Introduction to the Ontology Alignment Evaluation 2005

Jérôme Euzenat
INRIA Rhône-Alpes
655 avenue de l'Europe
38330 Monbonnot, France
Jerome.Euzenat@inrialpes.fr

Heiner Stuckenschmidt
Vrije Universiteit Amsterdam
De Boelelaan 1081a
1081 HV Amsterdam, The Netherland
heiner@cs.vu.nl

Mikalai Yatskevich
Dept. of Information and Communication Technology
University of Trento
Via Sommarive, 14
I-38050 Povo, Trento, Italia
yatskevi@unitn.it

The increasing number of methods available for schema matching/ontology integration suggests the need to establish a consensus for evaluation of these methods.
The Ontology Alignment Evaluation Initiative[1] is now a coordinated international initiative that has been set up for organising evaluation of ontology matching algorithms.
After the two events organized in 2004 (namely, the Information Interpretation and Integration Conference (I3CON) and the EON Ontology Alignment Contest [4]), this year one unique evaluation campaign is organised. Its outcome is presented at the Workshop on Integrating Ontologies held in conjunction with K-CAP 2005 at Banff (Canada) on October 2, 2005.
Since last year, we have set up a web site, improved the software on which the tests can be evaluated and set up some precise guidelines for running these tests. We have taken into account last year's remarks by (1) adding more coverage to the benchmarck suite and (2) elaborating two real world test cases (as well as addressing other technical comments).
This paper serves as a presentation to the 2005 evaluation campaign and introduction to the results provided in the following papers.

1. GOALS

Last year events demonstrated that it is possible to evaluate ontology alignment tools.

One intermediate goal of this year is to take into account the comments from last year contests. In particular, we aimed at improving the tests by widening their scope and variety. Benchmark tests are more complete (and harder) than before. Newly introduced tracks are more 'real-world' and of a considerable size.

[1] `http://oaei.inrialpes.fr`

K-CAP'05, Integrating ontologies workshop, October 2, 2005, Banff, Alberta, Canada.

The main goal of the Ontology Alignment Evaluation is to be able to compare systems and algorithms on the same basis and to allow drawing conclusions about the best strategies. Our ambition is that from such challenges, the tool developers can learn and improve their systems.

2. GENERAL METHODOLOGY

We present below the general methodology for the 2005 campaign. In this we took into account many of the comments made during the previous campaign.

2.1 Alignment problems

This year's campaign consists of three parts: it features two real world blind tests (anatomy and directory) in addition to the systematic benchmark test suite. By blind tests it is meant that the result expected from the test is not known in advance by the participants. The evaluation organisers provide the participants with the pairs of ontologies to align as well as (in the case of the systematic benchmark suite only) expected results. The ontologies are described in OWL-DL and serialized in the RDF/XML format. The expected alignments are provided in a standard format expressed in RDF/XML [2].

Like for last year's EON contest, a systematic benchmark series has been produced. The goal of this benchmark series is to identify the areas in which each alignment algorithm is strong and 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 directory real world case consists of alignming web sites directory (like open directory or Yahoo's). It is more than two thousand elementary tests.

The anatomy real world case covers the domain of body anatomy and consists of two ontologies with an approximate size of several 10k classes and several dozen of relations.

The evaluation has been processed in three successive steps.

2.2 Preparatory phase

The ontologies and alignments of the evaluation have been provided in advance during the period between June 1st and July 1st. This was the occasion for potential participants to send observations, bug corrections, remarks and other test cases to the organizers. The goal of this primary period is to be sure that the delivered tests make sense to the participants. The feedback is important, so all participants should not hesitate to provide it. The final test base has been released on July 4th. The tests did only change after this period for ensuring a better and easier participation.

2.3 Execution phase

During the execution phase the participants have used their algorithms to automatically match the ontologies of both part. The participants were required to only use one algorithm and the same set of parameters for all tests. Of course, it is regular to select the set of parameters that provide the best results. Beside the parameters the input of the algorithms must be the two provided ontology to align and any general purpose resource available to everyone (that is no resourse especially designed for the test). In particular, the participants should not use the data (ontologies and results) from other test sets to help their algorithm.

The participants have provided their alignment for each test in the Alignment format and a paper describing their results[2].

In an attempt to validate independently the results, they were required to provide a link to their program and parameter set used for obtaining the results.

2.4 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 the case of the real world ontologies only the organizers will do the evaluation with regard to the withheld alignments.

The standard evaluation measures are precision and recall computed against the reference alignments. For the matter of aggregation of the measures we have computed a true global precision and recall (not a mere average). We have also computed precision/recall graphs for some of the participants (see below).

Finally, in an experimental way, we will attempt this year at reproducing the results provided by participants (validation).

3. COMMENTS ON THE EXECUTION

We had more participants than last year's event and it is easier to run these tests (qualitatively we had less comments and the results were easier to analyse). We summarize the list of participants in Table 1. As can be seen, not all participants provided results for all the tests and not all system were correctly validated. However, when the tests are straightforward to process (benchmarks and directory), participants provided results. The main problems with the anatomy test was its size. We also mentioned the kind of results sent by each participant (relations and confidence).

We note that 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 analyse their results and improve their algorithms. On the one hand, this prevents having algorithms really tuned for the contests, on the other hand, this can be frustrating for the participants. We should try to allow more time for participating next time.

Complete results are provided on http://oaei.inrialpes.fr/2005/results/. These are the only official results (the results presented here are only partial and prone to correction). The summary of results track by track is provided below.

4. BENCHMARK

The benchmark test case improved on last year's base by providing new variations of the reference ontology (last year the test contained 19 individual tests while this year it contains 53 tests). These new tests are supposed to be more difficult. The other improvement was the introduction of other evaluation metrics (real global precision and recall as well as the generation of precision-recall graphs).

4.1 Test set

The systematic benchmark test set is built around one reference ontology and many variations of it. The participants have to match this reference ontology with the variations. These variations are focussing the characterisation of the behaviour 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.

Since the goal of these tests is to offer some kind of permanent benchmarks to be used by many, the test is an extension of last year EON Ontology Alignment Contest. Test numbering (almost) fully preserves the numbering of the first EON contest.

The reference ontology is based on the one of the first EON Ontology Alignment Contest. It is improved by comprising a number of circular relations that were missing from the first test. 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 (based on area, quality, etc.). We choose the one common among scholars based on mean of publications; as many ontologies below (tests #301-304), it is reminiscent to BibTeX.

The reference ontology is that of test #101. It contains 33

[2]Andreas Hess from the UCDublin has not been able to provide a paper in due time. Description of his system can be found in [3]

Name	System	Benchmarks	Directory	Anatomy	Validated	Relations	Confidence
U. Karlsruhe	FOAM	√	√			=	cont
U. Montréal/INRIA	OLA	√	√		√	=	cont
IRST Trento	CtxMatch 2	√	√			=, ≤	1.
U. Southampton	CMS	√	√	√		=	1.
Southeast U. Nanjin	Falcon	√	√	√	√	=	1.
UC. Dublin	?	√	√			=	cont
CNR/Pisa	OMAP	√	√			=	1.

Table 1: Participants and the state of the state of their submissions. Confidence is given as 1/0 or continuous values.

named classes, 24 object properties, 40 data properties, 56 named individuals and 20 anonymous individuals.

The reference ontology is put in the context of the semantic web by using other external resources for expressing non bibliographic information. It takes advantage of FOAF (http://xmlns.com/foaf/0.1/) and iCalendar (http://www.w3.org/2002/12/cal/) for expressing the People, Organization and Event concepts. Here are the external reference used:

- http://www.w3.org/2002/12/cal/#:Vevent (defined in http://www.w3.org/2002/12/cal/ical.n3 and supposedly in http://www.w3.org/2002/12/cal/ical.rdf)
- http://xmlns.com/foaf/0.1/#:Person (defined in http://xmlns.com/foaf/0.1/index.rdf)
- http://xmlns.com/foaf/0.1/#:Organization (defined in http://xmlns.com/foaf/0.1/index.rdf)

This reference ontology is a bit limited in the sense that it does not contain attachement to several classes.

Similarly the kind of proposed alignments is still limited: they only match named classes and properties, they mostly use the "=" relation with confidence of 1.

There are still three group 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 some features of the reference ontology. The considered features were (names, comments, hierarchy, instances, relations, restrictions, etc.). The tests are systematically generated to as to start from some reference ontology and discarding a number of information in order to evaluate how the algorithm behave when this information is lacking. These tests were largely improved from last year by combining all feature discarding.
- four real-life ontologies of bibliographic references (3xx) that were found on the web and left mostly untouched (they were added xmlns and xml:base attributes).

Table 5 summarize what has been retracted from the reference ontology in the systematic tests. There are here 6 categories of alteration:

Name Name of entities that can be replaced by (R/N) random strings, (S)ynonyms, (N)ame with different conventions, (F) strings in another language than english.
Comments Comments can be (N) suppressed or (F) translated in another language.
Specialization Hierarchy can be (N) suppressed, (E)xpansed or (F)lattened.
Instances can be (N) suppressed
Properties can be (N) suppressed or (R) having the restrictions on classes discarded.
Classes can be (E)xpanded, i.e., relaced by several classes or (F)latened.

4.2 Results

Table 2 provide the consolidated results, by groups of tests. Table 6 contain the full results.

We display the results of participants as well as those given by some very simple edit distance algorithm on labels (edna). The computed values here are real precision and recall and not a simple average of precision and recall. This is more accurate than what has been computed last year.

As can be seen, the 1xx tests are relatively easy for most of the participants. The 2xx tests are more difficult in general while 3xx tests are not significantly more difficult than 2xx for most participants. The real interesting results is that there are significant differences across algorithms within the 2xx test series. Most of the best algorithms were combining different ways of finding the correspondence. Each of them is able to perform quite well on some tests with some methods. So the key issue seems to have been the combination of different methods (as described by the papers).

One algorithm, Falcon, seems largely dominant. But a group of other algorithms (Dublin, OLA, FOAM) are computing against each other. While the CMS and CtxMatch currently perform at a lower rate. Concerning these algorithm, CMS

algo	edna		falcon		foam		ctxMatch2-1		dublin20		cms		omap		ola	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	0.96	1.00	1.00	1.00	0.98	0.65	0.10	0.34	1.00	0.99	0.74	0.20	0.96	1.00	1.00	1.00
2xx	0.41	0.56	0.90	0.89	0.89	0.69	0.08	0.23	0.94	0.71	0.81	0.18	0.31	0.68	0.80	0.73
3xx	0.47	0.82	0.93	0.83	0.92	0.69	0.08	0.22	0.67	0.60	0.93	0.18	0.93	0.65	0.50	0.48
H-means	0.45	0.61	0.91	0.89	0.90	0.69	0.08	0.24	0.92	0.72	0.81	0.18	0.35	0.70	0.80	0.74

Table 2: Means of results obtained by participants (corresponding to harmonic means)

seems to priviledge precision and performs correctly in this (OLA seems to have privileged recall with regard to last year). CtxMatch has the difficulty of delivering many subsumption assertions. These assertions are taken by our evaluation procedure positively (even if equivalence assertions were required), but since there are many more assertions than in the reference alignments, this brings the result down.

These results can be compared with last year's results given in Table 3 (with aggregated measures computed at new with the methods of this year). For the sake of comparison, the results of this year on the same test set as last year are given in Table 4. As can be expected, the two participants of both challenges (Karlsruhe2 corresponding to foam and Montréal/INRIA corresponding to ola) have largely improved their results. The results of the best participants this year are over or similar to those of last year. This is remarkable, because participants did not tune their algorithms to the challenge of last year but to that of this year (more difficult since it contains more test of a more difficult nature and because of the addition of cycles in them).

So, it seem that the field is globally progressing.

Because of the precision/recall trade-off, as noted last year, it is difficult to compare the middle group of systems. In order to assess this, we attempted to draw precision recall graphs. We provide in Figure 1 the averaged precision and recall graphs of this year. They involve only the results of all participants. However, the results corresponding to participants who provided confidence measures different of 1 or 0 (see Table 1) can be considered as approximation. Moreover, for reason of time these graphs have been computed by averaging the graphs of each tests (instead to pure precision and recall).

These graphs are not totally faithful to the algorithms because participants have cut their results (in order to get high overall precision and recall). However, they provide a rough idea about the way participants are fighting against each others in the precision recall space. It would be very useful that next year we ask for results with continuous ranking for drawing these kind of graphs.

4.3 Comments

A general comments, we remarks, that it is still difficult for participants to provide results that correspond to the challenge (incorrect format, alignment with external entities). Because time is short and we try to avoid modifying provided results, this test is still a test of both algorithms and their ability to deliver a required format. However, some teams are really performant in this (and the same teams generally have their tools validated relatively easily).

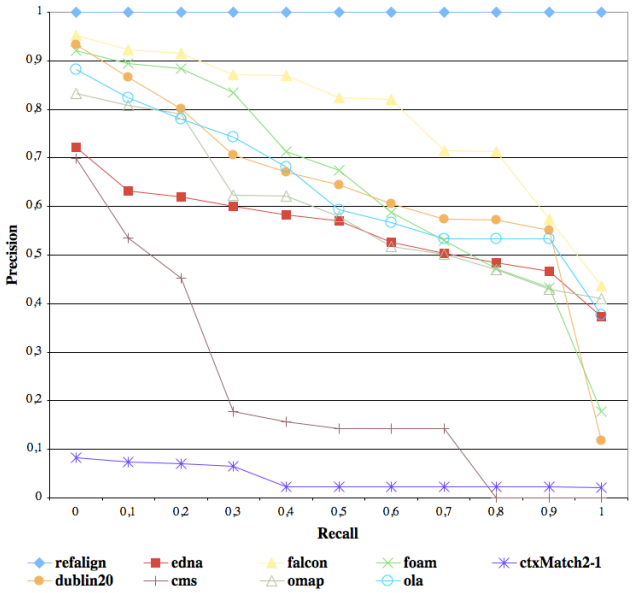

Figure 1: Precision-recall graphs

The evaluation of algorithms like ctxMatch which provide many subsumption assertions is relatively inadequate. Even if the test can remain a test of inference equivalence. It would be useful to be able to count adequately, i.e., not negatively for precision, true assertions like owl:Thing subsuming another concept. We must develop new evaluation methods taken into account these assertions and the semantics of the OWL language.

As a side note: all participants but one have used the UTF-8 version of the tests, so next time, this one will have to be the standard one with iso-latin as an exception.

5. DIRECTORY

5.1 Data set

algo	karlsruhe2		umontreal		fujitsu		stanford	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	NaN	0.00	0.57	0.93	0.99	1.00	0.99	1.00
2xx	0.60	0.46	0.54	0.87	0.93	0.84	0.98	0.72
3xx	0.90	0.59	0.36	0.57	0.60	0.72	0.93	0.74
H-means	0.65	0.40	0.52	0.83	0.88	0.85	0.98	0.77

Table 3: EON 2004 results with this year's aggregation method.

algo	edna		falcon		foam		ctxMatch2-1		dublin20		cms		omap		ola	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
1xx	0.96	1.00	1.00	1.00	0.98	0.65	0.10	0.34	1.00	0.99	0.74	0.20	0.96	1.00	1.00	1.00
2xx	0.66	0.72	0.98	0.97	0.87	0.73	0.09	0.25	0.98	0.92	0.91	0.20	0.89	0.79	0.89	0.86
3xx	0.47	0.82	0.93	0.83	0.92	0.69	0.08	0.22	0.67	0.60	0.93	0.18	0.93	0.65	0.50	0.48
H-means	0.66	0.78	0.97	0.96	0.74	0.59	0.09	0.26	0.94	0.88	0.65	0.18	0.90	0.81	0.85	0.83

Table 4: This year's results on EON 2004 test bench.

The data set exploited in the web directories matching task was constructed from Google, Yahoo and Looksmart web directories as described in [1]. The key idea of the data set construction methodology was to significantly reduce the search space for human annotators. Instead of considering the full mapping task which is very big (Google and Yahoo directories have up to $3*10^5$ nodes each: this means that the human annotators need to consider up to $(3*10^5)^2 = 9*10^{10}$ mappings), it uses semi automatic pruning techniques in order to significantly reduce the search space. For example, for the dataset described in [1] human annotators consider only 2265 mappings instead of the full mapping problem.

The major limitation of the current dataset version is the fact that it contains only true positive mappings (i.e., the mappings which tell that the particular relation holds between nodes in both trees). At the same time it does not contain true negative mappings (or zero mappings) which tell that there are no relation holding between pair of nodes. Notice that manually constructed mapping sets (such as ones presented for systematic tests) assume all the mappings except true positives to be true negatives. This assumption does not hold in our case since dataset generation technique guarantee correctness but not completeness of the produced mappings. This limitation allows to use the dataset only for evaluation of Recall but not Precision (since Recall is defined as ratio of correct mappings found by the system to the total number of correct mappings). At the same time measuring Precision necessarily require presence of the true negatives in the dataset since Precision is defined as a ratio of correct mappings found by the system to all the mappings found by the system. This means that all the systems will have 100% Precision on the the dataset since there are no incorrect mappings to be found.

The absence of true negatives has significant implications on the testing methodology in general. In fact most of the state of the art matching systems can be tuned either to produce the results with better Recall or to produce the results with better Precision. For example, the system which produce the equivalence relation on any input will always have 100% Recall. Therefore, the main methodological goal in the evaluation was to prevent Recall tuned systems from getting of unrealistically good results on the dataset. In order to accomplish this goal the double validation of the results was performed. The participants were asked for the binaries of their systems and were required to use the same sets of parameters in both web directory and systematic matching tasks. Then the results were double checked by organizers to ensure that the latter requirement is fulfilled by the authors. The process allow to recognize Recall tuned systems by analysis of systematic tests results.

The dataset originally was presented in its own format. The mappings were presented as pairwise relationships between the nodes of the web directories identified by their paths to root. Since the systems participating in the evaluation all take OWL ontologies as input the conversion of the dataset to OWL was performed. In the conversion process the nodes of the web directories were modelled as classes and classification relation connecting the nodes was modelled as rdfs:subClassOf relation. Therefore the matching task was presented as 2265 tasks of finding the semantic relation holding between pathes to root in the web directories modelled as sub class hierarchies.

5.2 Results

The results for web directory matching task are presented on Figure 2. As from the figure the web directories matching task is a very hard one. In fact the best systems found about 30% of mappings form the dataset (i.e., have Recall about 30%).

The evaluation results can be considered from two perspec-

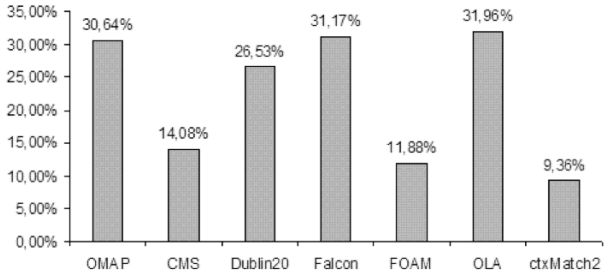

Figure 2: Recall for web directories matching task

tives. On the one hand, they are good indicator of real world ontologies matching complexity. On the other hand the results can provide information about the quality of the dataset used in the evaluation. The desired mapping dataset quality properties were defined in [1] as *Complexity*, *Discrimination capability*, *Incrementality* and *Correctness*. The first means that the dataset is "hard" for state of the art matching systems, the second that it discriminates among the various matching solutions, the third that it is effective in recognizing weaknesses in the state of the art matching systems and the fourth that it can be considered as a correct one.

The results of the evaluation give us some evidence for *Complexity* and *Discrimination capability* properties. As from Figure 2 TaxME dataset is hard for state of the art matching techniques since there are no systems having Recall more than 35% on the dataset. At the same time all the matching systems together found about 60% of mappings. This means that there is a big space for improvements for state of the art matching solutions.

Consider Figure 3. It contains partitioning of the mappings found by the matching systems. As from the figure 44% of the mappings found by any of the matching systems was found by only one system. This is a good argument to the dataset *Discrimination capability* property.

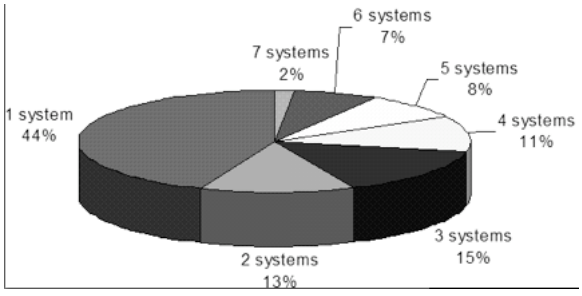

Figure 3: Partitioning of the mappings found by the matching systems

5.3 Comments

The web directories matching task is an important step towards evaluation on the real world matching problems. At the same time there are a number of limitations which makes the task only an intermediate step. First of all the current version of the mapping dataset provides correct but not complete set of the reference mappings. The new mapping dataset construction techniques can overcome this limitation. In the evaluation the mapping task was split to the the tiny subtasks. This strategy allowed to obtain results form all the matching systems participating in the evaluation. At the same time it hides computational complexity of "real world" matching (the web directories have up to 10^5 nodes) and may affect the results of the tools relying on "look for similar siblings" heuristic.

The results obtained on the web directories matching task coincide well with previously reported results on the same dataset. According to [1] generic matching systems (or the systems intended to match any graph-like structures) have Recall from 30% to 60% on the dataset. At the same time the real world matching tasks are very hard for state of the art matching systems and there is a huge space for improvements in the ontology matching techniques.

6. ANATOMY

6.1 Test set

The focus of this task is to confront existing alignment technology with real world ontologies. Our aim is to get a better impression of where we stand with respect to really hard challenges that normally require an enormous manual effort and requires in-depth knowledge of the domain.

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 are:

- Very large models: be prepared to handle OWL models of more than 50MB !
- Extensive Class Hierarchies: then 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 too much in the different model
- Clear Modelling Principles: The modelling principles are well defined and documented in publications about the ontologies

This implies that the task will be challenging from a technological point of view, but there is guidance for tuning matching approach that needs to be taken into account.

The ontologies to be aligned 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.

6.1.1 The Foundational Model of Anatomy

The Foundational Model of Anatomy is a medical ontology developed by the University of Washington. We extracted an OWL version of the ontology from a Protege database. The model contains the following information:

- Class hierarchy;
- Relations between classes;
- Free text documentation and definitions of classes;
- Synonyms and names in different languages.

6.1.2 The OpenGalen Anatomy Model

The second ontology is the Anatomy model developed in the OpenGalen Project by the University of Manchester. We created an OWL version of the ontology using the export functionality of Protege. The model contains the following information:

- Concept hierarchy;
- Relations between concepts.

The task is to find alignment between classes in the two ontologies. In order to find the alignment, any information in the two models can be used. In addition, it is allowed to use background knowledge, that has not specifically been created for the alignment tasks (i.e., no hand-made mappings between parts of the ontologies). Admissible background knowledge are other medical terminologies such as UMLS as well as medical dictionaries and document sets. Further, results must not be tuned manually, for instance, by removing obviously wrong mappings.

6.2 Results

At the time of printing we are not able to provide results of evaluation on this test.

Validation of the results on the medical ontologies matching task is still an open problem. The results can be replicated in straightforward way. At the same time there are no sufficiently big set of the reference mappings what makes impossible calculation of the matching quality measures.

We are currently developing an approach for creating such a set is to exploit semi-automatic reference mappings acquisition techniques. The underlying principle is that the task of creating such a reference alignment is fundamentally different from the actual mapping problem. In particular, we believe that automatically creating reference alignments is easier than solving the general mapping problem. The reason for this is, that methods for creating general mappings have to take into account both, correctness and completeness of the generated mappings. This is difficult, because allying very strict heuristics will lead to correct, but very incomplete mappings, using loose heuristics for matching nodes will create a rather complete, but often incorrect set of mappings. In our approach for generating reference alignments, we completely focus on the correctness. The result is a small set of reference mappings that we can assume to be correct. We can evaluate matching approaches against this set of mappings. The idea is that the matching approaches should at least be able to determine these mappings. From the result, we can extrapolate the expected completeness of a matching algorithm.

We assume that the task is to create a reference alignment for two a number of known conceptual models. In contrast to existing work [1] we do not assume that instance data is available or that the models are represented in the same way or using the same language. Normally, the models will be from the same domain (eg. medicine or business). The methodology consists of four basic steps. In the first step, basic decisions are made about the representation of the conceptual models and instance data to be used. In the second step instance data is created by selecting it from an existing set or by classifying data according to the models under consideration. In the third step, the generated instance data is used to generate candidate mappings based on shared instances. In the forth step finally, the candidate mappings are evaluated against a set of quality criteria and the final set of reference mappings is determined.

6.2.1 Step 1. Preparation

The first step of the process is concerned with data preparation. In particular, we have to transform the conceptual models into a graph representation and select and prepare the appropriate instance data to be used to analyze overlap between concepts in the different models. We structure this step based on the KDD process for Knowledge Discovery and Data Mining.

6.2.2 Step 2. Instance Classification

In the second step the chosen instance data is classified according to the different conceptual models. For this purpose, an appropriate classification method has to be chosen that fits the data and the conceptual model. Further, the result of the classification process has to be evaluated. For this step we rely on established methods from Machine Learning and Data Mining.

6.2.3 Step 3. Hypothesis Generation

In the third step, we generate hypothesis for reference mappings based on shared instances created in the first two steps. In this step, we prune the classification by removing instances that are classified with a low confidence and selecting subsets of the conceptual models that show sufficient overlap. We further compute a degree of overlap between concepts in the different models and based on this degree of overlap select a set of reference mappings between concepts with a significant overlap.

6.3 Step 4. Evaluation

In the last step, the generated reference mapping is evaluated against the result of different matching systems as described in **??** using a number of criteria for a reference mapping. These criteria include correctness, complexity of the mapping problem and the ability of the mappings to discriminate between different matching approaches.

We are testing this methodology using a data set of medical documents called OHSUMED. The data set contains 350.000 articles from medical journals covering all aspects of medicine. For classifying these documents according to the two ontologies of anatomy, we use the collexis text indexing and retrieval system that implements a number of automatic methods for assigning concepts to documents. Currently, we are testing the data set and the system on a subset of UMLS with known mappings in order to assess the suitability of the methodology. The generation of the reference mappings for the Anatomy case will proceed around the end of 2005 and we are hopeful to have thoroughly tested set of reference mappings for the 2006 alignment challenge.

6.4 Comments

We had very few participants able to even produce the alignments between both ontologies. This is mainly due to their inability to load these ontologies with current OWL tools (caused either by the size of the ontologies or errors in the OWL).

7. RESULT VALIDATION

As can be seen from the procedure, the results published in the following papers are not obtained independently. The results provided here have been computed from the alignment provided by the participants and can be considered as the official results of the evaluation.

In order to go one step further, we have attempted, this year, to generate the results obtained by the participants from their tools. The tools for which the results have been validated independently are marked in Table 1.

8. LESSON LEARNED

A) It seems that there are more and more tools able to jump in this kind of tests.

B) Contrary to last year it seems that the tools are more robusts and people deal with more wider implementation of OWL. However, this can be that we tuned the tests so that no one has problems.

C) Contrary to what many people think, it is not that easy to find ontological corpora suitable for this evaluation test. From the proposals we had from last year, only one proved to be usable and with great difficulty (on size, conformance and juridical aspects).

D) The extension of the benchmark tests towards more coverage of the space is relatively systematic. However, it would be interesting and certainly more realistic, instead of crippling all names to do it for some random proportion of them (5% 10% 20% 40% 60% 100% random change). This has not been done for reason of time.

E) The real world benchmarks were huge benchmarks. Two different strategies have been taken with them: cutting them in a huge set of tiny benchmark or providing them as is. The first solution brings us away from "real world", while the second one raised serious problems to the participants. It would certainly be worth designing these tests in order to assess the current limitation of the tools by providing an increasingly large sequence of such tests (0.1%, 1%, 10%, 100% of the corpus for instance).

F) Validation of the results are quite difficult to establish.

9. FUTURE PLANS

The future plans for the Ontology Alignement Evaluation Initiative are certainly to go ahead and improving the functioning of these evaluation campaign. This most surely involves:

- Finding new real world 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.

Of course, these are only suggestions and other ideas could come during the wrap-up meeting in Banff.

10. CONCLUSION

In summary, the tests that have been run this year are harder and more complete than those of last year. However, more teams participated and the results tend to be better. This shows that, as expected, the field of ontology alignment is getting stronger (and we hope that evaluation is contributing to this progress).

Reading the papers of the participants should help people involved in ontology matching to find what make these algorithms work and what could be improved.

The Ontology Alignment Evaluation Initiative will continue these tests by improving both test cases and test methodology for being more accurate. It can be found at:

`http://oaei.inrialpes.fr.`

11. ACKNOWLEDGEMENTS

We warmly thank each participant of this contest. We know that they worked hard for having their results ready and they provided insightful papers presenting their experience. The best way to learn about the results remains to read what follows.

Many thanks are due to the teams at the University of Washington and the University of Manchester for allowing us to use their ontologies of anatomy.

The other members of the Ontology Alignment Evaluation Initiative Steering committee: Benjamin Ashpole (Lockheed Martin Advanced Technology Lab.), Marc Ehrig (University of Karlsruhe), Lewis Hart (Applied Minds), Todd Hughes (Lockheed Martin Advanced Technology Labs), Natasha Noy (Stanford University), and Petko Valtchev (Université de Montréal, DIRO)

This work has been partially supported by the Knowledge Web European network of excellence (IST-2004-507482).

12. REFERENCES

[1] Paolo Avesani, Fausto Giunchiglia, and Michael Yatskevich. A large scale taxonomy mapping evaluation. In *Proceedings of International Semantic Web Conference (ISWC)*, 2005.

[2] Jérôme Euzenat. An API for ontology alignment. In *Proc. 3rd international semantic web conference, Hiroshima (JP)*, pages 698–712, 2004.

[3] Andreas Heß and Nicholas Kushmerick. Iterative ensemble classification for relational data: A case study of semantic web services. In *Proceedings of the 15th European Conference on Machine Learning*, Pisa, Italy, 2004.

[4] York Sure, Oscar Corcho, Jérôme Euzenat, and Todd Hughes, editors. *Proceedings of the 3rd Evaluation of Ontology-based tools (EON)*, 2004.

Montbonnot, Amsterdam, Trento, September 7th, 2005

#	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 (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 5: Structure of the systematic benchmark test-case

algo	edna		falcon		foam		ctxMatch2-1		dublin20		cms		omap		ola	
test	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
101	0.96	1.00	1.00	1.00	n/a	n/a	0.10	0.34	1.00	0.99	n/a	n/a	0.96	1.00	1.00	1.00
103	0.96	1.00	1.00	1.00	0.98	0.98	0.10	0.34	1.00	0.99	0.67	0.25	0.96	1.00	1.00	1.00
104	0.96	1.00	1.00	1.00	0.98	0.98	0.10	0.34	1.00	0.99	0.80	0.34	0.96	1.00	1.00	1.00
201	0.03	0.03	0.98	0.98	n/a	n/a	0.00	0.00	0.96	0.96	1.00	0.07	0.80	0.38	0.71	0.62
202	0.03	0.03	0.87	0.87	0.79	0.52	0.00	0.00	0.75	0.28	0.25	0.01	0.82	0.24	0.66	0.56
203	0.96	1.00	1.00	1.00	1.00	1.00	0.08	0.34	1.00	0.99	1.00	0.24	0.96	1.00	1.00	1.00
204	0.90	0.94	1.00	1.00	1.00	0.97	0.09	0.28	0.98	0.98	1.00	0.24	0.93	0.89	0.94	0.94
205	0.34	0.35	0.88	0.87	0.89	0.73	0.05	0.11	0.98	0.97	1.00	0.09	0.58	0.66	0.43	0.42
206	0.51	0.54	1.00	0.99	1.00	0.82	0.05	0.08	0.96	0.95	1.00	0.09	0.74	0.49	0.94	0.93
207	0.51	0.54	1.00	0.99	0.96	0.78	0.05	0.08	0.96	0.95	1.00	0.09	0.74	0.49	0.95	0.94
208	0.90	0.94	1.00	1.00	0.96	0.89	0.09	0.28	0.99	0.96	1.00	0.19	0.96	0.90	0.94	0.94
209	0.35	0.36	0.86	0.86	0.78	0.58	0.05	0.11	0.68	0.56	1.00	0.04	0.41	0.60	0.43	0.42
210	0.51	0.54	0.97	0.96	0.87	0.64	0.05	0.08	0.96	0.82	0.82	0.09	0.88	0.39	0.95	0.94
221	0.96	1.00	1.00	1.00	1.00	1.00	0.12	0.34	1.00	0.99	1.00	0.27	0.96	1.00	1.00	1.00
222	0.91	0.99	1.00	1.00	0.98	0.98	0.11	0.31	1.00	0.99	1.00	0.23	0.96	1.00	1.00	1.00
223	0.96	1.00	1.00	1.00	0.99	0.98	0.09	0.34	0.99	0.98	0.96	0.26	0.96	1.00	1.00	1.00
224	0.96	1.00	1.00	1.00	1.00	0.99	0.10	0.34	1.00	0.99	1.00	0.27	0.96	1.00	1.00	1.00
225	0.96	1.00	1.00	1.00	0.00	0.00	0.08	0.34	1.00	0.99	0.74	0.26	0.96	1.00	1.00	1.00
228	0.38	1.00	1.00	1.00	1.00	1.00	0.12	1.00	1.00	1.00	0.74	0.76	0.92	1.00	1.00	1.00
230	0.71	1.00	0.94	1.00	0.94	1.00	0.08	0.35	0.95	0.99	1.00	0.26	0.89	1.00	0.95	0.97
231	0.96	1.00	1.00	1.00	0.98	0.98	0.10	0.34	1.00	0.99	1.00	0.27	0.96	1.00	1.00	1.00
232	0.96	1.00	1.00	1.00	1.00	0.99	0.12	0.34	1.00	0.99	1.00	0.27	0.96	1.00	1.00	1.00
233	0.38	1.00	1.00	1.00	1.00	1.00	0.12	1.00	1.00	1.00	0.81	0.76	0.92	1.00	1.00	1.00
236	0.38	1.00	1.00	1.00	1.00	1.00	0.09	1.00	1.00	1.00	0.74	0.76	0.92	1.00	1.00	1.00
237	0.91	0.99	1.00	1.00	1.00	0.99	0.11	0.31	1.00	0.99	1.00	0.23	0.95	1.00	0.97	0.98
238	0.96	1.00	0.99	0.99	1.00	0.99	0.07	0.34	0.99	0.98	0.96	0.26	0.96	1.00	0.99	0.99
239	0.28	1.00	0.97	1.00	0.97	1.00	0.14	1.00	0.97	1.00	0.71	0.76	0.85	1.00	0.97	1.00
240	0.33	1.00	0.97	1.00	0.94	0.97	0.10	1.00	0.94	0.97	0.71	0.73	0.87	1.00	0.97	1.00
241	0.38	1.00	1.00	1.00	1.00	1.00	0.12	1.00	1.00	1.00	0.81	0.76	0.92	1.00	1.00	1.00
246	0.28	1.00	0.97	1.00	0.97	1.00	0.14	1.00	0.97	1.00	0.71	0.76	0.85	1.00	0.97	1.00
247	0.33	1.00	0.94	0.97	0.94	0.97	0.10	1.00	0.94	0.97	0.71	0.73	0.87	1.00	0.97	1.00
248	0.06	0.06	0.84	0.82	0.89	0.51	0.00	0.00	0.71	0.25	0.25	0.01	0.82	0.24	0.59	0.46
249	0.04	0.04	0.86	0.86	0.80	0.51	0.00	0.00	0.74	0.29	0.25	0.01	0.81	0.23	0.59	0.46
250	0.01	0.03	0.77	0.70	1.00	0.55	0.00	0.00	1.00	0.09	0.00	0.00	0.05	0.45	0.30	0.24
251	0.01	0.01	0.69	0.69	0.90	0.41	0.00	0.00	0.79	0.32	0.25	0.01	0.82	0.25	0.42	0.30
252	0.01	0.01	0.67	0.67	0.67	0.35	0.00	0.00	0.57	0.22	0.25	0.01	0.82	0.24	0.59	0.52
253	0.05	0.05	0.86	0.85	0.80	0.40	0.00	0.00	0.76	0.27	0.25	0.01	0.81	0.23	0.56	0.41
254	0.02	0.06	1.00	0.27	0.78	0.21	0.00	0.00	NaN	0.00	0.00	0.00	0.03	1.00	0.04	0.03
257	0.01	0.03	0.70	0.64	1.00	0.64	0.00	0.00	1.00	0.09	0.00	0.00	0.05	0.45	0.25	0.21
258	0.01	0.01	0.70	0.70	0.88	0.39	0.00	0.00	0.79	0.32	0.25	0.01	0.82	0.25	0.49	0.35
259	0.01	0.01	0.68	0.68	0.61	0.34	0.00	0.00	0.59	0.21	0.25	0.01	0.82	0.24	0.58	0.47
260	0.00	0.00	0.52	0.48	0.75	0.31	0.00	0.00	0.75	0.10	0.00	0.00	0.05	0.86	0.26	0.17
261	0.00	0.00	0.50	0.48	0.63	0.30	0.00	0.00	0.33	0.06	0.00	0.00	0.01	0.15	0.14	0.09
262	0.01	0.03	0.89	0.24	0.78	0.21	0.00	0.00	NaN	0.00	0.00	0.00	0.03	1.00	0.20	0.06
265	0.00	0.00	0.48	0.45	0.75	0.31	0.00	0.00	0.75	0.10	0.00	0.00	0.05	0.86	0.22	0.14
266	0.00	0.00	0.50	0.48	0.67	0.36	0.00	0.00	0.33	0.06	0.00	0.00	0.01	0.15	0.14	0.09
301	0.48	0.79	0.96	0.80	0.83	0.31	0.10	0.07	0.74	0.64	1.00	0.13	0.94	0.25	0.42	0.38
302	0.31	0.65	0.97	0.67	0.97	0.65	0.14	0.27	0.62	0.48	1.00	0.17	1.00	0.58	0.37	0.33
303	0.40	0.82	0.80	0.82	0.89	0.80	0.04	0.29	0.51	0.53	1.00	0.18	0.93	0.80	0.41	0.49
304	0.71	0.95	0.97	0.96	0.95	0.96	0.11	0.26	0.75	0.70	0.85	0.22	0.91	0.91	0.74	0.66
H-means	0.45	0.61	0.91	0.89	0.90	0.69	0.08	0.24	0.92	0.72	0.81	0.18	0.35	0.70	0.80	0.74

Table 6: Full results