# A Large Scale Dataset for the Evaluation of Matching Systems

Mikalai Yatskevich[1], Fausto Giunchiglia[1], and Paolo Avesani[2]

[1] Dept. of Information and Communication Technology,
University of Trento,
38050 Povo, Trento, Italy
{fausto,yatskevi}@dit.unitn.it
[2] ITC-IRST,
38050 Povo, Trento, Italy
avesani@itc.it

**Abstract.** Ontology matching is one of the biggest challenges of Semantic Web research. In the last years the number of matching techniques and systems has significantly increased, and this, in turn, has raised the issue of their evaluation and comparison. In this paper we present a mapping dataset extracted from the Google, Yahoo and Looksmart web directories. This dataset allows for the evaluation of both Precision and Recall, and it is an order of magnitude larger than the state of the art datasets with the same capabilities. We have evaluated this dataset on nine state of the art matching solutions. The evaluation results highlight the fact that the dataset has three key properties, namely it is error-free, it is hard to solve, and it can discriminate among systems.

## 1 Introduction

Match is a critical operator in many applications. It takes two graph-like structures, e.g., lightweight ontologies, such as Google [3] and Looksmart [4], or business catalogs, such as UNSPSC [5] and eCl@ss [6], and produces a mapping between the nodes that correspond semantically to each other. Many diverse solutions to the matching problem have been proposed so far, see for example [2, 19, 18, 16, 9, 6, 13]. These solutions can be classified as implementing syntactic or semantic matching, depending on how mapping elements are computed and on the kind of similarity relation used (see [11] for in depth discussion). In syntactic matching the idea is to compute a syntactic (very often string based) similarity between the labels of nodes. Similarity in this case is typically represented as a [0, 1] coefficient, which is often considered as equivalence relation with a certain level of plausibility or confidence (see, e.g., [13, 8]). In semantic matching the idea is to compute semantic relations between concepts (not labels) at nodes (see [10, 11]). The possible semantic relations are: equivalence (=); more general or generalization ($\sqsupseteq$); less general or specification ($\sqsubseteq$) mismatch ($\perp$); overlapping ($\cap$).

[3] http://www.google.com/Top/
[4] http://www.looksmart.com/
[5] http://www.unspsc.org/
[6] http://www.eclass.de/

Unfortunately all the matching solutions suffer from the lack of evaluation. Until very recently there was no comparative evaluation and it was quite difficult to find two systems which were evaluated on the same dataset. On top of this, when existing, the evaluation efforts were mostly concentrated on datasets artificially synthesized under questionable assumptions or on “toy” examples. One noticeable example was the large scale dataset called *TaxME* described in [1]. This dataset is constructed from the mappings extracted from real web directories and contains thousands of mappings. However, this dataset contains only an incomplete set of positive mappings, and this inherently limits its use, in that it allows only for the evaluation of Recall. However, Recall can be easily maximized at the expense of a poor Precision, for instance by returning all possible correspondences, i.e., the cross product of the input graphs. In order to overcome this kind of problems a sophisticated evaluation methodology was exploited in [7]. The key idea was to validate the systems’ results on another dataset of much smaller size, where both Precision and Recall could be estimated. However, this opened a range of problems related to the comparability of the results obtained on two different datasets, and a general solution for the problem still does not exists.

In this paper we present a new large scale mapping dataset called *TaxME 2*. *TaxME 2* extends *TaxME*, it contains about 4500 mappings and it allows for the evaluation of both Precision and Recall. We have evaluated *TaxME 2* using nine state of the art solutions to the matching problem. The evaluation shows that *TaxME 2* satisfies the key important properties of *Complexity* and *Discrimination capability*, as introduced in [1]. A dataset is complex if it is hard to solve even for state of the art matching systems, while it is discriminating if different sets of mappings taken from the dataset are hard for different systems.

The rest of the paper is organized as follows. Section 2 presents a short introduction to the notions of matching and matching evaluation. Section 3 extends the results presented in [1] and discusses the features and properties of *TaxME*. Section 4 illustrates how *TaxME 2* has been constructed. Section 5 presents the results of our experiments and shows that *TaxME 2* satisfies the described requirements. Section 6 concludes the paper.

## 2 Matching evaluation

In order to motivate the matching problem and illustrate one of the possible situations which can arise in the data integration task let us use the (parts of the Google and Yahoo) directories depicted in Figure 1. Suppose that the task is to integrate these two directories. The first step in the integration process is to identify the matching candidates. For example, $Shopping_{O1}$ can be assumed equivalent to $Shopping_{O2}$, while $Board\_Games_{O1}$ is less general than $Games_{O2}$. Hereafter the subscripts designate the directory (either O1 or O2) of the node considered.

We think of a *mapping element* as a 4-tuple $\langle ID_{ij}, n1_i, n2_j, R\rangle$, $i = 1, ..., N_1$; $j = 1, ..., N_2$; where $ID_{ij}$ is a unique identifier of the given mapping element; $n1_i$ is the i-th node of the first graph, $N_1$ is the number of nodes in the first graph; $n2_j$ is the j-th node of the second graph, $N_2$ is the number of nodes in the second graph; and $R$ specifies a similarity relation of the given nodes. We define *matching* as the process of

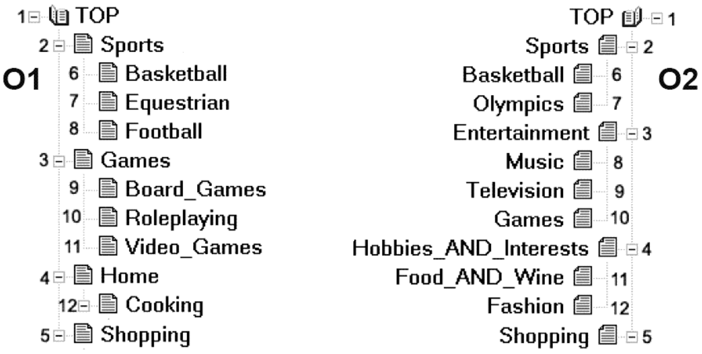

**Fig. 1.** Parts of Google and Yahoo directories

discovering mappings between two graph-like structures through the application of a matching algorithm.

A quantitative matching evaluation is based on the well known in information retrieval measures of relevance, namely *Precision* and *Recall*. Consider Figure 2; the calculation of these measures is based on the comparison between the mappings produced by a matching system (the area inside the circle labelled $S$ in Figure 2) and a complete set of reference mappings $H$ considered to be correct (the area inside the dotted circle in Figure 2). $H$ is usually produced by humans. Here and further we refer to the set of all possible mappings (i.e., cross product of two input graphs) as $M$. Finally, the correct mappings found by the system are the *true positives*, $TP = S \cap H$, the incorrect mappings found by the system are the *false positives*, $FP = S - S \cap H$, the correct mappings missed by the system are the *false negatives*, $FN = H - S \cap H$, and the incorrect mappings not returned by the system are the *true negatives*, $TN = M - S \cap H$. Further we call $H$ the "*golden standard*", the mappings in $H$ *positive mappings*, and the mappings in $N = M - H = TN + FP$ *negative mappings*.

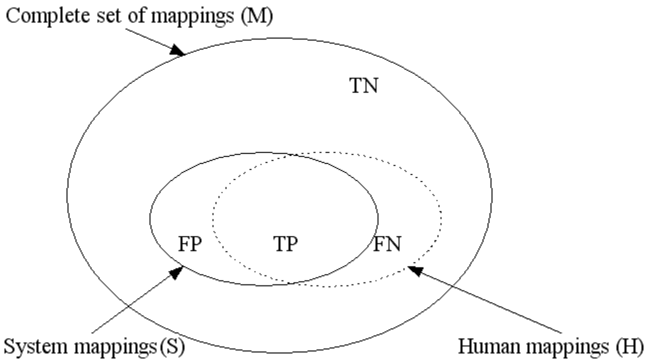

**Fig. 2.** Basic sets of mappings

Precision is a correctness measure which varies from [0, 1]. It is calculated as

$$Precision = \frac{|TP|}{|TP + FP|} = \frac{H \cap S}{S} \quad (1)$$

Recall is a completeness measure which varies from [0, 1]. It is calculated as

$$Recall = \frac{|TP|}{|TP + FN|} = \frac{H \cap S}{H} \quad (2)$$

However, neither Precision nor Recall alone can accurately evaluate the match quality. In particular, Recall can easily be maximized at the expense of a poor Precision by returning all possible correspondences, i.e. the cross product of two input graphs. At the same time, a high Precision can be achieved at the expense of a poor Recall by returning only few (correct) correspondences. Therefore, it is necessary to consider both measures or a combined measure.

F-measure is a global measure of the matching quality. It varies from [0, 1] and calculated as a harmonic mean of Precision and Recall:

$$F - Measure = \frac{2 * Recall * Precision}{Recall + Precision} \quad (3)$$

Notice that the golden standard $H$ must be known in advance in order to calculate both Precision and Recall. This opens a problem of how to acquire it. The problem is that the construction of $H$ is a manual process which, in the case of matching, is quadratic with respect to the size of the graphs to be matched. This process turns to be unfeasible for large datasets. For instance, in the dataset we have exploited in this work, namely the Google, Yahoo and Looksmart web directories, each structure has the order of $10^5$ nodes. This means that construction of $H$ would require the manual evaluation of $10^{10}$ mappings.

## 3 A dataset for evaluating Recall

A semiautomatic method for an approximation of the golden standard $H$ was proposed in [1] and it was applied to the Google, Yahoo and Looksmart web directories. The key idea was to rely on a reference interpretation for nodes, constructed by looking at their use. The assumption is that the semantics of nodes can be derived from their pragmatics, namely from analyzing which documents are classified under which nodes. In particular, in the work described in [1] the authors have argued that two nodes are 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.

Consider the example presented in Figure 3. 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. The set of documents $A_2$ denotes the contents classified in the ancestor node of $N_2$; the set of documents $C_1$ denotes the contents classified in the children nodes of $N_1$.

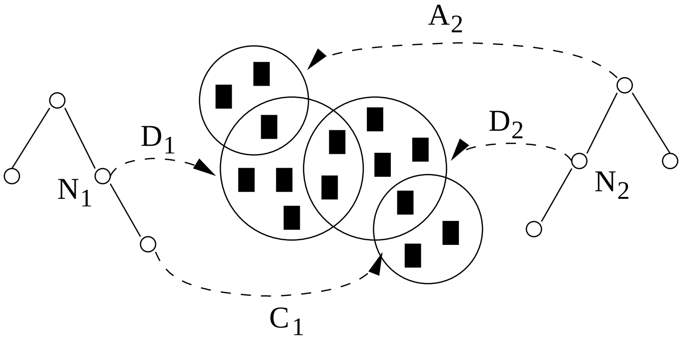

**Fig. 3.** *TaxME*. Illustration of a document-driven similarity assessment.

The *equivalence* measure we use, as defined in [1], is

$$Eq(N_1, N_2) = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2| - |D_1 \cap D_2|} \tag{4}$$

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 is $Eq(N_1, N_2)$ with $Eq(N_1, N_2)$ becoming infinite with $D_1 \equiv D_2$. Following what described in [1] $Eq(N_1, N_2)$ is normalized to [0,1]. The special case of $D_1 \equiv D_2$ is approximated to 1.

Eq. 4 can be extended/modified to model also more generality and less generality. The basic intuition is to revise Eq. 4 taking advantage of the contextual encoding of knowledge in terms of the hierarchy of categories. For instance, less generality can be defined by looking at the overlapping of the sets of documents classified in the descendants of $N_1$ ($C_1$ in Figure 3) and the ancestors of $N_2$ ($A_2$ in Figure 3 ).

*TaxME* is computed starting from three main web directories: 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. The following table summarizes the total amount of processed data.

**Table 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 by which *TaxME* was constructed.

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

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

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

**Step 4** A manual selection was performed with the goal to restrict the assessment of the similarity metric to the subtrees concerning the same topic. 50 pairs of sub trees were selected.

**Step 5** For each of the subtree pairs selected, an exhaustive assessment of correspondences holding between nodes was performed. This was done by exploiting equivalence metric defined by Eq. 4 and the corresponding metrics for less and more generality. The TaxME similarity metric was computed to be the biggest out of the three metrics, namely

$$Sim_{TaxME} = max(Eq(N_1, N_2), Lg(N_1, N_2), Mg(N_1, N_2)) \tag{5}$$

where $Lg$ and $Mg$ denote less and more generality metrics respectively.

The distribution of mappings constructed using $Sim_{TaxME}$ is depicted in Figure 4.

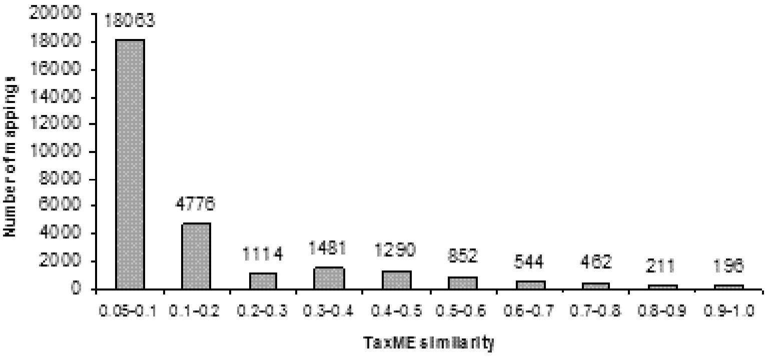

**Fig. 4.** Distribution of mappings according to TaxME similarity metric

Notice that the *TaxME* dataset is very robust to the change of its metric. The number of mappings is in fact very stable and it grows substantially, of two orders of magnitude, only with a very small value of the metric (less than 0.1). As a pragmatic decision, the mappings with TaxME similarity metric above 0.5 were taken to constitute the golden standard $H$. This results in a set of 2265 reference mappings, half of which are equivalence relationships and half are generalization relationships. As depicted in Figure 5, *TaxME* is incomplete in the sense it contains only part of the mappings holding between the graph structures. The key difference with Figure 2 is the fact that a complete golden standard (the area inside the dotted circle in Figure 5) is simulated by exploiting an incomplete one (the area inside the dashed circle in Figure 5).

However, if we assume that *TaxME* is a good representative of $H$ we can use Eq. 2 for an estimation of Recall. In order to ensure that this assumption holds a set of requirements to be satisfied by *TaxME* can be defined [1]: [7]

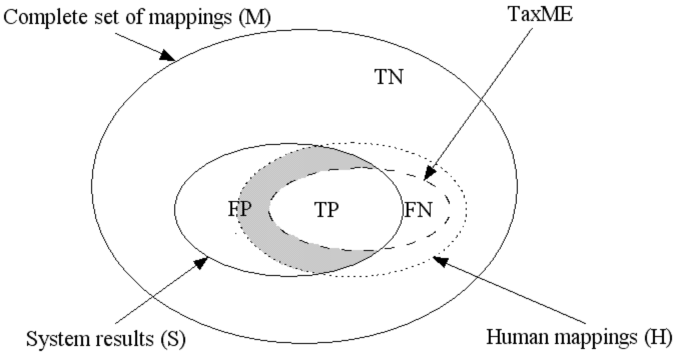

**Fig. 5.** Mapping comparison using *TaxME*. $TP$, $FN$ and $FP$ stand for true positives, false negatives and false positives in respect with *TaxME*

1. *Correctness*, namely the fact that $TaxME \subset H$ (modulo annotation errors).
2. *Complexity*, namely the fact that state of the art matching systems experience difficulties when run on *TaxME*.
3. *Discrimination Capability*, namely the fact that different sets of mappings taken from *TaxME* are hard for the different state of the art systems.

As discussed in [1], *TaxME* satisfies these requirements.

## 4 A dataset for evaluating Precision

As from Eq 1 in order to evaluate Precision we need to know $FP$, which in turn requires that we know $H$. However, as from Section 2, computing $H$ in the case of a large scale matching task requires an implausible human effort. Notice also that we can not either use an incomplete golden standard composed only from positive mappings, e.g. $TaxME$. In fact, as shown in Figure 5, FP can not be computed. This is the case because $FP_{unknown} = S \cap (H - TaxME)$, marked as a gray area in Figure 5, is not known (we do not know how to compute $H$).

[7] [1] introduced also the *Incrementality* property. This property states that a dataset allows for the incremental discovery of the weaknesses of the tested systems. We do not consider this property here because it is irrelevant to our goals.

Our proposal in this paper is to construct a golden standard for the evaluation of both Recall and Precision, let us call it $TaxME\,2$, as follows:

$$TaxME\,2 = TaxME \cup N_{T2} \tag{6}$$

where $N_{T2}$ is an incomplete golden standard composed only of negative mappings (i.e., $N_{T2} \subset M - H$ see Figure 5). Of course $TaxME\,2$ must be a good representative of $M$ and therefore satisfy the three 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 mapping 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 $TaxME$ (i.e., $|N_{T2}| \geq |TaxME|$).

We construct $N_{T2}$ in two steps. In the first step, as depicted in Figure 7, a subset $M'$ of $M$ is selected so that $M'$ contains a big number of "hard" negative mappings. Intuitively a "hard" negative mapping is the mapping with high value of similarity measure which is incorrect according to manual annotation. Consider again the web directories used to construct *TaxME* and $Sim_{TaxME}$. We have randomly selected 100 mappings ranging over various $Sim_{TaxME}$ values and manually evaluated their correctness. Notice that this results in a relatively small amount of manual work as there are only about one thousand of mappings to be analyzed. The results are presented in Figure 6.

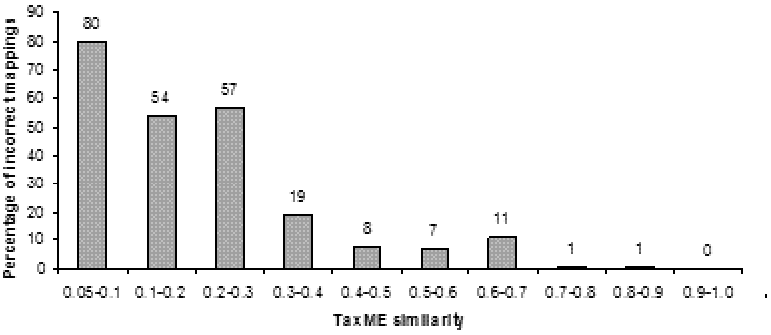

**Fig. 6.** Distribution of incorrect mappings. Each column is calculated evaluating 100 randomly selected mappings

The evaluation shows that *TaxME* is very robust as:

- it is very stable with a small percentage of incorrect mappings for a very large range [0.3,1];
- the number of incorrect mappings becomes substantial for very small values of TaxME similarity metric, namely with threshold less than 0.1.

Taking in account the requests of complexity and scalability we have selected the mappings with TaxME similarity in the 0.05-0.2 range. As from Figure 4, this allowed us to obtain 18063+4776=22836 candidate mappings.

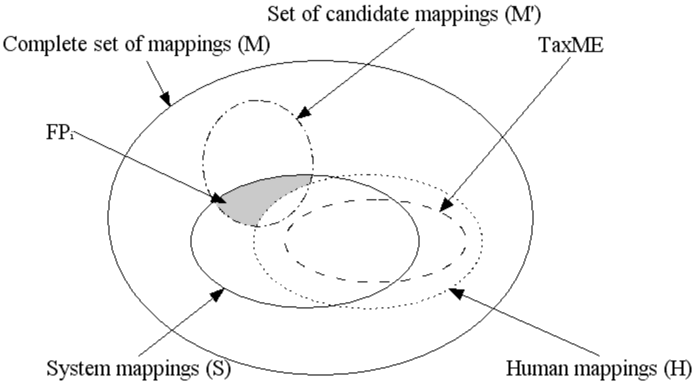

**Fig. 7.** Mapping sets in *TaxME 2*. The gray area stands for $FP_i$ a set of FP produced by the i-th matching system on $M'$

Notice that at this point it is unclear whether $M'$ contains a large enough number of negative mappings. This will be shown in the second step, where the subset $N_{T2}$ of $M'$ is selected. This is done according to the following requirements:

1. Construct $N_{T2}$ from the FPs computed by running state of the art matching systems on $M'$. This ensures that $N_{T2}$ will be hard for all existing systems. Notice that determining whether a mapping produced by a matching system is in FP requires human annotation.
2. Select heterogeneous matching systems, namely systems which make mistakes on different sets of mappings.
3. The selected systems should be representatives of the different classes of the existing matching techniques. This should prevent $N_{T2}$ from being biased towards a particular class of matching solutions.
4. Construct $N_{T2}$ as $N_{T2} = \bigcup_i FP_i$, where $FP_i$ is the FP produced by i-th matching system on $M'$, as depicted in Figure 7. This ensures that $N_{T2}$ is hard for each of the systems and it is also discriminative.
5. The number of FPs produced by each of the systems should be comparable in order to prevent bias towards a particular class of matching solutions.

We have selected three matching systems COMA [13], Similarity Flooding (SF) [17] and S-Match (SM) [11]. The first, as from [1, 11], is arguably the best syntactic matching system. The matching process proposed in COMA has been further extended in [5] and parts of it have been reused in the number of matching systems including [15]. SF utilizes a matching algorithm based on the ideas of similarity propagation. SF computes an initial mapping exploiting a string based matcher. Then the mapping is refined using fix-point computation and filtered according to some predefined criteria. The idea of similarity propagation have been further reused in [9] where the fix point algorithm is exploited for solving the system of linear equations. The SF mapping filtering techniques have been further reused in the system described in [12]. S-Match [8] [11] differs from SF and COMA as it implements semantic matching approach, as described in Section 1. Other semantic matching systems, similar to S-Match, are [3, 4].

We have manually evaluated the mappings found by running COMA, SF and S-Match on $M'$ and computed FP. Notice that here when evaluating the matching quality we have not distinguished among different semantic relations. For example, the mappings $A \sqsubseteq B$ 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 human judgement.

Table 2 provides a quantitative description of the content of *TaxME 2*, and of the effort needed to build it. As from the first row of Table 2 the total number of annotated mappings is 2553+2163+2151=6867. Notice that this is 6 orders of magnitude lower than the number of mappings to be considered in the case of complete golden standard. Notice also that the number of mappings per system is very balanced, as required. Figure 8 how the FPs found by the systems are partitioned.

**Table 2.** Total number of mappings and number of FP on $M'$

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

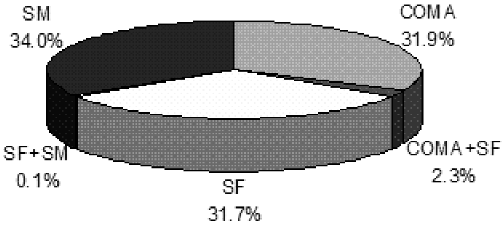

**Fig. 8.** Partitioning of the FPs on $M'$

[8] In the evaluation discussed in this paper we have used the basic version of S-Match and not the enhanced version described in [1].

As from Figure 8, there are no FPs found by SM, COMA and SF together, or by SM and COMA together. There are the small intersections between the FPs produced by SM and SF (0.1%) or by COMA and SF (2.3 %). These results justify our assumption that all 3 systems belong to different classes.

The final result is that $N_{T2}$ consists of 2374 mappings. Notice that the size of $N_{T2}$ is not equal to the sum of the FPs reported in the second row of Table 2 since there is, as from Figure 8, some intersection among these sets. The union of $N_{T2}$ with $TaxME$ has allowed us to compute a golden standard $TaxMe\ 2$, which can be used for the evaluation of both Recall and Precision, of 2265+2374=4639 mappings.

## 5 Evaluating the dataset

This evaluation is designed in order to assess the Complexity and Discrimination Capability of $TaxMe$ 2. This evaluation is done exploiting 6 state of the art systems (Falcon [14], Apfel[6], CMS[15], ctxMatch2[4], OLA [9]and OMAP [20]). For all the systems we use the default settings or, if applicable, the settings provided by the authors for the OAEI-2005 [7] evaluation. We compare these results with the results obtained by the systems exploited in the dataset construction process (COMA, SF and SM). The evaluation results, in terms of TP and FP, are presented in Table 3.

**Table 3.** Number of FP and TP on *TaxME 2* dataset

| | Falcon | Apfel | CMS | ctxMatch 2.2 | OLA | OMAP | COMA | SF | SM |
|---|---|---|---|---|---|---|---|---|---|
| FP | 1313 | 670 | 367 | 299 | 1356 | 1113 | 870 | 776 | 781 |
| TP | 706 | 269 | 319 | 298 | 724 | 694 | 876 | 218 | 669 |

### 5.1 Complexity

Figure 9 presents the Precision, Recall and F-Measure of the systems. As from the figure the maximum Precision results are about 0.5, a value which is significantly lower than the results obtained with the previous datasets. For example, the average Precision demonstrated by Falcon, FOAM, CMS and OMAP on the real world part of the systematic tests (problems 301, 302, 303, 304) in the OAEI-2005 evaluation [7] was in the 0.91-0.93 range.

The Recall results mostly replicate the results presented in [1, 7]. The F-Measure results are more interesting since they demonstrate the aggregated matching quality. As from Figure 9, the best F-Measure is 0.44 what is much lower than the previously reported values for the systems taking part in the evaluation, previously reported in other papers. The other interesting observation is that on dataset construction process (COMA,SF,SM) demonstrate a performance which is comparable with the other systems. In fact all evaluated systems have experienced the same problems as COMA, SF and SM. This justifies the claim that $TaxME\ 2$ is very hard for the state of the art matching systems.

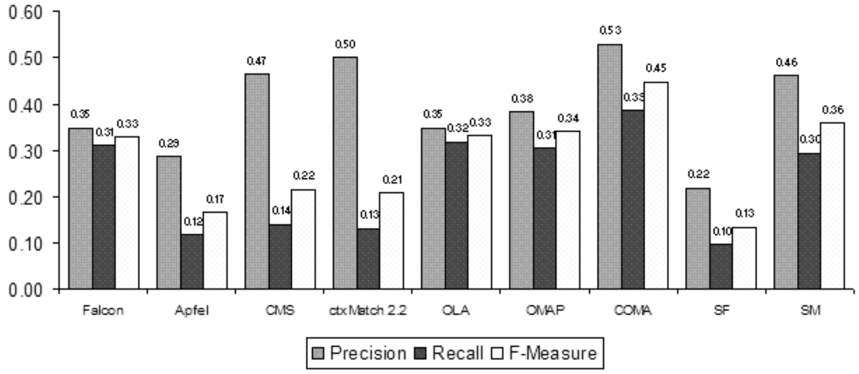

**Fig. 9.** Evaluation results. Precision, Recall and F-Measure on *TaxME 2*

### 5.2 Discrimination Capability

Consider Figure 10. It describes how the FPs in *TaxME 2* are partitioned according to the results of Falcon, Apfel, CMS, ctxMatch2, OLA and OMAP. Various systems experience difficulties on various parts of the dataset and only 3% of the mappings are computed as FP by all six matching systems. This shows that $TaxME$ 2 is difficult for the different systems in different ways.

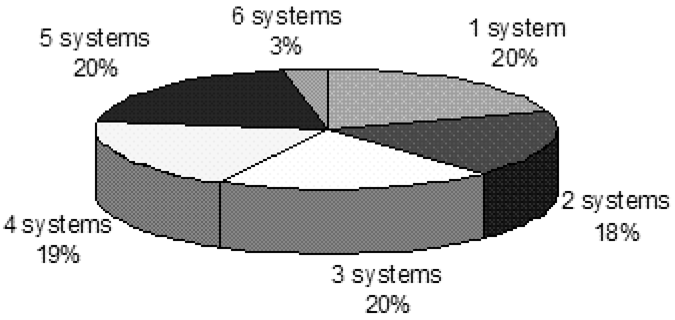

**Fig. 10.** Partitioning of the FPs found by the 6 matching systems according to the number of systems which have found them

## 6 Conclusion and Future Work

In this paper we have presented a large scale mapping dataset constructed starting from the Google, Yahoo and Looksmart web directories. The dataset allows for the evaluation of Precision and Recall. Nine state of the art matching solutions were evaluated using $TaxME$ 2. The evaluation results highlight the fact that the dataset posses the key important properties of Correctness, Complexity and Discrimination capability.

As a future work we are going to investigate the mapping dataset construction process in the case of ontologies which are more complex than simple taxonomies. The other promising direction of research is devoted to the further automation of the mapping dataset construction process. The ultimate goal in this direction is to minimize the human effort and increase the size of the datasets.

## Acknowledgment

This work has been partially supported by the European Knowledge Web network of excellence (IST-2004-507482) and by the STREP Open Knowledge (www.openk.org).

## References

1. P. Avesani, F. Giunchiglia, and M. Yatskevich. A large scale taxonomy mapping evaluation. In *Proceedings of International Semantic Web Conference (ISWC)*, 2005.
2. S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, (28(1)):54–59, 1999.
3. P. Bouquet, L. Serafini, and S. Zanobini. Semantic coordination: a new approach and an application. In *Proceedings of 2nd international semantic web conference (ISWC 2003)*, Sanibel Island, Florida, 20-23 October 2003.
4. P. Bouquet, L. Serafini, and S. Zanobini. Bootstrapping semantics on the web: Meaning elicitation from schemas. In *Proceedings of 15nd international World Wide Web conference (WWW 2006)*, Edinburgh, UK, 23-26 May 2006.
5. M. Ehrig and S. Staab. QOM - quick ontology mapping. In *Proceedings of the Third International Semantic Web Conference*, pages 683–697, Hiroshima, Japan, November 2004.
6. M. Ehrig, S. Staab, and Y. Sure. Bootstrapping ontology alignment methods with apfel. In *Proceedings of International Semantic Web Conference (ISWC)*, 2005.
7. J. Euzenat, H. Stuckenschmidt, and M. Yatskevich. Introduction to the ontology alignment evaluation 2005. In *Proceedings of K-CAP 2005 Workshop on Integrating Ontologies*, 2005.
8. J. Euzenat and P. Valtchev. An integrative proximity measure for ontology alignment. In *Proceedings of Semantic Integration workshop at International Semantic Web Conference (ISWC)*, 2003.
9. J. Euzenat and P. Valtchev. Similarity-based ontology alignment in OWL-lite. In *Proceedings of European Conference on Artificial Intelligence (ECAI)*, pages 333–337, 2004.
10. F. Giunchiglia and P. Shvaiko. Semantic matching. *The Knowledge Engineering Review Journal*, (18(3)):265–280, 2003.
11. F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-Match: an algorithm and an implementation of semantic matching. In *Proceedings of 1st european semantic web symposium (ESWS'04)*.
12. A. Hess. An iterative algorithm for ontology mapping capable of using training data. In *Proceedings of the Third European Semantic Web Conference*, Budva, Montenegro, 2006.
13. H.H.Do and E. Rahm. COMA - a system for flexible combination of schema matching approaches. In *Proceedings of Very Large Data Bases Conference (VLDB)*, pages 610–621, 2001.
14. N. Jian, W. Hu, G. Cheng, and Y. Qu. FalconAO: Aligning ontologies with Falcon. In *Proceedings of K-CAP 2005 Workshop on Integrating Ontologies*, 2005.

15. Y. Kalfoglou and B. Hu. Crosi mapping system (cms). In *Proceedings of K-CAP 2005 Workshop on Integrating Ontologies*, 2005.
16. D. L. McGuinness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. In *Proceedings of International Conference on the Principles of Knowledge Representation and Reasoning (KR)*, pages 483–493, 2000.
17. S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm. In *Proceedings of International Conference on Data Engineering (ICDE)*, pages 117–128, 2002.
18. P. Mitra, N.F. Noy, and A.R. Jaiswal. Ontology mapping discovery with uncertainty. In *Proceedings of International Semantic Web Conference (ISWC)*, 2005.
19. N. Noy and M. A. Musen. Anchor-prompt: Using non-local context for semantic matching. In *Proceedings of workshop on Ontologies and Information Sharing at International Joint Conference on Artificial Intelligence (IJCAI)*, pages 63–70, 2001.
20. U. Straccia and R. Troncy. omap: Combining classifiers for aligning automatically owl ontologies. In *Proceedings of 6th International Conference on Web Information Systems Engineering (WISE'05)*, 2005.