A LIGHTWEIGHT ELEMENT MATCHING METHOD FOR INDUSTRIAL TERMINOLOGY HARMONIZATION

Exploiting Minimal Semantics based on Naming Conventions

Pekka Aarnio, Seppo Sierla and Kari Koskinen
Department of Automation and Systems Technology, Aalto University, Otaniementie 17, Espoo, Finland

Keywords: Industrial terminology standards, Standard harmonization, Ontology matching, Concept naming rules.

Abstract: Harmonization of terminologies used in industrial standards has been widely understood to be necessary for better interoperability of industrial information systems. Partial automation of the terminology comparison and matching phases of this process is necessary, in order to reduce the workload of human experts. Terminology dictionaries have been developed by various national or international organizations and for different contexts, so their taxonomy structure and lexical content can be very different. Further, because they cannot be considered as true ontologies, advanced ontology matching techniques are not directly applicable. The goal of our research was to develop a lightweight element matching approach based on structural similarities of concept names. This approach is applicable, when similar naming conventions and rules have been applied during the development of both terminology dictionaries. In this paper, we present a new ElemMatcher method and demonstrate its application to the harmonization of PSK standards with ISO 15926-4. PSK Standardisation is an association of Finnish industry.

1 INTRODUCTION

Interoperability of industrial systems and applications during the entire lifecycle of a plant is a fundamental issue when targeting internal and external efficiency, reliability and flexibility of production and high quality products and services. One of its main aspects concerns information integration and data exchange between internal and external applications. The benefits of better interoperability are obvious and widely understood, and several industrial standards have been developed to manage these issues.

Several national and international plant model or product model standards with overlapping context exist, yet they use their own terminology dictionaries, resulting in interoperability problems. Industrial terminology (dictionary) standards provide terminology for product model standards. They identify and specify names of classes and properties and meaning in definitions, descriptions and remarks. Terminology standards are often extended as classification standards, i.e. taxonomies. The harmonization of several overlapping standards as deeply as possible and selecting a wide scope international terminology dictionary standard as a common shared integration standard has been proposed as a solution by the European industrial network Orchid (CEN Orchid Roadmap).

Our research goal is to develop a lightweight matching method ElemMatcher (EM) for industrial cases, in which source and target terminology standards are very dissimilar and contain very little semantic information that could be exploited by advanced ontology matchers. As a proof of concept, EM was applied to the harmonization of Finnish national PSK standards (PSK Standardisation) with ISO 15926-4, Reference data, (ISO 15926-4). This harmonization process has been started, following Orchid roadmap (CEN Orchid Roadmap) guidelines, in PSK Standardisation, which is an association of Finnish industry closely co-operating with the official Finnish Standards Association SFS.

The EM matching method applies structuring rules derived from the general naming principles and rules specified in two terminology work standards (ISO 860:2007), (ISO 704:2000) and in a metadata registries standard (ISO/IEC 11179-5:2010).

Terminology work standards document principles that should be followed in the formation of concept names. The main idea behind the EM

Aarnio P., Sierla S. and Koskinen K..
A LIGHTWEIGHT ELEMENT MATCHING METHOD FOR INDUSTRIAL TERMINOLOGY HARMONIZATION - Exploiting Minimal Semantics based on Naming Conventions.
DOI: 10.5220/0003640803900395
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 390-395
ISBN: 978-989-8425-80-5

approach has been crystallized by the *transparency* principle: *"a concept name (designation) is considered transparent when the concept it designates can be inferred, at least partially, without a definition or an explanation. In other words, the meaning of a name can be deduced from its parts."* (ISO 704:2000)

In the ISO/IEC 11179-5 "Naming and identification principles" standard (ISO/IEC 11179-5:2010), *naming convention* has been defined as a set of rules for creating names and their associations. Prescriptive conventions should be documented by semantic, syntactic, lexical and uniqueness rules. This standard defines also the basic general structure and different term parts of a concept name.

The rest of the paper is organized as follows. Section 2 presents briefly different ontology matching techniques and related work. Section 3 describes the EM approach. The results of a matching case are presented in Section 4 and Section 5 reports some conclusions.

2 RELATED WORK

A harmonization process includes a comparison and matching phase that can be at least partially automated. Several different methods have been developed for entity matching originating from database schema matching techniques. Today, state-of-the-art technology is called *ontology matching*. The goal of ontology matching is to find the relationships between entities expressed in different ontologies (Euzenat 2007).

Most matching approaches exploit at least element-level lexical information and calculate string distances of entity (concept or property) labels. Structural methods try to extract similarity features from hierarchy structures (taxonomies) or from the attribute structures of concepts (product models). In addition, advanced approaches exist that can exploit also other kinds of common semantic information. Language based methods can be applied, if textual data is available. Extrinsic methods exploit some external knowledge base (e.g. WordNet) in order to reveal additional semantic information concerning the entities to be matched. Extensional methods try to find similarities between instance data. A survey of several matching approaches is presented in (Rahm & Bernstein, 2001), (Shvaiko & Euzenat, 2005).

Semantic techniques are perhaps the most recent direction of matching approaches. It requires that concept systems to be matched are true ontologies or that semantics have been represented formally. Examples of state-of-the-art approaches that include semantic methods in their method suite are ASMOV (Jean-Marya, 2009) and S-Match (Giunchiglia, 2007). S-Match has been categorized as a schema-based semantic matching approach. It can apply element level string-based methods, structural methods and extrinsic methods that can exploit external dictionaries. As a result, it is applicable not only for ontology but also for lightweight ontology and schema matching (Giunchiglia, et.al. 2009).

Applications of ontology matching techniques in industry are still uncommon. One of the main reasons for this might be that *"there is no integrated solution that is clear success, which is robust enough to be the basis for future development, and which is usable by non expert users"* (Shvaiko 2008, p. 1165). Furthermore, studies made by (Lauser, et. al. 2008) reveal that the success of matching techniques is largely case dependent.

Recent research on industrial applications of ontology matching (Uslar & Rohjans, 2009), (Fiorentini, et. al. 2009), (Zan, et. al. 2010) targets industrial standards in ontology form or information and product model standards for which ontology matching techniques are applicable, whereas our goal is the harmonization of industrial terminology standards.

3 THE EM APPROACH

The EM approach exploits the minimal common semantic information hidden in the structure of concept names. This implicit information, when decoded, can provide a hint of the relative position of a concept in a generic hierarchy, which further enables to infer possible relations between different concepts. The EM matching process includes five main steps:

1. Normalization
2. Quasi-synonymization
3. Equivalence matching
4. Name structuring
5. Hierarchy relationship search

3.1 Normalization

Before actual matching algorithms can be applied, some pre-processing is needed. In this case, hierarchy structures need to be first reduced into a flat list of concept names followed by a normalization phase. The purpose of the

normalization phase is to transform the entity strings into a common normal form in order to eliminate syntactic differences between entities in different lists. The other normalization operations applied are a combination of those proposed in (Euzenat 2007) and (Leukal 2006).

3.2 Quasi-synonymization

The next processing step is quasi-synonymization (q-syn). A term word of a source name can be replaced with its q-syn, after which this modified copy of the name element is added into the source list. The prefix quasi has been used, because these common term word pairs are only potential synonyms in an industrial context. The look-up table of q-syn term pairs is filled beforehand in an ad-hoc manner. For instance, this table contained the following term pairs: *maximum – upper limit; size – diameter; operation – operating; electrical – electric*.

3.3 Equivalence Matching

The actual matching phase begins with a search of equivalence relationships between source and target concepts. The most obvious argument supporting entity equivalence is the full string equality of their names. A simple string-based matching method was adequate in this case, since only full similarity of strings was accepted as proof of concept equivalence. The produced alignment set is labelled with "A=B".

3.4 Name Structuring

In order to find hierarchy relationships between entity names, the EM method analyzes the inherent structure of the names. The applied naming conventions determine this structure. According to the general naming rules that have been defined in terminology work standards (see Chapter 2), the names of class and property concepts can be composed of more than one *term word*.

The most important name part is the *root term* carrying the main meaning of the concept. When the target concept is an equipment class, the root term is of type *object class term* and in the case of an equipment property concept it is of type *property term*. Besides, the root term can be preceded by, one or more *qualifier terms* that specialize or constrain the basic meaning declared by the root term. In addition, *a representation term* is a word, or a combination of words, that semantically represent the data type (value domain) of a data element. (ISO/IEC 11179-5:2010)

One extra category has been defined for the EM method. The names of properties can have an additional object class term that represents a constraint for the domain scope of that property. If such a term is the last term in a name phrase, separated from the root term by one of predefined prepositions, it is categorized as a *scope qualifier term*. Altogether, five term part categories are considered by EM method:

1. object class terms
2. property terms
3. qualifier terms
4. scope terms
5. representation terms

The following list contains some examples of concept name structuring. The first two are equipment class names and the last two are property names. Qualifier terms have been separated by curly brackets and representation terms by square brackets. A slash separates the preposition and scope qualifiers from the root term (underlined).

1. {Tube heat} exchanger
2. {Piston} compressor
3. {Bearing inlet} pressure /of oil
4. {Manufacturer} [name]

3.5 Hierarchy Relationships

A necessary condition for finding any correspondence relations is that both names under comparison have the same root term, i.e. *root term equality*. If this condition is true, the type of the found correspondence relationship can be defined based on the structuring information of the both entity names. The following rules were applied (the source and target names are assumed to be otherwise similar, apart from the differences stated by the rules):

- If the source name has one preceding qualifier more than the target name, the target entity is a direct *super class* of the source entity (label "A<B").
- If the source name has one preceding qualifier less than the target name, the target entity is a direct *sub-class* of the source entity (label "A>B") .
- If the names differ only in their scope term (representation term) parts, an *equality with scope* (*equality with representation)* relationship will be assumed (label: "As=B", "Ar=B).
- If both names have the same number of qualifier terms, but the first terms are different, a *sibling* rela-

tionship is assumed (label "A--B").

Figure 1 presents three different correspondence relations that can be found for the source concept *"centrifugal pump"* applying equality matching and the simple rules above. Table 1 lists some relationships found for property concepts using the above rules.

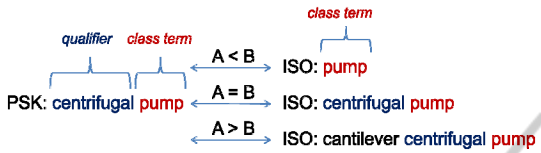

Figure 1: A concept name structuring and three possible semantic correspondence relations.

Table 1: Correspondence relations between properties.

PSK:Property	Rel.	ISO:Property
Diameter of drum	As=B	Diameter
Frequency of vibrator	As=B	Frequency
Bearing inlet pressure of oil	As<B	Inlet pressure
Manufacturer	A=Br	Manufacturer name

4 MATCHING RESULTS

At the beginning of the harmonization process, two PSK standards were selected to be harmonized with the initial set of *ISO 15926-4 Reference Data (2010)*. The first of them, *PSK5965 Equipment Classes and Subclasses,* (321) was matched against the equipment class sets of ISO 15926-4 (5465) and the second standard, *PSK5980 Data Element Dictionary,* (353properties) against the property sets of ISO 15926-4 (1986).

The source and target entity sets have many differences. ISO 15926-4 (target set) contains in total ten times more entities than the PSK standards, and those entities have been categorized into several context collections. The taxonomy structure is also different. ISO 15926-4 has a deep generalization hierarchy at least up to eight levels, whereas the number of levels in the PSK5965 standard has been limited to two (main class – subclass). Furthermore, only ISO 15926-4 includes textual descriptions for equipment classes and properties.

4.1 Comparing Alignment Results

The matching tasks were first carried out using the EM tool. In addition, the S-Match tool (S-M) was used for the same tasks, in order to evaluate the capabilities of an extrinsic method using general language WordNet dictionary (WordNet) and a structural method in this special case. The S-M approach applies a name structuring algorithm, in which it represents a concept node as a logical expression of the meanings of its term words. The local meaning of an atomic concept is its natural language meaning defined in WordNet (Giunchiglia, et. al. 2007).

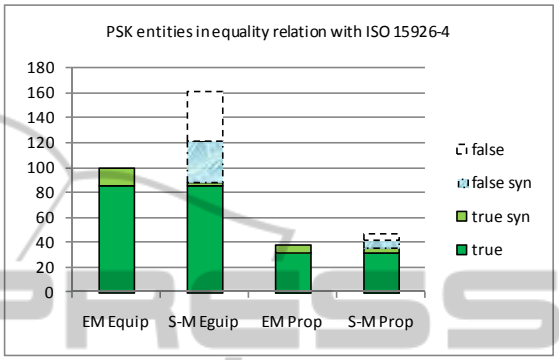

Figure 2: Number of PSK equipment classes (equip) and properties (prop) for which an equality match with ISO 15926-4 was found by EM and S-M tools.

4.1.1 Equality Relations

Figure 2 presents the number of source (PSK) concepts that were found to have an equality (eq) relation with at least one target (ISO) concept. It indicates that all matches (100;39) found by the EM tool were correct ones, i.e. true positives (dark and light green solid). S-M tool found only a few more correct matches (compared with the EM method without q-syn step) by browsing WordNet synsets; in total for (89;36) PSK entities.

However, the S-M tool found also several false positives. Post-analysis of the results revealed that about one half of those mismatches were actually close synonyms (blue texture dashed) in a natural language context (WordNet), but cannot be considered equal in an industrial context.

Furthermore, there were a large set of source classes (40;6) for which S-M found matches that were clearly incorrect (white dashed). Some of these false positive results are related to the internal representation of the concepts as logical expressions. The order of the term words in a multiword concept does not have any impact on interpretation of this expression. In contrast, the algorithm used in EM takes into account the order of the term words in a concept name so that this kind of mismatch is not possible.

4.1.2 Matching Quality Measures

Matching quality can be indicated by quality measures. The discovered alignment sets are compared against a reference set containing all the correct correspondence relations. In this case, the reference set is the final refined matching set produced by the industrial expert group. The matching correctness measure is *precision,* defined as the percentage of discovered correct correspondences in the entire extracted alignment set. The completeness measure of matching is *recall*, defined as the percentage of discovered correct correspondences in the reference alignment set. The *F-measure* is computed as a harmonic mean of precision and recall. All these measures vary in the [0,1] range. (Euzenat, 2007).

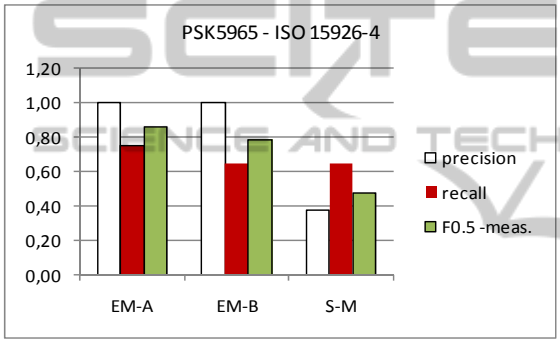

Figure 3: Matching quality measures of the equality alignment sets produced by EM method (EM-A), EM method without q-syn (EM-B) and S-M method (S-M).

These quality measures for equality alignment sets are presented in Figure 3. The results of S-M tool should be compared with the EM results without the ad-hoc q-syn step (EM-B). This comparison reveals that the recall measures (middle bar) are almost the same (0.65; 0.67), but the precision (left bar) of the S-M method is much lower (1.0; 0.38).

A separate test case was generated for evaluating the feasibility of a structural method for a case, in which the hierarchy structures are very different. The S-M tool was used in matching the source file (PSK5965) with one of the target category files (ISO Rotating), both in their full hierarchy format. The quality measures of this test case (S-M Hier) are presented in Figure 4, together with flat structure test cases (EM B, S-M Flat). These results indicate that the extra structural information did not provide any advantage in this special case. On the contrary, the recall measure was notably decreased.

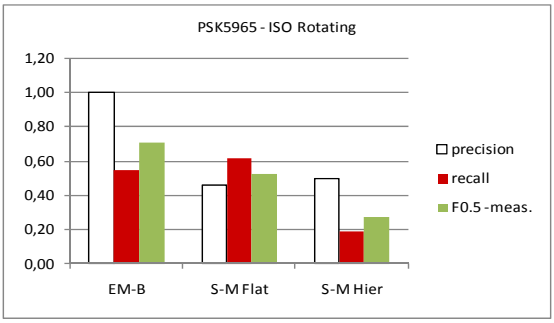

Figure 4: Matching quality measures of the equality alignment sets produced by EM method (EM-B), S-M method with flat hierarchy (S-M flat) and S-M method with full hierarchy (S-M Hier).

4.1.3 Super Class and Subclass Relations

The primary purpose of the EM name structuring algorithm is to find hierarchy correspondence relations with good precision. These relations include direct super class (A<B), direct sub-class (A>B) and sibling class (A- -B) relations. The following Table 2 lists the total number of some of these correspondence relations (alignments).

Table 2: Total numbers of alignments found by the EM method and S-M method.

PSK5965 - ISO 15926-4 Alignment Sets				
	A=B	A<B	A>B	A - - B
ElemMatch	100	174	361	922
	A=B	A<<<B	A>>>B	
S-Match	226	2651	8281	-

The results of the EM tool are compared with those produced by S-M tool. The S-M approach deduces semantic relations between entities by logical inference. Therefore, these relations are: equivalence (A=B), more general (A<<<B), more specific (A>>>B) and disjoint (⊥) (Giunchiglia, et.al., 2007).

These alignment sets have not been fully checked for validity by the industrial expert group. However, a partial analysis indicated that the spot checked alignments found by EM are true positives, whereas S-M has also found many clearly incorrect ones. The main reason for the high precision of the EM alignment sets is due to the fact that the name structuring rules comply with the naming conventions applied during terminology development.

5 CONCLUSIONS

The harmonization of terminologies used in industrial standards has been widely understood to be necessary for better interoperability of industrial information systems. Partial automation of the comparison and matching phases of the harmonization process is considered necessary, in order to reduce the workload of human experts and to speed up the process. However, advanced ontology matching methods are not directly applicable, because terminology dictionaries are not true ontologies and may differ greatly in their taxonomy structure and lexical content. We have developed a lightweight element level matching approach to address this problem. It is based on general concept name structuring rules defined in terminology work standards. This approach is applicable, when similar naming conventions have been applied.

This ElemMatcher approach was applied to an industrial terminology matching case in the first phase of the PSK - ISO 15926-4 harmonization process. The matching results indicate high matching precision for the equality alignment set and good precision of the other alignment sets. Additional experiments using advanced structural and extrinsic methods that exploit only general purpose dictionaries showed that no advantage was gained in this case study of industrial terminology standards harmonization.

REFERENCES

CEN Orchid Roadmap – Standardising information in the plant engineering supply chain. Parts 1-3. [referenced 2011-04-15]. Available at: http://www.cen.eu/CEN/sectors/sectors/isss/workshops/Pages/workshop orchid.aspx).

Euzenat, J. Shvaiko, P., 2007. Ontology Alignment. *Springer*. ISBN 978-3-540-49611-3.

Fiorentini, X., Rachuri, S., Ray, S., Sriram, R., 2009. Towards a method for harmonizing information standards. *5th Annual IEEE Conference on Automation Science and Engineering*. Bangalore, India, August 22-25, 2009.

Giunchiglia, F., Yatskevich, M., Shvaiko, P., 2007. *Semantic Matching: Algorithms and Implementation*. Technical Report # DIT-07-001, Department of Information and communication Technology, University of Trento.

Giunchiglia, F., Dutta, B., and Maltese, V., 2009. Faceted Lightweight Ontologies. Conceptual Modeling: Foundations and Applications. Alex Borgida, Vinay Chaudhri, Paolo Giorgini, Eric Yu (Eds.) *LNCS,*Vol. 5600, Springer, pp 36-51.

ISO 704:2000,Terminology work — Principles and methods. [referenced 2011-04-15]. Available at: http://www.iso.org/iso/store.htm

ISO 860:2007 Terminology work – Harmonization of concepts and terms. [referenced 2011-04-15]. Available at: http://www.iso.org/iso/store.htm

ISO 15926:2007, Industrial automation systems and integration — Integration of life-cycle data for process plants including oil and gas production facilities. [referenced 2011-04-15]. Available at: http://www.iso.org/iso/iso_catalogue/catalogue_tc

ISO 15926-4:2010, Industrial automation systems and integration — Integration of life-cycle data for process plants including oil and gas production facilities - Part 4: Initial reference data. [referenced 2011-04-15]. (ed1 files available at: http://ng.tc184-sc4.org/index.cfm?PID=804&FID=56498&r=/)

ISO/IEC 11179-5:2010 Information technology - Metadata registries. Part 5: Naming and identification principles. [referenced 2011-04-15]. Available at: http://metadata-stds.org/11179/

Jean-Marya, Y., Shironoshitaa, E., Kabuka, M., 2009. Ontology matching with semantic verification. *Web Semantics: Science, Services and Agents on the WorldWideWeb 7*. Elsevier. 235–251

Lauser, B., et al., (2008), 'Comparing human and automatic thesaurus mapping approaches in the agricultural domain'. In *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications*, pp. 43–53.

Leukel, J., 2006. Controlling Property Growth in Product Classification Schemes: A data management Approach. *Enterprise Information Systems: 8th International Conference*, ICEIS 2006.

PSK Standardisation. [website]. [referenced 2011-04-15]. Available at: http://www.psk-standardisointi.fi/

Rahm, E., Bernstein, P., 2001. A survey of approaches to automatic schema matching. *T.he VLDB Journal 10*, Springer. pp. 334–350

Shvaiko, P., Euzenat, J. 2005. A Survey of Schema-Based Matching Approaches *LNCS 3730*, Springer. pp. 146–171.

Shvaiko, P., Euzenat, J., 2008. Ten Challenges for Ontology Matching. *LNCS 5332*, Springer. pp. 1164–1182.

Uslar, M., Rohjans, S., 2009. Ontology-based Integration of the Heterogeneous Standards IEC 61970 and 61850. In: *Proceedings of International ETG-Kongress 2009*. VDE Verlag Gmbh. Paper 1.59.

WordNet - A Lexical Database for English. Princeton University. [referenced 2011-04-15] http://wordnet.princeton.edu/

Zhan, P., Jayaram, U., Kim, O., Zhu, L., 2010. Knowledge Representation and Ontology Mapping Methods for Product Data in Engineering Applications. *Journal of Computing and Information Science in Engineering*. June 2010, Vol. 10.