Relevance-Based Evaluation of Alignment Approaches: the OAEI 2007 food task revisited

Willem Robert van Hage[1,2], Hap Kolb[1], and Guus Schreiber[2]

[1] TNO Science & Industry, Stieltjesweg 1, 2628CK Delft, the Netherlands,
hap.kolb@tno.nl
[2] Vrije Universiteit Amsterdam, de Boelelaan 1081a, 1081HV Amsterdam, the Netherlands,
wrvhage@few.vu.nl, schreiber@cs.vu.nl

Abstract. Current state-of-the-art ontology-alignment evaluation methods are based on the assumption that alignment relations come in two flavors: correct and incorrect. Some alignment systems find more correct mappings than others and hence, by this assumption, they perform better. In practical applications however, it does not only matter *how many* correct mappings you find, but also *which* correct mappings you find. This means that, apart from correctness, relevance should also be included in the evaluation procedure. In this paper we demonstrate how to incorporate relevance in sample evaluation of alignment approaches by using high relevancy to a set of prototypical search tasks as a selection criterion when drawing sample mappings. We expand the sample-based evaluation of the OAEI 2007 *food task* with relevance-based evaluation and compare the results of this new evaluation method to the existing results. This leads to new insights on the performance of the participating ontology-alignment systems in practice.

1 Introduction

In recent years ontology alignment has become a major field of research [3, 6, 11]. Especially in the field of digital libraries it has had a great impact. Many libraries have made the transition to offer access to their resources through the web. This has made it possible to access multiple collections at the same time. Different libraries have different indexing schema's and protocols. This complicates federated access. Alignment offers a way to bridge the semantic gap between the indexing schema's so that users can profit from their joint coverage.

Good evaluation of alignment approaches is important. In past decades, research communities that focus on other complex computer-science subjects, such as natural-language processing and information retrieval, have developed suitable evaluation methods. Some of their methods in these communities are applicable to ontology alignment and have been adopted in recent years by evaluation efforts such as the Ontology Alignment Evaluation Initiative (OAEI). The main contribution of this paper is to improve the evaluation methodology of alignment to better capture the performance of alignment approaches in actual applications. We introduce a simple evaluation method, *relevance-based evaluation*, that remedies some of the shortcomings of existing methods by using a sampling technique that takes the needs of users into account. We apply this method to the data of the OAEI 2007 *food task* [2].

In section 2 we discuss existing evaluation methods. In section 3 we describe our new method, relevance-based evaluation. In section 4 we describe the procedure we followed to apply relevance-based evaluation to alignment in the agricultural domain. In section 5 we go into detail on every step of this procedure and the data sets that were involved. In section 6 we show the results of relevance-based evaluation on the OAEI 2007 *food task* data and compare them to the existing results. We test how these new results, based on our "second opinion", differ from the old results and we draw conclusions about the validity of the OAEI 2007 *food task* results.

2 Alignment Evaluation

Nearly all existing evaluation measures used to determine the quality of alignment approaches are based on counting mappings [1, 2]. For instance, in the context of ontology alignment, the definition of Recall is defined as the number of correct mappings a system produces divided by the total number of correct mappings that can possibly be found (*i.e.* that are desired to be part of the result). Regardless of their differences, most of these measures have one thing in common: They do not favor one mapping over the other in order to give an objective impression of system performance. Any mapping could prove to be important to some application. Therefore, they can only tell us *how many* mappings are found on average by a system, but not *which* mappings are found and whether the mappings that are found are those that are useful for a certain application. Whenever someone wants to decide which alignment approach is best suited for his application (*e.g.* [8]) he will have to reinterpret average expected performance in the light of his own needs. This can be a serious obstacle for users.

A solution to this problem is to incorporate the importance of mappings (*i.e.* relevance) into the evaluation result. This solution immediately raises two new problems:

1. How to come up with suitable importance weights
2. How to define a simple and intuitive way to use these weights

With respect to problem 1, there are many sensible ways to weigh the importance of mappings. One possibility is to assign weights that are independent of how often the mappings are used, but dependent on the size of the logical implication of a mapping, *cf.* Semantic Precision and Semantic Recall [1]. The intuition underlying this method is that mappings with a greater logical consequence have more benefit to users, because more implications can be made using these mappings. However, a mapping might have a large logical consequence while it is never used in a specific application. In this paper we do not account for logical implications. Another possibility is to assign weights according to how often a mapping can be expected to be used [4]. This method makes the assumption that each concept has an equal probability of being used as a query in an application. Under this assumption, Hollink et al. estimate the frequency a mapping will be used based on the distance of a mapping the query concept. In this paper we do not assume a uniform query distribution.

Likewise, with respect to problem 2, there are many sensible ways to incorporate mapping importance into an evaluation method. They can, for example, be used as coëfficients in a linear equation. (*cf.* Kekäläinen's approach to including varying degrees of relevance in information-retrieval evaluation in [7]) Or, in the case of sample evaluation, they can be used to weigh sample sets as a whole (*cf.* the *Alignment Sample Evaluation* method described in [12]).

Another related evaluation approach is described in [9], where the ontology construction process is guided by its effect on end-to-end task performance.

3 Relevance-Based Evaluation

The evaluation method we propose in this paper consists of two steps:

1. **Gather Relevant Mappings** Depending on the application, we determine which mappings are directly involved in achieving the user's goals (*e.g.* finding documents of special importance). These mappings are considered relevant, the rest is considered irrelevant. We gather a set of relevant mappings that reflects typical usage scenarios.
2. **Apply Sample Evaluation of Relevant Mappings** We assume that the selected relevant mappings are representative of all mappings that are useful to the application. We calculate performance scores on the sample of relevant mappings using existing sample-evaluation methods (*e.g.* [12]).

As opposed to existing methods to account for the relevance of mappings that include it as a variable in an evaluation measure, we use relevance to steer the sample-selection process. Instead of randomly selecting mappings for the evaluation of alignment approaches (*cf.* the *food* and *environment tasks* described in [2]) we select *only* those that are relevant to an application. This way we can use existing and well-understood evaluation metrics, like Precision and Recall, to measure performance on important tasks as opposed to fictive average-case performance. The advantages of not adapting the evaluation measure, but influencing the drawing of samples are the following:

- The evaluation measure can remain simple. This makes it easier to interpret what scores mean.
- Using the same evaluation measure for relevance-based as for non-relevance-based evaluation allows us to easily explore how performance in specific applications differs from average-case performance, because only the samples differ.
- Existing experiments can be easily extended to account for new use cases. Additional samples can be added to compensate for the underrepresentation of certain usage scenarios.
- Different sources of relevance estimates can be used besides each other, because the estimation is not part of the evaluation measure.

4 Experimental Set-up

We demonstrate how relevance-based evaluation works by applying it to the existing results of the OAEI 2007 *food task*, which did not take relevance into account. We determine relevance for the mappings based on hot topics related to this task, like global warming and increasing food prices, which we obtain by means of query-log analysis, expert interviews, and news feeds. For the original OAEI 2007 *food task*, Recall was measured on samples that represent the frequency of topics in the vocabularies. For example, if 60% of the concepts in the vocabularies were animal or plant species names, then sample mappings from and to animal and plant species names determined 60% of the end result. In this paper we will repeat the measurement of Recall on samples that represent the relevance of mappings to finding documents on hot topics. For example, most species names, except 'Oryza sativa' (the rice plant) are probably irrelevant to the hot topic of rising rice prices. On the other hand, topics that are covered by few concepts in the vocabularies[3] might prove to be vital to hot topics. Without a specific application it is uncertain which mappings turn out to be important.

The application we choose for this second-opinion evaluation of the systems participating in the OAEI 2007 *food task* is finding documents about prototypical agricultural topics in the one collection using the indexing vocabulary of the other. A similar approach was used for the evaluation of the OAEI 2007 *library task* [5], except that in this case relevancy of mappings was not taken into account.

We implemented the two steps described in section 3 as follows:

Gather Relevant Mappings

1. **Gather topics that represent important use cases.** In this step we research which topics are currently "hot" in agriculture. We gather topics from the query log files of the FAO AGRIS/CARIS search engine, the FAO newsroom website, and interviews with experts from the FAO's David Lubin library and the TNO Quality of Life food-safety group. We manually construct search-engine queries for each topic. Further elaboration can be found in section 5.1.
2. **Gather documents that are highly relevant to the topics.** In this step we ascertain which documents would be sufficient for the hot topics. We gather suitable candidate documents from the part of the FAO AGRIS/CARIS and USDA AGRICOLA reference databases that overlaps. We use a free-text search engine[4] and manually filter out all irrelevant documents, see section 5.2.
3. **Collect the meta-data describing the subject of these documents and align the concepts that describe the subject of the documents to concepts in the other thesaurus.** In this step we determine which mappings are necessary to find these documents. We collect values of the Dublin Core subject field from the AGRIS/CARIS and AGRICOLA reference databases. These values come from subject vocabularies, respectively AGROVOC and the NAL Agricultural Thesaurus. We manually align each concept to the most similar concept in the other vocabulary, see section 5.3. The resulting mappings make up our sample set of relevant mappings.

Apply Sample Evaluation on Relevant Mappings

4. **Count how many of these mappings have been found by ontology alignment systems and compare system performance based on these counts.** We re-calculate

[3] This refers to the number of concepts in the thesaurus on a given subject, not the number of times they are used to index a document.

[4] http://www.fao.org/agris/search

Recall for the top-4 systems of the OAEI 2007 *food task*, following the same procedure as described in [2, 12], but use the new set of relevant mappings. The details and results can be found in section 6.

5 Sample Construction

5.1 Topics

In order to get a broad overview of current affairs in the agricultural domain we gathered topics from three sources: AGRIS/CARIS search log analysis, topics in the “Focus on the issues” section of the FAO Newsroom, and interviews with a food-safety expert at TNO Quality of Life and a reference librarian at the David Lubin Memorial Library of the FAO. A long description of the topics that resulted from these three sources can be found at http://www.few.vu.nl/ wrvhage/om2008/topics.html.

Log analysis The FAO AGRIS/CARIS search engine is used by a broad range of people all around the world: Information scientists at agricultural research facilities, farmers in search of new techniques for their profession, internal FAO information officers, people involved in development and education, and the occasional data mining bot. This means the query log is very heterogeneous. After simple syntactic preprocessing of the queries we sorted them by frequency and selected four topics that were represented by multiple, easily interpretable queries amongst the most frequent queries of the log. Amongst the top queries are many query-syntax mistakes, single-letter queries (*e.g.* `M`, perhaps the initial of an author), stray boolean operators (*e.g* `AND` without actual terms), and spelling mistakes (*e.g.* `babanas`). Many of the most frequent terms are clearly not related to hot topics, like `University` or `title`. For most queries it is impossible to reconstruct the original meaning without unreasonable guessing. For example, the query `rice` does not reveal which aspect of rice was intended. In such cases we searched for queries that contained `rice`, like `paddy rice and fertilizers` or `rice fish system`. When this yielded a connection to current affairs we added it to the list of hot topics. An important reason to practice rice/fish cultivation, for example, is the great reduction of pesticide that it permits.[5]

The hot topics that were selected based on evidence mainly from query log analysis were the following: [6] Avian influenza, Malaria in Africa, Genetic modification of soy, Cattle traceability.

The FAO Newsroom One of the main tasks of the FAO is to disseminate information about agriculture (*i.e.* agronomy, forestry, and fishery) to the world. The Newsroom[7] is one of the channels the FAO uses to reach people around the world. The Newsroom has

[5] see: http://www.fao.org/newsroom/en/news/2005/102401

[6] Detailed descriptions of the topics can be found at http://www.few.vu.nl/ wrvhage/om2008/topics.html.

[7] http://www.fao.org/newsroom

a section about current events.[8] We used this section as a source of hot topics and to verify evidence from query-log analysis and interviews.

The hot topics that were selected based on evidence mainly from the FAO newsroom were the following: [6] Rice and pesticides, The role forestry can play in climate change, Plants and advancing desertification, Biofuels and their effect on corn prices, Biofuels and their effect on water supply.

Expert Opinions Information officers and reference librarians at the FAO in Rome and food-safery researchers at the Netherlands Organisation for Applied Scientific Research (TNO) deal with questions from journalists on a daily basis. Apart from consulting tangible sources of topics we have also consulted these domain experts. Besides confirming the topics we obtained from the other two sources they mentioned these additional issues: [6] Acrylamide found in fried foods, Benzene found in food or drink, Dioxins found in food or drink, The effect of bee extinction on pollination, The effect of fish farming and antibiotics use on wild fish.

5.2 Documents

Per topic we retrieved the top-100 hits of a full-text search on the AGRIS/CARIS search engine limited to the set of documents that is shared between the AGRICOLA and AGRIS/CARIS collections.[9] From these 1500 documents we selected only the ones that are relevant to our topics and that have been assigned Dublin Core subject terms in both collections. This left 52 documents. How many suitable double-annotated documents we found per topic and how many of these were also relevant is shown in table 1. For four of the topics we found no documents that were both relevant and indexed in both collections: Cattle traceability, both topics about biofuels, and the effect of antibiotics in fish farming on wild fish. The reason for this is that these topics are all very new issues. The greatest overlap between the AGRIS/CARIS and AGRICOLA collections exists for documents published between 1985 and 1995. The total number of documents that was imported from AGRICOLA to AGRIS/CARIS per year is shown in figure 1. After the year 2000 no documents have been imported and thus it is hard to find relevant documents for new issues. We assume that the 52 double-annotated relevant documents are representative of the set of all relevant documents with subject meta-data, *i.e.* also the documents with only annotations in one of the two collections. These are the documents for which alignment could make the biggest difference. This is a reasonable assumption, because the indexing process of both collections is regulated by a protocol. The indexing protocol of both libraries differ quite a lot, but within each collection annotations are quite stable. For both libraries it goes that not all documents are indexed, but those that are were indexed by the same protocol.

[8] http://www.fao.org/newsroom/en/focus

[9] This can be accomplished by limiting the search to data from the USDA data center by adding `+center:(US)` to the search query.

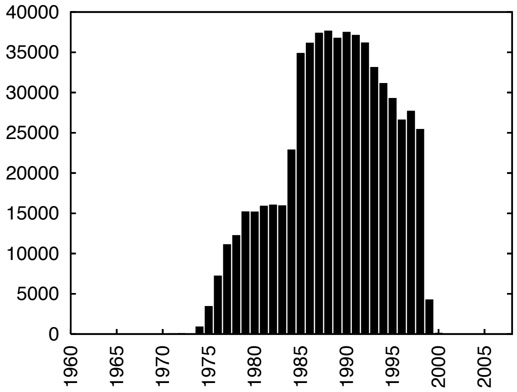

Fig. 1. The number of documents imported by the FAO from the USDA AGRICOLA collection into the AGRIS/CARIS collection per year.

	suitable	suitable	indexed with # concepts		
topic	documents	and relevant	NALT	AGROVOC	mappings
avian influenza	48	9	52	35	72
malaria in africa	51	12	75	67	112
genetic modification of soy	40	1	8	10	12
cattle traceability	34	0	-	-	-
rice and pesticides	41	3	10	5	12
climate change and forestry	21	4	29	25	36
desertification	47	8	21	23	33
biofuels and corn price	35	0	-	-	-
biofuels and water	5	0	-	-	-
acrylamide in fried foods	31	5	20	13	25
benzene in food	13	5	28	24	38
dioxins in food	9	4	26	15	31
bee extinction	62	2	15	5	18
fish farming and antibiotics	1	0	-	-	-

Table 1. Statistics per topic. Shown are the number of double-annotated documents in the top-100 of the AGRIS/CARIS search engine, the number of relevant documents amongst these, the number of indexing terms used for these documents, and the number of mappings this led to for the relevance-based reference alignment.

5.3 Mappings

Now that we have established which documents are potentially important to find, we will decide which mappings will be of most benefit to someone who wants to find them. This can be done with a search engine that employs mappings. There are many possible ways in which such a search engine works. Each retrieval method has strong and weak points. Some methods that apply mappings during retrieval work well with an incomplete set of mappings, others do not. Some make use of the extra synonyms that are made available through mappings, others are geared towards exploiting extra hierarchical relations. To maximize the generalizability of our work, we avoid having to choose a specific retrieval method by making two assumptions about the retrieval methods that will be used.

1. We assume that retrieval methods work best if each concept (in both vocabularies) is aligned to the most similar concept in the other vocabulary.[10]
2. We assume that relevant mappings are all equally important during the retrieval process and that irrelevant mappings are all equally unimportant for retrieval.

The first assumption corresponds exactly to the protocol that was used by human experts to create the reference alignments for the OAEI *food* and *environment tasks*. The second merely states that we use boolean weights for relevancy or, specifically, that we will create a sample set of only relevant mappings.

Given this, the set of mappings that works best for finding the 53 relevant documents is the set that aligns each of the describing concepts with its most similar counterpart. For example, if a document is indexed with the concepts agrovoc:chickens and agrovoc:frying in AGRIS/CARIS and with nalt:chickens and nalt:fried_foods in AGRICOLA then the ideal set of mappings for this document is:

```
agrovoc:chickens  skos:exactMatch  nalt:chicken  .
agrovoc:frying  skos:exactMatch  nalt:frying  .
agrovoc:foods  skos:narrowMatch  nalt:fried_foods  .
```

In this way we manually mapped the 266 NALT concepts and 212 AGROVOC concepts, see table 1, to their counterpart in the other thesaurus with the help of thesaurus experts at the FAO and USDA, Gudrun Johannsen and Lori Finch. This led to a sample reference alignment consisting of 347 mappings[11]: 74 broadMatch / narrowMatch and 273 exactMatch (79%). 11 concepts had no exact, broader or narrower counterpart. This is a higher percentage of exactMatch mappings than we expected based on our experiences with the OAEI *food task*. For the *food task*, arbitrary subhierarchies of the AGROVOC and NAL thesaurus were drawn and manually aligned with the other thesaurus. Most of the resulting mappings were equivalence relations. The sample sets, the percentage of equivalence mappings in the reference alignment (*i.e.* the desired equivalence relations) varied between 54% and 71%.

[10] As opposed to alignments consisting mainly of, for example, rdf:type or partitive relations.

[11] Adding up the number of mappings per topic leads to a total of 373 mappings. The lower total is due to overlap between the topics.

Table 2 gives an overview of the kinds of mappings in the new reference alignment and the kinds submitted to the OAEI 2007 *food task* by the participants. The sample reference alignments of the OAEI 2007 *food task* focussed much more on taxonomical terms and less on biological and chemical terms. Common categories of mappings that were not recognized as such in the OAEI 2007 food evaluation were: scientific methods, anatomy, and production or processing techniques (*e.g.* for crops or natural resources).

	submitted to OAEI 2007 food	required for hot topics
taxonomical	55%	14%
biological/chemical	9%	20%
geographical	3%	8%
miscellaneous	33%	58%

Table 2. The relative size of topics in the sets of mappings found by the participants of the OAEI 2007 *food task* and in the set of mappings that is necessary to find documents on hot topics.

6 Sample Evaluation Results

Having constructed a new sample reference alignment we can use it to measure the performance of alignment approaches. We choose to reiterate the evaluation of Recall[12] on the OAEI 2007 *food task* for two reasons: It allows us to show the effect of relevance-based evaluation as opposed to non-relevance-based evaluation by referring to known results; and it offers a second opinion to test the validity of the evaluation method used for the OAEI *food tasks*. The latter is important, because the evaluation of Recall under the open-world assumption is inherently tricky business (*i.e.* an unsolved subject of research). If the results of relevance-based evaluation differ significantly from the results of independent evaluation then we should wonder whether non-relevance-based evaluation as it is performed in all OAEI tasks is a suitable evaluation method.

For the sake of simplicity we calculate Recall scores of the top-4 of the systems that participated in the OAEI 2007 *food task*. The results are shown in table 3. There are a

	Falcon-AO	RiMOM	DSSim	X-SOM
OAEI 2007 food, only exactMatch (54% of total)	0.90	0.77	0.37	0.11
hot topics, only exactMatch (79% of total)	0.96 ↑	0.60 ↓	0.16 ↓	0.07 ↓
OAEI 2007 food, exact, broad, narrowMatch	0.49	0.42	0.20	0.06
hot topics, exact, broad, narrowMatch	0.75 ↑	0.47 ↑	0.12 ↓	0.05 ≈

Table 3. Recall of alignment approaches measured on sample mappings biased towards relevance to hot topics in agriculture and on impartial, non-relevance-based sample mappings from the OAEI 2007 *food task*.

[12] “the whole truth” as opposed to “nothing but the truth”, $|Correct \cap Found| \,/\, |Correct|$.

number of striking points to note about these results.

If we look at the difference between rows labeled "OAEI 2007 food" and those labeled "hot topics" in table 3 we can see that for most systems there is a significant positive or negative difference. Falcon-AO performs 6% better on only exactMatch mappings for hot topics than it did in the OAEI 2007 *food task*, while DSSim performs 21% worse on hot topics, a very large relative difference.

Overall, the difference with non-relevance-based evaluation is quite great. In the second row of table 3, we can see that for exactMatch relations performance in general is lower for relevance-based evaluation than for non-relevance-based evaluation, with the exception of Falcon-AO, although the relative difference is small. However, even though there is a clear difference, the ranking of the alignment approaches is left unchanged. The results of relevance-based evaluation seem to exaggerate the differences between the performance of the approaches. This can be explained by the relatively high number of obvious matches (93%) in the set of mappings on hot topics. None of the approaches was able to find a substantial number of difficult mappings, but the best approaches were good at finding all obvious mappings before resorting to speculation about the harder mappings. The relatively high number of easy matches significantly boosts the scores of approaches that find the obvious matches. We expect that the reason why so many of the relevant mappings are easy is that the indexers at the USDA and FAO attempt to help users by using the most obvious words. (*cf.* the debated *basic level* described by Eleanor Rosch et al. in [10])

Another thing we can note is that the best two systems, Falcon-AO and RiMOM performed relatively good for all relation types, the last row of table 3. This has nothing to do with their ability to find particular relation types, because they found no broadMatch and narrowMatch relations. It is due to the kind of exactMatch relations they *did*, which were mostly of the obvious kind (*i.e.* literal matches), which was exactly the kind that was needed most for the hot topics. The high percentage of exactMatch relations in the set on hot topics accentuates their behavior. The converse goes for DSSim, which found a relatively low number of obvious mappings.

Fewer broadMatch and narrowMatch mappings seem to be needed than one would expect from the non-relevance-based evaluation method. Compare the percentage in the OAEI 2007 Recall set, 54%, to the percentage based on hot topics, 78.6%. Although there is a large part of the AGROVOC and NALT vocabularies that does not have a counterpart in the other vocabulary, the portion that is actually used suffers less than one would expect from this mismatch. Apparently, indexers mainly pick their terms from a limited set, which shows a greater overlap. (After all, why needlessly complicate things?) On one hand this means that approaches that can only find equivalence mappings perform better in practice than was expected. On the other hand it confirms the expectation that a large part (*more than 20%*) of the mappings that are needed for federated search over AGRIS/CARIS and AGRICOLA consists of other relations than equivalence relations. Also, one can conclude that systems that are incapable of finding a substantial number of equivalence relations can only play a marginal role.

7 Discussion

By using relevance as a sample criterion we avoid having to come up with an artificial approximation of importance. We can simply explore the performance difference on samples consisting of relevant mappings and samples consisting of irrelevant mappings. This has a few advantages. We can use existing evaluation measures without adaptation, therefore results using this method are easily comparable to existing results. Due to the simplicity of this method results are easy to interpret. Linear weighing of the mappings by some real value representing relevance as in [7], for example, can make it difficult to see whether an alignment approach found many marginally relevant mappings or a few reasonably relevant mappings. If you use the weights for the drawing of the samples you can save the sample for later use. We can easily extend existing experiments. For instance, to investigate a new use case.

Under minimal assumptions we avoid having to choose a specific retrieval method while retaining the the character of an end-to-end evaluation. (*cf.* the *End-to-end Evaluation* method described in [12]) This saves us the effort of extensive user studies while not ignoring the behavior of alignment approaches in real-life situations.

Considering the fact that AGROVOC and NALT are two of the most widely used agricultural ontologies, and that they are prototypical examples of domain thesauri in their design we conclude the following. From the point of view of a developer of a federated search engine in the agricultural domain that needs an alignment we can conclude that at the moment the Falcon-AO is a good starting point. For use cases similar to the prototypical set-up described in this paper, Falcon-AO can be expected to find three quarters of the mappings. Demands change through time, and hence, current thesauri, current hot topics, and perhaps current alignment techniques will be outdated.

Another thing to note, which is besides the main message of this paper, is that this empirical study has shown that at least 20% of the required mappings to solve the typical federated-search problem are hierarchical relations. Even though this is a smaller fraction than we initially expected it is still a large part.

Acknowledgments

We would like to thank the FAO and NAL for allowing us to use their collections and the participants of the OAEI 2007 *food task*. Specific thanks go to Edgar Meij for his thoughts on log analysis, Fred van de Brug and Patricia Merrikin for their input of topics, Andrew Bagdanov and Stefano Anibaldi for allowing us to access the AGRIS/CARIS search log files, Lori Finch and Gudrun Johannson for creating part of the reference alignment, Tuukka Ruotsalo and Ruud Stegers for their comments, and Laura Hollink, Mark van Assem, Mike Frame, and Gail Hodge for valuable discussions. This work has been partially funded by the Dutch BSIK project Virtual Laboratory for e-science (VL-e).

References

1. Jérôme Euzenat. Semantic precision and recall for ontology alignment evaluation. In *Proceedings of IJCAI 2007*, pages 348–353, 2007.

2. Jérôme Euzenat, Malgorzata Mochol, Pavel Shvaiko, Heiner Stuckenschmidt, Ondřej Šváb, Vojtěch Svátek, Willem Robert van Hage, and Mikalai Yatskevich. Results of the ontology alignment evaluation initiative, 2007.
3. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007. ISBN 978-3-540-49611-3.
4. Laura Hollink, Mark van Assem, Shenghui Wang, Antoine Isaac, and Guus Schreiber. Two variations on ontology alignment evaluation: Methodological issues. In *Proceedings of 5th European Semantic Web Conference 2008 (ESWC 2008)*, 2008.
5. Antoine Isaac, Henk Matthezing, Lourens van der Meij, Stefan Schlobach, Shenghui Wang, and Claus Zinn. Putting ontology alignment in context: usage scenarios, deployment and evaluation in a putting ontology alignment in context: usage scenarios, deployment and evaluation in a library case. In *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, 2008.
6. Yannis Kalfoglou and Marco Schorlemmer. Ontology mapping: the state of the art. *The knowledge engineering review*, 18(1):1–31, march 2003.
7. Jaana Kekäläinen. Binary and graded relevance in ir evaluations–comparison of the effects on ranking of ir systems. *Information Processing and Management*, 41(5):1019–1033, 2005.
8. Malgorzata Mochol, Anja Jentzsch, and Jérôme Euzenat. Applying an analytic method for matching approach selection. In *Proceedings of the 1st International Workshop on Ontology Matching (OM-2006)*, pages 37–48, 2006.
9. Robert Porzel and Rainer Malaka. A task-based approach for ontology evaluation. In *Proceedings of the ECAI Workshop on Ontology Learning and Population: Towards Evaluation of Text-based Methods in the Semantic Web and Knowledge Discovery Life Cycle*, 2004.
10. Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, July 1976.
11. Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. *Journal on Data Semantics*, 3730:146–171, 2005.
12. Willem Robert van Hage, Antoine Isaac, and Zharko Aleksovski. Sample evaluation of ontology-matching systems. In *Proceedings of the 5th International Evaluation of Ontologies and Ontology-based Tools Workshop (EON 2007)*, 2007.