D1.5 – Thesauri KOS analysis and selected thesaurus mapping methodology on the project case-study

Final Version

S. Faro, E. Francesconi, V. Sandrucci

ITTIG-CNR – Institute of Legal Information Theory and Techniques

First Review Meeting - Luxembourg, October 16th, 2007
Overview

- Introduction to the Thesaurus Interoperability problem
- Analysis of the thesauri for the project case study
- Overview of Schema/Ontology Mapping methodologies
- The proposed approach for thesaurus mapping
- Standards for implementing the proposed methodology
Problem of accessing heterogeneous data sources in a distributed environment;

Terminological resources (thesauri or ontologies) can guarantee a better quality in document indexing and retrieval;

Cross-collections retrieval:
- providing queries from a single interface using a specific thesaurus as support (where available), and retrieving pertinent documents from different collections.

Quality of retrieval in single collections
- linked to availability of specific thesauri

Quality of retrieval in cross-collections
- linked to interoperability among thesauri
Using a particular thesaurus for querying a collection

Mapping such a thesaurus:
- to thesauri in other languages
- to more specialized vocabularies
- to different versions of the thesaurus

to obtain a retrieval from different collections which is coherent to the original query
Interoperability among Thesauri: the case study

- **EUROVOC** the main EU thesaurus considering issues of specific and common interest for the EU and its Member States

- **ECLAS** the European Commission Central Libraries thesaurus

- **GEMET** GEneral Multilingual Environmental Thesaurus

- **UNESCO Thesaurus** developed by the United Nations Educational, Scientific and Cultural Organisation

- **European Training Thesaurus (ETT)** a thesaurus providing support to indexing and retrieval vocational education and training documentation in the European Union
Two objects:

- **Thesauri Structure**
  in order to provide meaningful guidelines to thesauri representation for the interoperability test

- **Content (subjects coverage) analysis**
  in order to provide useful information for choosing and selecting data set for the interoperability test

- Only the English version of each thesaurus was considered
Thesauri Structure

All the Thesauri of interest are built up according to ISO standard 2788-1986 “Guidelines for the establishment and development of monolingual thesauri”

Considered issues:

- Hierarchical structuring
- Relationships
- Scope notes and definitions
- Mono/poly hierarchy relating to fields and NT/BT relationship
- Form of terms (number)
Hierarchical structuring

<table>
<thead>
<tr>
<th>EUROVOC</th>
<th>ECLAS</th>
<th>UNESCO</th>
<th>GEMET</th>
<th>ETT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fields/Domains</td>
<td>04 POLITICS</td>
<td>09 - Trade</td>
<td>1. Education</td>
<td>ATMOSPHERE (air, climate)</td>
</tr>
<tr>
<td>Microthesauri (MT)/Chapters</td>
<td>0406 political framework</td>
<td>09.01 - Demand. Market. Consumption</td>
<td>1.05. Education sciences and environment</td>
<td>1. Training</td>
</tr>
<tr>
<td>Sub-chapters</td>
<td>-</td>
<td>09.01.01 - Trade. Service industry</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Layers of the hierarchical structure are often marked not only by titles but also by numerical codes
The level of deepness of the hierarchical relationship among descriptors differs from one thesaurus to another (e.g. EUROVOC goes up to NT5, ETT to NT4).
### Descriptors and non-descriptors in mapping

<table>
<thead>
<tr>
<th>Eurovoc</th>
<th>Eclis</th>
<th>Unesco</th>
<th>Gemet</th>
<th>ETT</th>
</tr>
</thead>
<tbody>
<tr>
<td>MT 4421 personnel management and staff remuneration</td>
<td>13.07 - Wages. Wage incentives</td>
<td>MT 6.80 Personnel management</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>13.07.00 - Wages. Severance pay</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BTI remuneration of work</td>
<td>BT Personnel management</td>
<td>BT fee</td>
<td>BT working conditions</td>
<td></td>
</tr>
<tr>
<td><strong>Pay</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13.07.00.0850 - Wages</td>
<td>Wages</td>
<td>Wage system</td>
<td>Wage</td>
<td></td>
</tr>
<tr>
<td>SN Payments made for work performed</td>
<td>SN Payment made for work performed</td>
<td>System which compensates the employees with a fixed sum per piece, hour, day or another period of time, covering all compensations including salary</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>UF</strong></td>
<td><strong>Remuneration</strong></td>
<td><strong>Salaries</strong></td>
<td><strong>Salaries</strong></td>
<td><strong>UF</strong></td>
</tr>
<tr>
<td><strong>Salary</strong></td>
<td></td>
<td></td>
<td></td>
<td>basic wage</td>
</tr>
<tr>
<td><strong>Wages</strong></td>
<td></td>
<td></td>
<td></td>
<td>efficiency wage</td>
</tr>
<tr>
<td>NT! deduction at source</td>
<td></td>
<td></td>
<td></td>
<td>personal income</td>
</tr>
<tr>
<td>hourly wage</td>
<td></td>
<td></td>
<td></td>
<td>remuneration salary</td>
</tr>
<tr>
<td>low pay</td>
<td></td>
<td></td>
<td></td>
<td>wage differential</td>
</tr>
<tr>
<td>piece work pay</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RT</td>
<td>tax on employment income wage cost (4026) wage earner (4411)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UF</td>
<td>Fees</td>
<td>UF Wage determination</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remuneration</td>
<td></td>
<td>NT Fringe benefits</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Salaries</td>
<td></td>
<td>NT Minimum wage</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>NT Wage policy</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RT Income</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RT Labour</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RT Labour economics</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RT Performance appraisal</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In all the thesauri **scope notes** are present and are referred to a part of descriptors (e.g. in EUROVOC are 759, in ETT 210 and about 600 in UNESCO)

GEMET together with scope note also has a consistent number of **definitions**: more than 4,000 definition are available, which provide a useful glossary function where the semantic of the thesaurus structure might not be completely caught
<table>
<thead>
<tr>
<th>EUROVOC</th>
<th>ECLAS</th>
<th>UNESCO</th>
<th>GEMET</th>
<th>ETT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mono</td>
<td>Mono</td>
<td>Mono</td>
<td>Poly</td>
<td>Mono</td>
</tr>
</tbody>
</table>

*but Certain descriptors in fields “72 Geography” and “76 International Organizations” are polyhierarchical

*but Polyhierarchy exists for Country names*
<table>
<thead>
<tr>
<th>EUROVOC</th>
<th>ECLAS</th>
<th>UNESCO</th>
<th>GEMET</th>
<th>ETT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td>Plural</td>
<td>Singular/Plural</td>
<td>Singular</td>
<td>Singular</td>
</tr>
<tr>
<td><em>plural when the singular does not correspond to normal usage</em></td>
<td><em>singular where there is a distinction of meaning</em></td>
<td></td>
<td><em>a limited number of terms in plural form, to prevent change of meaning</em></td>
<td><em>plural where it seems more appropriate</em></td>
</tr>
<tr>
<td>eurovoc</td>
<td>eclas</td>
<td>unesco</td>
<td>gemet</td>
<td>ETT</td>
</tr>
<tr>
<td>-----------------------</td>
<td>--------------------------------------</td>
<td>--------------------------</td>
<td>--------------------------------------------</td>
<td>----------------------------</td>
</tr>
<tr>
<td>MT 2831 culture and religion</td>
<td>05.04 - Ethics. Religions</td>
<td>MT 3.20 Religion</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>05.04.04 - Religious institutions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT1 Religion</td>
<td>BT Religious institutions</td>
<td>BT Religious buildings</td>
<td>BT community facility</td>
<td>BT society</td>
</tr>
<tr>
<td>church</td>
<td>05.04.04.0100 - Catholic church</td>
<td>church</td>
<td>churches</td>
<td>religion</td>
</tr>
<tr>
<td></td>
<td>05.04.04.0150 - Christian churches</td>
<td></td>
<td>A building for religious activities</td>
<td></td>
</tr>
<tr>
<td></td>
<td>05.04.04.0175 - Orthodox church</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>05.04.04.0185 - Protestant churches</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NT1 ecclesiastical council</td>
<td>NT Church and State</td>
<td>RT Church and State</td>
<td></td>
<td>UF church</td>
</tr>
<tr>
<td></td>
<td>Religious communities</td>
<td>RT Religious communities</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RT Religious movements</td>
<td>RT Religious movements</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RT Religious sects</td>
<td>RT Christianity Monuments</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Content analysis

- general pictures of the distribution of descriptors per field in each thesaurus

- comparative analysis of the thesauri subject domains
  - for each EUROVOC field a possible correspondence has been identified with
    - one or more ECLAS chapters
    - one or more UNESCO MT
    - one or more GEMET groups or part of a group
    - one or more ETT fields or part of a field
Comparative analysis

Law

Employment and working conditions

Social questions
Thesaurus Mapping

Definition

The process of identifying terms, concepts and hierarchical relationships that are approximately equivalent between thesauri.

The problem is moved to concept equivalence definition.
Concept equivalence

Definition (*Instance-based equivalence*)
Two concepts are deemed to be equivalent if they are associated with, or classify the same set of objects.

Definition (*Schema-based equivalence*)
Two concepts are deemed to be equivalent if there exists a similarity among their features.
Thesauri heterogeneity

- Different word use
  - due to the variety of languages used to build thesauri;

- Different coverage
  - thesauri exist in different state of development, scope and specificity of user needs.

- Different semantics
  - thesauri of the same domain derive from different conceptualisations, due to different languages or classifications criteria;

- Different semantic relations
  - it can derive by the use of single hierarchies or multiple hierarchies.
The Semantics of Thesaurus Mapping

Notion of “equivalence expressions” given in the ISO Guidelines for the development of multilingual thesauri (ISO5964);

- Partial equivalence
- Exact equivalence
- Inexact equivalence
- Single to multiple equivalence

ISO5964 extensions based on Boolean expressions and ISO2788 relations (BT, NT, RT, USE, UF) (Doerr and Fundulaki (1998))

- “partial equivalence” becomes “broader equivalence” (is subset of)
- “exact equivalence” is interpreted as “same set as”
- “inexact equivalence” is interpreted as “overlaps with”
- “single to multiple equivalence” becomes a “compound” (Boolean expression of target terms with AND, OR, NOT) of “exact”, “broader” or “narrower equivalence”
Notion of “equivalence expressions” given in the ISO Guidelines for the development of multilingual thesauri (ISO5964):

- Partial equivalence
- Exact equivalence
- Inexact equivalence
- Single to multiple equivalence

ISO5964 extensions based on Boolean expressions and ISO2788 relations (BT, NT, RT, USE, UF) (Doerr and Fundulaki (1998))

- “partial equivalence” becomes “broader equivalence” (is subset of) or “narrower equivalence” (is superset of);
- “exact equivalence” is interpreted as “same set as”
- “inexact equivalence” is interpreted as “overlaps with”
- “single to multiple equivalence” becomes a “compound” (Boolean expression of target terms with AND, OR, NOT) of “exact”, “broader” or “narrower equivalence”
Thesaurus Mapping and Schema/Ontology Matching

- Thesaurus Mapping can be considered a sub-problem of the more general problem of schema/ontology matching.

**Definition**

Schema/ontology matching is the problem of producing a mapping between elements of two schemas that correspond semantically to each other. 

[Rahm and Bernstein, 2001]
A mapping element can be defined as a 5-uple: \((id, e, e', sim, R)\)

- **id**: unique identifier of the given mapping element;
- **e** and **e'**: entities (e.g. tables, XML elements, properties, classes) of the first and the second schema/ontology respectively;
- **sim**: confidence (similarity) measure of the correspondence between the entities in some mathematical structure (e.g. \([0,1]\) range);
- **R**: relation between the entities (e.g. equivalence (\(\equiv\)); more general (\(\sqsubseteq\)); disjointness (\(\perp\)); overlapping (\(\sqcap\))) \(e\) and \(e'\).
Definition

An alignment $A$ is a set of mapping elements of a pair of schemas/ontologies ($o$ and $o'$).

The matching process can be characterized by:

1. the use of an input alignment $A$ which is to be completed by the process;
2. the matching parameters $p$ (e.g., weights, thresholds);
3. external resources $r$ used by the matching process (e.g., thesauri, semantic networks, ontologies).
Overview of Schema/Ontology Matching approaches
Classification of Schema/Ontology Matching Approaches

[S. Faro, E. Francesconi, V. Sandrucci]

Further criteria:
- Match cardinality
- Auxiliary information used

Sample approaches
Instance-based Matching Approaches

- Characterized by data instances giving contents and meanings to schema elements

- Two main classes according to the type of information available:
  - plain text:
    - information retrieval techniques for document querying or term co-occurrence evaluation to verify term relations [van Hage et al., 2005];
    - automatic document classification
  - structured data: constraint-based characterization (numerical value ranges, character patterns)
Schema-based Matching Taxonomy

[Shvaiko and Euzenat, 2005]
Element-level matching: entities are analyzed in isolation, ignoring relations with other entities

Structure-level techniques: entities are analyzed in conjunction with other entities in a structure
Terminological methods

- string-based (considering the terms as sequences of characters)
- linguistic (based on the interpretation of these terms as linguistic objects)
Schema-based Matching Taxonomy: Kind of Input

- **Terminological methods**
  - string-based (considering the terms as sequences of characters)
  - linguistic (based on the interpretation of these terms as linguistic objects)

- **Structural methods**
  - attributes and their types (considering the internal structure of entities);
  - relational (considering the relation of entities with other entities);
Schema-based Matching Taxonomy: Kind of Input

- **Terminological methods**
  - string-based (considering the terms as sequences of characters)
  - linguistic (based on the interpretation of these terms as linguistic objects)

- **Structural methods**
  - attributes and their types (considering the internal structure of entities);
  - relational (considering the relation of entities with other entities);

- **Semantic methods** (based on semantic interpretation of the ontology to deduce the correspondences)
Element-level techniques

- String-based techniques:
  - string match similarity measure (prefix/suffix match, edit distance (Levenshtein distance), n-gram match)

- Language-based techniques: use of linguistic properties of words in some natural language
  - Tokenization: parsing into single words
  - Lemmatization: reduction to word morphological root (stemming)
  - Elimination of very frequent terms (stopwords elimination)

- Linguistic-based techniques: use of external linguistic resource (WordNet)

- Alignment reuse techniques: use of previous schemas/ontologies alignments

- Upper level formal ontologies: use of external knowledge resources (SUMO [Niles and Pease, 2001], DOLCE [Gangemi et al., 2003], etc.)
Element-level techniques

- **String-based techniques:**
  - string match similarity measure (prefix/suffix match, edit distance (Levenshtein distance), n-gram match)

- **Language-based techniques:** use of linguistic properties of words in some natural language
  - Tokenization: parsing into single words
  - Lemmatization: reduction to word morphological root (stemming)
  - Elimination of very frequent terms (stopwords elimination)

- **Linguistic-based techniques:** use of external linguistic resource

- **Alignment reuse techniques:** use of previous schemas/ontologies alignments

- **Upper level formal ontologies:** use of external knowledge resources (SUMO [Niles and Pease, 2001], DOLCE [Gangemi et al., 2003], etc.)
Element-level techniques

• String-based techniques:
  – string match similarity measure (prefix/suffix match, edit distance (Levenshtein distance), n-gram match)

• Language-based techniques: use of linguistic properties of words in some natural language
  – Tokenization: parsing into single words
  – Lemmatization: reduction to word morphological root (stemming)
  – Elimination of very frequent terms (stopwords elimination)

• Linguistic-based techniques: use of external linguistic resource (WordNet)
Element-level techniques

- **String-based techniques:**
  - string match similarity measure (prefix/suffix match, edit distance (Levenshtein distance), n-gram match)

- **Language-based techniques:** use of linguistic properties of words in some natural language
  - Tokenization: parsing into single words
  - Lemmatization: reduction to word morphological root (stemming)
  - Elimination of very frequent terms (stopwords elimination)

- **Linguistic-based techniques:** use of external linguistic resource (WordNet)

- **Alignment reuse techniques:** use of previous schemas/ontologies alignments
Element-level techniques

- **String-based techniques:**
  - string match similarity measure (prefix/suffix match, edit distance (Levenshtein distance), n-gram match)

- **Language-based techniques:** use of linguistic properties of words in some natural language
  - Tokenization: parsing into single words
  - Lemmatization: reduction to word morphological root (stemming)
  - Elimination of very frequent terms (stopwords elimination)

- **Linguistic-based techniques:** use of external linguistic resource (WordNet)

- **Alignment reuse techniques:** use of previous schemas/ontologies alignments

- **Upper level formal ontologies:** use of external knowledge resources (SUMO [Niles and Pease, 2001], DOLCE [Gangemi et al., 2003], etc.)
Structure-level techniques

- Graph-based techniques
  - input as labeled graphs (Graph match, Children/Leaves nodes similarity, Relation similarity)

- Taxonomy-based techniques
  - works according to is-a relations only

- Repository structures-based techniques
  - use already computed and stored similarities between schemas/ontologies (reduction of computational burden)

- Model-based algorithms
  - input described by symbolic rules and use of deductive methods and reasoning techniques [Giunchiglia et al., 2004] [Sotnykova et al., 2005]
Structure-level techniques

- **Graph-based techniques**
  - input as labeled graphs (Graph match, Children/Leaves nodes similarity, Relation similarity)

- **Taxonomy-based techniques**
  - they works according to *is-a* relations only
Structure-level techniques

- **Graph-based techniques**
  - input as labeled graphs (Graph match, Children/Leaves nodes similarity, Relation similarity)

- **Taxonomy-based techniques**
  - they work according to *is-a* relations only

- **Repository structures-based techniques**
  - use already computed and stored similarities between schemas/ontologies (reduction of computational burden)
Structure-level techniques

- **Graph-based techniques**
  - input as labeled graphs (Graph match, Children/Leaves nodes similarity, Relation similarity)

- **Taxonomy-based techniques**
  - they work according to *is-a* relations only

- **Repository structures-based techniques**
  - use already computed and stored similarities between schemas/ontologies (reduction of computational burden)

- **Model-based algorithms**
  - input described by symbolic rules and use of deductive methods and reasoning techniques [Giunchiglia et al., 2004] [Sotnykova et al., 2005]
Types of matching systems (architectural-centric view) [Rahm and Bernstein, 2001]

- **Individual matchers**: mapping based on single matching criterion;

- **Combining matchers**: mapping based on multiple matching criteria,
  - *hybrid matcher*: matchers are used in sequence;
  - *composite matcher*: matchers are used in parallel the results combined (by considering the average, the maximum, etc.).
Types of matching systems (user-centric view)
[Shvaiko and Euzenat, 2005]

- **Alignments as solutions**
  - alignment as a solution to the matching problem (continuous or discrete optimization problem [Euzenat and Valtchev, 2004] [Melnik et al., 2002]);

- **Alignments as theorems**
  - systems which rely on semantics and require the alignment to satisfy it (sub-category of alignments as solutions, expressed in semantic terms) [Giunchiglia et al., 2004];

- **Alignments as likeness clues**
  - algorithms producing indications for a user to select the alignment providing probability measure for a given proposed match [Do and Rahm, 2001] [Madhavan et al., 2001].
## Available Tools

<table>
<thead>
<tr>
<th>Available Tools</th>
<th>Element-level</th>
<th>Structure-level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Syntactic</strong></td>
<td><strong>External</strong></td>
<td><strong>Syntactic</strong></td>
</tr>
<tr>
<td><strong>Similarity</strong></td>
<td>string-based; data types; key properties</td>
<td>–</td>
</tr>
<tr>
<td><strong>Artemis</strong></td>
<td>domain compatibility; language-based</td>
<td>commonn thesaurus (CT); synonyms, broader, related terms</td>
</tr>
<tr>
<td><strong>Cupid</strong></td>
<td>string-based; language-based; data types; key properties</td>
<td>auxiliary thesauri; synonyms; hypernyms; abbreviations</td>
</tr>
<tr>
<td><strong>COMA</strong></td>
<td>string-based; language-based; data types</td>
<td>auxiliary thesauri; synonyms; hypernyms; abbreviations; alignment reuse</td>
</tr>
<tr>
<td><strong>NOM FOAM/QOM</strong></td>
<td>string-based; domains and ranges</td>
<td>application-specific vocabulary</td>
</tr>
<tr>
<td><strong>Anchor-PROMPT</strong></td>
<td>string-based; domains and ranges</td>
<td>–</td>
</tr>
<tr>
<td><strong>OLA</strong></td>
<td>string-based; language-based; data types</td>
<td>WordNet</td>
</tr>
<tr>
<td><strong>S-Match</strong></td>
<td>string-based; language-based</td>
<td>WordNet sense-based, gloss-based</td>
</tr>
</tbody>
</table>
The proposed approach for the Thesaurus Mapping problem
Phases to approach the **Thesaurus Matching** problem

1. **Identification of the problem specificities wrt Schema/Ontology Matching problem**
2. **Proposal of a possible formal characterization of the problem**
3. **Identification of**
   - methodology class
   - basic techniques
for the project case study
Identification of the problem characteristics

- Thesaurus mapping for the project case study is a problem of term alignments, where only schema information is available.

- It is not a problem of instance classification with respect to a predefined schema of classes (Instance-based matching).

- It is a problem where to measure the conceptual / semantic similarity between a term (simple or complex) in the source thesaurus and candidate terms in a target thesaurus (Schema-based matching).
Our proposal for Thesaurus Mapping formal characterization

- We propose to characterize the problem of Thesaurus Mapping ($\mathcal{T}\mathcal{M}$) as a problem of Information Retrieval ($\mathcal{IR}$)

- In $\mathcal{IR}$ the aim is to find the documents, in a document collection, better matching the semantics of a query

- Similarly, in $\mathcal{T}\mathcal{M}$ the aim is to find the terms, in a term collection (target thesaurus), better matching the semantics of a given term in a source thesaurus
Our proposal for Thesaurus Mapping formal characterization

- We propose to characterize the problem of Thesaurus Mapping ($TM$) as a problem of Information Retrieval ($IR$).

- In $IR$ the aim is to find the documents, in a document collection, better matching the semantics of a query.

- Similarly, in $TM$ the aim is to find the terms, in a term collection (target thesaurus), better matching the semantics of a given term in a source thesaurus.

<table>
<thead>
<tr>
<th>$TM$</th>
<th>$IR$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Term in source thesaurus</td>
<td>Query</td>
</tr>
<tr>
<td>Term in target thesaurus</td>
<td>Document</td>
</tr>
</tbody>
</table>
Definition

IR is 4-upla \([D, Q, F, R(q_i, d_j)]\) [Baeza-Yates and Ribeiro-Neto, 1999]

- \(D\) is a set of the possible representations (logical views) of a document in a collection
- \(Q\) is a set of the possible representations (logical views) of the documents with respect to the user information needs (queries)
- \(F\) is a framework for modeling document representations, queries, and their relationships
- \(R(q_i, d_j)\) is a ranking function, which associates a real number with \((q_i, d_j)\) where \(q_i \in Q\), \(d_j \in D\). Such ranking defines an ordering among the documents with respect to the query \(q_i\)
Our $\mathcal{T} \mathcal{M}$ formal characterization

**Definition**

We propose to characterize $\mathcal{T} \mathcal{M}$ as a 4-upla $[D, Q, F, R(q_i, d_j)]$ where:

- $D$ is the set of possible representations (logical views) of a term in a target thesaurus (in IR documents in a collection).
- $Q$ is the set of the possible representations (logical views) of a term in a source thesaurus (in IR queries to be matched with documents of the collections).
- $F$ is the framework of term representations in source and target thesauri.
- $R(q_i, d_j)$ is the ranking function, which associates a real number to a $(q_i, d_j)$ where $q_i \in Q$, $d_j \in D$, giving an order of relevance to the terms in a target thesaurus $d_j$ with respect to a term of the source thesaurus $q_i$. 
Isomorphism between $TM$ and $IR$
Instantiation of $\mathcal{TM}$ for the project purposes

1. Term pre-processing

2. Identification of suitable frameworks $F$ for $\mathcal{TM}$, namely the representations of terms in source ($Q$) and target ($D$) thesauri

   - for better representing the semantics of the terms in the thesauri;
   - in a way amenable for computation;

3. Identification of the ranking function $R(q_i, d_j)$ between source and target terms providing a measure of similarity between their semantics
Thesaurus term pre-processing

Pre-processing steps:

- to increase the statistical quality of terms
- to reduce the computational complexity of the problem

Typical pre-processing steps in document classification and information retrieval:

1. digit characters can be represented using a special character;
2. non alphanumeric characters can be represented using a special character;
3. lemmatization: reducing words to their morphological root (stemming)
4. elimination: elimination of very frequent terms (articles, prepositions, conjunctions) (stopwords elimination)

The efficacy of such pre-processing steps has to be proved in this work
Different terms can be used to identify the same concept

- in the same language (e.g. ‘pollution’, ‘contamination’, ‘discharge of pollutants’);
- in different language (e.g. EUROVOC EN term ‘water’ and IT term ‘acqua’)

$TM$ should aim at matching term meanings (the semantics of the terms) rather than formal (lexical) manifestations

Hypothesis

The more terms in source and target thesauri are semantically characterized, the more the system will be able to match them according to their meanings, enhancing mapping reliability
The proposed Logical Views of terms in source (Q) and target (D) thesauri

Each term (simple or complex) has to be described to capture its semantics.

The semantics of a term is conveyed:
1. by its morphological characteristics
2. by the context in which the term is used
3. by the relations with other terms

We propose to represent the semantics of a term in a thesaurus, according to an ascending degree of expressiveness, by:
1. its Lexical Manifestation
2. its Lexical Context
3. its Lexical Network
The proposed Logical Views of terms in source ($Q$) and target ($D$) thesauri

**Lexical Manifestation** of a thesaurus term
- a *string* of characters, in case normalized according to pre-processing steps

**Lexical Context** of a thesaurus term
- a *vector* of binary/weighted terms composed by the term itself, relevant terms in its definition and linked terms

**Lexical Network** of a thesaurus term
- a *graph* where nodes are terms along with related ones, and the labeled edges are semantically characterized relations between terms
A Lexical Manifestation

(Stemmed variation)

Parliamentary committees → Parliament$ committee$
A Lexical Context is a vector $\vec{d}$ of binary/weighted terms $[w_1, \ldots, w_{|T|}]$, where $T$ is the dimension of a target thesaurus vocabulary.
A Lexical Network is a direct graph \( g = (V, E, \alpha, \beta, L_V, L_E) \), where

- \( V \) denotes a finite set of nodes;
- \( E \subseteq V \times V \) is a finite set of edges;
- \( \alpha : V \rightarrow L_V \) is a node labelling function;
- \( \beta : E \rightarrow L_E \) is an edge labelling function;
- \( L_V \) is a set of node labels;
- \( L_E \) is a set of edge labels.
The proposed Framework $F$

- **Lexical Manifestations**, $F$ is composed of strings representing terms and the standard operations on strings.

- **Lexical Contexts**, $F$ is composed of t-dimensional vectorial space and linear algebra operations on vectors.

- **Lexical Networks**, $F$ is composed by graphs (described by nodes, edges and related labels) and the algebra operations on graphs.

Such frameworks provide also the intuition for constructing a ranking function $R$ for each chosen representation of the space elements (terms).
The proposed Ranking Functions ($R$)

- $R$ has to provide a similarity measure between terms in a source thesaurus and target thesaurus.

- Extended to a set of target terms $R$ provide a matching order among terms.

- $R$ may provide matching measures:
  - as values $\{0, 1\}$ (exact algorithms, compute the absolute solution to a problem)
  - in the range $[0, 1]$ (approximate algorithms, sacrifice exactness to performance)
The proposed Ranking Functions ($R$) for term *Lexical Manifestations*

### Definition

**Edit distance** (Levenshtein distance): minimum number of operations (insertion, deletion, or substitution of a single character) needed to transform one string into another

- applied on pre-processed strings through language-based techniques (Tokenization, Lemmatization (*Stemming*) and Elimination (*Stopword elimination*))

### Example (Edit distance $lev$)

$d_1 = \text{Politics}$ and $d_2 = \text{Political}$

- n. insertions 2 + n. deletions 1 = 3

$lev(d_1, d_2) = \frac{3}{9} = 0.33$

But: $d_1 = \text{score}$ $d_2 = \text{store}$

$lev(d_1, d_2) = \frac{1}{5} = 0.2$ (different concepts!)
The proposed Ranking Functions ($R$) for term *Lexical Context*

**Definition**

**Cosine Distance**: given $\vec{q}$ and $\vec{d}_j$ two *Lexical Contexts* of binary/weighted terms $[w_1, \ldots, w_{|T|}]$ in source and target thesaurus respectively, the Cosine Distance $\text{sim}$ between $\vec{q}$ and $\vec{d}_j$ is

$$\text{sim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \times \vec{q}}{|\vec{d}_j| \cdot |\vec{q}|}$$

where $|\vec{d}_j|$ and $|\vec{q}|$ are the norms of the vectors representing terms in target and source thesauri.
The proposed Ranking Functions ($R$) for Lexical Network

Being a Lexical Network a direct graph $g = (V, E, \alpha, \beta, L_V, L_E)$, the similarity between Lexical Networks is reduced to a problem of graph isomorphism.

Definition

Optimal Error-correcting graph matching / Graph Edit Distance

Minimum number of node deletions/insertions/substitutions and edge deletions/insertions/substitutions to transform a graph $g_1$ into a graph $g_2$
Using the notion of maximum common subgraph $mcs(g)$ two measure can be defined:

$$\delta(g_1, g_2) = 1 - \frac{|mcs(g_1, g_2)|}{\max(|g_1|, |g_2|)}$$

Using the maximum common subgraph, such a measure relates the size of the common parts of $g_1, g_2$ to the size of the graphs.
Alternatives to Graph Edit Distance

Considering two graphs $g_1$ and $g_2$, let $g_c = mcs(g_1, g_2)$, [y Gómez et al., 2000] defines the similarity between them as a combination of two values: their conceptual similarity $s_c$ and their relational similarity $s_r$.

**Definition**

The **conceptual similarity** $s_c$ expresses how many concepts the two graphs $g_1$ and $g_2$ have in common:

$$s_c = \frac{2n(g_c)}{n(g_1) + n(g_2)}$$

**Definition**

The **relational similarity** $s_r$ indicates how similar the relations between the same concepts in both graphs are:

$$s_r = \frac{2m(g_c)}{m_{g_c}(g_1) + m_{g_c}(g_2)}$$
Considering a graph $g$ to be matched with a graph $g_T$ as reference, a possible similarity measure can be [Cai et al., 2004]:

**Definition**

\[
\text{sim} = \frac{N_c(g, g_T) + E_c(g, g_T)}{N(g_T) + E(g_T)}
\]

- $N_c(g, g_T)$ is the number of nodes shared by graph $g$ and $g_T$
- $E_c(g, g_T)$ is the number of edges common to $g$ and $g_T$
- $N(g_T)$ is the number of nodes in graph $g_T$
- $E(g_T)$ the number is of edges in $g_T$
Standards used for implementing the proposed mapping methodology as well as link storage and management
RDF SKOS (Simple Knowledge Organisation System) standards will be used to represent terms and relationships.

- **SKOS Core**
  - defines classes and properties to represent the common features of a standard thesaurus
  - primitive objects are not terms, but abstract concepts represented by terms defined as an RDF resource

- **SKOS Mapping**
  - vocabulary to express matching (exact or fuzzy) of concepts from one concept scheme to another

- **SKOS Extension**
  - declare relationships between concepts with more specific semantics than the simple “broader-narrower”, such as class-instance or partitive relationships.

\[1\] http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/
SKOS Mapping aims to provide a vocabulary to express matching (exact or fuzzy) of concepts from one concept scheme to another

- **broadMatch**: the resources indexed against A is a subset of the resources indexed against B;
- **exactMatch**: the resources indexed against A is identical to the set of resources indexed against B;
- **majorMatch**: the resources indexed against A $\geq 50\%$ resources indexed against B;
- **mappingRelation**: the super-property of all properties expressing information about how to create mappings;
- **minorMatch**: $0 <$ the resources indexed against A $< 50\%$ resources indexed against B
- **narrowMatch**: the resources indexed against A is a superset of the resources indexed against B.
The part of SKOS used for this project

- The semantics provided by SKOS is able to analytically describe mapping relationships.

- For practical purposes, especially in automatic system where documents are not available, it is extremely difficult to state to what extent such concepts are equivalent (i.e., more or less than 50%): only exactMatch, broadMatch and narrowMatch will be used.

- For similar reasons, possible extensions or composition of SKOS Mapping relations using boolean operators, as well as SKOS Extensions properties, will not be considered.
*Modern Information Retrieval.*
Addison Wesley.

Address extraction: a graph matching and ontology-based approach to conceptual information retrieval.

Do, H. H. and Rahm, E.
Coma - a system for flexible combination of schema matching approaches.
*In Proceedings of the Very Large Data Bases Conference (VLDB).*

Similarity-based ontology alignment in owl-lite.
Sweetening wordnet with dolce.

S-match: an algorithm and an implementation of semantic matching.
In *Proceedings of the European Semantic Web Symposium (ESWS)*, pages 61–75.

*Graph Matching – Filtering Databases of Graphs Using Machine Learning Techniques*.

Generic schema matching with cupid.


