Methodology for the definition of a glossary in a collaborative research project and its application to a European Network of Excellence

Paola Velardi[1], Raúl Poler[2], José Vicente Tomás[2]

[1] University of Rome "La Sapienza", Via Salaria 113, 00198 Rome, Italy
{velardi@di.uniroma1.it}

[2] Polytechnic University of Valencia, Camino de Vera, s/n, 46002 Valencia, Spain
{rpoler, jotomi@cigip.upv.es}

Abstract. The aim of this paper is to describe the methodology of creation of a Glossary in a collaborative research project and its application to the Network of Excellence IST-508011 INTEROP "Interoperability Research for Networked Enterprise Applications and Software" for the definition of a glossary in the area of interoperability of enterprise applications and software. The proposed methodology is based on an adaptation of a method of the University of Rome for the semiautomatic acquisition of terms and definitions starting from a source of documents related to the research areas of a collaborative project.

1 Introduction

Knowledge Management has been gaining significant importance within organisations and is considered an important success factor in enterprise operation. For some time, there have been many techniques to model processes and other elements of the enterprise in order to capture the explicit knowledge. Modelling in this context means creating an explicit representation, usually computable, for the purposes of understanding the basic mechanics involved.

But knowledge can mean different things to different people and companies must spend some time looking for an appropriate mechanism to avoid misunderstanding in knowledge transmission. One mechanism to avoid this problem is to build a Glossary. The goal is to make accessible the organizational knowledge by unifying the language used in representing explicit knowledge. The semantic unification is a key factor for the success of the knowledge dissemination and diffusion through an organization.

2 Methodology to obtain a glossary

A general method for constructing a glossary is: collect a vocabulary, collect definitions, establish format rules, establish rules for writing definitions, examine definitions for multiple meanings, write basic definitions, review and compare for

consistency of meaning, classify, select preferred words, group words, review and finalize the glossary. In practice, some of these steps are omitted, while other steps are developed to considerable depth, depending on the final objective of the glossary. The complexity of detail for any phase of the glossary depends upon the scope of the glossary, the size of the vocabulary, and the number of persons participating in the project. This is understandable because a large working group reflecting a wide range of subjects, introduces more words for consideration and supplies multiple meanings relative to different backgrounds. Starting from a wide list of terms with the objective of building a whole glossary with inputs of several researchers could consume a great amount of time and effort. Therefore, a specific methodology was defined:

- **1st stage**: to define the purpose of the glossary.
- **2nd stage**: to built an initial list of terms and definitions using a semi-automatic glossary acquisition.
- **3rd stage**: to set up the collaborative glossary online module to support the sharing and extension of the glossary.
- **4th stage:** to complete the glossary by means of manual inputs and reviews, that is, the extension of the glossary.

2.1. 1st Stage: Purpose of the Glossary

In a collaborative environment, all participants must have in mind what is the main purpose of the glossary and the benefits they are going to obtain as well. Because of this, the project representatives must meet in order to clarify objectives and possible applications of the glossary. This fact becomes critical when the research project has different related research areas. There are several benefits in creating a durable glossary facility:

- *Semantic unification*: the glossary represents an informal, but shared view of relevant concepts. This activity will let semantics *emerge* naturally from applications and collaborative work.
- *Classification/retrieval* of documents: glossary terms may be used as meta-data for indexing documents and databases.
- *Integration of competences* from different research areas.

2.2. 2nd Stage: Semi-automatic glossary acquisition

The second stage of the methodology will lead to obtain the first version of the Glossary. This stage is based on a semi-automatic tool for ontology building called OntoLearn. This first version of the Glossary must be addressed as a preliminary stage for the generation of the final glossary. The extension and the diffusion between the research community are strictly required to meet the projected requirements.

Main steps of OntoLearn semi-automatic procedure. Figure 1 provides a snapshot of the OntoLearn ontology learning methodology.

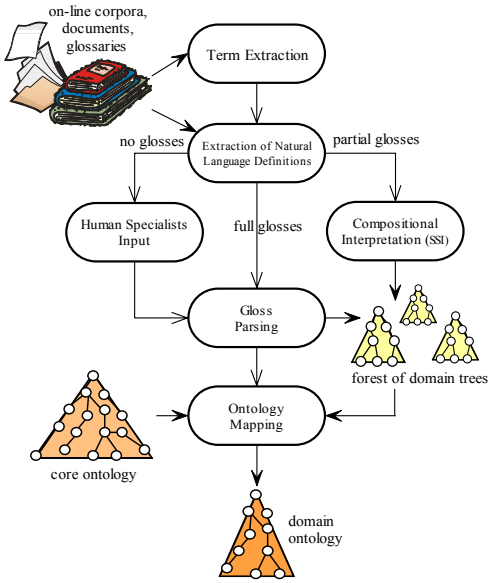

Fig. 1. An outline of the ontology learning phases in the OntoLearn system

The following steps are performed by the system:

Step 1. Extract pertinent domain terminology. Simple and multi-word expressions are automatically extracted from domain-related corpora, like enterprise interoperability (e.g. *collaborative work*), hotel descriptions (e.g. *room reservation*), computer network (e.g. *packet switching network*), art techniques (e.g. *chiaroscuro*). Statistical and natural language processing (NLP) tools are used for automatic extraction of terms [3]. Statistical techniques are specifically aimed at simulating *human consensus* in accepting new domain terms. Only terms uniquely and consistently found in domain-related documents, and not found in other domains used for contrast, are selected as candidates for the domain terminology.

Step 2. Search on the web for available natural language definitions from glossaries or documents. Available definitions are searched on the web using on-line glossaries or extracting definitory sentences in available documents. A context free (CF) grammar is used to extract definitory sentences. An excerpt is:

S → PP ‘,’ NP SEP
NP → N1 KIND1
KIND1 → MOD1 NOUN1
MOD1 → Verb | Adj | Verb ‘,’ MOD1 | Adj ‘,’ MOD1
NOUN1 → Noun
N1 → Art | Adj
SEP → ‘,’ | ‘.’ | Prep | Verb | Wh
PP → Prep NP

In this example, *S*, *NP* and *PP* stand for sentence, noun phrase and prepositional phrase, respectively. KIND1 captures the portion of the sentence that identifies the kind, or genus, information in the definition. This grammar fragment identifies (and analyses) definitory sentences like e.g.: "[In a programming language]$_{PP}$, [an *aggregate*]$_{NP}$ [that consists of data objects with identical attributes, each of which may be uniquely referenced by subscription]$_{SEP}$", which is a definition of *array* in a computer network domain. The grammar is tuned for high precision, low recall. In fact, certain expressions (e.g. *X is an Y*) are overly general and produce mostly noise when used for sentence extraction.

Step 3. IF definitions are found:

Step 3.1. Filter out non relevant definitions. Multiple definitions may be found on the internet, some of which may be not pertinent to the selected domain (e.g. in the interoperability domain *federation* as "the forming of a nation" rather than "a common object model, and supporting Runtime Infrastructure."). A similarity-based filtering algorithm is used to prune out "noisy" definitions, with reference to a domain. Furthermore, an extension of the CF grammar of step 2 is used to select[1], when possible, "well formed" definitions. For example, definitions with *genus(kind-of)* and *differentia (modifier)*, like the *array* example in step 2, are preferred to definitions by example, like: *Bon a Tirer* "When the artist is satisfied with the graphic from the finished plate, he works with his printer to pull one perfect graphic and it is marked "Bon a Tirer," meaning "good to pull". These definitions can be pruned out since they usually do not match any of the CF grammar rules.

Step 3.2. Parse definitions to extract kind-of information. The CF grammar of step 3.1 is again used to extract kind-of relations from natural language definitions. For example, in the *array* example reported in step 2, the same grammar rule can be used to extract the information (corresponding to the KIND1 segment in the grammar excerpt): $array \xrightarrow{kind-of} aggregate$

Step 4. ELSE IF definitions are not found:

Step 4.1. ***IF definitions are available for term components*** (e.g. no definition is found for the compound *integration strategy* but *integration* and *strategy* have individual definitions).

Step 4.1.1. Solve ambiguity problems. In technical domains, specific unambiguous definitions are available for the component terms, e.g.: *strategy*: "a series of planned and sequenced tasks to achieve a goal" and *integration*: "the ability of applications to share information or to process independently by requesting services and satisfying service requests" (interoperability domain). In other domains, like tourism, definitions of component terms are often extracted from general purpose dictionaries (e.g. for

[1] The grammar used for analysing definitions is a superset of the grammar used to extract definitions from texts. The analysed sentences are extracted both from texts and glossaries, therefore expressions like X is an Y must now be considered.

housing list, no definitions for *list* are found in tourism glossaries, and in generic glossaries the word *list* is highly ambiguous). In these cases, a word sense disambiguation algorithm, called SSI[2] [4] [5], is used to select the appropriate meaning for the component terms.

Step 4.1.2. Create definition compositionally. Once the appropriate meaning components have been identified for a multi-word expression, a generative grammar is used to create definitions. The grammar is based on the presumption (not always verified, see [5] for a discussion) that the meaning of a multi-word expression can be generated compositionally from its parts. According to this compositional view, the syntactic head of a multi-word expression represents the *genus* (kind-of), and the other words the *differentia* (modifier). For example, *integration strategy* is a *strategy* for *integration*. Generating a definition implies, first, to identify the *conceptual relations* that hold between the complex term components[3], and then, to compose a definition using segments of the components' definitions. For example, given the term *integration strategy*, the selected underlying conceptual relation is *purpose*:

$$Strategy \xrightarrow{purpose} Integration$$

and the grammar rule for generating a definition in this case is:

<MWE> = **a kind of** <H>**,** <HDEF>**, for** <M>, <MDEF> . (**1**)

Where MWE is the complex term, H is the syntactic head, HDEF is the main sentence of the selected definition for H, M is the modifier of the complex term and MDEF is the main sentence of the selected definition for M.

For example, given the previous definitions for *strategy* and *integration,* the following definition is generated by the rule (1): *integration strategy*: a **kind of** *strategy*, a series of planned and sequenced tasks to achieve a goal, **for** *integration*, the ability of applications to share information or to process independently by requesting services and satisfying service requests. As better discussed in [5] this definition is quite verbose, but has the advantage of showing explicitly the sense choices operated by the sense disambiguation algorithm. A human supervisor can easily verify sense choices and reformulate the definition in a more compact way.

Step 4.2. ELSE ask expert. If it is impossible to find even partial definitions for a multi-word expression, the term is submitted to human specialists, who are in charge of producing an appropriate and agreed definition.

Step 5. Arrange terms in hierarchical trees. Terms are arranged in forests of trees, according to the information extracted in steps 3.2 and 4.1.1.

---

[2] The SSI algorithm (Structural Semantic Interconnections) is one of the novel and peculiar aspects of the OntoLearn system. SSI recently participated to an international challenge, Senseval-3, obtaining the 2nd best score in a word sense disambiguation task.

[3] Machine learning techniques are used to assign appropriate conceptual relations.

Step 6. Link sub-hierarchies to the concepts of a Core Ontology. The semantic disambiguation algorithm SSI (mentioned in step 4.1.1) is used to append sub-trees under the appropriate node of a Core Ontology. In our work, we use a general purpose wide-coverage ontology, WordNet. This is motivated by the fact that sufficiently rich domain ontologies are currently available only in few domains (e.g. medicine).

Step 7. Provide output to domain specialists for evaluation and refinement. The outcome of the ontology learning process is then submitted to experts for corrections, extensions, and refinement. In the current version of OntoLearn, the output of the system is a taxonomy, not an ontology, since the only information provided is the kind-of relation. However, extensions are in progress, aimed at extracting other types of relations from definitions and on-line lexical resources.

2.3. 3rd Stage: Setting up a glossary online collaborative platform

Once completed the first list of terms and definitions using a semi-automatic glossary acquisition, the procedure selected to extend the glossary is the use of a Glossary Collaborative Online Module (GCOM). At the same time, this tool allows the sharing and utilization of the glossary. A methodology to implement the GCOM is defined: i) GCOM requirements definition: data, safety and interfaces, ii) Existing Glossary based tools analysis and iii) Selection of the solution to be implemented.

2.4. 4th Stage: Glossary extension and sharing

The last stage of the methodology comprises the extension and validation of the glossary by means of the GCOM. The semi-automatic glossary acquisition procedure generates a set of interrelated terms inside a domain starting from a group of documents of that same domain. Although this procedure generates an important number of definitions, it is common that some terms belonging to the research domain may be excluded, either because they don't appear in enough number in the evaluated documents or because they have appeared in later dates to the development of the stage 2. Based on this, the project researchers must extend the glossary terms to complete the final version of the glossary. Likewise, the project researchers must unify their approaches regarding the generated definitions. Therefore, this stage consists on a combined process of sharing-extension-validation using the newest ICT and developed by all the researchers of the project. The stage may be split up in:

- **Step 1. Glossary sharing:** The glossary must be uploaded in the GCOM in order to share the definitions between the research community.
- **Step 2. Glossary extension:** Then, the project researchers will extend the glossary with new definitions.
- **Step 3. Glossary validation***:* The project researchers must check each term and definition in terms of clarity and coherency.

3 Application of the methodology

3.1. Introduction

INTEROP is a Network of Excellence (NoE) whose primary goal is to sustain European research on interoperability for enterprise applications and software. The originality of the project lies in the multidisciplinary approach merging different research areas which support the development of interoperable systems: architectures and platforms, enterprise modelling, and ontology.

To speed up the glossary definition, the Department of Computer Science of the University of Roma made available a battery of ontology building algorithms and tools, the OntoLearn system [3] [2]. The OntoLearn system builds a domain ontology relating domain terms to the concepts and conceptual relations of the WordNet[4] lexicalised ontology. The final ontology is therefore an extended and trimmed version of WordNet. In OntoLearn, WordNet acts as a "general purpose" upper ontology, but other more specialised upper ontologies can be used, if available. Since the use of WordNet as a reference ontology is not a current choice of the INTEROP project, and since for the glossary acquisition task some additional feature was foreseen, we conceived a partly new experiment, using some of the tools and algorithms provided by the OntoLearn system, and some new feature that we developed for the purpose of the task at hand. In this chapter the methodology and the results of this experiment are described, that led to the acquisition of a hierarchically structured glossary of about 380 interoperability terms, subsequently evaluated by a team of 6 domain experts.

3.2. 1st Stage: Purpose of the glossary for the INTEROP NoE

Semantic unification is needed in order to facilitate the existing knowledge exchange within the NoE. The creation of a glossary is then a critical task for the project. Several INTEROP working groups have ascertained the need of identifying a glossary of interoperability terms for a variety of tasks, e.g.:

- Building a knowledge map and classifying knowledge domains.
- Classifying partner's competences for the INTEROP mobility matrix.
- Providing a list of relevant domain concepts for educational objectives.
- More in general, indexing with a set of common meta-data the various deliverables, state of art, scientific papers and databases.

Finally, the glossary will be used as main information source to build an Ontology on Interoperability. This ontology will allow structuring the knowledge all over the NoE, facilitating the information retrieval and clustering on the collaborative platform.

[4] http://www.cogsci.princeton.edu/~wn/

3.3. 2nd Stage: Application of the semi-automatic glossary acquisition procedure on the interoperability domain. The INTEROP experiment

For the purpose of the INTEROP glossary acquisition task, step 6 of the proposed methodology has been omitted, since an interoperability Core Ontology was not available, and the adoption of an available reference ontology (like WordNet) is not agreed in the project. The preliminary objective in this phase of INTEROP was to obtain a sort of partially structured glossary, rather than an ontology, i.e. a forest of term trees, where, for each term, the following information has to be provided: *definition* of the term, *source* of the definition, *kind-of* relation.

Step 1. Term extraction: The first step of the INTEROP glossary procedure was to derive an initial list of terms using the evidence provided by interoperability-related documents. The INTEROP collaborative workspace was used to collect from all partners the relevant documents, among which, the proceedings of INTEROP workshops and deliverables, highly referenced scientific papers, partners' papers, tutorials, etc. The OntoLearn TermExtractor module [3] extracted from these documents 517 terms. A generic computer science glossary was used to remove overly general technical terms and the list was then quickly reviewed manually to delete clearly identifiable spurious terms. The final list included 376 terms.

Step 2. Generation of definitions: Given the list of terms, we activated step 2 of the automatic glossary acquisition procedure. During this step, 28 definitions were not found, 22 were generated compositionally, and the remaining terms were extracted either from glossaries or from available documents. For each definition, we kept track of the source (URL of the web page). For some term, more than one definition survived the well-formedness and domain similarity criteria (step 3.1 of the OntoLearn semi-automatic procedure), therefore the total number of definitions submitted to the experts for revision was 358.

Step 3. Evaluation by experts: Six domain experts[5] in INTEROP were asked to review and refine the glossary. Each expert could review (*rev*), reject (*rej*), accept (*ok*) or ignore (*blank*) a definition, acting on a shared database. The experts added new definitions for brand-new terms, but they also added new definitions for terms that may have more than one sense in the domain. There have been a total of 67 added definitions, 33 substantial reviews, and 26 small reviews (only few words changed or added). Some term (especially the more generic ones, e.g. *business domain, agent, data model*) was reviewed by more than one expert who proposed different judgements (e.g. ok and rev) or different revised definitions. In order to harmonise the results, a first pass was conducted automatically, according to the following strategy: If a judgement is shared by the majority of voters, then select that judgement and ignore the others (e.g. if a definition receives two *ok* and one *rev*, then, ignore *rev* and accept the definition as it is). If the only judgement is *rej*(ect), then delete the definition. If a definition has a *rej* and one (or more) reviewed versions, then, ignore the reject and keep the reviews. This step led to a final glossary including 425

[5] The experts have been chosen according to their expertise in the three INTEROP domains.

definitions, 23 of which with a surviving ambiguity that could not be automatically conciliated. Therefore a second, short manual pass was necessary, involving this time only three reviewers. After resolving ambiguity, one definition (the most representative) per term was selected. Final glossary has 283 terms and definitions.

Step 4. Speed-up factors: The objective of the methodology is to *speed-up* the task of building a glossary by a team of experts. Evaluating whether this objective has been met is difficult, since no studies are available for a comparison. We consulted several sources, finally obtaining the opinion of a very experienced professional lexicographer[6] who has worked for many important publishers. He outlined a three-steps procedure for glossary acquisition including: i) internet search of terms, ii) production of definitions, and iii) harmonization of definitions style. He evaluated the average time spent in each step in terms of 6 minutes, 10 min. and 6 min. per definition, respectively. He also pointed out that conducting this process with a team of experts could be rather risky in terms of time[7], however he admits that in very new fields the support of experts could be necessary. Though the comparison is not fully possible, the procedure described in this paper has three phases in which man-power is requested:

After term extraction (step 1), to prune non-terminological and non-domain relevant strings. This requires 0.5 minutes per term. After the extraction of definitions (step 2), to evaluate and refine definitions. We asked each expert to declare the time spent on this task, and we came out with an average of 4 minutes per definition. Since some definition was examined by more than one expert, this amount must be increased to 6 min. approximately. In a second-pass review, to agree on the conflicting judgements. This depends on the number of conflicts, that in our case was less than 10%, mostly solved automatically (section 3.3). Overestimating, we may still add 1 minute per definition. The total time is then 7.5 minutes per definition, against the 16 declared by the lexicographer for steps 1 and 2 of his procedure. In this comparison we exclude the stylistic harmonisation (step 3 of the lexicographer), which is indeed necessary to obtain a good quality glossary, but has not been conducted in the case of the INTEROP experiments. However, since this phase would be necessarily manual in both cases, it does not influence the computation of the speed-up factor. The above evaluation is admittedly very questionable, because on one side we have an experienced lexicographer, on the other side we have a team of people that are certainly experts of a very specific domain, but have no lexicographic skills. Our intention here was only to provide a very rough estimate of the manpower involved, given that no better data are available in literature. Apparently, a significant speed-up is indeed obtained by our procedure.

Generation of domain sub-trees. As remarked in the introduction, the glossary terms must have some kind of hierarchical ordering, leading eventually to a formal ontology. A hierarchical structure simplifies the task of document annotation, and is a

[6] We thank Orin Hargraves for his very valuable comments.
[7] To cite his words: “no commercial publisher would subject definitions to a committee for fear of never seeing them in recognizable form again”

basis for further developments such as automatic clustering of data (e.g. for document classification), identification of similarities (e.g. for researchers mobility), etc. In other words, it is a first step towards semantic annotation. To arrange terms in term trees, we used the procedure described in steps 3.2 and 4.1.1 of the OntoLearn semi-automatic procedure. The definitions have been parsed and the word, or complex term, representing the hyperonym (genus) has been identified. Given the limited number of definitions, we verified this task manually, obtaining a figure of 91,76 % precision, in line with previous evaluations that we did on other domains (computer networks, tourism, economy). Contrary to the standard OntoLearn algorithm, we did not attached sub-trees to WordNet, as motivated in previous sections. Overall, the definitions were grouped in 125 sub-trees, of which 39 including only 2 nodes, 43 with 3 nodes, and the other with >3 nodes. Examples of two term trees are shown:

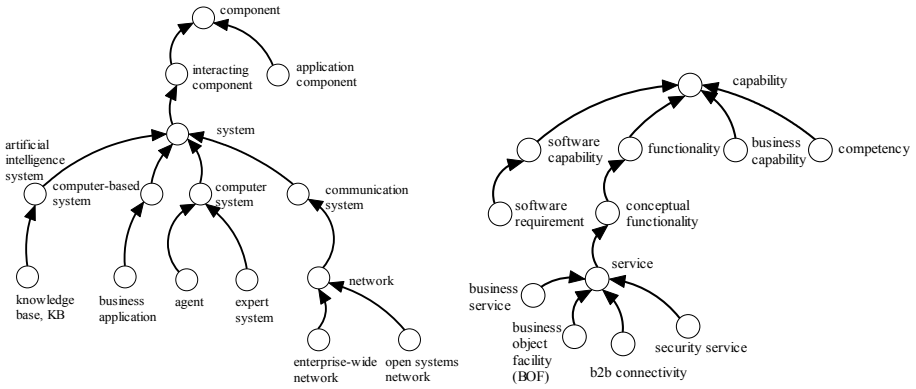

Fig 2. Sub-trees extracted from the Interoperability domain

In figure 2, the collocation of the term *system* might seem inappropriate, since this term has a very generic meaning. However, the definition of *system* in the interoperability glossary is quite specific: "*a set of interacting components for achieving common objectives*", which justifies its collocation in the tree. A similar consideration applies to *service* in the bottom tree. An interesting paper [1] provides an analysis of typical problems found when attempting to extract (manually or automatically) hyperonymy relations from natural language definitions, e.g. attachments too high in the hierarchy, unclear choices for more general terms, or-conjoined heads, absence of hyperonym, circularity, etc. These problems are more or less evident – especially over-generality – when analysing the term trees forest extracted from the glossary. However, our purpose here is not to overcome problems that are inherent with the task of building a domain concept hierarchy: rather, we wish to automatically *extract, with high precision, hyperonymy relations* embedded in glossary definitions, just as they are: possibly over-general, circular, or-conjoined. The target is, again, to speed up the task of ontology building and population, extracting and formalizing domain knowledge expressed by human specialists in an unstructured way. Discrepancies and inconsistencies can be corrected later by the human specialists, who will verify and rearrange the nodes of the forest.

Conclusion. As already remarked, the glossary provides a first set of shared terms to be used as metadata for annotating documents and data in the INTEROP platform. Several features/improvements are foreseen to improve this initial result, both on the interface/architecture and the methodological side. For example, annotation tools must be defined and integrated in the INTEROP platform. The taxonomic structuring of the glossary must be manually reviewed in the light of a core ontology to be defined, and methods to include new terms must be provided. Finally, the use of terms for document access, clustering and retrieval must be implemented and evaluated.

3.4. 3rd Stage: Setting up the INTEROP Glossary Web Module

Extending, accessing and using the glossary are collaborative and sharing activities that need to be supported by specific tools. Furthermore, the spreading of the glossary requires tools that can have a wide access by the research community. Currently, web environments have proved to be a suitable solution to address interaction based applications between several actors in different locations. The main features of web environments are the use of standard interfaces, the ease of implementation, the Worldwide access and the advanced interaction capabilities. These features provide the required functionality in order to allow a controlled and validated development of the future INTEROP Glossary. A methodology has been defined in order to facilitate the definition of the technical and operational specifications of the Glossary Module. Furthermore, this methodology is also aimed to select the software to support the INTEROP Glossary Web Module (GWM in what follows). The stages of this methodology are:

- **Stage 1. Define INTEROP GWM requirements**: A set of requirements will be defined related to the data and graphical user interface specifications.
- **Stage 2. Analysis of existing Glossary Web based modules**.
- **Stage 3. Selection of the solution**: Based on the previous analysis, the software to support the INTEROP GWM will be selected.

Stage 1 was done based on the requirements defined by the project researchers. Concerning to stages 2 and 3 some decisions were taken. Based on the general specifications, the Glossary module must be integrated within the INTEROP collaborative platform (PLONE-based system[8]) in order to take profit of the benefits of the mutual existence of a glossary and a set of resources (documents, papers, etc.). A search of glossary tools in the market has been performed. There exist some glossary building tools and some OpenSource e-learning platforms that provide glossary facilities, but none of them are integrated in PLONE. These solutions are discarded. Furthermore, there does not exist any commercial software based on PLONE to support the building of a glossary. This fact leads to consider the possibility to develop by an external company an *ad hoc* software on PLONE to support the Glossary extension and spreading.

[8] http://www.plone.org

3.5. 4th Stage: Extending the INTEROP Glossary

The INTEROP Project has foreseen 5 tasks to be developed in the next year:

- **Glossary Format and structuring:** Currently, the glossary is a "flat" list of terms and textual definitions. This flat structure may be inadequate for many tasks. A first activity is to structure the terms in taxonomic order. Taxonomic structuring of keywords is a first step towards concept-based search.
- **Glossary extension and updating:** In this sub-task the procedures and software tools for updating and extending the glossary will be defined.
- **Glossary Usages:** New usages must be specified in the working groups.
- **Implementation of defined glossary-based applications.**
- **Evaluation and assessment:** Finally, an evaluation will be carried out to check the consistency of the tasks developed.

4 Conclusions

A 4-stage methodology to create a glossary in a collaborative research project has been defined. The new proposed methodology is based on an adaptation of a methodology of the University of Rome for the semiautomatic acquisition of terms and definitions starting from a source of documents related to the research areas of the collaborative project. Based on this methodology, a first glossary of terms related to the interoperability domain has been obtained inside the IST-508011 INTEROP Network of Excellence. The requirements of a web tool to share and to permanently enlarge the INTEROP glossary have been defined. An analysis of the existing glossary tools has been carried out. As conclusion, an *ad hoc* tool will be developed.

References

1. Ide N. and Véronis J. (1993). Refining Taxonomies extracted from machine readable Dictionaries, In Hockey, S., Ide, N. *Research in Humanities Computing*, 2, Oxford University Press.
2. Navigli R., Velardi P and Gangemi A. (2003). Ontology Learning and its Application to Automated Terminology Translation. *IEEE Intelligent Systems*, vol. 18, pp. 22-31
3. Navigli R. and Velardi P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, vol. 50 (2).
4. Navigli R. and Velardi P. (2004b). Structural Semantic Interconnection: a knowledge-based approach to Word Sense Disambiguation, Proc. *3rd Workshop on Sense Evaluation*, Barcelona.
5. Navigli R., Velardi P., Cucchiarelli A. and Neri F. (2004). Quantitative and Qualitative Evaluation of the OntoLearn Ontology Learning System. *Proc. 20th COLING 2004*, Geneva.