# Ontology Alignment - Karlsruhe

Marc Ehrig and York Sure

Institute AIFB, University of Karlsruhe

## 1 Introduction

This short paper accompanies the EON Ontology Alignment Contest at the International Semantic Web Conference 2004 in Hiroshima. First, it describes the system as used in Karlsruhe. The results of the actual experiments follow. Then we have a short discussion and interpretation of the results. In the end we provide a link to the actual results.

## 2 System

### 2.1 Alignment Process

We introduce the process of ontology alignment we use for the experiments here. The general process and an efficient algorithm for it have been presented in [ES04b,ES04a]. The presented approach is only capable of extracting semantical identity. Subsumption or complex alignments are currently not supported.

To find out whether two entities can be aligned we rely on rules which indicate support for an alignment. A rule consists of two main parts: an entity feature, and a comparison function. The result is a value between 0 and 1.

An entity is described through features which are represented in the ontology. A feature can be as simple as the label of entity. It can include intensional structural elements such as super- or sub-concepts for concepts, or domain and range for relations. Further we use extensional elements: a concept is described through its instances. Instance features can be instantiated attributes. According to our assumption we restrict our focus on these simple features, where simple means that only directly connected relations are respected, i.e. we do not consider the relations of a sub-concept of a super-concept of a concept. The following example will clarify the used features.

```
<owl:Class rdf:about=''automobile''/>
<owl:Class rdf:about=''car''/>
<rdf:Description rdf:about=''porsche''>
    <rdf:type rdf:resource=''automobile''>
    <rdf:type rdf:resource=''car''>
    <auto:speed rdf:resource=''fast''>
</rdf:Description>
<rdf:Description rdf:about=''mercedes''>
    <rdf:type rdf:resource=''automobile''>
    <auto:speed rdf:resource=''fast''>
</rdf:Description>
```

**Example 1.** Two entities with the similar super-concepts.

In Example 1 both `Porsche` and `Mercedes` have the feature `type`. Extracting the instantiation of the feature we receive `automobile` and `car`, respectively only `automobile`. Further as stated before also domain-ontology features can be included through this step e.g. `auto:speed`, which is `fast` for both the first and second entity.

Next, the comparison of objects can normally be expressed through exact logical rules. Equality can easily be checked by comparing the unique resource identifiers (URI); for sets one can use the Dice-coefficient [CAFP98]. According to our assumption this is not suitable for the alignment case. We need inexact heuristical ways of comparing objects. These are functions expressing similarity on strings [Lev66], object sets [CC94], checks for inclusion or inequality, rather than exact logical identity. The heuristics can also make use of additional background knowledge e.g. the use of WordNet [Fel99] when checking the identity of labels of entities. Returning back to our example we use a similarity function based on the instantiated results, i.e. we check whether the two concept sets parent concepts of `Porsche` (`automobile` and `car`) and of `Mercedes` (only `automobile`) are the same. In the given case this is true to a certain degree, effectively returning a similarity value of 0.5. Extending the approach of previous work, we do not only compare the same features of two entities, but also allow different ones, e.g. comparing sub-concepts of one concept with super-concepts of another one (which have to be disjunct to support the equality of two entities).

Finally the individual feature-heuristic combinations indicating an alignment are combined to calculate the overall support function. This can be achieved through a simple averaging step, but also through complex aggregation functions. Thereafter the value is interpreted and the algorithm decides whether the entity pair actually aligns i.e. the two entities are semantically identical. The weights for the combination and the interpretation threshold are set manually once i.e. they are the same for every mapping run.

Further we iterate over the whole process in order to bootstrap the amount of structural knowledge. In the given approach we perform five rounds.

### 2.2 System

The presented approach has been implemented in Java using the capabilities of the KAON[1]-framework. The framework uses its own KAON-ontology format, but can handle RDFS imports. Using KAON allows for easy use and extension of ontology technology.

### 2.3 Adaptations for the Contest

There were no substantial adaptations to the actual system for the contest. Minor adaptations were needed for input and output.

The input ontologies had to be transformed into a plain RDF(S)-ontology-format only. For the given case this meant the following steps:

- Transformation of owl:Class into rdfs:Class

[1] http://kaon.semanticweb.org

- Replacement of owl:Restriction through rdfs:Class. This meant that every restriction is treated as an own (anonymous) class.
- Transformation of owl:ObjectProperty and owl:DatatypeProperty into rdf:Property.
- Further, blank nodes originating from restrictions or unions were materialized. Otherwise they would have not been internally alignable, thus reducing the alignment quality of derived classes.

For the output the saving method was adjusted in the following way:

- Deletion of KAON internal constructs, mostly language identifiers.
- Export of alignments conforming to the given alignment format.

## 3 Results

In this section we present the results of the individual alignment experiments. As the system can neither handle OWL constructs nor alignments beyond identity, we expect to lose quality in comparison with systems that do handle both.

### 3.1 Experiments

Please note, that we have not performed every ontology pair of the experiment. Considering the presented algorithm and the restrictions of RDF(S), the skipped experiments were either identical to others or only provided very marginal differences. Additionally the used KAON system adds own language constructs which were filtered out for the evaluation.

### 3.2 Measures

We use standard information retrieval metrics to assess the different ontology pairs (cf. [DMR02]):

Precision $p = \frac{\#correct\_found\_alignments}{\#found\_alignments}$

Recall $r = \frac{\#correct\_found\_alignments}{\#existing\_alignmentss}$

F-Measure $f_1 = \frac{2pr}{p+r}$, the harmonic mean of precision and recall

Depending on the application scenario different evaluation values can be of importance. The following results therefore always indicate the results for the best possible precision, the best possible f-measure, and the best possible recall.

Nevertheless to allow strict comparability across the EON tests and with other algorithms we provide the numbers for a fixed similarity threshold across all tests, in our case this was 0.148. The threshold has been determined through a big number of tests beyond this EON experiment. If no other information about the application is given these numbers can be seen as the most objective and comparable ones.

### 3.3 101 concept test: ID

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 1.0 | 1.0 |
| best f-measure | 1.0 | 1.0 | 1.0 |
| best recall | 1.0 | 1.0 | 1.0 |
| fixed threshold | 1.0 | 1.0 | 1.0 |

The system loads the ontologies. If the URIs of two entities are identical, which is the case here, this automatically means that the entities are identical. No alignment is required for this.

### 3.4 102 concept test: ?

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | undefined | undefined | undefined |
| best f-measure | undefined | undefined | undefined |
| best recall | undefined | undefined | undefined |
| fixed threshold | undefined | undefined | undefined |

No alignments were retrieved by the system. Having no outcome it is not possible to calculate the presented evaluation measures. This is the expected outcome when trying to align two ontologies of completely different domains.

### 3.5 201 systematic: no names

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.2826 | 0.4407 |
| best f-measure | 0.9057 | 0.5217 | 0.6621 |
| best recall | 0.9057 | 0.5217 | 0.6621 |
| fixed threshold | 0.9524 | 0.4348 | 0.5971 |

This represents one of the most difficult runs. The entry point of our process is identification of equal entities through similar labels. Only then further alignments can be extracted via the structure. This is especially a problem for our system, because apart from the labels we do not consider further documentation or comments.

### 3.6 204 systematic: naming conventions

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 1.0 | 1.0 |
| best f-measure | 1.0 | 1.0 | 1.0 |
| best recall | 1.0 | 1.0 | 1.0 |
| fixed threshold | 1.0 | 1.0 | 1.0 |

Smaller differences in label naming are either already filtered out by the used syntactic similarity of strings based on the edit distance. Apart from this entity pairs experience massive structural identity support from neighboring entity pairs, who have previously been identified as equal.

### 3.7 205 systematic: synonyms

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.1196 | 0.2136 |
| best f-measure | 0.8730 | 0.5978 | 0.7097 |
| best recall | 0.8261 | 0.6196 | 0.7125 |
| fixed threshold | 0.8710 | 0.5870 | 0.7012 |

Synonyms are more challenging than the inexact naming of the previous data set. Our system can only identify alignments based on similar labels at first. Currently we do not use any dictionary. Having completely different words means no matches in the first place.

### 3.8 206 systematic: foreign names

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.0109 | 0.0215 |
| best f-measure | 0.9677 | 0.6522 | 0.7792 |
| best recall | 0.9677 | 0.6522 | 0.7792 |
| fixed threshold | 0.98 | 0.5326 | 0.6901 |

There is basically no difference for the system of not knowing a synonym term or not knowing a term because it is written in a different language. It seems that French expressions are more similar to the basic English ontology than the English synonyms are.

### 3.9 223 systematic: expanded hierarchy

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.9783 | 0.9890 |
| best f-measure | 1.0 | 0.9783 | 0.9890 |
| best recall | 0.9891 | 0.9891 | 0.9891 |
| fixed threshold | 0.9891 | 0.9891 | 0.9891 |

The additional classes in the hierarchy don’t affect the system very much. Obviously the labeling is strong enough to not be distracted by additional classes.

### 3.10 224 systematic: no instances

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 1.0 | 1.0 |
| best f-measure | 1.0 | 1.0 | 1.0 |
| best recall | 1.0 | 1.0 | 1.0 |
| fixed threshold | 1.0 | 1.0 | 1.0 |

Again the labels seem to be strong enough themselves. Removing instances doesn’t affect the outcome. However, it will though when the labels become less reliable as in the synonym or foreign language case.

### 3.11 230 systematic: flattening entities

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 1.0 | 1.0 |
| best f-measure | 1.0 | 1.0 | 1.0 |
| best recall | 1.0 | 1.0 | 1.0 |
| fixed threshold | 1.0 | 1.0 | 1.0 |

Removing some entities doesn't affect the evaluation. Besides the fact of having less alignments, all aligning is still correct.

### 3.12 301 real ontology: BibTeX/MIT

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.0806 | 0.1471 |
| best f-measure | 0.9231 | 0.3871 | 0.5455 |
| best recall | 0.9231 | 0.3871 | 0.5455 |
| fixed threshold | 0.9231 | 0.3871 | 0.5455 |

The real interesting evaluations are, whether the system can automatically align two different ontologies covering the same domain. Unfortunately the results are not too promising. We do not receive high levels of quality. Even if we remove the special alignment subsumption, which our system can't extract, the results are not too good. Obviously the good labels are superposed by the big differences in structure. One consequence could be to reduce the impact of structure. People describe their world through natural language rather than through intensional structures. As another consequence one could raise the question: Despite the same labels, are the entities actually the same, as they have a very different intensional structure?

### 3.13 302 real ontology: BibTeX/UMBC

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.2 | 0.3333 |
| best f-measure | 1.0 | 0.2 | 0.3333 |
| best recall | 1.0 | 0.2 | 0.3333 |
| fixed threshold | 1.0 | 0.2 | 0.3333 |

The same comments as for the previous experiment still hold. Having a look at the UMBC ontology one can see that they have many bibliographic properties pointing from publications to literals/Strings. With the label being different and the range of a property being different the system can not identify these relations as equal. The classes on the other hand were well identified. Interesting for this one is the very high precision with a very low recall. Our system obviously prefers losing information for the sake of keeping the error number small. If the alignments are used for further mapping with inferencing high error rates can seriously harm the results.

### 3.14 303 real ontology: Karlsruhe

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.24 | 0.3871 |
| best f-measure | 0.9512 | 0.78 | 0.8571 |
| best recall | 0.9512 | 0.78 | 0.8571 |
| fixed threshold | 0.9286 | 0.78 | 0.8478 |

Karlsruhe and INRIA seem to have a closer understanding of the bibliographic world than MIT and UMBC with INRIA.

### 3.15 304 real ontology: INRIA

| | precision | recall | f-measure |
|---|---|---|---|
| best precision | 1.0 | 0.7692 | 0.8670 |
| best f-measure | 0.9867 | 0.9487 | 0.9673 |
| best recall | 0.9494 | 0.9615 | 0.9554 |
| fixed threshold | 0.9867 | 0.9487 | 0.9673 |

Comparing one INRIA ontology with another INRIA one yields very good results, as one can expect.

## 4 General Comments

### 4.1 Lessons Learnt

From the experiments one can generally say that labels have a high value for the alignment process. If these labels are the same the probability that they actually indicate the same is very high. Humans tend to express the same things through the same natural language; structure is only modelled in the second place. When having two ontologies where many labels are the same it might make sense to stop at this point or at least fix the alignments before continuing with the structure.

On the other hand if the labels differ, just a little or even considerably, the structural elements become more important. Through structural similarity the differences in labelling can be overcome to a certain degree.

### 4.2 Comments on the test cases

It is always difficult to create good test cases. The cases covered a wide range of discrepancies occurring when having two ontologies. For the cases always one feature was explicitly changed and the alignment algorithm had to cover this. However in a real world scenario, as presented by the last four test cases, the ontologies vary in many features at the same time. There will always be an optimal algorithm to circumvent one feature of ontology mismatch. The real challenge is to manage many differences at the same time. Focus should therefore lie more on real world ontologies.

### 4.3 Comments on measures

The measures always reflect the goal one wants to reach. In this paper the measures were purely qualitative. And even there it is difficult to define a *good* result. For querying retrieving more false answers is tolerable, as long as the correct answers are included (high recall). If further inferencing is the goal, every mistake seriously changes the whole result. Here we need a high precision. The used f-measure is a compromise to handle both scenarios adequately. It might also make sense to measure several points rather than just the f-measure.

Besides the qualitative measures it can also be interesting to measure scalability and efficiency (e.g. for on the fly alignment in peer-to-peer systems).

## 5 Raw Results

The results presented in this paper can be downloaded at http://www.aifb.uni-karlsruhe.de/WBS/meh/mapping/

Not all alignments in the result files are actually interpreted as an alignment; too low values are removed through the cut-off measure. Nevertheless, for better interpretability the whole files were put online.

For a better overview the results of all the tests are shown once more in one table, each time using the fixed similarity threshold:

| | precision | recall | f-measure |
|---|---|---|---|
| 101 ID | 1.0 | 1.0 | 1.0 |
| 102 ? | undefined | undefined | undefined |
| 201 no names | 0.9524 | 0.4348 | 0.5971 |
| 204 naming conventions | 1.0 | 1.0 | 1.0 |
| 205 synonyms | 0.8710 | 0.5870 | 0.7012 |
| 206 foreign names | 0.98 | 0.5326 | 0.6901 |
| 223 expanded hierarchy | 0.9891 | 0.9891 | 0.9891 |
| 224 no instances | 1.0 | 1.0 | 1.0 |
| 230 flattening entities | 1.0 | 1.0 | 1.0 |
| 301 BibTeX/MIT | 0.9231 | 0.3871 | 0.5455 |
| 302 BibTeX/UMBC | 1.0 | 0.2 | 0.3333 |
| 303 Karlsruhe | 0.9286 | 0.78 | 0.8478 |
| 304 INRIA | 0.9867 | 0.9487 | 0.9673 |

## 6 Conclusion

We can draw the following conclusion from the experiment. The Karlsruhe process performs differently on the various kinds of ontology pairs. Even though the results were not too bad in general, some optimization, possibly by performing different actions for different alignment steps. Nevertheless ontology alignment will never yield 100% correct results, and any application relying thereon should be aware of this.

## References

[CAFP98] S. V. Castano, M. G. De Antonellis, B. Fugini, and C. Pernici. Schema analysis: Techniques and applications. *ACM Trans. Systems*, 23(3):286–333, 1998.

[CC94] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman and Hall, 1994.

[DMR02] H. Do, S. Melnik, and E. Rahm. Comparison of schema matching evaluations. In *Proceedings of the second int. workshop on Web Databases (German Informatics Society)*, 2002.

[ES04a] Marc Ehrig and Steffen Staab. QOM - quick ontology mapping. In *Proceedings of the Third International Semantic Web Conference (ISWC-2004)*, Hiroshima, Japan, 2004.

[ES04b] Marc Ehrig and York Sure. Ontology mapping - an integrated approach. In *Proceedings of the First European Semantic Web Symposium, ESWS 2004*, volume 3053 of *Lecture Notes in Computer Science*, pages 76–91, Heraklion, Greece, May 2004. Springer Verlag.

[Fel99] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. 1999.

[Lev66] I. V. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 1966.