

# THESAURUS REVIEW SYSTEM FOR COLLABORATION

**András FÖRHÉCZ**  
**Advisor: György STRAUSZ**

## **I. Introduction**

In the European Union citizens are welcome to study or work abroad. We should take advantage of the integrated European labour market, but to take part in such an international cooperation is somewhat difficult. One of the most important factors for both parties is to clearly understand the expected skills and competencies.

In recent years EU institutions have developed tools to establish the transparency of qualifications, like the Europass CV [1]. But all these suffer from the same weakness: the core concepts of these tools – terms for skills and competencies – are neither standardised nor internationally compatible, thus they can only offer limited support.

## **II. Project background**

The Leonardo da Vinci Pilot Project DISCO (European Dictionary on Skills and Competencies, [2]) intends to fill this gap by providing a common terminology, a multilingual thesaurus on vocational skills and competencies. Having strong international background, the project has operational partners from Austria, the United Kingdom, Germany, France, Belgium, the Czech Republic, Lithuania and Hungary. Most of the countries already have national thesauri or at least a compilation of possible qualifications. I took part in this project by supporting the development of the methodology and creating a thesaurus review system for partner collaboration.

## **III. Project goals**

The main result should be a translation tool for skills and competencies, which is missing from European transparency tools at the moment. The translation tool consists of a multilingual thesaurus and an online tool for individuals, which offers querying, browsing and supporting automatic translation of qualifications.

The common terminology means not only a standardized English thesaurus which describes vocational skills and competencies, but it contains the national translations as well. By translation the terms' meaning may slightly change, and although it cannot be completely avoided, it has to be kept down as much as possible. Certainly, the semantic distortions should be documented.

We expect to have approximately 2.000–3.000 terms and 1.000–5.000 synonyms per language. Besides different translations, terms should be mapped to concepts of existing thesauri related to qualifications. There are two tightly related standard collections. ISCED [3] is established by UNESCO for the compilation of national and international education statistics. It also contains the hierarchy of the fields of education. ISCO [4] organises occupations in a hierarchical framework. Although it has a looser connection with qualifications, the higher levels should be mapped to DISCO, in order to assist in searching required skills for a given occupation.

## **IV. Mapping and merging**

As mentioned earlier, the purpose of the project is to complete a multilingual thesaurus on the basis of already available national collections in the domain of skills and competencies.

“Thesauri are designed as an agreement and compromise on a set of shared terms for common concepts.” [5] The project partners already have their more or less strict conception about the assignment between terms and objects, now we have to achieve interoperability between these terminologies.

Thesaurus mapping is the process of identifying approximately equivalent terms and hierarchical relationships from one thesaurus to another. It has limited interoperability, depending on the number of found equalities. For example, if a field of interest is not mentioned in a foreign terminology, we will not be able to translate it at all.

Full integration can be achieved with thesaurus merging: construction of a common thesaurus with all concepts involved, and mapped from the source terminologies. Certainly, this task is substantially complex: besides finding similarities, we also have to resolve differences.

## V. Methodology

There was a lot of effort made developing collaboration tools, especially in the field of ontology compilation. These tools make it possible for different users to modify the same knowledge base at the same time (e.g. Ontolingua, KAON Engineering Server).

The project does not gain as much from a real collaboration tool as it would seem at first. The size and complexity of the thesaurus implies that it must be composed in a centralized manner. Otherwise, it would be impossible to guarantee consistency (e.g. to avoid duplicates).

After translating national thesauri into English, a charged partner tries to merge them into a central English terminology – only one sector at a time. Certainly, this cannot be performed without errors, because the partner may not be aware of all national specialities, possibly whole fields specific to a country. The final terminology is developed through alternating merging and review phases.

Review phases are performed with a custom online tool, which is able to import and track thesaurus changes. This way the review phase is separated from the engineering tools, the partner merging the thesaurus can use advanced software of own choice. All partners can suggest modifications and discuss accurate semantics for individual terms on the review interface.

## VI. Recent and future work

The project is just entering the review phase. A specification of the review system has been written and accepted based on the described methodology and former discussions between the project partners. I have developed the review system, but we have not got yet any feedback on its usability.

If all partners start to comment on the current pilot thesaurus, we will see how suitable the concept of this methodology is, and how hard it is to fulfil the partners’ requests to merge content. It can happen that the growing size of the thesaurus forces us to refine the applied methodology, and also improve the review system in parallel.

In the meantime I can start the development of a public user interface. End users with no former experience in this field need a more intuitive tool to explore the content and use for translation.

## References

- [1] The Europass Curriculum Vitae, URL: <http://europass.cedefop.eu.int/>
- [2] European Dictionary on Skills and Competencies, URL: <http://disco.youtrain.net/>
- [3] International Standard Classification of Education, ISCED 1997, URL: [http://www.unesco.org/education/information/nfsunesco/doc/isced\\_1997.htm](http://www.unesco.org/education/information/nfsunesco/doc/isced_1997.htm)
- [4] International Standard Classification of Occupations, ISCO-88, URL: <http://www.warwick.ac.uk/ier/isco/isco88.html>
- [5] M. Doerr, “Semantic Problems of Thesaurus Mapping,” *Journal of Digital Information*, 1(8), Apr. 2001. URL: <http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/>