MultiFarm: A Benchmark for Multilingual Ontology Matching

Christian Meilicke[c], Raúl García-Castro[d], Fred Freitas[a], Willem Robert van Hage[b], Elena Montiel-Ponsoda[d], Ryan Ribeiro de Azevedo[a], Heiner Stuckenschmidt[c], Ondřej Šváb-Zamazal[e], Vojtěch Svátek[e], Andrei Tamilin[f], Cássia Trojahn[g], Shenghui Wang[b]

[a] *Universidade Federal de Pernambuco*
[b] *Vrije Universiteit Amsterdam*
[c] *University of Mannheim*
[d] *Universidad Politécnica de Madrid*
[e] *University of Economics, Prague*
[f] *Fondazione Bruno Kessler, Trento*
[g] *INRIA, Grenoble*

Abstract

In this paper we present the MultiFarm dataset, which has been designed as a benchmark for multilingual ontology matching. The MultiFarm dataset is composed of a set of ontologies translated in different languages and the corresponding alignments between these ontologies. It is based on the OntoFarm dataset, which has been used successfully for several years in the Ontology Alignment Evaluation Initiative (OAEI). By translating the ontologies of the OntoFarm dataset into eight different languages – Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish - we created a comprehensive set of realistic test cases. Based on these test cases, it is possible to evaluate and compare the performance of matching approaches with a special focus on multilingualism.

Keywords: Ontology Matching, Benchmarking, Multilingualism, Data Integration.

1. Motivation

Ontologies have been introduced in computer science as a means for solving the problem of interoperability between different knowledge sources [1]. In the context of the Semantic Web, it became clear that ontologies do not really solve the problem of semantic interoperability but rather lift it to a higher level of representation. As an answer to this, ontology matching has been established as a field of research concerned with the development of methods for determining equivalent elements in different ontologies [2]. One of the insights of this new field of research is that there is not a single best solution to the problem, but that the performance of a matching method depends on the nature of the ontologies to be matched. Thus, the systematic evaluation of matching methods is an important task. It can reveal strengths and weaknesses of existing methods and guide the selection of the most appropriate method for a given task.

In the past six years, the OAEI has carried out systematic evaluation of ontology matching technology, providing many important insights [3]. While the OAEI features a variety of different benchmark datasets covering a wide range of typical matching problems, almost all datasets considered so far assume that the ontologies to be aligned use English as a common language for naming and describing concepts and relations. This assumption is significant as virtually all matching methods are based on a lexical matching step in which the names of elements are compared, providing an initial estimate of the likelihood that two elements refer to the same real world phenomenon [2].

The increased awareness of the usefulness of ontologies for practical applications has lead to a situation where an increasing number of ontologies actually used in real world applications do not use English as a base language. As argued by Fu et al. [4], such ontologies are an important link between the information available on the Semantic Web and the individual user that prefers to have information presented in his or her local language. The existence of such multilingual ontologies pushes the ontology matching problem to a new level as the basic step used by most matching algorithms has to be completely revised. However, currently there have only been a few attempts to tackle the problem of multilingual ontology matching (e.g. [5, 6, 7, 8]).

We think that further progress in this area is hindered by the lack of a commonly accepted benchmark dataset with a special focus on multilingualism. This view is supported by the observation that existing publications on the topic always rely on a very specific dataset for evaluation that has been created for the purpose of the publication and that have serious shortcomings which are described in more details below. The existence of a carefully engineered and commonly accepted benchmark dataset would be an important enabler fostering progress in multilingual ontology matching in the same way, as the current OAEI datasets have fueled research in monolingual ontology matching.

In this paper, we attempt to solve the problems described above by proposing a comprehensive benchmark dataset for multilingual ontology matching. This dataset has been jointly created by the authors on the basis of an existing dataset from the OAEI campaigns. The proposed benchmark consists of seven ontologies for which mutual reference alignments have been created manually. Each of the ontologies has been translated into eight different languages other than English – Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish. Each combination of ontologies and languages establishes a test case for multilingual ontology matching summing up to roughly 1500 test cases.

The rest of the paper is structured as follows. We first describe characteristics to be taken into account when defining a multilingual dataset (Section 2). We discuss existing multilingual datasets and evaluations pointing to problems that limit the validity of these datasets for evaluation purposes (Section 3). We present the MultiFarm dataset providing details about generating translated ontologies as well as creating the reference (gold standard) alignments between the ontologies (Section 4). Then, we focus on some decisions we made while creating the dataset both in terms of language-independent and language-specific aspects (Section 5). In a preliminary series of experiments, we evaluated current state-of-the-art matching systems against the dataset (Section 6). Finally, we comment on the availability of the dataset and conclude with a discussion of remaining shortcomings and future possible improvements (Section 7).

2. Characteristics of a multilingual dataset

In this section, we present different characteristics to be taken into account when defining a multilingual dataset, since they can affect the results of an ontology matcher. Most of the listed features could also influence a monolingual alignment task, as they are mainly related with the Natural Language (NL) descriptions associated to ontology elements, and the ontology structure *per se*. The identified characteristics have been distributed into three levels: (a) Format or encoding level; (b) Lexical and terminological level; and (c) Ontology structure level. Without claiming to be exhaustive, the set of characteristics accounted for in this section covers those aspects of the NL descriptions associated to ontologies as well as ontology expressiveness. They are currently supported by the most commonly used ontology formalisms. We argue that the presence or absence of these ontological features will contribute in the success of the alignment task.

2.1. Format or encoding level

This level includes those characteristics related to the encoding in which the ontology is serialized, the alphabet used in the labels or NL descriptions associated to ontology elements, and the format used for labels.

- **Encoding**. The character encoding in which the ontology is serialized (e.g., UTF-8) can affect the alignment task, as some tools can process multiple encodings, whereas others cannot.

- **Diacritics**. This feature specifies whether diacritics are used in labels or any other type of NL descriptions associated to ontologies. In some languages, the same word written with or without accent can have a different meaning (e.g., in Spanish 'río' means river, whereas 'rio' without accent refers to the first person singular of the verb 'laugh').

- **Language tags**. In specific syntaxes, such as RDF/XML, one can restrict the scope of a particular label (or any other type of NL description related to the ontology) to a certain natural language (e.g., '@en' for English). At a multilingual level, such a language tag may also contribute to avoid errors, since certain groups of languages with common roots share the same words with different meanings (e.g., 'nombre' in Spanish means 'name' and in French 'number').

- **NL description placement**. This characteristic has to do with the place where NL descriptions of ontology elements appear: in URIs, in labels (using *rdfs:label*, *skos:preflabel*, etc.), in both places, or in an external linguistic model created for that purpose (see LIR[1], LexInfo, lemon[2]). Identifiers in URIs suffer from some restrictions of the URI naming scheme (e.g., some characters such as white spaces cannot be part of URIs).

- **Word separation**. Specifies the way used to separate words in multiple-word terms in URIs or as label annotations (e.g., CamelCase, hyphen, white space). A correct identification of the multiple words that compose a term is necessary to avoid mistakes (e.g., 'hasVAT' consists of the verb 'has' and the acronym 'VAT').
- **Capitalization**. Specifies how capital letters are used in labels or terms (only first word capitalized, all words capitalized, etc.). In some cases, capitalization may lead to incorrect matchings (e.g., 'white house' vs. 'White House').
- **Punctuation**. When showing up in NL descriptions (mostly, compound words or complex Noun Phrase constructions), punctuation marks may signalize the several components that make up a term (e.g. 'Acquisitions trough business combinations, intangible assets').

In Section 5 we explain which of these features appears in MultiFarm.

2.2. *Lexical and terminological level*

This level includes those characteristics concerning the linguistic descriptions that may be related to ontology elements. The amount and type of linguistic descriptions range from labels and comments (as supported by the RDF/XML syntax) or terminological variants (such as the ones enabled by SKOS properties), to more complex linguistic descriptions.

- **Terms as ontology labels**. Specifies whether terms are provided for naming ontology elements. We understand terms as words or expressions that have a precise meaning in a certain domain. When dealing with general knowledge ontologies, we could talk about lexical entries.
- **Definitions**. Specifies whether labels (terms, lexical entries) are accompanied by definitions or glosses in natural language. These definitions can be used in the alignment task to disambiguate the meaning of terms, as they usually provide contextual information (e.g. reference to the superclass, specific properties of the term, etc.).
- **Linguistic variants**. This feature refers to the inclusion of synonyms (e.g., 'programme brochure' - 'programme flyer') or terminological variants (such as acronyms, abbreviations, short forms, full forms, transliterations, etc., e.g., 'PC' - 'Programme Chair') that further describe and complement ontology labels. The availability of this sort of linguistic descriptions (e.g., WordNet synsets) may leverage the possibilities of finding alignments.
- **Multilingual labels**. Specifies whether the ontology contains terms in more than one language. A multilingual ontology provides additional information that can be exploited in the matching process. For example, if the ontology is already available in German and French, when trying to align it to an ontology in Russian, we can use the multilingual labels in the original ontology.
- **Other linguistic descriptions**. This feature accounts for additional linguistic information that may further describe terms or lexical entries. Here, we include several types of linguistic descriptions such as basic lexical properties (part-of-speech, gender, number, etc.), morpho-syntactic decomposition of terms or lexical entries (inflected forms, phrase structure, syntactic properties, etc.), representation of multi-word expressions, etc. If the ontologies to be aligned contain some of these additional linguistic descriptions, they can increase the probabilities of finding correspondences. Such additional information can be provided nowadays by models such as LIR, LexInfo or lemon.

We have decided to add no additional information that goes beyond the information expressed in a single label. This makes our test cases difficult and similar to many real-life ontologies.

2.3. *Ontology structure level*

This level includes those characteristics related to the ontology structure.

- **Type of components**. Specifies the types of ontology components used in ontologies (e.g., concept, relation, individual). Identifying correspondences between the different types of ontology components may require different approaches.

[1] `http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/downloads/63-lir`

[2] LexInfo and lemon are available from `http://lexinfo.net/`

- **Ontology expressiveness**. Reasoning mechanisms can support the matching process as argued by Niepert et al. [9].
- **Ontology structure**. Differences in the structure of the ontologies to be aligned and their individual structure (e.g., flat or a deep hierarchy) affect the matching process.

In Section 4 we describe the OntoFarm dataset that was used as starting point to construct MultiFarm. Within this section we talk about structure and expressiveness of OntoFarm and resulting MultiFarm test cases.

3. Generating Multilingual Datasets

A simple benchmark for ontology matching consists of two ontologies and a reference alignment between them. Standard evaluation techniques are based on compliance-based metrics such as precision and recall [2]. These measures are used to analyze in how far the generated alignment is complete and correct compared against the reference alignment. In the following, we review some related work on generating datasets, i.e., ontologies and reference alignments for multilingual ontology matching. In particular, we distinguish between three different approaches depending on the source from which the dataset was generated.

3.1. Datasets from Monolingual Ontologies

One approach for generating a multilingual dataset is to pick an ontology O_{l_0} written in language l_0 and to translate it into other languages $l_1, \ldots, l_n$. As a result, we have n ontologies O_{l_0} to O_{l_n}. Keeping track of the original concepts and their translated counterparts results in a correct and complete reference alignment for all pairs of ontologies. Such an approach has been applied by Trojahn et al. [5] and again by Fu et al. [6].

However, a dataset created in such a way is only useful to a limited degree, as the high structural similarity between the multilingual ontologies dominates the dataset. Thus, it is hard to single out the positive effects of a specific multilingual technique opposed to the positive effects of techniques that simply exploit structural characteristics. We present experimental results that support this claim in Section 6.

3.2. Datasets from Multilingual Ontologies

Another obvious approach which avoids this problem requires much more effort by the creator of the dataset. The approach starts with several ontologies concerned with the same domain, which are specified in different languages. In this approach, the reference alignments are missing. They have to be generated for each pair of ontologies in the chosen set of ontologies.

Our experience on manually creating, verifying and extending reference alignments taught us that it is a laborious and time-consuming task. Moreover, several persons have to be involved to ensure a high quality in the alignment. In the case of a missing reference alignment, alternatively, a small sample from the alignment to be evaluated can be used to estimate alignment precision.

This approach has been used for the very-large-cross-lingual-resources track at OAEI 2008 [10]. This bilingual dataset contains three large SKOS subject heading lists for libraries that have to be aligned to each other: the thesaurus of the Netherlands Institute for Sound and Vision (GTAA), written in Dutch, and the WordNet and DBpedia lists, written in English. The DBpedia list, which contains labels in both languages, can therefore be used as a mediator between the two other lists.

In order to alleviate the effort of creating the alignments between the multilingual ontologies, Jung et al. [11] proposed an approach for the indirect composition of multilingual alignments between ontologies. The idea is to use existing intermediary alignments between ontologies to compose new alignments. For instance, an alignment between French and Portuguese ontologies can be generated if intermediary alignments between these two ontologies and a third one, English for instance, are available. This general idea is also used in our approach, which is explained in Section 4.

3.3. Datasets from Multilingual Resources

A third approach uses existing non-ontological resources and their available representations in different languages l_1 to l_n (i.e, the English and Japanese representations of the web directories Yahoo!). Generating a multilingual dataset from such a given resource requires to convert each of its representations in a given language l_i into an ontology o_i, keeping trace of the correspondences inter-representations.

Such an approach has been conducted in the context of OAEI 2008 [10] for both the multilingual directory track (a bilingual dataset that contains web directory ontologies in English and Japanese) and partially for the very-large-cross-lingual-resources track. However, the approach suffers from the fact that the resulting ontologies are weakly structured taxonomies of limited expressivity.

With respect to large multilingual datasets, one limitation is the lack of reference alignments between their ontologies. Examples are the Financial Accounting Standards (FAS) datasets,[3] which has been exploited in [8], where only a subset of manually created alignment are available.

4. The MultiFarm Benchmark

As a basis for generating our multilingual benchmark for ontology matching we have chosen the *OntoFarm* dataset. In the following, we describe this dataset and motivate its choice. We explain how the features of the extended OntoFarm help to avoid the problems of benchmarks mentioned in the previous section.

4.1. The OntoFarm Dataset

The motivation for initiating the creation of the OntoFarm collection (in Spring 2005 [12]) was the lack of a material for testing ontology engineering (especially, matching) techniques that would exhibit at the same time high *OWL expressivity*, *comprehensibility* to broad audience, *grounding in reality* and *natural heterogeneity*. We will briefly explain these requirements. Regarding the first requirement, all ontologies were build natively in OWL, by people with at least minimal training in OWL-DL basics (either graduate students of a semantic web course or researchers familiar with this field). By consequence, most ontologies were equipped with DL axioms of various kinds, which opened the way to use semantic matching techniques [13] and reasoning-based matching approaches [9]. Second, conference organisation was chosen as a familiar domain for people from academic life. The third requirement, grounding in reality, was achieved by deriving each model from a real-world resource. Finally, the requirement of natural heterogeneity was addressed by the diverse nature of those underlying resources, which belonged to three different categories (see below). The designers were also not given any specific rules or guidelines regarding the modelling style or design methodology, neither they interacted among themselves.

4.1.1. The OntoFarm Ontologies

Currently, there are sixteen ontologies within the OntoFarm dataset. The ontologies differ in numbers of classes, properties, and in their DL expressivity. Overall, the ontologies have a high variance with respect to structure and size, which makes the matching process harder. They were based upon three types of resources (cf. Table 1):

- actual conferences and their web pages (type 'Web'),
- actual software tools for conference organisation support (type 'Tool'), and
- experience of people with personal participation in organisation of actual conferences (type 'Insider').

During 2006 and 2007, the participants of the OAEI were asked to freely explore the OntoFarm dataset. This effort materialized in usual alignments as well as in interesting individual correspondences that were hard to detect (referred to as 'nuggets'), aggregated statistical observations and/or implicit design patterns. Additionally, in the next three years, when reference alignments were already available, the participants were also asked to find all correct correspondences (equivalence, and for OAEI 2009 also subsumption correspondences). Meanwhile, the collection grew until the current size.

4.1.2. Reference Alignments

Reference alignments were gradually built in the course of three years. First, during autumn 2008, an initial collection of ten reference alignments was created. Second, during summer 2009 the mutual (non-directional) alignments were built between all pairs of seven ontologies, thus yielding 21 reference alignments. Each of them had between 4 to 25 correspondences. Finally, in 2010, minor corrections have been applied[4].

In the first two years, there have been three evaluators involved in the process. Their effort has been eased with a reasoning-based support tool [14]. The process of building reference alignments basically had four steps: (1) evaluation of an initial set of alignments by each participant; (2) discussion about disagreements in order to achieve consensus; (3) evaluation of additionally added alignments by each participant; and (4) discussion about disagreements in order to achieve consensus.

The initial set of alignments was taken from all the available results of the OAEI conference track. Participants evaluated those alignments independently and used the support tool. This tool enabled evaluators to explore logical conflicts between correspondences of an alignment. Thanks to this setting, the evaluators could only spot those subset of correspondences involved in conflicts without analyzing all correspondences step by step. After the first phase, the evaluators tried to arrive at a consensus via discussion. This resulting reference

[3] http://www.xbrl.org/

[4] Reference alignments available for three last years of the OAEI are at: `http://nb.vse.cz/~svabo/oaei20[08|09|10]/reference-alignment.zip`

Name	Type	#C	#DP	#OP	DL
Ekaw	Insider	74	0	33	$\mathcal{SHIN}$
Sofsem	Insider	60	18	46	$\mathcal{ALCHIF}$
Sigkdd	Web	49	11	17	$\mathcal{ALEI}$
Iasted	Web	140	3	38	$\mathcal{ALCIN}$
ConfTool	Tool	38	23	13	$\mathcal{SIN}$
Cmt	Tool	36	10	49	$\mathcal{ALCIN}$
Edas	Tool	104	20	30	$\mathcal{ALCOIN}$

Table 1: Seven original ontologies from the OntoFarm dataset. The columns #C, #DP, and # OP refer to the number of classes, datatype properties and object properties respectively.

alignment was already featuring high precision. In order to increase recall, further alignments have been evaluated, discussed and added in 2008 and 2009.

4.2. Extending the OntoFarm Dataset

We have reported above about the problem of generating multilingual datasets for ontology matching. Our approach has been to generate the dataset from a set of monolingual ontologies for which we already have reference alignments by translating these ontologies into different languages. In this way, the approach avoids the infeasible effort to create a large set of reference alignments as we explain in the following.

We start with an existing monolingual dataset for ontology matching (the OntoFarm dataset), which comprises a set of different ontologies as well as reference alignments between them. Seven ontologies, the ones for which we have reference alignments, have been chosen as starting point for the MultiFarm benchmark. These ontologies are those listed in Table 1.

We translated the original ontologies of the dataset into different languages, in particular, Chinese, Czech, Dutch, French, German, Portuguese, Russian, and Spanish. These translations were performed by different groups of people, each group composed of native speakers with knowledge about conference organisation. The process has not been supported by any automatic machine translation techniques, but was conducted completely manually. Our approach allows us to derive cross-lingual reference alignments from existing monolingual reference alignments. Moreover, these reference alignments connect differently modelled ontologies. Hence, they are not at all trivial. Figure 1 illustrates this by an example. In the OntoFarm dataset, we have a reference alignment between the CMT$_{en}$ and EKAW$_{en}$ ontologies (bold arrow). By translating these ontologies into Portuguese and Spanish, we obtain translation alignments (normal arrow) that can be used with the previous alignments to derive non-trivial multilingual reference alignments (dashed arrow) to all other language variants.

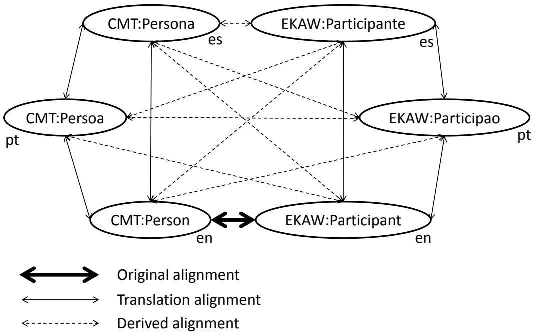

Figure 1: Alignments in the MultiFarm dataset.

In particular, considering that we translated the ontologies into 8 languages (+1 for the original English ontologies), we have 36 pairs of languages. The original dataset has a reference alignment between all 21 undirected pairs of seven ontologies. There is no distinction between the matching task CMT-EKAW and EKAW-CMT in the OntoFarm dataset, while there is a difference between CMT$_{en}$-EKAW$_{de}$ and CMT$_{de}$-EKAW$_{en}$. Thus, we can derive 42 reference alignments for each language pair. We can also construct new reference alignments for matching each ontology on its translation (e.g., CMT$_{en}$-CMT$_{de}$) resulting in seven more reference alignments per language pair. In all, we have 36 × 49 matching tasks.

For each ontology we generated a translated variant instead of generating one single multilingual ontology. This allows to define matching tasks specific to a particular pair of languages and to hide any additional information from the matching system.

5. Specifics of the Translation

Finally, we had to make some decisions with respect to those features listed in Section 2, which have not been determined by our choice of extending OntoFarm. These are discussed in the following two subsections. Moreover, during the translation process and the generation of the test cases, we identified several interesting issues, which are reported afterwards.

5.1. URI identifier versus Label

When analysing the OntoFarm dataset we realized that concept and property names have been encoded as fragments of the URIs that identify those resources (e.g., `http://cmt#Person`). This modelling style is not recommended, even if it is quite often used. It is more appropriate to express the human-readable names via an annotation property (i.e., as the value of `rdfs:label`). See [15] for an interesting discussion on pros and cons of the use of URI's local names vs. labels for describing ontologies. Thus, we decided to encode translation results as labels. We applied the same transformation to the original English ontologies.

Moreover, the use of the `rdfs:label` annotation avoids having to perform a pre-processing (tokenizing) of the local names to obtain natural, fluent labels that can be then further analysed or translated. We also added language tags like '@en'.

5.2. Diacritics and Encoding

We expect that some matching systems might have problems with diacritics while others might have no problems and can in addition exploit them to resolve ambiguity in the translation (see Section 2). Finally, we decided to conduct all translations taking diacritics into account. This means that we translated, for example, *ProgramCommitteeChair* to the Czech phrase *předseda programového výboru*. Furthermore, we ensured that all ontologies are encoded in UTF-8, which allows to represent phrases from all languages, including Russian and Chinese, correctly.

5.3. Randomizing URI identifiers

Matching systems try to benefit from any kind of information available in the ontologies. For that reason, we had to ensure that no information encoded originally in the ontologies remained usable distorting the result for the multilingual matching task. This holds for English identifiers encoded as part of the URIs. Thus, we replaced each URI identifier by a random string like `http://cmt_es#c-0796534-0846894` and modified the reference alignments in accordance.

In addition, we also used different URI identifiers in two translations of the same ontology. This allows, for example, to use CMT$_{cn}$ and CMT$_{pt}$ as a test case where we have full structural agreement, without giving an advantage to matching systems that compare the URI identifiers with a lexical similarity measure. Such a test case can be used to verify whether a matching system can exploit this kind of information.

5.4. Cultural Differences and Test Difficulty

All ontologies from the MultiFarm dataset have been translated from English to other languages. Thus, the dataset does not contain cases in which the conceptual hierarchy differs due to different cultural backgrounds. However, languages are products from cultures; therefore, we had to take into account differences of cultural situations, with their contexts, typical relations, protocols and behaviours while translating a label of a concept into another language. Taking into account the resulting and sometimes subtle problems explains also the difficulty of the MultiFarm test cases. We describe two types of problems exemplarily.

- Literal translations without meaning: many literal translations bear no actual meaning in some languages. For instance, the term `Camera_Ready_event` could be literally translated to Portuguese as `Evento_das_Versões_Finais`. However, in this language it would not be clear that this event describes the delivery of final versions. Therefore, the proper translation is `Evento_de_Entrega_das_Versões_Finais`, which includes the concept of delivery (`entrega`).
- Literal translations with another meaning: in Czech there is a common equivalent of `poster` (`plakát`), which however has the connotation of something commercial or, at least, non-scientific. In the context of scientific conferences, the English term is usually borrowed for this specific meaning by Czechs. Thus, the label `poster` is also used for the Czech translations.

The organizers of previous OAEI campaigns [3] report about an average F-measure between 35% and 55% for a standard monolingual matching system, where top systems can reach up to 65%. We expect it will be much harder to match translated variants instead of the original ontologies for the reasons listed above.

6. Preliminary Evaluation Results

In a first set of preliminary experiments, we have been running a subset of the MultiFarm dataset against the matching systems participating in OAEI 2011. These matching systems are designed as systems for solving monolingual matching tasks. It can thus be argued that they will not be able to generate any meaningful results for the MultiFarm dataset. The subset that we have chosen excluded for that reason the languages Chinese and Russian. Note we had to execute some of the tools for several days to generate results for all test cases.

As explained in Section 4.2, the dataset can be divided in those test cases where the same original ontology has been translated in different languages and in

matcher	different ontologies				same ontologies			
	size	p	r	f	size	p	r	f
CIDER [16]	1433	0.42	0.12	0.18	1090	0.66	0.06	0.12
CODI [17]	923	0.43	0.08	0.13	7056	0.77	0.48	**0.59**
LogMap [18]	826	0.39	0.06	0.11	469	0.71	0.03	0.06
MapSSS [19]	2513	0.16	0.08	0.10	6008	0.97	0.51	**0.67**
LogMapLt	826	0.26	0.04	0.07	387	0.56	0.02	0.04
MaasMatch [20]	558	0.24	0.03	0.05	290	0.56	0.01	0.03
CSA [21]	17923	0.02	0.06	0.03	8348	0.49	0.36	**0.42**
YAM++ [22]	7050	0.02	0.03	0.03	4779	0.22	0.09	**0.13**
Aroma- [23]	0	-	0.00	-	207	0.54	0.01	0.02
Lily [24]	0	-	0.00	-	11	1.00	0.00	0.00

Table 2: Precision(p), recall(r), and f-measure(f) aggregated per matching system.

those test cases that have been derived from existing reference alignments. We have argued that the latter should be in the focus of multilingual ontology matching, while test cases that use the same ontologies as input suffer from a high structural similarity that dominates the evaluation results (see Section 3.1).

Table 2 shows the results of our experiments. We have ordered the systems according to the f-value we measured for those test cases built on matching different ontologies. These results are depicted on the left side of the table. The best results are achieved by CIDER followed by LogMap and MapSSS. CIDER has both better precision and recall scores than any other system. Compared to the top-results that have been reported for the original conference dataset (f-value ≥ 60%) the test cases of the MultiFarm dataset are obviously much more difficult and it seems that state-of-the-art matching systems cannot generate good results on MultiFarm.

The results we measured for test cases based on matching the same ontologies in different languages differ significantly. In particular, the results of MapSSS are a surprise compared to the results for test cases based on different ontologies. MapSSS can exploit the structural equivalence of the matched ontologies to achieve an f-measure of 67%. This system can leverage the structural information to cope with the problem of matching labels expressed in different languages. Similar to MapSSS, we also observe a higher f-measure for CODI, CSA and YAM++. Note that all these systems have an f-measure of at least 5 times higher than the f-measure for the harder test cases that are based on matching different ontologies. For all other systems we observe a slightly decreased f-measure.

Comparing these results with the results measured for the OAEI 2011 benchmark track, it turns out that all systems listed in the previous paragraph have been among the top five systems of this track. All test cases of this track have a similar property, namely, their reference alignments contain for each entity of the smaller ontology exactly one counterpart in the larger ontology. An explanation for this can be that these systems have been developed or at least configured to score well for the benchmark track. For that reason they generate the good results reported in the right half of Table 2, while results on the left side are less good. We recommend to take this distinction into account for further OAEI evaluation campaigns.

A detailed analysis of these experiments is presented by Meilicke et al. in [25], where the authors also discuss differences between different language pairs. Our main focus in this paper is related to the lessons learned for the design of multilingual ontology matching testcases. We draw the conclusion that it is important to match *different ontologies described in different languages*. Otherwise, it is very hard to single out whether good results are related to some multilingual matching technique or caused by exploiting the structure and the specifics of the matching task that is not related to the problem of multilingual ontology matching.

7. Conclusion

Multilingual ontology matching is a crucial task towards realizing a multilingual Semantic Web. This paper has presented the MultiFarm dataset, a systematically generated dataset for multilingual ontology matching based on a subset of the OntoFarm dataset. Our main aim was to overcome the lack of multilingual benchmarks for evaluating matching systems. Although the OAEI campaigns have proposed some multilingual datasets, they have not been successful in terms of participation, especially due to the fact that (a) they are limited to a few languages (English and Japanese, for instance), (b) they only have partial reference alignments, which makes the full evaluation of the systems difficult; and (c) some of the datasets are not publicly available. Our dataset overcomes all these issues.

In the future, we plan to include the MultiFarm dataset as a new track in the OAEI campaigns. Furthermore, we intend to extend the dataset, including new translations. Besides, there is a plan to extend the original OntoFarm collection with further reference alignments. Thus, we will then extend the dataset for including these new reference alignments.

Furthermore, many matching systems exploit the `rdfs:comment` annotation as a source for their algorithms. However, the original OntoFarm dataset does not include such kind of annotations and one possible improvement in the MultiFarm dataset could be extending the current translations in order to consider comments as well. This allows for better expressing the context and semantic of the translations.

The current version of the dataset is available at `http://web.informatik.uni-mannheim.de/multifarm/`. Furthermore, we made the whole dataset available as a set of 36 test suites via the SEALS platform [26], one test suite per language pair. The SEALS platform, developed in the context of the SEALS project [5], is an infrastructure that allows for automated evaluation of semantic technologies, including ontology matching. Publishing a dataset via the SEALS platform makes the dataset (and its versions) available over time.

Acknowledgements. Ondřej Šváb-Zamazal and Vojtěch Svátek have been partially supported by the CSF grant no. P202/10/0761. Christian Meilicke, Heiner Stuckenschmidt, Cássia Trojahn dos Santos, and Raúl García-Castro have been partially supported by the SEALS European project (IST-2009-238975). Elena Montiel-Ponsoda has been partially supported by the Monnet European project (FP7-248458), and the BabeLData Spanish project (TIN2010-17550).

[1] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator, W. Swartout, Enabling technology for knowledge sharing, AI Magazine 12 (3) (1991) 36–56.

[2] J. Euzenat, P. Shvaiko, Ontology Matching, Springer-Verlag, Berlin, Heidelberg, 2007.

[3] J. Euzenat, C. Meilicke, H. Stuckenschmidt, P. Shvaiko, C. Trojahn, Ontology alignment evaluation initiative: six years of experience, Journal on Data Semantics XV (2011) 158–192.

[4] B. Fu, R. Brennan, D. O'Sullivan, Ontology mapping and its use on the multilingual semantic web, in: Cross-Lingual In Proceedings of the 1st Workshop on the Multilingual Semantic Web at the 19th International World Wide Web Conference (WWW 2010), Vol. Vol.571 of CEUR Proceedings, Raleigh, USA, 2010.

[5] C. Trojahn, P. Quaresma, R. Vieira, A framework for multilingual ontology mapping, in: 6th edition of the Language Resources and Evaluation Conference (LREC-2008), 2008.

[6] B. Fu, R. Brennan, D. O'Sullivan, Cross-lingual ontology mapping - an investigation of the impact of machine translation, in: Proceedings of the 4th Annual Asian Semantic Web Conference (ASWC 2009), Vol. 5926 of LNCS, 2009.

[7] S. Wang, A. Isaac, B. A. C. Schopman, S. Schlobach, L. van der Meij, Matching multi-lingual subject vocabularies, in: Proceedings of the 13th European Conferene on Digital Libraries (ECDL 2009), Vol. 5714 of Lecture Notes in Computer Science, Springer, 2009, pp. 125–137.

[8] D. Spohr, L. Hollink, P. Cimiano, Multilingual and cross-lingual ontology matching and its application to financial accounting standards, in: Proceedings of the 10th International Semantic Web Conference (ISWC2011), 2011.

[9] M. Niepert, C. Meilicke, H. Stuckenschmidt, A probabilistic-logical framework for ontology matching, in: M. Fox, D. Poole (Eds.), Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI Press, 2010, pp. 1413–1418.

[10] C. Caracciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilicke, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Šváb-Zamazal, V. Svátek, Results of the ontology alignment evaluation initiative 2008, in: Proceedings of the 3rd International Workshop on Ontology Matching (OM-2008) collocated with the 7th International Semantic Web Conference (ISWC-2008), CEUR-WS, 2008, pp. 73–119.

[11] J. J. Jung, A. Håkansson, R. Hartung, Indirect alignment between multilingual ontologies: A case study of korean and swedish ontologies, in: Proceedings of the 3rd KES International Symposium on Agent and Multi-Agent Systems, Springer, Berlin, Heidelberg, 2009, pp. 233–241.

[12] O. Šváb, V. Svátek, P. Berka, D. Rak, P. Tomášek, Ontofarm: Towards an experimental collection of parallel ontologies, in: Poster Track of ISWC 2005, 2005.

[13] F. Giunchiglia, M. Yatskevich, P. Shvaiko, Semantic matching: Algorithms and implementation, Journal on Data Semantics IX (2007) 1–38.

[14] C. Meilicke, H. Stuckenschmidt, O. Šváb-Zamazal, A reasoning-based support tool for ontology mapping evaluation, in: European Semantic Web Conference (ESWC-09), 2009.

[15] E. Montiel-Ponsoda, D. Vila-Suero, B. Villazón-Terrazas, G. Dunsire, E. Escolano, A. Gómez-Pérez, Style guidelines for naming and labeling ontologies in the multilingual web, in: Proceedings of the DCMI International Conference on Dublin Core and Metadata Applications (DC-2011), The Hague, The Netherlands, 2011.

[16] J. Gracia, J. Bernad, E. Mena, Ontology matching with CIDER: evaluation report for OAEI 2011, in: Proceedings of the ISWC 2011 workshop on Ontology Matching, Bonn, 2011.

[17] J. Huber, T. Sztyler, J. Noessner, C. Meilicke, Codi: Combinatorial optimization for data integration: results for OAEI 2011, in: Proc. 6th ISWC workshop on ontology matching (OM), Bonn, 2011.

[18] E. Jimenez-Ruiz, A. Morant, B. C. Grau, LogMap results for OAEI 2011, in: Proceedings of the ISWC 2011 workshop on Ontology Matching, Bonn, 2011.

[19] M. Cheatham, MapSSS results for OAEI 2011, in: Proceedings of the ISWC 2011 workshop on Ontology Matching, Bonn, 2011.

[20] F. Schadd, N. Roos, Maasmatch results for OAEI 2011, in: Proc. 6th ISWC workshop on ontology matching (OM), Bonn, 2011.

[21] Q.-V. Tran, R. Ichise, B.-Q. Ho, Cluster-based similarity aggregation for ontology matching, in: Proc. 6th ISWC workshop on ontology matching (OM), Bonn, 2011.

[22] D. Ngo, Z. Bellasene, R. Coletta, YAM++ – results for OAEI 2011, in: Proc. 6th ISWC workshop on ontology matching (OM), Bonn, 2011.

[23] J. David, Aroma results for OAEI 2011, in: Proc. 6th ISWC workshop on ontology matching (OM), Bonn, 2011.

[24] P. Wang, Lily results on SEALS platform for OAEI 2011, in: Proc. 6th ISWC workshop on ontology matching (OM), Bonn, 2011.

[25] C. Meilicke, C. Trojahn, O. Šváb-Zamazal, D. Ritze, Multilingual ontology matching evaluation - a first report on using multifarm, in: Proceedings of the Second International Workshop on Evaluation of Semantic Technologies, Heraklion, Greece, 2012.

[26] R. García-Castro, M. Esteban-Gutiérrez, A. Gómez-Pérez, Towards an infrastructure for the evaluation of semantic technologies, in: eChallenges e-2010 Conference Proceedings, IIMC International Information Management Corporation, 2010.

[5] `http://www.seals-project.eu/`