Zimmermann on the material of Chapter 2.

We thank Songmao Zhang (Chinese Academy of Sciences) and Olivier Bodenreider (U.S. National Library of Medicine) for allowing us to reuse some material of [Zhang and Bodenreider, 2007b] concerning the cross-validation of the alignments in the anatomy test case of Chapter 5.

We warmly thank each participant of the OAEI-2006 campaign. We know that they have worked hard for obtaining their results. We are grateful to the teams of the University of Washington and the University of Manchester for allowing us to use their ontologies of anatomy. Finally, we thank the other members of the Ontology Alignment Evaluation Initiative Steering committee for the support in the organization of the OAEI-2006 campaign: Lewis Hart (AT&T, USA), Tadashi Hoshiai (Fujitsu, Japan), Todd Hughes (DARPA, USA), Yannis Kalfoglou (University of Southampton, UK), John Li (Teknowledge, USA), Natasha Noy (Stanford University, USA), York Sure (University of Karlsruhe, Germany) Raphaël Troncy (CWI, Amsterdam, The Netherlands), and Petko Valtchev (Université du Québec à Montréal, Canada).

We also thank all the providers of test sets that have been used in the evaluation: the University of Washington, the University of Manchester Google, Yahoo and Looksmart, the Food and Agriculture Organization of the United Nations, the United States National Agricultural Library, and World Wide Jobs GmbH.

# Bibliography


[Ashpole *et al.*, 2005] Benjamin Ashpole, Marc Ehrig, Jérôme Euzenat, and Heiner Stuckenschmidt, editors. *Proc. K-CAP Workshop on Integrating Ontologies*, Banff (CA), 2005.

[Avesani *et al.*, 2005] Paolo Avesani, Fausto Giunchiglia, and Mikalai Yatskevich. A large scale taxonomy mapping evaluation. In *Proc. 4th International Semantic Web Conference (ISWC)*, volume 3729 of *Lecture notes in computer science*, pages 67–81, Galway (IE), 2005.

[Bouquet *et al.*, 2003] Paolo Bouquet, Fausto Giunchiglia, Frank van Harmelen, Luciano Serafini, and Heiner Stuckenschmidt. C-OWL – contextualizing ontologies. In *Proc. 2nd International Semantic Web Conference (ISWC)*, volume 2870 of *Lecture notes in computer science*, pages 164–179, Sanibel Island (FL US), 2003.

[Castano *et al.*, 2006] Silvana Castano, Alfio Ferrara, and Gianpaolo Messa. Results of the HMatch ontology matchmaker in OAEI 2006. In *Proc. 1st ISWC International Workshop on Ontology Matching (OM)*, pages 134–143, Athens (GA US), 2006.

[Do and Rahm, 2002] Hong-Hai Do and Erhard Rahm. COMA – a system for flexible combination of schema matching approaches. In *Proc. 28th International Conference on Very Large Data Bases (VLDB)*, pages 610–621, Hong Kong (CN), 2002.

[Do *et al.*, 2002] Hong-Hai Do, Sergei Melnik, and Erhard Rahm. Comparison of schema matching evaluations. In *Proc. Workshop on Web, Web-Services, and Database Systems*, volume 2593 of *Lecture notes in computer science*, pages 221–237, Erfurt (DE), 2002.

[Ehrig and Euzenat, 2005] Marc Ehrig and Jérôme Euzenat. Relaxed precision and recall for ontology matching. In *Proc. K-CAP Workshop on Integrating Ontologies*, pages 25–32, Banff (CA), 2005.

[Ehrig *et al.*, 2005] Marc Ehrig, Steffen Staab, and York Sure. Bootstrapping ontology alignment methods with APFEL. In *Proc. 4th International Semantic Web Conference (ISWC)*, volume 3729 of *Lecture notes in computer science*, pages 186–200, Galway (IE), 2005.

[Euzenat and Shvaiko, 2007] Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer, Heidelberg (DE), 2007.

[Euzenat, 2007] Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 248–253, Hyderabad (IN), 2007.

[Gangemi, 2005] Aldo Gangemi. Ontology design patterns for semantic web content. In *Proc. 3rd International Semantic Web Conference (ISWC)*, volume 3298 of *Lecture notes in computer science*, pages 262–276, Hiroshima (JP), 2005.

[Ghidini and Serafini, 1998] Chiara Ghidini and Luciano Serafini. Distributed first order logics. In *Proc. 2nd Conference on Frontiers of Combining Systems (FroCoS)*, pages 121–139, Amsterdam (NL), 1998.

[Ghidini and Serafini, 2006] Chiara Ghidini and Luciano Serafini. Reconciling concepts and relations in heterogeneous ontologies. In *Proc. 3rd European Semantic Web Conference (ESWC)*, volume 4011 of *Lecture notes in computer science*, pages 50–64, Budva (ME), 2006.

[Giunchiglia *et al.*, 2004] Fausto Giunchiglia, Pavel Shvaiko, and Mikalai Yatskevich. S-Match: an algorithm and an implementation of semantic matching. In *Proc. 1st European Semantic Web Symposium (ESWS)*, volume 3053 of *Lecture notes in computer science*, pages 61–75, Hersounisous (GR), 10-12 May 2004.

[Hu *et al.*, 2006] Wei Hu, Gong Cheng, Dongdong Zheng, Xinyu Zhong, and Yuzhong Qu. The results of Falcon-AO in the OAEI 2006 campaign. In *Proc. 1st ISWC International Workshop on Ontology Matching (OM)*, pages 124–133, Athens (GA US), 2006.

[Langlais *et al.*, 1998] Philippe Langlais, Jean Véronis, and Michel Simard. Methods and practical issues in evaluating alignment techniques. In *Proc. 17th International Conference on Computational Linguistics (CoLing)*, pages 711–717, Montréal (CA), 1998.

[Li *et al.*, 2006] Yi Li, Juanzi Li, Duo Zhang, and Jie Tang. Result of ontology alignment with RiMOM at OAEI-06. In *Proc. 1st ISWC International Workshop on Ontology Matching (OM)*, pages 181–190, Athens (GA US), 2006.

[Mao and Peng, 2006] Ming Mao and Yefei Peng. PRIOR system: Results for OAEI 2006. In *Proc. 1st ISWC International Workshop on Ontology Matching (OM)*, pages 173–180, Athens (GA US), 2006.

[Maßmann *et al.*, 2006] Sabine Maßmann, Daniel Engmann, and Erhard Rahm. COMA++: Results for the ontology alignment contest OAEI 2006. In *Proc. 1st ISWC International Workshop on Ontology Matching (OM)*, pages 107–114, Athens (GA US), 2006.

[McCray *et al.*, 1994] Alexa T. McCray, Suresh Srinivasan, and Allen C. Browne. Lexical methods for managing variation in biomedical terminologies. In *Proc. 18th Annual Symposium on Computer Applications in Medical Care (SCAMC)*, pages 235–239, Washington (US), 1994.

[Meilicke *et al.*, 2006] Christian Meilicke, Heiner Stuckenschmidt, and Andrei Tamilin. Improving automatically created mappings using logical reasoning. In *Proc. 1st ISWC International Workshop on Ontology Matching (OM)*, pages 61–72, Athens (GA US), 2006.

[Melnik *et al.*, 2002] Sergey Melnik, Hector Garcia-Molina, and Erhard Rahm. Similarity flooding: a versatile graph matching algorithm. In *Proc. 18th International Conference on Data Engineering (ICDE)*, pages 117–128, San Jose (CA US), 2002.

[Rauch and Šimunek, 2005] Jan Rauch and Milan Šimunek. An alternative approach to mining association rules. In T. Y. Lin, S. Ohsuga, C. J. Liau, and S. Tsumoto, editors, *Data Mining: Foundations, Methods, and Applications*, pages 211–232. Springer, 2005.

[Rector *et al.*, 2004] Alan L. Rector, Nick Drummond, Matthew Horridge, Jeremy Rogers, Holger Knublauch, Robert Stevens, Hai Wang, and Chris Wroe. OWL pizzas: Practical experience of teaching OWL-DL: Common errors & common patterns. In *Proc. 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, volume 3257 of *Lecture notes in computer science*, pages 63–81, Whittlebury Hall (UK), 2004.

[Serafini and Tamilin, 2005] Luciano Serafini and Andrei Tamilin. DRAGO: Distributed reasoning architecture for the semantic web. In *Proc. 2nd European Semantic Web Conference (ESWC)*, volume 3532 of *Lecture notes in computer science*, pages 361–376, Hersounisous (GR), May 2005.

[Shvaiko *et al.*, 2006] Pavel Shvaiko, Jérôme Euzenat, Natalya Noy, Heiner Stuckenschmidt, Richard Benjamins, and Michael Uschold, editors. *Proc. 1st ISWC International Workshop on Ontology Matching (OM)*, Athens (GA US), 2006.

[Sun and Lin, 2001] Aixin Sun and Ee-Peng Lin. Hierarchical text classification and evaluation. In *Proc. 1st International Conference on Data Mining (ICDM)*, pages 521–528, San Jose (CA US), 2001.

[Sure *et al.*, 2004] York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proc. 3rd ISWC Workshop on Evaluation of Ontology-based tools (EON)*, Hiroshima (JP), 2004.

[van Rijsbergen, 1975] Cornelis Joost (Keith) van Rijsbergen. *Information retrieval*. Butterworths, London (UK), 1975. http://www.dcs.gla.ac.uk/Keith/Preface.html.

[Šváb *et al.*, 2007] Ondřej Šváb, Vojtěch Svátek, and Heiner Stuckenschmidt. A study in empirical and 'casuistic' analysis of ontology mapping results. In *Proc. 4th European Semantic Web Conference (ESWC)*, Innsbruck (AU), 2007.

[Zhang and Bodenreider, 2007a] Songmao Zhang and Olivier Bodenreider. Experience in aligning anatomical ontologies. *International Journal on Semantic Web and Information Systems*, 3(2), 2007.

[Zhang and Bodenreider, 2007b] Songmao Zhang and Olivier Bodenreider. Reconciling concepts and relations in heterogeneous ontologies. In *Proc. 12th International Health (Medical) Informatics Congress (Medinfo)*, Brisbane (AUS), 2007. to appear.

# Related deliverables

There are several Knowledge Web deliverables that are related to this one:

| Project | Number | **Title** and relationship |
|---|---|---|
| KW | D2.1.1 | **Survey of scalability techniques for reasoning with ontologies** provided an in-depth discussion about benchmarking techniques that have been mentioned in this deliverable. |
| KW | D2.1.4 | **Specification of a methodology, general criteria, and test suites for benchmarking ontology tools** provided a general framework for defining a benchmarking test. |
| KW | D2.2.1 | **Specification of a common framework for characterizing alignment** provided a framework for defining the benchmarking actions. |
| KW | D2.2.3 | **State of the art in ontology alignment** provided a panorama of many of the techniques the results of evaluation of which has been reported in the current deliverable. |
| KW | D2.2.4 | **Alignment implementation and benchmarking results** presented the two first OAEI campains and improved methodologies. |
| KW | D1.2.2.2.1 | **Case-based recommendation of alignment tools and techniques** shows how to exploit the results of evaluations as given here. |