Supporting Scientific Collaboration in a Network of Excellence Through a Semantically Indexed Knowledge Map

Paola Velardi[1], Alessandro Cucchiarelli[2] and Michaël Petit[3]

1 Introduction

Automatic building of a thesaurus and its application to the management of a distributed research community has been one of the initial achievements of a European network of excellence, the INTEROP NoE, whose main objective is to support scientific advancements and dissemination actions in the field of enterprise and software interoperability. Two INTEROP work packages have contributed to this result.

The work-package WP1 "*Knowledge Map*" (KMap) aims at drawing a picture of the status of research in interoperability, and to keep this picture up-to-date in the future. The main objective of the KMap is to perform a periodic diagnostic of the extent of research collaboration and coordination among INTEROP partners. This diagnostic will support the formulation of recommendations to strengthen this collaboration and the better guidance of future research efforts of the various NoE partners. The work-package WPG "*Interoperability Glossary*" aims at building a thesaurus of interoperability terms, reflecting the three main INTEROP knowledge domains: ontology, enterprise modeling, software architectures and platforms. There are several benefits in creating a durable INTEROP glossary and associated facility:

1. *Semantic unification*: the thesaurus represents a mid-formal, shared view of the relevant domain concepts. This activity will let the semantics naturally *emerge* from applications and collaborative work;
2. *Classification/retrieval* of documents: glossary terms may be used as metadata for indexing documents and databases;
3. *Integration of competences*: a taxonomically structured set of terms can support the diagnostic task of the KMap, through the identification of

[1] Department of Computer Science, University of Roma "La Sapienza", via Salaria 113, 00198 Roma, Italy, velardi@di.uniroma1.it

[2] DIIGA, Università Politecnica delle Marche, via Brecce Bianche 12, 60131 Ancona, Italy, cucchiarell@diiga.univpm.it

[3] Department of Computer Science, University of Namur, rue Grandgagne 21, B-5000 Namur, Belgium, mpe@info.fundp.ac.be

semantic similarity patterns between the descriptions of partner competences collected in the KMap database.

In the following we present results related to the third objective: we show how the integration of competences in a distributed research community, like the INTEROP NoE, can be facilitated through the use of a KMap semantically indexed with a domain taxonomy. Section 2 summarizes the objectives of the KMap. Section 3 shortly describes the automatic taxonomy acquisition procedure. Section 4 presents a measure of semantic similarity of partner competences, based on the INTEROP taxonomy and computed on the KMap data. Finally, Section 5 presents a visualization tool to semantically navigate the KMap.

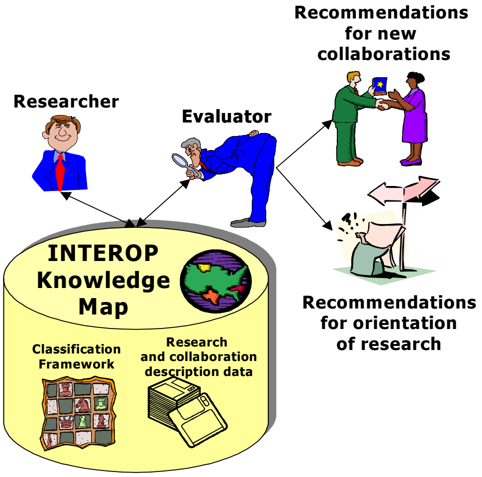

Recommendations for new collaborations
Researcher
Evaluator
INTEROP Knowledge Map
Classification Framework
Research and collaboration description data
Recommendations for orientation of research

Fig. 1 The INTEROP Kmap.

2 Building a Knowledge Map of Competences in a Network of Excellence

One of the major objectives of the INTEROP project, as all other Networks of Excellence funded by the European Community, is to tackle the problem of research fragmentation and to have a long-term integration effect on research in the field of interoperability. One of the tools that INTEROP is setting up to reach this objective is the INTEROP Knowledge Map (KMap) that aims at drawing and maintaining an up-to-date picture of the status of research in interoperability. The KMap will support a periodic diagnostic of the extent of research collaboration among INTEROP partners in order to support the formulation of recommendations to strengthen their collaboration and better orient the future research.

The KMap has the form of an application based on a database containing data about research activities, results and collaboration within INTEROP (see Figure 1).

It contains knowledge about existing research and industry institutions, groups and individuals with skills and competencies in interoperability. In order to meaningfully compare elements of this knowledge, one must classify them according to the different interoperability *domains of interest* they address. This framework took the form of a taxonomy and was defined through a specific INTEROP Work Package (WP Interoperability Glossary).

3 Automatic Learning of the Interoperability Taxonomy

The method used in the WP "Interoperability Glossary" aims ultimately at the semi-automatic creation of a *domain ontology* on interoperability, though currently it produces a structured (taxonomically ordered) set of concepts with attached natural language definitions, i.e. a *thesaurus*. The procedure is articulated in the following steps (A=automatic, M=manual):

1. (M) Starting from available documents exchanged among or published by the members of a web community, incrementally identify a larger set of related web accessible documents;
2. (A) Extract a domain lexicon *L* from the selected documents, i.e. a list of relevant terms in the subject domain;
3. (A) For each term, search extensively on the web for sentences that are candidate definitions for that term;
4. (A) Filter sentences to reduce noise (sentences that are not domain-pertinent, or non-definitions);
5. (M) Manually validate the obtained glossary *G* (add, correct, or exclude definitions) through a web interface[4];
6. (A) Parse definitions to extract the hypernym (kind of) information;
7. (M) Use extracted hypernyms as well as other available information (e.g. on-line linguistic ontologies) to arrange terms in a forest of taxonomically ordered sub-trees *T*, the *thesaurus*.

Phases from 1 to 5 have been already described in [1,2] and will not be further discussed here. In the following, details on phases 6 and 7 are provided.
Once a validated glossary is available, the subsequent step is to structure the glossary in a taxonomic order. This task can be facilitated through the use of automatic procedures described in this section. Discovering taxonomic relations between terms is accomplished first, by parsing definitions with a part of speech and syntactic parser, then, applying regular expressions [3] to the parsed sentences. We write regular expressions that impose constraints on a sentence structure at the

[4] The automatic process is not meant to fully replace humans, but simply to speed up the complex and time consuming task of creating a glossary in a new domain, like INTEROP. In [1,2], we have already performed an evaluation of the glossary extraction task, with encouraging results.

lexical, *part of speech* and *syntactic* level using an available parser, the TreeTagger[5].. Usually, well-formed definitions are provided in terms of *genus* (the kind, or *hypernym*, to which an entity belongs) and *differentia* (what differentiates the entity wrt the more general class), e.g. "*enterprise information integration* ***is the process*** *of integrating structured data from any relevant source for the purpose of presenting an intelligent, real-time view of the business to a business analyst or an operational application."* In this definition, the phrase that identifies the genus is marked in bold.

Not all definitions are well-formed in the abovementioned sense, e.g. "*component integration is obtained by composing the component's refinement structures together, resulting in (larger) refinement structures which can be further used as components*". Therefore, the objective of regular expressions is first, to verify well-formedness, and then to extract the "main" phrase, the one specifying the hypernym information.

An example of a regular expression used to verify the well-formedness criterion is the following: *r* = "^(PP)?(NP)+". This regular expression prescribes a sentence structure (for a *definition*) composed of a facultative prepositional phrase (^(PP)?) followed by the *main noun phrase* (NP), followed by anything else (+)". An example of sentence matching *r* is:

"domain model: [in the traditional software engineering perspective]$_{PP}$*,[a precise representation]*$_{NP}$ *of specification and implementation concepts that define a class of existing systems".*

An additional example of regular expression is the following:

p1="^(Refers|Referring)\\sto\\s(((a|the)\\s)?(type|kind)\\sof\\s)?(.*)"

The regular expression *p1* applies only *lexical constraints* and detects the presence of "cue" words like *refers*, *is a type of*, etc.

If a sentence is selected as being well formed, additional regular expressions are used to extract the *kind_of* (*hypernym)* information from the main NP.

For example, consider the regular expression r_1 = "^(A|D)?((V|C|,|J|N|R)*)(N)" that imposes *part of speech* constraints on a sentence fragment. Symbols in r_1 are part of speech (POS) tags, e.g. article (A), verb (V), adjective (J), noun (N) etc.

The previous definition of *domain model* matches both *r* and *r1*. When parsing this sentence with the TreeTagger we obtain:

Syntactic Analysis: (PP **NP** PP CNP RVP NP PP)

POS Analysis: (PAJNNN AJ**N** PNCNNWVANPJN)

The bold POS (**N**) represents the fragment selected as the hypernym returned by the matching of r_1, namely "*representation*", allowing us to conclude that:

$$domain\_model \xrightarrow{kind\_of} representation$$

Or-conjoined hypernyms are also handled, e.g. "The systematic *format* and technical *structure* that...", a sentence that returns two hypernyms, *format* and *structure*.

[5] TreeTagger is available at http://www.ims.unistuttgarrt.de/projekte/corplex/TreeTagger/Decision/TreeTagger.html. We augmented the TreeTagger with named entity recognition of Persons, Organizations and Locations.

Parsing definitions allows it to structure the terms in *L* in taxonomic order. However, ordering terms according to the hypernyms extracted from definitions has well-known drawbacks[6]. To reduce these problems, we proceeded as follows:

1. First, we arrange the terms in the lexicon *L* taxonomically according to simple *string inclusion*. String inclusion is a very reliable indicator of a taxonomic relation, though it does not capture all possible relations. This step produces a forest of sub-trees. Let ST_i be one of such trees, for example:

```
integration
      representation integration
      model integration
            enterprise model integration
      schema integration
      ontology integration
      knowledge integration
      data integration
      information integration
      application integration
      service integration
```

2. Then, we use hypernymy information extracted from definitions to capture additional taxonomic relations between terms *at the same level of generality* (e.g. in the example above: *representation, model, schema, ontology, knowledge, data, information, application,service*).
3. If terms have more than one selected definition, or have or-conjoined heads in the main NP, more than one hypernym is extracted by the algorithm previously described. However, we select only hypernyms *belonging to the set of domain relevant words in the domain* (see [1,2]). Hence for example, *knowledge integration* has the following extracted hypernyms: *information*, *fact-and-relationship* and *meaning*. Only the first is selected.
4. After step 3, component terms of the sub-trees ST_i have one or more hypernym associated. Given a term t: $t_l\, t_r$ (where t_l and t_r are left and right components of t, e.g. t=*enterprise application integration*, t_l=*enterprise application*, t_r=*integration*) we verify whether there is a multi-word term t' : $t'_l t'_r$ in the taxonomy such that $t_r = t'_r$ and either $t'_l \xrightarrow{kind\_of} t_l$ or $t_l \xrightarrow{kind\_of} t'_l$ (e.g. if t=*service integration* and t'=*application integration*, it holds that $service \xrightarrow{kind\_of} application$, and therefore $service\_integration \xrightarrow{\text{kind\_of}} application\_integration$).

Based on step 4, the taxonomy ST_i of the previous example, is re-arranged as follows:

[6] In [4] an analysis is provided of typical problems found when attempting to extract (manually or automatically) hypernymy relations from natural language definitions, e.g. attachments too high in the hierarchy, unclear choices for more general terms, or-conjoined hypernyms, absence of hypernym, circularity, etc.

```
knowledge integration
    representation integration
          model integration
                enterprise model integration
                schema integration
                ontology integration
    application integration
          service integration
    information integration
          data integration
```

From an initial set of 1566 extracted definitions, the procedure produced a forest of 621 sub-trees. The root nodes of these sub-trees were manually linked to a "core" taxonomy inspired by the Enterprise Ontology[7] and enriched with WordNet[8] concepts, to fill the gap between the core EO concepts and the 621 roots[9].

Though string inclusion was the main mechanism used to identify *kind-of* relations, taxonomic relations that are not explicit as string inclusions were identified by the algorithm. Overall, 254 (24%) of the total automatically detected taxonomic relations in ST are not of the string inclusion type. These taxonomic relations are 97% correct wrt the parsed sentence, i.e. the system correctly discards not well-formed sentences and correctly identifies the word representing the hypernym in the sentence.

However, the problem is *how well this automatic procedure approximates the task of taxonomically ordering concepts by a group of human specialists* in a given domain, considering e.g. the limitations outlined in [4]. A fine-grained evaluation will be performed by INTEROP partners soon, but in the meantime a preliminary evaluation has been conducted on a different domain. In that experiment: we considered a well-established on-line thesaurus, the AAT art and architecture thesaurus[10]. We applied our methodology to 814 glosses from the *Visual Works* sub-tree of the AAT thesaurus. The remarkable result (wrt analogous evaluations in literature e.g. [5,6,7]) is that in 34% of the cases the automatically extracted hypernym is the same as in AAT, and in 26% of the cases, either the extracted hypernym is more general than the one defined in AAT, or the contrary. Overall, in 60% of the cases there is a clear relation between the task performed by a highly qualified team of specialists and our automatic procedure. We may conclude that there is good hope that, indeed, the automatically acquired taxonomic relations will considerably speed-up the task of creating a domain taxonomy in INTEROP.

4 Computing the Semantic Similarity Measure

Starting from data stored in the INTEROP Knowledge Map (KMap) the first goal was the extraction of a set of research competences for each INTEROP partner and

[7] www.aiai.ed.ac.uk/project/enterprise/enterprise/ontology.html.

[8] www.wordnet.princeton.edu.

[9] The final taxonomy, though not yet validated by partners, is visible through the glossary validation interface at: http://interop-noe.org/workspaces/wpg/ps_interop/ps_glossary_view _terms?taxo=1&depth=1.

[10] http://www.getty.edu/research/conducting_research/ vocabularies/aat/.

the definition of a similarity measure able to express, in a single figure, the degree to which competences are common for any pair of partners. We used the preliminary version of the INTEROP taxonomy T, acquired according to the procedure outlined in Section 3.

The following information extracted from the KMap for each partner was used: (1) *Domains of interest* expressed by the partner by choosing one or more terms from a subset S of the INTEROP taxonomy *T*[11], (2) *Additional domains of interest* defined by partner (not present in the initial list S), (3) free *text descriptions of additional domains* given by partner, (4) *Titles* and *abstracts of publications* cited by the partner, (5) *Titles* and *descriptions* of *projects* in which the partner participate and (6) Short description of *software products* used by or produced by the partner.

All information related to points (3)-(6), collected in the KMap as English sentences, was parsed to extract additional terms $t \in T$ representing partners' competences not explicitly included in (1) and (2).

At the end of this phase, the following data were available for each partner: (1) a set of concepts $C \subset T$ representing the partner's domains of interest (as remarked, these are a subset of the WPG taxonomy T), resulting from the merge of terms entered in point 1 and 2 of the previous list and terms extracted by the linguistic processor; (2) a set of *generalised concepts* $C' \subset T$, i.e. the elements of the kind_of taxonomy T *reachable through a single generalization step* starting from terms in set C[12]. Topmost concepts are filtered out to avoid over generality. Let *D* be the total set of concepts (collected for all partners) in $C \cup C'$, and let V be the arity of *D*.

All this information has been mapped for each partner into a binary vector with dimension V. Given a partner P_A his associated vector VA can be seen as composed by two sub-vectors: VA_T representing terms indicated by the partner (in an explicit way or extracted by indirect information) and VA_C representing the generalised concepts related to terms in VA_T. The similarity measure between each pair of partners, $Sim(P_A,P_B)$, has been defined by considering:

a. *Direct matches*: the matches between terms in VA_T, e VB_T,
b. *Indirect matches*, first type: the matches between terms in VA_T not used in step (a) and generalized concepts related to VB_T and vice versa (two terms in VA_T, e VB_T are related by a kind_of link in T, e.g.

$$\textit{application integration} \xleftarrow{kind\_of} \textit{service integration}).$$

[11] Only a subset S of terms from the taxonomy, selected by the partners of WP1, were used (about 500). The INTEROP terminology includes over 1600 terms and, for practical reasons, it seemed unrealistic to ask partners to select from such a long list.

[12] An encouraging result was that free terms inserted by partners turned out to belong to the difference set T-S, except for one term, even if partners were not aware of T, but only of S.

c. *Indirect matches*, second type: the matches between generalized concepts related to terms in VA_T and VB_T, not used in steps (a) and (b) (two terms in VA_T, e VB_T have a common kind in T e.g:
ontology evaluation $\xrightarrow{kind\_of}$ *assessment* $\xleftarrow{kind\_of}$ *computer-aided assessment*)

For each case, the contribution to *Sim(*P_A*,*P_B*)* has been defined by using the *cosine similarity measure* (well known in the information retrieval field) applied to the terms' vectors of P_A and P_B. Given two vectors A and B with the same number of elements, the cosine similarity between them is defined as:

$$\cos(A,B) = \frac{\sum_i a_i b_i}{\sqrt{\sum_i a_i^2}\sqrt{\sum_i b_i^2}}$$

where a_i (b_i) is the i-th element of P_A (P_B).

By calling ***SIM1*** the result of this measure for case (a), ***SIM2*** and ***SIM3*** for case (b) and ***SIM4*** for case (c), we define the measure of *Sim(*P_A*,*P_B*)* as:

$$Sim(P_A, P_B) = \alpha SIM1 + \beta\left(\frac{SIM2 + SIM3}{2}\right) + \chi SIM4$$

where: $\alpha > \beta + \chi$

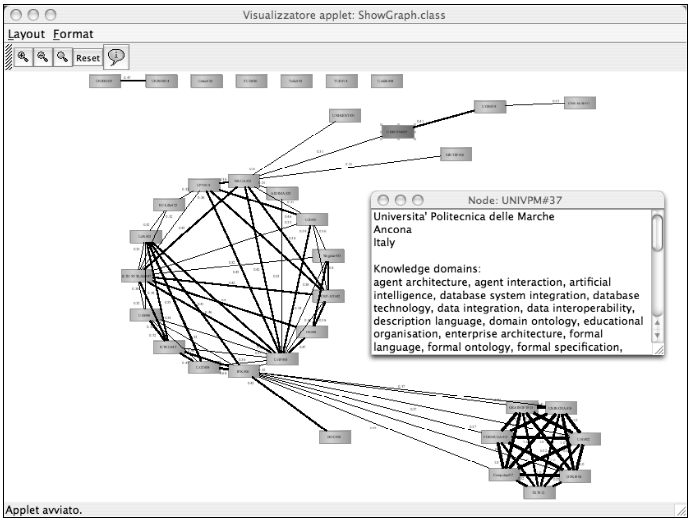
In our experiments, the parameters α, β and γ have been experimentally tuned to 1.0, 0.2 and 0.2 respectively.

5 Visualization of Partner Competences

As result of the similarity computation between each pair of INTEROP partners, we have a graph data structure, with nodes representing partners, and the generic edge between nodes P_A and P_B expressing the competence similarities *Sim(*P_A*,* B_B*)* between the pair of partners. A graphic interface has been developed to visualize this information in a comprehensible way, and is based on yFiles[13], a commercial Java class library. The interface uses the thickness of edges to reflect the value of the semantic similarity *Sim(*P_A*,* B_B*)* (see figure 2).

The user can select the layout to apply to graph display. For example, the circular layout draws circles formed by highly interconnected nodes. The latter has been especially useful for detecting clusters of partners sharing the same domains of competence. This is clearly shown in figure 2, where nodes are arranged as two circles, each composed of partners with high domain competences similarity. The figure also displays the information related to a node, in terms of partner's name and domains of competence.

[13] http://www.yWorks.com.



Node: UNIVPM#37
Universita' Politecnica delle Marche
Ancona
Italy

Knowledge domains:
agent architecture, agent interaction, artificial intelligence, database system integration, database technology, data integration, data interoperability, description language, domain ontology, educational organisation, enterprise architecture, formal language, formal ontology, formal specification,

Fig. 2. Competence Map of INTEROP partners.

This interface provides some initial diagnostics of the INTEROP K-Map that can be easily extended in the near future. For example, it is possible to identify partners with similar competences, and use this information to organise at best the cooperation between members. It is also possible to identify the concepts that are most "popular", i.e. those appearing on the highest number of edges and similarly, those for which there is limited competence in the network.

6 Conclusion

In this paper we presented an algorithm to automatically build a thesaurus of terms in a novel research domain, and an application in which taxonomic ordering betwen terms is exploited to improve the diagnostic of a Knowledge Map, aimed at facilitating and supporting synergy and collaboration in a distributed research community, the INTEROP NoE.

The proposed approach allows to speed-up the development of the thesaurus and requires less effort than with traditional manual approaches[14]. The obtained thesaurus might also be used in the INTEROP project for various other uses such as document indexing and semantic searching.

[14] With the help of figures provided by an experienced lexicographer, who developed several glossaries for publishing companies, we estimated a speed-up factor of more than 50%. See also [4] for details. We thank Orin Hargraves for providing us this information.

Though a preliminary evaluation aimed at eliminating clear errors of the automatic procedure has been already performed by a restricted team, a more fine-grained evaluation will be conducted by INTEROP partners in the continuation of the Interoperability Glossary Work Package.

Envisaged extensions to this work include: improving the thesaurus definition methodology by allowing the incorporation of additional information during the discovering of "kind-of" relationships; definition and use of other graphical views for the diagnostic of the KMap, such as a view showing connectedness among research domains (displaying domains as nodes and showing links thickness depending on e.g. existing collaborations or projects dealing with these two domains).

7 Acknowledgment

This work has been supported by the INTEROP Network of Excellence under contract n. IST-2003- 508011. Many thanks to Pierluigi D'Amadio and Francesca Neri who contributed to software developments.

8 References

[1] Navigli, R., Velardi, P.: "Automatic Acquisition of a Thesaurus of Interoperability Terms", Proc. of *16th IFAC World Congress* (http://www.ifac.cz), Praha, Czech Republic, July 4-8th, 2005.

[2] Velardi, P., Pòler, R., Tomàs, J.V.: "Methodology for the definition of a glossary in a collaborative research project and its application to a European network of Excellence" First Int. Conf. on Interoperability of Enterprise Software and Applications, February 23-25, Geneva Switzerland, 2005.

[3] Friedl, J.E.F: "Mastering Regular Expressions" O'Reilly eds., ISBN: 1-56592-257-3, First edition January 1997.

[4] Ide, N., Véronis, J.: "Refining Taxonomies extracted from machine readable Dictionaries."http://www.up.univ-mrs.fr/~veronis/pdf/1994rhc2.pdf, 1994.

[5] Fiszman, M., Rindflesch, T., Kilicoglu, H.: "Identifying Hypernymic Propositions in an Online Medical Encyclopedia", Proceedings of American Medical Informatics Association, 2003.

[6] Kashyap, V., Ramakrishnan, C., Rindflesch, T.: "Toward (Semi)-Automatic Generation of Bio-medical Ontologies", Proceedings of American Medical Informatics Association, 2003.

[7] Montemagni, S., Vanderwende, L.: "Large-scale resources: Structural patterns vs. string patterns for extracting semantic information from dictionaries" Proceedings of the 14th conference on Computational Linguistics - Volume 2, August 1992.