Deliverable D2.2

Interoperability Assessment Environment
Final version

E. Francesconi, S. Faro, E. Marinai, V. Sandrucci
Institute of Legal Theory and Techniques
Italian National Research Council
Florence, Italy

Phase: PH2 - Interoperability

Abstract
This document contains the description of the interoperability assessment environment, in terms of the components foreseen to test the thesaurus mapping approach, the workflow for automatic thesaurus mapping, the modalities of mapping assessment. Finally a possible software architecture for thesaurus interoperability environment editorial activity is sketched.

Version 1.0, November 14th, 2007
Eurovoc Studies - LOT2
Institute of Legal Theory and Techniques
Italian National Research Council
via de’ Barucci 20
50127 Firenze
Italy

tel: +39 055 43995
fax: +39 055 4399605
http://www.ittig.cnr.it

Corresponding author:
Enrico Francesconi
tel: +39-055-4399665
francesconi@ittig.cnr.it
http://www.ittig.cnr.it
Contents

1 Introduction 4

2 Selected data samples for interoperability assessment 4

3 Mapping example for interoperability assessment: the “gold standard” 5
   3.1 Criteria to create the “gold standard” 5
   3.2 Tools to construct the “gold standard” 7

4 The proposed Thesaurus Mapping workflow 13

5 Testing the proposed approaches for thesaurus mapping 15

6 A proposal of a software architecture for Thesaurus Interoperability
   Environment editorial activity 17
   6.1 Deployment diagram: Architecture 17
   6.2 Class diagram: Conceptual Domain Model 17
   6.3 Use Case diagrams: Actors 18
   6.4 Use Case diagrams: User use cases 19
   6.5 Use Case diagrams: Project Manager use cases 19
   6.6 Activity diagram: Project Management activity 20
   6.7 Use case diagrams: Thesaurus Mapper use cases 21
   6.8 Use case diagrams: Mapping Reviewer use cases 21
   6.9 Use case diagrams: Mapping Manager use cases 22
   6.10 Use case diagrams: System use cases 23
1 Introduction

The thesaurus mapping methodologies identified and described in Deliverable D1.5 – “Thesauri KOS analysis and selected thesaurus mapping methodology on the project case-study” are expected to be tested on selected samples of the thesauri of interest, identified in the first phase of the project.

In this report the interoperability assessment environment is described. In particular the data set selected to test the interoperability approach (Section 2) as well as a tool, developed within the project to produce a thesaurus mapping “gold standard” on the selected data set are described (Section 3). Moreover the workflow for automatic thesaurus mapping (Section 4) and the modalities of mapping assessment (Section 5) are discussed. Finally a possible software architecture for Thesaurus Interoperability Environment editorial activity is sketched (Section 6).

Part of these arguments have been introduced in D1.5; in this deliverable they are revised and expanded according to the most recent achievements of the project.

2 Selected data samples for interoperability assessment

In this study the proposed methodologies for thesaurus mapping are expected to be tested on a meaningful data sample selected from each thesaurus. A sample is considered “meaningful” if it allows to cover all the relations types within thesauri, and mapping relations, between thesauri, expected to be considered providing reliable statistics.

Following the results of the comparative analysis shown in D1.5, considering as starting point the EUROVOC fields, two areas of interest have been chosen to assess the interoperability approach:

1. Law
2. Employment and Working Conditions

In case the previous domains will not give meaningful results, the Social Questions area will be considered as well.

Such domains are able to fulfill the requirements for test:

• firstly they are domains covered by almost all the given thesauri;
• secondly they contains a number of descriptors and relations allowing to obtain meaningful statistics;
• thirdly they contain types of relations within each thesaurus and may establish mapping relations between thesauri able to cover, from a statistical point of view, all the mapping relations to be tested.

These area of interest and related term instances will be used to build up the “gold standard” of thesaurus term mapping examples and to validate mapping predictions of the system.
3 Mapping example for interoperability assessment: the “gold standard”

Interoperability between thesauri will be specifically assessed on data sample selected from the thesauri of interest. In order to evaluate the performance of automatic mapping algorithms an intellectual activity is needed to create a “gold standard”, namely a groundtruth file of thesauri term mapping examples (one for each couple of thesauri having EUROVOC as pivot) which represents the ideal set of expected correct mappings. It is aimed at:

1. tuning system heuristics: similarity measure threshold values will be tuned to obtain the best results with respect to the gold standard (performance convergence);

2. evaluating the performances of automatic mapping algorithms, comparing the ideal set of mapping examples with system predictions.

The “gold standard”, as well as the mapping strategies to be applied, will be involving thesauri descriptors, using English as pivot language.

For the purpose of comparison between the “gold standard” and the algorithms prediction, mapping relations of the “gold standard” will be described using SKOS Mapping, limited to the exactMatch, broadMatch and narrowMatch relations.

According to the project requirements of evaluating interoperability between EUROVOC and the other thesauri of interest (namely EUROVOC is to considered a pivot), the “gold standard” will contain relationships between EUROVOC and the other thesauri, and vice-versa.

3.1 Criteria to create the “gold standard”

Specific dispositions will be given to the experts in order to harmonize their criteria of creation of the “gold standard”. In particular, a part from the use of the English language as a pivot one to evaluate relations, criteria to establish specific SKOS mapping relations between terms of different thesauri will be given.

**Exact match** When a concept in EUROVOC corresponds exactly to one or more concepts in a target thesaurus according to the expert judgment, the relation is an exact match. Unlike the relations of broad and narrow match, this one is always applied when a correspondence can be identified (see for example Tab. 1).

**Broad/Narrow match** A broad equivalence is one where the source term is more specific in some sense than the target term. Similarly a narrow equivalence is one where the source term is more general in some sense than the target term or expression [1]. Following [2], when it is possible, mappings that are at least complete, and ideally optimal, have to be established.

A complete mapping [1] is one where a source term, having no exact equivalence in the target, is matched to at least one target term that is semantically broader and at least one target term that is semantically narrower. This means that all the
Table 1: Examples of exactMatch correspondence between terms in different thesauri. Row number 1 represents exactMatch involving identical descriptors included in all the thesauri; rows number 2 and 3 represent exactMatch involving descriptors which may be different or not included in all the thesauri but all referring to the same concept.

<table>
<thead>
<tr>
<th>EUROVOC</th>
<th>ECLAS</th>
<th>UNESCO thesaurus</th>
<th>GEMET</th>
<th>ETT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parliament</td>
<td>Parliament</td>
<td>Parliament</td>
<td>Parliament</td>
</tr>
<tr>
<td>2</td>
<td>Pay</td>
<td>Wages</td>
<td>Wages</td>
<td>Wage system, Pay policy</td>
</tr>
<tr>
<td>3</td>
<td>Freedom of movement</td>
<td>Free movement</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

Table 2: Examples of broadMatch and narrowMatch correspondences between terms in different thesauri. Row 1 represents a broadMatch correspondence between a descriptor of EUROVOC and descriptors of all the other thesauri. Row 2 represents a narrowMatch correspondence between an EUROVOC descriptor and descriptors of two other thesauri.

<table>
<thead>
<tr>
<th>EUROVOC</th>
<th>ECLAS</th>
<th>UNESCO thesaurus</th>
<th>GEMET</th>
<th>ETT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Regional parliament</td>
<td>Parliament</td>
<td>Parliament</td>
<td>Parliament</td>
</tr>
<tr>
<td>2</td>
<td>Internal law of religions</td>
<td>Islamic law</td>
<td>Islamic law</td>
<td>–</td>
</tr>
</tbody>
</table>

The following rules are given to the experts:

1. Relations have to be established only between descriptors (i.e. only the layer highlighted in the following table is involved in mapping implementation)
2. As discussed, mapping is to be complete and optimal, as a consequence:

(a) A term in a source thesaurus may have more terms in a target thesaurus with which sharing an exactMatch relation. An inexact or partial equivalent concept may be accepted as equivalent if this is considered to provide acceptable retrieval results in most cases;

(b) A term in a source thesaurus may have more terms in a target thesaurus with which sharing a narrowMatch relation. The criterion to be followed is to identify the term in a target thesaurus that is the broadest (the less specific) one among the narrower (the nearest narrower), then the relation can be extended to sibling terms;

(c) A term in a source thesaurus is to have only one term in a target thesaurus with which sharing a broadMatch relation: this term has to be the narrowest (the more specific) one among the broader (the nearest broader).

3.2 Tools to construct the “gold standard”

The “gold standard” construction intellectual activity will be carried out by experts of the specific chosen domains, using tools to make their work easier.

Some solutions have been evaluated to implement the “gold standard”, in particular Protégé PROMPT merging tool, Vocabulary Integration Environment (VINE), AIDA (see Deliverable D1.5 - Section 6.2).

The chosen solution  All these software have been tested. For different reasons they have been considered too problematic and not enough user-friendly (see also Deliverable D1.5 - Section 6.2 for a detailed analysis).

Therefore for the purpose of creating a term mapping “gold standard” for this project the more appropriate solution seems to be the development of a specific application, using tools familiar to the users, that is able to provide a user-friendly access to thesauri and simple functionalities to establish term relations for mapping.

Contrary to what was proposed in deliverable D1.5 - Section 6.2, the application for thesauri alignment will be based on a relational database. This choice has been preferred to the one based on Open-Office Calc, mainly because of the short time available within the project. The use of a relational database in fact allows to develop in a short time advanced functionalities for the users, using SQL facilities.

For the project tight schedule, the application has been developed on an MSAccess relational database: this solution guarantees to meet the expected deadline,
considering also the expertise allocated on this task; moreover it does not interfere with the end-user requirements whose systems are mainly based on Windows platform. The availability of such tool on other platforms (Linux, MacOS) can be foreseen for the future without any impact on data formats, since SQL guarantees interoperability among relational database applications.

Such a THesauri ALigning ENvironment (THALEN) will give the users the following basic functionalities in order to produce the “gold standard”:

- login/logout
- thesaurus loading
- parallel view of two thesauri
- search modalities:
  - browsing
  - term searching (using specific properties (descriptors, fields, etc.) or full text searching)
- mapping (term selection, choice of mapping relations, etc.)
- established term mapping summary
- RDF SKOS mapping relation exporting

Moreover this application can be used for the human validation of the automatic mapping as produced by the algorithms identified in D1.5 and implemented for this project, in case they will be used on large scale on the whole thesauri of interest and the involvement of an editorial staff for the mapping validation.

In Figs. 1 to 9 some screenshots of THALEN as developed for the project are shown.
Figure 1: The “about” of THALEN, the tool developed within the project for thesauri alignment.

Figure 2: Choosing an EUROVOC field and a target thesaurus for mapping.
Figure 3: Choosing an EUROVOC field and a field or group of the target thesaurus for mapping.

Figure 4: Visualisation of an EUROVOC microthesaurus and another thesaurus (or a subset of it). Possible functionalities: 1) textual searching on any information associated to single node (Descriptor); 2) expanding and collapsing nodes; 3) showing all the information associated to the selected nodes (Descriptor (prefLabel), Synonyms (altLabel), NT, BT, RT, Scope note); 4) selecting nodes for establishing mapping relations.
**Figure 5:** Visualisation of the thesauri full-text search results (terms are highlighted); search on specific descriptors properties (node information) can be carried out as well, by selecting specific check-boxes.

**Figure 6:** On-demand visualisation (using a specific button) of the node information.
Figure 7: Visualisation of the nodes already included in mapping relations.

Figure 8: Window used to establish mapping relations: using the buttons indicated by the arrows, the selected descriptors/nodes are transferred in the mapping window, then a mapping relation can be chosen and assigned to a couple of terms. A note explaining the reasons for the chosen mapping relation can be inserted.
4 The proposed Thesaurus Mapping workflow

According to the project specification specifications, a mapping between EUROVOC and the other thesauri of interest are expected. Other different mappings in fact might not be meaningful since some of them pertain to different domains (ex: GEMET focused on environmental issues, while ETT on learning issues).

Therefore in the proposed mapping strategy, EUROVOC will be acting as a reference, and the mapping strategies will be tested to and from EUROVOC terms. Nevertheless, this approach allows also to achieve interoperability among the other thesauri having EUROVOC as a mapping pivot. This technique reduces the computational complexity of the problem of multi-thesaurus interoperability (N-to-N mapping), since it works in a framework similar to the one considered for the problem of multilingual translation, where a pivot language is used for mapping. The use of a “pivot” language in a N-language environment allows the reduction of the number of bilingual thesauri from a factor $N^2$ to a factor $N$.

The proposed approach will be applied according to the typical divide et impera criterion. Thesaurus mapping solution here proposed will be firstly implemented at the level of fields and microthesauri or corresponding sub-structures (such as domains, chapters or groups). The basic assumption is that the methods for mapping, previously described and proposed for this study, give better results on semantically correlated sub-domains. Each field or microthesaurus is identified by a term or group of terms which are a lexical representation of a concept and a specific domain of interest.

The basic mapping methodologies will be applied to descriptors within corresponding microthesauri in their English version as a pivot language.

The system workflow, sketched also in Fig. 10, consists therefore in the following steps:

![Figure 9: Table where all the established mapping relations are shown.](image-url)
1. **SKOS Core transformation of each thesaurus:**
   it will be performed using XSLT techniques developing specific RDF SKOS Core one;

2. **Thesaurus term pre-processing:**
   thesaurus terms will be normalized so that digit characters and non-alphabetic characters (is any) will be represented by a special character, then other operation as *stemming* or the use of word stoplists (*stopwords elimination*) will be performed;

3. **Thesaurus term representation:**
   the three representations of a thesaurus term introduced in D1.5 - Section 5.2.2 will be implemented:
   - *Lexical Manifestation*;
   - *Lexical Context*;
   - *Lexical Network*

   The choice of the term connection degree to be used for a *Lexical Context* or a *Lexical Network* description (descriptors, non-descriptors, micro-thesaurus terms, descriptors definition, etc.) will be one of the aim of the test mapping.

   Some criteria can be given in advance:

   - in case of mapping implementation at the level of fields and microthesauri or corresponding sub-structures (such as domains, chapters or groups), terms representing the related sub-structured will not be inserted in Lexical
Contexts or Lexical Networks representation, since this information would be redundant;
• in case of mapping implementation at the level of the whole thesauri, terms representing sub-structures (ex. fields or microthesauri) will provide relevant information and they will be considered for Lexical Contexts or Lexical Networks representation.

4. Thesaurus term candidate selection for mapping:
the ranking functions described in D1.5 - Section 5.2.4 will be evaluated; according to the related term representation they are:
• Levenshtein Distance (for the Lexical Manifestation);
• Cosine Distance (for the Lexical Context);
• Graph Edit distance (or variants, for the Lexical Network).

5. Ranking among candidate terms and mapping implementation:
terms of the target thesaurus, represented according to one of the foreseen models, will be matched with the chosen term in a source thesaurus, represented with the same model, using a proper similarity distance.

Candidates terms of the target thesaurus will be ranked according to the similarity measure values \( sim \in [0, 1] \) and the semantics to the mapping relation will be assigned using proper heuristic threshold values (\( T_1, T_2 \in [0, 1] \))

- if \( sim < T_1 \) \( \Rightarrow \) exactMatch
- if \( T_1 < sim < T_2 \) \( \Rightarrow \) partial match (broadMatch or narrowMatch)
- if \( T_2 < sim \) \( \Rightarrow \) No Match

The problem of predicting the type of partial match in terms of broadMatch or narrowMatch can be solved using external linguistic resources as WordNet.

Possible alternatives for the mapping strategies are the mapping implementation: 1) at the level of the whole thesauri; 2) at the level of sub-structures (corresponding fields, microthesauri, etc.). The basic assumption is that the methods for mapping may give better results on semantically correlated sub-domains.

6. Representation of the semantics of mapping in SKOS Mapping:
the established relations between thesaurus terms will be described and stored using RDF SKOS Mapping standard.

5 Testing the proposed approaches for thesaurus mapping
The “gold standard” produced by experts will be used to assess the proposed methodologies for thesaurus mapping.

Given a couple of thesauri and chosen a term in a source thesaurus, among those selected to build the “gold-standard”, a system implementing the proposed methodologies for mapping will be able to provide a mapping prediction, proposing a ranking of possible matching terms according to a similarity measure value.
As discussed in Section 4, the semantic values of mapping will be assigned using specific threshold values. Such values will be properly tuned on the basis of the output of the system and fixed to the values which produce the best mapping performances with respect to the “gold standard”.

The system mapping performances will be assessed with respect to the “gold standard” as regards each single mapping relation type, using the typical Precision and Recall measures.

Given two thesauri and the related “gold standard” including the following SKOS Mapping types of relations $C = \{\text{exactMatch, broadMatch, narrowMatch}\}$:

- let $N_i$, $i \in C$ the number of the exactMatch, broadMatch or narrowMatch relations established by experts in the “gold standard”;
- let $N^t_i$, $i \in C$ the number of correct exactMatch, broadMatch or narrowMatch relations predicted by the system;
- let $N^f_i$, $i \in C$ the number of wrong exactMatch, broadMatch or narrowMatch relations predicted by the system.

the performances of the mapping methodologies will be measured in terms of

$$
\text{Precision}_i = \frac{N^t_i}{N^t_i + N^f_i}, \quad \text{Recall}_i = \frac{N^t_i}{N_i}
$$

where $i$ represents the semantic label of a SKOS Mapping relation ($i \in C$).

The global performances of the mapping system may be evaluated in two different ways [3]:

- **microaveraging**: precision and recall are obtained by summing over all individual decisions:

$$
\text{Precision}^\mu = \frac{1}{|C|} \sum_{i}^{C} \text{Precision}_i, \quad \text{Recall}^\mu = \frac{1}{|C|} \sum_{i}^{C} \text{Recall}_i
$$

where $\mu$ indicates microaveraging;

- **macroaveraging**: precision and recall are first evaluated “locally” for each SKOS Mapping relation type, and then “globally” by averaging over the results of the different relation types:

$$
\text{Precision}^M = \frac{1}{|C|} \sum_{i}^{C} \text{Precision}_i, \quad \text{Recall}^M = \frac{1}{|C|} \sum_{i}^{C} \text{Recall}_i
$$

where $M$ indicates macroaveraging.
6 A proposal of a software architecture for Thesaurus Interoperability Environment editorial activity

In this Section a proposal of a possible collaborative environment able to facilitate expert of different domains in carrying on an editorial activity on thesaurus mapping is described. This tool is designed to work in a distributed environment and it aims at validating thesaurus mapping automatic predictions, as well as at implementing a reliable thesaurus mapping knowledge base using semantic Web technologies.

The project of such an environment has been defined using UML diagrams on the basis of possible user requirements, and consists of:

- Deployment diagrams
- Class diagrams
- Use Cases diagrams
- Activity diagrams

Hereinafter the details of such diagrams as related to the whole software architecture as well as to specific categories of users are reported, along with a short description of the foreseen actions.

6.1 Deployment diagram: Architecture

![Architecture Diagram]

The designed architecture is a typical three tier web application, where data and software reside on a Server, while Clients acts in a distributed environment.

6.2 Class diagram: Conceptual Domain Model

In this diagram the main entities dealing with the application are sketched, in particular

- Data (thesauri and mapping relations to be implemented)
- Participants
- Mapping activities
6.3 Use Case diagrams: Actors

Going in a deeper detail, this diagram summarizes the possible typologies of the actors involved in the thesaurus mapping editorial activities

- System
- System Administrator
- Project Manager
- Project Users: Mapping Manager, Thesaurus Mapper, Mapping Reviewer
6.4 Use Case diagrams: User use cases

End-users, in this case the domain experts charged to implement the thesaurus mapping, share some basic activities which are summarized by this diagram:

- Login and/or Registration
- Create/Retrieve/Update/Delete profile
- User communications

6.5 Use Case diagrams: Project Manager use cases

The Project Manager activities are more related to the coordination of the activities without being involved in the mapping activities, therefore his main activities can be summed up as follows:

- Create/Close project
• Assign and authorize participants
• Create and assign activities

6.6 Activity diagram: Project Management activity

This diagram sketches the communication activities between the project manager and a user of the system.
6.7 Use case diagrams: Thesaurus Mapper use cases

In this use case diagram the specific activities of a system end-user (a thesaurus mapper) are illustrated. They can be summed up into:

- Mapping thesauri implementation and justifications
- Asking for reviewers
- Mapping deployment

6.8 Use case diagrams: Mapping Reviewer use cases

The Reviewer activities are sketched in this use case diagram, they basically consist in

- Concept exploration
• Mapping feedback

In particular the last activity aims at giving to the end-user motivations of possible mapping errors or more specific criteria to be used for other similar critical cases.

6.9 Use case diagrams: Mapping Manager use cases

This diagram illustrates the activities of the Mapping Manager, whose main tasks are

• Collecting mapping user predictions
• Solving mapping ambiguity
• Results publishing
6.10 Use case diagrams: System use cases

Finally the diagram of the system use cases is here reported, where the main functionalities of the system as regards the automatic thesaurus mapping implementation are shown:

- SKOS RDF Validation
- Automatic mapping methodologies implementation:
  - String matching
  - Vector matching
  - Graph matching
References

