Enrichment of Ontology-Based Competence Profiles with Semistructured Wiktionary Data

Vladimir Tarasov[1], Andrew Krizhanovsky[2]

[1] School of Engineering, Jönköping University, P.O. Box 1026, 55111 Jönköping, Sweden
[2] St.Petersburg Institute for Informatics and Automation RAS, 14th Line, 39, 199178 St.Petersburg, Russia
vladimir.tarasov@jth.hj.se, andrew.krizhanovsky@gmail.com

Abstract. Competence supply methods can contribute to information supply solutions by providing information on available competences and making this information accessible to the decision-makers. When competence profiles are stored in a machine-readable form, competence demand can be translated into a formal query to find matching profiles. However, this requires to use exactly the same skill names as the ones from the profile representations. This paper proposes to enrich ontology-based competence profiles with the help of Wiktionary data accessible via a SPARQL endpoint. The endpoint makes it possible to get lexicographic information (definitions, translations, synonyms) from Wiktionary. The procedure of enrichment with synonyms and system architecture supporting it are developed and tested in an experiment with profiles of software developers. The experiment showed the viability of the approach.

Keywords: Competence supply, competence profile, ontology, Wiktionary, synonyms

1 Introduction

The field of competence supply and management offers concepts and approaches which can be applied to identify, describe and search for competences at both organisational and individual levels. For example, for flexible supply networks it is important to describe what the different suppliers can contribute in order to find required production capability or services [1]. Providing information on available competences and making this information accessible to the decision-makers contributes to intelligent information supply solutions in an enterprise.

When competences of workers are represented as machine-readable profiles, a competence demand description can be translated into a formal query to find matching profiles. However, it is necessary to utilize exactly the same skill name as the ones used in the profile representation. This makes specification of competence demand very rigid because people would normally use different names of the same skill. This paper proposes an approach for enrichment of competence profiles with synonyms extracted from the Wiktionary thesaurus. Competence profiles are stored in

an OWL ontology, which can be enhanced based on the data of the English Wiktionary accessible via a SPARQL endpoint.

The paper is structured as follows. Section 2 introduces ontology-based competence profiles, Wiktionary relational database, and SPARQL client via D2R server. Section 3 describes the enrichment method and system architecture. Section 4 presents an example of using the Wiktionary data via SPARQL for enrichment of competence profiles. Section 5 summarizes the work.

2 Background

2.1 Ontology-Based Competence Profiles

To represent competence profiles in a machine-readable form we use ontologies. An example of an ontology describing collaborative competences for engineering design can be found in [2]. The ontology is built to include three major perspectives: general competence, cultural competence, and occupational competence. The first part represents general competences. They are subdivided into problem solving competence, planning and designing competence, and competence for team work. The cultural competences are composed of language competence and intercultural sensitivity to take into consideration abilities to act in a multicultural environment.

The last part is occupational competence that reflect competences in the field of engineering in question and different technical competences in this field. The person's education is represented with educational fields, which reflect educational areas studied by him/her and relevant for engineering design. The educational field class has several sub-classes representing broad fields, narrow fields and detailed fields. The work experience is represented with occupational groups, which shows the person's present and past jobs that are of significance for engineering design. The occupational group class has several sub-classes representing major groups, sub-major, minor and unit groups consecutively. This part of the competence ontology is shown in Fig. 1.

The constructed ontology includes several competence profiles that model persons who have worked in collaborative software design teams in short-term software development projects. Each profile describes the skills and abilities possessed by the developers and consists of instances belonging to the classes from the ontology.

2.2 Machine-Readable Wiktionary

Wiktionary[1] is a multilingual and multifunctional dictionary that is freely available and contains huge database of words with translations to many languages. The Machine-Readable Wiktionary is a project[2] intended for extraction of different types

[1] http://www.wiktionary.org
[2] http://code.google.com/p/wikokit/

of lexicographic data stored in the Wiktionary: definitions, thesaurus and translations. Only the thesaurus of the English Wiktionary is used in this paper. There are several types of semantic relations (paradigmatic relations) extracted from the Wiktionary: synonyms, antonyms, hypernyms, hyponyms, meronyms, holonyms and troponyms. An example of the first two types of semantic relations is given in Fig. 2.

The dump of the English Wiktionary (as of October 30, 2010) was the source data for our experiments. The created database of the machine-readable Wiktionary contains[3]:

- 1 731 784 total entries,
- 297 902 English entries,
- 227 437 total semantic relations (i.e. pairs of words, e.g. synonyms, antonyms, etc.) between English words, French words, Japanese words, etc.,
- 72 587 semantic relations between English words.

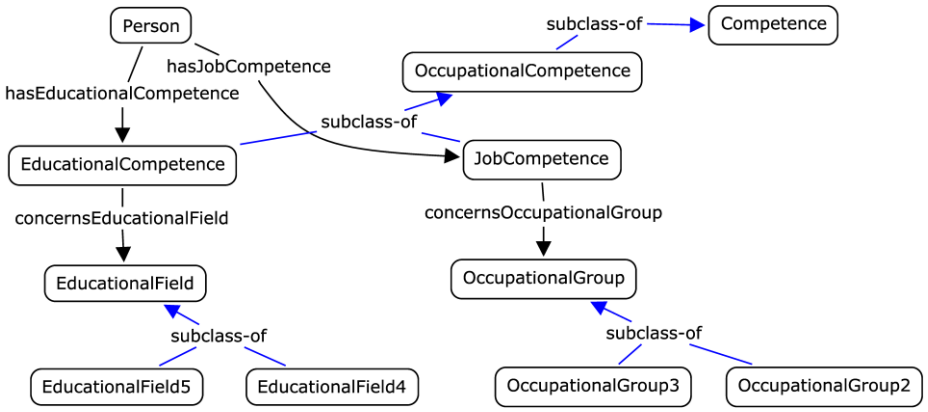

Fig. 1. A fragment of the competence model showing occupational competence

2.3 Wiktionary via SPARQL

With the help of D2R server [3] the data in the machine-readable Wiktionary are exposed as an RDF store. Thus, lexicographic information extracted from the Wiktionary is accessible via SPARQL queries. D2R server uses RDF and SPARQL languages in order to provide access to the relational database [3]. D2R accepts SPARQL queries from the web and rewrites them to SQL queries by using a specially prepared file (a D2RQ mapping file). The D2RQ mapping file has to be created only once. After that it is possible to access the relational database via SPARQL. SPARQL queries will be automatically translated on-the-fly into SQL by D2RQ platform. Therefore there is no need to replicate the database into RDF store.

[3] http://bit.ly/ixRmm5

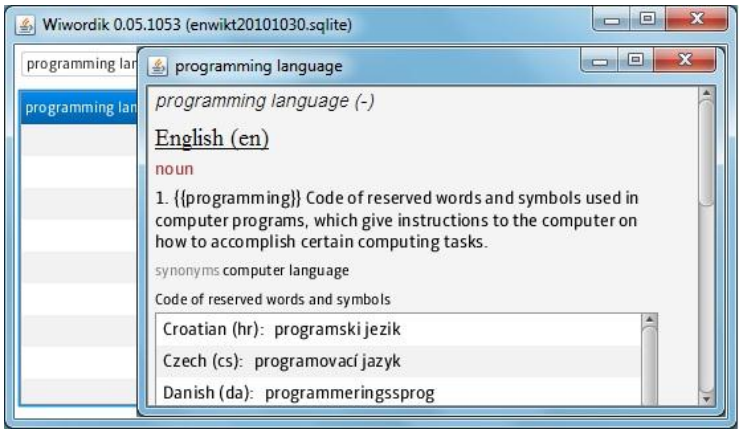

Fig. 2. The ”programming language” Wiktionary entry with semantic relations (synonyms)

The direct way to find synonyms in the database of the parsed English Wiktionary[4]: is from the entry to words, which are listed in the section Synonyms (e.g. programming language → computer language in Fig. 2). The following code contains an example of a long[5] SPARQL request for the machine-readable Wiktionary, which uses the ”direct way” to find synonyms. Input data for this request are (i) a language code (with value ”*en*”, i.e. English language in accordance with ISO 639-3 used in wiki projects), (ii) a Wiktionary entry (”*programming language*”), (iii) a type of relation (”*synonyms*”).

```
SELECT ?relationWord
WHERE {
  ?lang wikpa:lang_code ”en”; wikpa:lang_id ?langId.
  ?page wikpa:page_page_title ”programming language”;
    wikpa:page_id ?pageId.
  ?lang_pos wikpa:lang_pos_page_id ?pageId;
    wikpa:lang_pos_lang_id ?langId; wikpa:lang_pos_id
?langPosId.
  ?meaning wikpa:meaning_id ?meaningId;
    wikpa:meaning_lang_pos_id ?langPosId.
  ?relation_type wikpa:relation_type_name ”synonyms”;
    wikpa:relation_type_id ?relationTypeId.
  ?relation wikpa:relation_meaning_id ?meaningId;
    wikpa:relation_relation_type_id ?relationTypeId;
    wikpa:relation_wiki_text_id ?wikiTextIdRel.
  ?wiki_text wikpa:wiki_text_id ?wikiTextIdRel;
    wikpa:wiki_text_text ?relationWord.
 }
```

[4] http://code.google.com/p/wikokit/wiki/d2rqMappingSPARQL

[5] The query is long because the scheme of the machine-readable dictionary is inherently complex with a big number of tables and relations between them.

3 Enrichment of Competence Profiles

As noted in Sect. 1, using exactly the same skill name, as used in the competence profiles ontology, in a SPARQL query describing competence demand can be problematic. To be able to use synonyms of the skill name in a SPARQL query, it is necessary to add synonyms to individuals (instances of classes) representing skills/abilities. Sect. 3.1 describes the procedure of adding synonyms from Wiktionary to ”skill” individuals in the ontology, while Sect 3.2 presents the architecture of a system supporting such enrichment of competence profiles. After implementing the procedure and architecture, one can utilize either the initial skill name or its synonyms in SPARQL queries specifying competence demand.

3.1 Enrichment with Synonyms

The proposed method of profile enrichment is based on adding synonyms to individuals representing skills/abilities. The choice of only synonyms is based on the fact that the synonyms constitute more than 80 % of the 59103 semantic relations in Wiktionary[6]. Hence, using other types of relations will give little improvement right now. Individuals representing skills/abilities are those used to describe competences of people (e.g. `EducationField` class with its subclasses shown in Fig. 1). First, it is necessary to add the datatype property `skillName` to all such individuals in the ontology. Initially, this property contains only the original skill name. Then, the enrichment procedure is executed to search Wiktionary for synonyms and add them using the datatype property `skillName`} if found. The following pseudocode iterates over ”skill” individuals, which are first grouped into superclasses like `EducationField` by employing a classifier.

```
Input Ontology, SkillSets; {sets of individuals (classes) representing skills}
Output Ontology; {with modified sets of skills}
begin
  reasoner.classify(Ontology);
  for class∈ SkillSets do
    IndividualsSet := class.listIndividuals();
    for individual ∈ IndividualsSet do
      skill := individual.getPropertyValue(”skillName”);
      SynonymSet := WiktionaryClient.getRelatedWordsByEntry(skill);
      for synonym ∈ SynonymSet do
        individual.addLiteral(”skillName”, synonym);
    end-for
  end-for
end
```

[6] http://en.wiktionary.org/wiki/User:AKA_MBG/Statistics:Semantic_relations

3.2 Architecture of the System

To support the procedure of enrichment of competence profiles with synonyms (see Sect. 3.1), we propose a system architecture shown in Fig. 3. The *competence profile finder* component searches for profiles matching against competence demand queries with the help of Jena Framework[7] used to access the competence profiles ontology. The *Wiktionary SPARQL client* tries to retrieve synonyms of skill names. The *D2R server* provides access to the machine-readable Wiktionary via a SPARQL endpoint (see Sect. 2.3). The *D2RQ engine* translates SPARQL queries into SQL ones using Jena Framework and the mapping file. The *Wiktionary relational database* stores data from the Wiktionary dump.

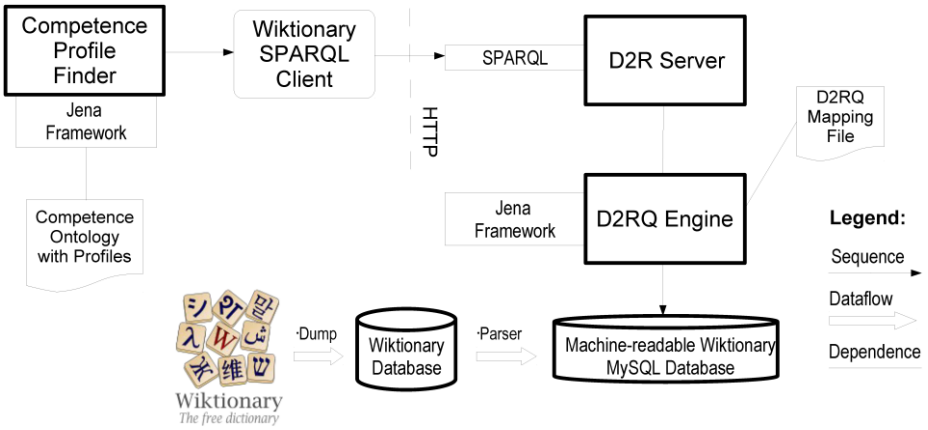

Fig. 3. Architecture of the platform integrating the competence profile finder with the machine-readable Wiktionary accessible via SPARQL queries

4 Experiment

This section presents results of experimental implementation of the enrichment method and system architecture described in Sect. 3. The goal of this experiment was to show viability of the proposed integration of the competence profile finder system, SPARQL and Wiktionary. The implementation was written in Java and we used Jena Framework for OWL[8] ontology management and SPARQL query processing as well as the D2RQ platform for accessing SQL databases via SPARQL. In the experiment we used the collaborative design ontology that was described in Sect. 2.1. The only difference is that the datatype property *skillName* was added to all the individuals representing skills.

[7] http://jena.sourceforge.net
[8] http://www.w3.org/TR/owl2-overview

Let us consider an example of competence demand. This example is deliberately simplified in order to fit the corresponding SPARQL query in the paper and make the query more understandable. We need to find a software engineer with knowledge of computer science and programming languages as well as work experience of systems analysis. This demand can be specified in the form of the following SPARQL query:

```
SELECT DISTINCT ?name
WHERE {
  ?person rdfs:label ?name .
  ?person cdo:hasJobCompetence ?job_comp .
    ?job_comp cdo:concernsOccupationalGroup ?occupGroup .
    ?occupGroup cdo:skillName "systems analyst" .
  ?person cdo:hasEducationalCompetence ?educ_comp1 .
    ?educ_comp1 cdo:concernsEducationalSubject ?educSubject1.
    ?educSubject1 cdo:skillName "computer language" .
  ?person cdo:hasEducationalCompetence ?educ_comp2 .
    ?educ_comp2 cdo:concernsEducationalSubject ?educSubject2.
    ?educSubject2 cdo:skillName "computing" .
}
```

This query uses the synonyms "computer language" for the term "programming language" and "computing" for "computer science". If this query is run against the ontology, no profiles will match because they initially use only the original terms. To extend the competence profiles with synonyms, the Wiktionary SPARQL client is run to obtain a list of semantically related terms from the English Wiktionary (see Sect. 2.2). This database is accessed via the SPARQL query described in Sect. 2.3. As a result a list of synonyms is retrieved if found (for the entry "programming language" see Fig. 2). After the enrichment, a matching competence profile is found. The updated "skill" individuals together with a fragment of the matched profile are shown in Fig. 4. During the enrichment other synonyms were found and added to the profiles but this is not shown for brevity.

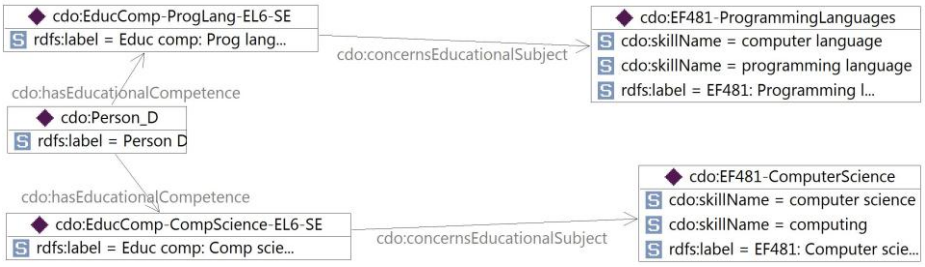

Fig. 4. A fragment of person D's competence profile enriched with Wiktionary related terms

5 Conclusions

In this paper we have proposed the method for enrichment of ontology-based competence profiles with synonyms extracted from Wiktionary. The enrichment procedure and system architecture supporting it were developed. During the

experiment, the database of the parsed English Wiktionary was used. The possibility of using synonyms extracted from Wiktionary via SPARQL for profile enrichment was successfully verified. The proposed approach can be employed for enrichment of other types of profiles, e.g. profiles of users of digital libraries (an example of such profiles is given in [4]).

However, the use of SPARQL can lead to problems. A server which provides a SPARQL endpoint could be easily broken or overloaded by poor, heavy or erroneous requests. This may slow down the enrichment procedure. To investigate this problem, we need to test our approach on bigger collections of competence profiles, which will require retrieval of more synonyms.

Another interesting problem to study is when the enrichment procedure should be executed. In our experiment we called it after no matching profiles had been found. An alternative way could be to call the procedure on a regular schedule. More experimentation is needed to test it.

To improve results, several Wiktionaries can be integrated into one machine-readable dictionary, since different Wiktionaries contain both overlapping and unique data (see the analysis of English Wiktionary and Russian Wiktionary in [5]). Using all the types of semantic relations simultaneously during enrichment can allow for result improvement as well. The use of several languages can be beneficial for competence supply and other applications such as digital libraries.

Acknowledgments. This work was financed by the Swedish Institute within the project CoReLib by grant # 00760/2010, by the Russian Foundation for Basic Research by grants # 09-07-00066, # 09-07-00436 and # 11-01-00251, and by the Russian Academy of Sciences within the research program ”Intelligent information technologies, mathematical modelling, system analysis and automation”.

References

1. Sandkuhl, K., Tarasov, V.: Modelling competence demand for exible supply networks. In: 13th IFAC symposium on Information Control Problems in Manufacturing, IFAC (2009) 103-108
2. Tarasov, V., Lundqvist, M.: Modeling collaborative design competence with ontologies. International Journal of e-Collaboration 3(4) (2007) 46-62
3. Cyganiak, R., Bizer, C.: A semantic web front-end to existing relational databases. Poster at BXMLT2006, Berliner XML Tage (2006) http://richard.cyganiak.de/2008/papers/d2r-server-bxmlt2006.pdf.
4. Sandkuhl, K., Smirnov, A., Mazalov, V., Vdovitsyn, V., Tarasov, V., Krizhanovsky, A., Lin, F., Ivashko, E.: Context-based retrieval in digital libraries: Approach and technological framework. In: Digital Libraries: Advanced Methods and Technologies, Digital Collections: Proceedings of the XIth All-Russian Research Conference RCDL'2009. (2009) 151-157
5. Krizhanovsky, A.A.: The comparison of Wiktionary thesauri transformed into the machine-readable format. CoRR abs/1006.5040 (2010) http://arxiv.org/abs/1006.5040.