GRISP: A Massive Multilingual Terminological Database for Scientific and Technical Domains

Patrice Lopez, Laurent Romary*

*INRIA & HUB - IDSL
Berlin, Germany
patrice_lopez@hotmail.com, laurent.romary@inria.fr

Abstract

The development of a multilingual terminology is a very long and costly process. We present the creation of a multilingual terminological database called GRISP covering multiple technical and scientific fields from various open resources. A crucial aspect is the merging of the different resources which is based in our proposal on the definition of a sound conceptual model, different domain mapping and the use of structural constraints and machine learning techniques for controlling the fusion process. The result is a massive terminological database of several millions terms, concepts, semantic relations and definitions. This resource has allowed us to improve significantly the mean average precision of an information retrieval system applied to a large collection of multilingual and multidomain patent documents.

1. Introduction

Technical and scientific documents aim at supporting specialist communication and are thus written in specialist language, 30-80% of which is composed of terminology (Ahmad, 1996). Terminology is the main vehicle by which technical and scientific units of knowledge are represented and conveyed.

A vast range of applications related to technical and scientific knowledge requires semantic and terminological descriptions covering multiple domains. For instance, Biosis[1] from Thomson Scientific is a terminological database of more than 2 millions terms used for classifying and indexing life science scientific articles at large (i.e. biology, medicine, genetics, agriculture, etc.). The multilingual terminology of the European Union, IATE[2], contains 8,4 million terms in 23 languages covering EU specific terminology as well as multiple fields such as agriculture or information technology. The development and the maintenance of such large terminological resources is an extremely long and difficult process requiring continuous human expertize from multiple domains.

Many domain specific resources exist, often well curated and, sometimes, freely available. The present work addresses the following question: Is it possible to exploit these heterogeneous resources, even the less constrained ones, such as Wikipedia, for creating a unique terminological resource covering multidomain technical and scientific content?

The TermScience portal (Khayari et al., 2006) is a first step toward the combination of heterogeneous multilingual scientific terminological resources, but does not address the problem of controlling and realizing appropriate fusions. The problem of merging resources from different terminology has been identified. However, as the main goal of authors was to investigate the problem of modeling, they described a solution for encoding the heterogeneity of the sources, and not a solution for controlling and realizing appropriate fusions.

The issue of merging different semantic resources have been well studied in the context of the fusion of ontologies, in particular with the popularity of the semantic web framework (McGuinness et al., 2000; Madhavan et al., 2001; Doan et al., 2001; Gal et al., 2005). Since the ontologies usually remain relatively small, some proposals rely on semi-automatic techniques as (McGuinness et al., 2000). Fully automatic methods exploit structural and linguistic matching (Madhavan et al., 2001) or machine learning techniques using different aspects of an ontology, such as concepts and properties (Doan et al., 2001). To avoid the problem of lack of training data, fuzzy logic methods have been proposed (Gal et al., 2005). To our knowledge, however, automatic merging techniques for heterogeneous terminologies has not been yet investigated. Terminologies follow different design principles than semantic web ontologies. They contain much richer textual content, they do not rely on formal and axiomatic organization of concepts and do not model facts and assertions.

For classification purposes, Digital Libraries (DL) also require descriptions across multiple domains and raise the issue of merging heterogeneous knowledge sources. (Wang et al., 2008), for instance, proposes a technique for merging multilingual subject heading lists from different classification scheme based on heuristics. As these heuristics rely on specific DL metadata, it does not appear possible to exploit them for terminologies.

Focusing on terminological resources, this paper presents the creation of a massive multilingual and multidomain terminology called **GRISP** (**G**eneral **R**esearch **I**nsight in **S**cientific and technical **P**ublications) from freely available resources for the purpose of computer applications. We describe first the conceptual terminological model which allowed us to represent in a common scheme the existing resources. We then present and evaluate how we controlled the correctness of the merging of the differ-

[1] http://thomsonreuters.com/products_services/science/ science_products/life_sciences/biology/biosis

[2] http://iate.europa.eu

ent sources. Finally, we describe how this resource was successfully used for a large scaled patent retrieval task.

2. Common Framework

2.1. Objective of the present work

Our main goal is to create a terminological resource able to support automatic text processing applications. In traditional terminology, natural language is viewed as an obstacle to objectification which should be constrict. Several principles of traditional terminology needed for the human design of terminological resources aim at reducing the impact of natural language. One example is standardisation of terminology. Standardisation is a strife for univocity (Temmerman, R., 1997). Following the Würsterian principle, one concept is referred to by one term (no synonymy) and one term can only refer to one concept (no polysemy).

Since we focus here on computer applications having to process natural language, we relaxed basic traditional terminological principles. For instance, for the purpose of computer applications, terms do not need canonical forms, and enumerating as many terms variants as possible appears useful for automatic concept annotation. Term ambiguity within a given main domain need to be allowed for covering actual data. Similarly, instead of providing one good definition, providing different definitions corresponding to different views of the same concept can be more appropriate for tasks requiring robustness. Our objective is closer to build a linguistic knowledge base than a terminology for the purpose of human-driven curation and standardization.

2.2. Terminological Conceptual Model

Contrary to dictionaries which are word-based, terminologies (which may also include non-linguistic items such as formulae, codes, symbols and graphics) are fundamentally concept-based, reflecting the fact that the terms which they contain map out an area of specialist knowledge in which encyclopedic information plays a central role. The goal of terminological modeling is to represent the vocabulary, the definitions and the essential properties of concepts. In addition, for maximizing the exploitation of a terminological resource, it appears crucial: (i) to be independent from any particular applications, (ii) to support multiple languages, (iii) to follow standards and best practices for interoperability.

In order to set up a common framework able to represent multiple terminologies, a generic model able to cover a variety of terminological resources is necessary. We organized our terminological database according to the principles of ISO 16642 (TMF – Terminological Markup Framework) (Romary, 2001). Based on a generic semasiological (sense to word) model, TMF ensures that each elementary field is both attached to the appropriate level of description, (e.g. Terminological Entry, Language Section or Term Section) and possibly refined with local meta data. Such local metadata are particularly relevant in compiled databases since they allow tracking the source and responsibility for any piece of information, but also permit the creation of views, virtually reconstructing coherent subsets within a given domain or originated from the same source (e.g. all MeSH-based entries; cf. (Khayari et al., 2006)).

TMF provides comprehensive and consistent representations for elementary linguistic features which can appear at different levels of description depending on a specific resource. Representing an existing terminological resource into the TMF framework supposes the identification of these standard units of representation in the source terminology and their mapping into the TMF elementary fields. The TMF elementary units of description are strictly defined and follow well-formedness constraints to facilitate an unambiguous structural mapping of data. Although terminological notions such as *term, concept, conceptual relation* or *definition* can vary from one resource to another one, depending on their level of description and purposes, they are well defined and controlled in TMF. By using this framework, we fulfill at the same time the need of having a general model suited to existing heterogeneous terminologies and the requirements of standardization.

2.3. Domains

Following (Bentivogli et al., 2004), a domain can be characterized by the name of a discipline where a certain specialist knowledge area is developed (e.g. chemistry) or by the specific object of the knowledge area (e.g. food).

A key property of traditional specialist terminology is the unambiguous semantic of the term given a domain. MeSH for instance distinguishes 129 main domains covering different aspects of the medical field, such as anatomy, genetics or biology, but also other domains such as computer science, sociology or geography. A given term can realize different concepts but never more than one concept per domain.

For building the present multidomain terminology, we use a set of 76 basic domains derived from the technical and scientific domains of WordNet Domains (Magnini and Cavaglià, 2000), itself derived from the Dewey decimal classification[3]. This set of domains is organized in a hierarchy and follows a degree of granularity which makes general domain mapping between different terminology source easy and well adapted to text processing tasks.

2.4. Resource mappings

Terminological resources normally include a division into categories for describing the fields under which the concepts are organized. In the present work, the correspondence between the resources is first given by the basic domains. Each upper level resource specific domain/category have been manually mapped to the relevant basic domains.

The realization of this mapping is normally relatively straightforward and easy to handle manually. However, hierarchies from different source vocabularies do not always map correctly, resulting in conflicting positioning of some concepts in the semantic network of the basic domains.

2.5. Concept Merging

We call *aggregation* the addition of different terminological sources into the same conceptual model. Obviously, different sources frequently overlap semantically. The real value

[3] http://www.oclc.org/dewey

Applied_Science	**Pure_Science**	**Social_Science**
Agriculture	Astronomy	Health
Animal_Husbandry	Biology	Body_Care
Food	Biochemistry	Military
Home	Anatomy	Pedagogy
Architecture	Physiology	School
Town_Planning	Genetics	University
Buildings	Animals	Publishing
Furniture	Plants	Sociology
Computer_Science	Environment	Artisanship
Engineering	Chemistry	Commerce
Mechanics	Earth	Industry
Astronautics	Geology	Transport
Electrotechnology	Meteorology	Aviation
Hydraulics	Oceanography	Vehicles
Telecommunication	Paleontology	Nautical
Post	Geography	Railway
Telegraphy	Mathematics	Economy
Telephony	Geometry	Enterprise
Medicine	Statistics	Finance
Dentistry	Physics	Insurance
Pharmacy	Acoustics	Tax
Psychiatry	Atomic_Physic	Administration
Radiology	Electricity	Politics
Surgery	Electronics	
	Gas	
	Optics	

Table 1: Basic Domains of GRISP

of the combination of different terminologies is the ability to identify common concepts to obtain consistent and enriched semantic representations. Beyond a simple aggregation of terminological resources, the crucial problem of the present work is the correct merging of concepts having different origins.

Conflicting domain mapping, high polysemy of term variants and incorrectly positioned concepts can cause two problems: (1) to incorrectly merge two concepts sharing common terms and common domains, (2) to lose precision in term descriptions when merging concepts.

A traditional terminology is based on the principle that one designation corresponds to one concept. As this univocal relationship does not occur in practice, subject fields are used to avoid polysemy, each subject field being considered as a closed domain (Cabré et al., 1999). From this principle, i.e. a term is not polysemous in a given domain, we specify a first merging rule:

Merging Rule 1: *If two concepts belong to the same domain and share a common term, the two concepts are merged.*

The univocity principle is well followed in a single traditional terminology or ontology, but not in a resource as Wikipedia. Many variant terms can be single word terms and abbreviations which are highly polysemic. In addition, as mentioned in the previous section, the division into domains is not consistent from one source to another. A more restrictive rule can, therefore, be introduced:

Merging Rule 2: *If two concepts belong to the same domain and their preferred terms are the same, the two concepts are merged.*

However, as a more general design, the merging of concepts can be rather viewed as an overall balance of evidences related to structural and property (terms, definition, etc.) similarity. We thus also propose and compare the usage machine learning techniques for refining concept merging decision which will be presented in section 4 and compared with the two introduced merging rules.

For all the approaches, an additional constraint is required to avoid the merging by transitivity of two concepts initially separated in one common source: For instance, if two concepts c_1 and c_2 originate from the source S_1, and c_3 from S_2, if c_1 is merged with c_3, resulting in the concept c_4=$c_1 \oplus c_3$, c_4 cannot be further merged with c_2 since the two concepts c_1 and c_2 were separated in S_1. The following invariant is thus introduced:

Source-Conformance Invariant: *Two concepts having at least one source in common cannot be merged.*

This invariant ensures that the precision in term description in one resource is kept in the merged terminology. In our example, the concept c_3 can be merged both with c_1 and c_2, but the merged concepts $c_1 \oplus c_3$ and $c_2 \oplus c_3$ will be kept separated. The Source-Conformance Invariant is, however, relevant only for standard terminology.

3. Resources

In this section, we give an overview of the resources used in the present work and their integration:

MeSH: The Medical Subject Heading[4] is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure, for a total of approx 650.000 terms. As MeSH already includes a conceptual organization, its integration in the GRISP conceptual model is straightforward.

The Specialist Lexicon[5] is an open source lexicon containing approx. 400.000 lexical entries from the biomedical field. It is also released as part of UMLS. It has been used to enrich our list of acronyms and term variants.

The Gene Ontology[6] is a major open resource providing a controlled vocabulary of approx. 28.500 terms for gene product attributes (The Gene Ontology Consortium, 2000).

ChEBI[7] is a freely available dictionary of molecular entities developed at the European Bioinformatics Institute (Degtyarenko and al., 2008). ChEBI is a valuable source of chemical vocabulary with approx. 42.000 concepts, 97.000 terms, 28.000 semantic relations and multilingual terms in 5 languages.

WordNet, WOLF and SUMO: WordNet is a linguistic knowledge base describing general language. We used WordNet Domains (Magnini and Cavaglià, 2000) in order to restrict the set of synsets to those related to technical and scientific domains. The resulting terms capture general technical and scientific vocabulary for to all considered domains. This restriction corresponds to approx. 22.000 synsets. For these synsets, the existing mapping to SUMO (the Suggested Upper Merged Ontology) has been imported for the purpose of interoperability of GRISP with ontologies, as well as the existing French terms present in WOLF (Sagot and Fišer, 2008).

IPC: The International Patent Classification[8] is a hierarchical classification of approx. 70.000 subdivisions distributed by the WIPO (World Intellectual Property Organization). It contains approx. 4.000 illustrations (mostly chemical compounds) and so-called catch words.

Wikipedia:[9] The collaborative encyclopedia is an extremely rich, multilingual and multidomain source of specialist vocabulary. It is, however, also very noisy, in the sense that the categorization (based on more than 140.000 categories) and the term variants (redirections) are defined without any constraints. The Wikipedia dump XML files have been processed with a slightly modified version of Wikiprep[10] able to extract multilingual relations in addition to usual structure and text information. Similarly as (Gabrilovich and Markovitch, 2007), we interpreted an article as a concept, the title of the article being the preferred term and the disambiguation redirections being variant terms realizing this concept. The first paragraph of an article has been used as definition. 170 Wikipedia top level categories corresponding to technical and scientific domains were mapped to the 76 basic GRISP domains.

Our experiments also include **UMLS** (Unified Medical Language System[11]) resources to complete the coverage of MeSH with an addition of approx. 850.000 terms. UMLS, however, is not a free resource and requires a specific license. For the multilingual resources (Wikipedia, ChEBI), we considered only the English, French and German languages. As WordNet, the IPC and Wikipedia cannot be considered as standard terminologies, the Source-Conformance Invariant was not considered for these sources.

4. Learning to Merge Concepts

4.1. Learning Model

Merging of concepts can be expressed as a machine learning problem, more precisely as a binary classification. When expressed as a regression problem, the regression model can provide a merging score which can be used with a threshold for selecting more or less aggressive merging strategies. We experimented SVM (Support Vector Machine) and MLP (Multi-Layer Perceptron) binary classification models based respectively on libSVM (Chang and Lin, 2001) and the WEKA toolkit (Witten and Frank, 2005) to decide if two concepts from different sources should be merged or not. The merging decision is applied recursively until a minimal number of merging per iteration is reached.

4.2. Feature definition

For capturing structural and content-based similarity between concepts having different origins, we introduce the features summarized on Table 2.

(f1-2)	Identification of the two involved sources
(f3)	Number of common domains between the two concepts
(f4)	Number of same source-specific categorizations
(f5)	Boolean indicating if both preferred terms are identical
(f6)	Boolean indicating if both preferred terms are identical after stemming
(f7)	Ratio of identical terms given all terms
(f8)	Similarity measure of the definition texts, after stemming and based on negative KL divergence
(f9)	Number of domains of the merged concept
(f10)	Number of words of the longest common terms

Table 2: List of features for machine learning merging.

[4] http://www.nlm.nih.gov/mesh

[5] http://lexsrv3.nlm.nih.gov/SPECIALIST

[6] http://www.geneontology.org

[7] http://www.ebi.ac.uk/chebi

[8] http://www.wipo.int/classifications/ipc/

[9] http://download.wikimedia.org

[10] http://sourceforge.net/projects/wikiprep

[11] http://www.nlm.nih.gov/research/umls

4.3. Training

We use two sources of training data providing examples of merging decisions.

- The first source is based on the existing MeSH mappings present in Wikipedia infobox templates for medical entities. For a large number of entries in the medical and biochemistry domains, the Wikipedia articles provide the corresponding MeSH concept identifier. This information can be used to evaluate the merging of Wikipedia with MeSH and UMLS concepts and, by generalization, the merging of Wikipedia with standard terminologies. We extracted from Wikipedia, a total number of 1.657 merging decisions.

- The second source is based on a multidisciplinary terminology for scientific and technical domains called PASCAL which was kindly provided by the INIST[12] in the framework of TermSciences (Khayari et al., 2006). A concept in PASCAL containing two terms that belong to two different concepts in the simple aggregation, and when at least one domain is shared, can be used as example of correct merging. We extracted from this source a set of 2.230 merging decisions.

5. Evaluation

We first present quantitatively the resulting terminological database and, second, evaluate the concept merging between resources.

5.1. Resulting Database

Table 3 gives a quantitative view of the resulting terminological database depending on the merging approach. The *aggregation* method corresponds to no merging at all.

Merger	Concepts	Terms	Sem. Rel.
Aggregation	1.503.818	3.140.726	970.864
Merg. Rule 1	1.457.538	3.157.179	1.022.303
Merg. Rule 2	1.476.508	3.114.711	971.218
SVM	1.450.688	3.195.118	1.088.446
MLP	1.451.710	3.192.325	1.081.955

Table 3: GRISP volume statistics following the different merging strategies.

In addition, GRISP contains 596.865 definitions, 1.321.988 source specific categorizations of concepts, approx. 20.000 acronyms, 14.268 chemical formulas and 12.375 chemical structure identifiers. We can observe that the merging of concepts concerns a relative small proportion of the whole set of concepts. This is due to the fact that many concepts and terms corresponds to product names, such as medical or chemical entities, which are not candidate for any merging. However, the merging are relevant for concepts which are more generic and frequently used.

5.2. Merging Accuracy

We present on Table 4 an evaluation using a reference set corresponding to a random subset of 10% of the merging examples extracted from Wikipedia/MeSH mappings and from the PASCAL terminology. The coverage (*cov.*) (nb of expected merging found / nb of merging to be found) is evaluated automatically based on the evaluation set. The accuracy (*acc.*) involves a limited manual evaluation for judging further merging found after the expected merging, but not in the evaluation set. The Merging Rule 2 produces almost perfect merging but with a very low coverage. Rule 1 extends the coverage at the price of a relatively high rate of merging error. The Machine Learning approaches further extend the coverage while maintaining a high precision. Using the MeSH/Wikipedia mappings as evaluation appear, however, relatively biased since it is clear that many MeSH terms have been added in the corresponding Wikipedia articles at the same time as the MeSH descriptor ID.

Merger	Wiki/MeSH	PASCAL
Merging Rule 1	cov. 0.6464 acc. 0.9497	cov. 0.5358 acc. 0.9371
Merging Rule 2	cov. 0.3607 acc. 0.9949	cov. 0.2735 acc. 0.9916
SVM	cov. 0.8642 acc. 0.9698	cov. 0.6203 acc. 0.9522
MLP	cov. 0.8607 acc. 0.9748	cov. 0.6178 acc. 0.9515

Table 4: Evaluation of the merging strategies.

6. Tool and Encoding

6.1. The GRISP browser

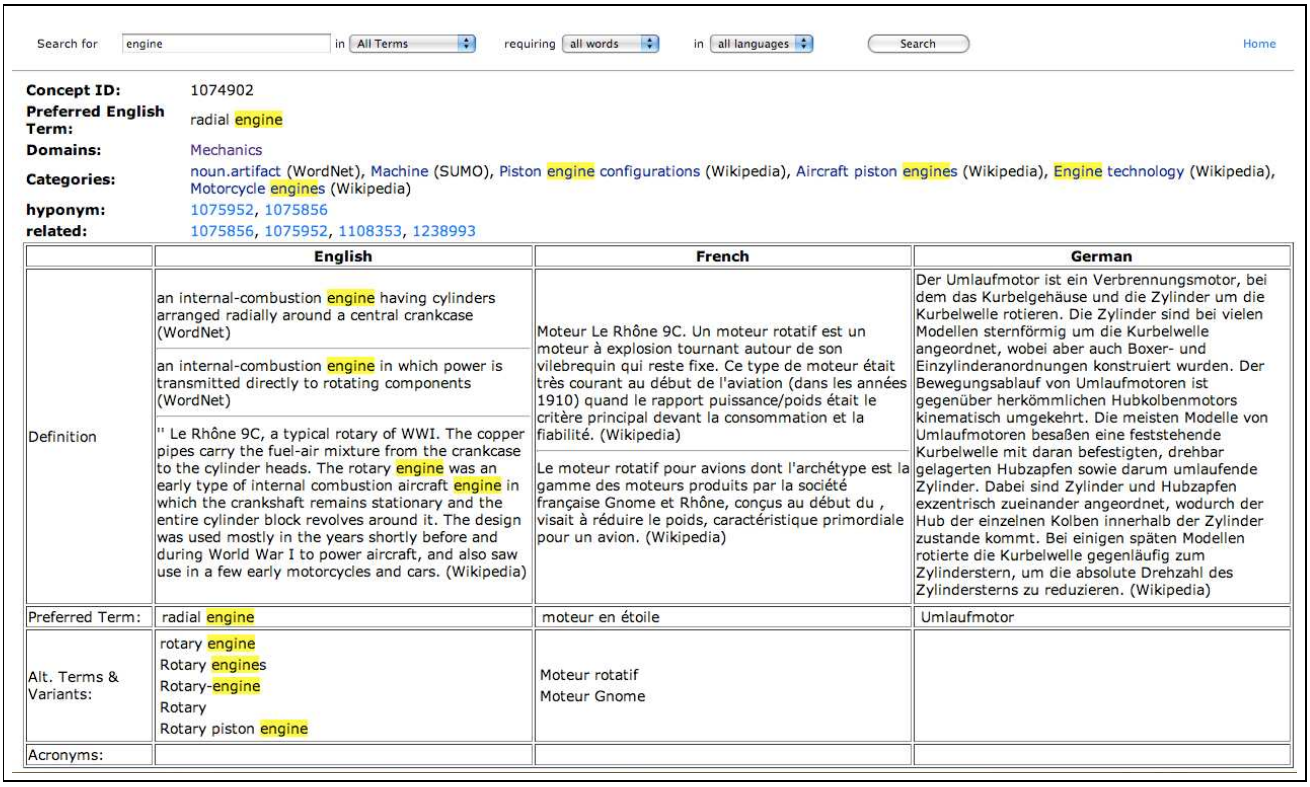
The GRISP terminology is currently stored in a MySQL database following a relational model implementing the conceptual model of ISO 16642.

We developed a web application for querying and browsing the resulting terminological database. Figure 1 illustrates the view of the concept corresponding to the star term *radial engine* as displayed by the GRISP browser. Although our primary goal is to create a terminological resource for computer application, we believe that this tool can also be used as a multilingual terminological database for supporting specialized manual translation and as a technical knowledge base. For these purposes, the browser exploits the encyclopedic entries of the GRISP and can display molecules or illustrations extracted from the primary sources, as illustrated by Figure 2.

6.2. TMF encoding

Using the data model based on ISO 16642 allow us to exploit Data Category Registry (DCR) following the ISO 12620 standard for facilitating the implementation of filters and converters between different terminology instances and to produce a Generic Mapping Tool (GMT) representation, i.e. a canonical XML representation.

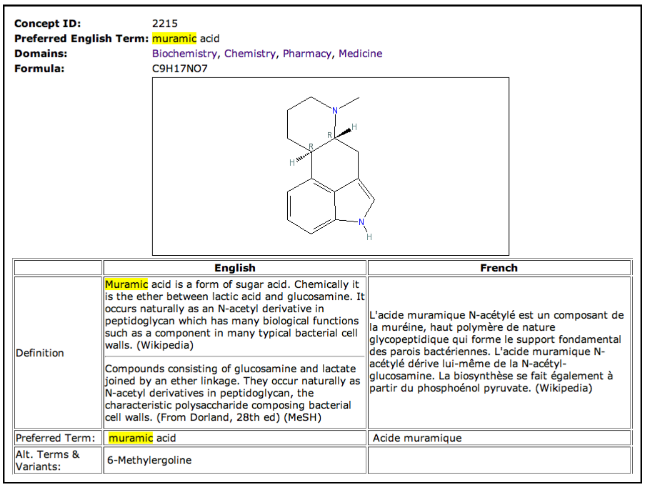
[12] The French National Institute for Technical and Scientific Information.



Concept ID: 1074902
Preferred English Term: radial engine
Domains: Mechanics
Categories: noun.artifact (WordNet), Machine (SUMO), Piston engine configurations (Wikipedia), Aircraft piston engines (Wikipedia), Engine technology (Wikipedia), Motorcycle engines (Wikipedia)
hyponym: 1075952, 1075856
related: 1075856, 1075952, 1108353, 1238993

	English	French	German
Definition	an internal-combustion engine having cylinders arranged radially around a central crankcase (WordNet) an internal-combustion engine in which power is transmitted directly to rotating components (WordNet) '' Le Rhône 9C, a typical rotary of WWI. The copper pipes carry the fuel-air mixture from the crankcase to the cylinder heads. The rotary engine was an early type of internal combustion aircraft engine in which the crankshaft remains stationary and the entire cylinder block revolves around it. The design was used mostly in the years shortly before and during World War I to power aircraft, and also saw use in a few early motorcycles and cars. (Wikipedia)	Moteur Le Rhône 9C. Un moteur rotatif est un moteur à explosion tournant autour de son vilebrequin qui reste fixe. Ce type de moteur était très courant au début de l'aviation (dans les années 1910) quand le rapport puissance/poids était le critère principal devant la consommation et la fiabilité. (Wikipedia) Le moteur rotatif pour avions dont l'archétype est la gamme des moteurs produits par la société française Gnome et Rhône, conçus au début du , visait à réduire le poids, caractéristique primordiale pour un avion. (Wikipedia)	Der Umlaufmotor ist ein Verbrennungsmotor, bei dem das Kurbelgehäuse und die Zylinder um die Kurbelwelle rotieren. Die Zylinder sind bei vielen Modellen sternförmig um die Kurbelwelle angeordnet, wobei aber auch Boxer- und Einzylinderanordnungen konstruiert wurden. Der Bewegungsablauf von Umlaufmotoren ist gegenüber herkömmlichen Hubkolbenmotors kinematisch umgekehrt. Die meisten Modelle von Umlaufmotoren besaßen eine feststehende Kurbelwelle mit daran befestigten, drehbar gelagerten Hubzapfen sowie darum umlaufende Zylinder. Dabei sind Zylinder und Hubzapfen exzentrisch zueinander angeordnet, wodurch der Hub der einzelnen Kolben innerhalb der Zylinder zustande kommt. Bei einigen späten Modellen rotierte die Kurbelwelle gegenläufig zum Zylinderstern, um die absolute Drehzahl des Zylindersterns zu reduzieren. (Wikipedia)
Preferred Term:	radial engine	moteur en étoile	Umlaufmotor
Alt. Terms & Variants:	rotary engine Rotary engines Rotary-engine Rotary Rotary piston engine	Moteur rotatif Moteur Gnome	
Acronyms:			

Figure 1: View of the multilingual terminological database for the concept corresponding to the term *Radial Engine*. This concept is obtained after multiple merging of concepts from WordNet and Wikipedia, resulting in a richer semantic and terminological description.



Concept ID: 2215
Preferred English Term: muramic acid
Domains: Biochemistry, Chemistry, Pharmacy, Medicine
Formula: C9H17NO7

	English	French
Definition	Muramic acid is a form of sugar acid. Chemically it is the ether between lactic acid and glucosamine. It occurs naturally as an N-acetyl derivative in peptidoglycan which has many biological functions such as a component in many typical bacterial cell walls. (Wikipedia) Compounds consisting of glucosamine and lactate joined by an ether linkage. They occur naturally as N-acetyl derivatives in peptidoglycan, the characteristic polysaccharide composing bacterial cell walls. (From Dorland, 28th ed) (MeSH)	L'acide muramique N-acétylé est un composant de la muréine, haut polymère de nature glycopeptidique qui forme le support fondamental des parois bactériennes. L'acide muramique N-acétylé dérive lui-même de la N-acétyl-glucosamine. La biosynthèse se fait également à partir du phosphoénol pyruvate. (Wikipedia)
Preferred Term:	muramic acid	Acide muramique
Alt. Terms & Variants:	6-Methylergoline	

Figure 2: View of the multilingual terminological database for the concept corresponding to the term *Muramic Acid*. The concept results from the merging of the corresponding MeSH, English and French Wikipedia and ChEBI entries.

6.3. Resource Life Cycle

As the aggregation and merging process is fully automated, the maintenance/curation of the individual sources over time can be integrate continuously in the existing merged terminological database.

7. Application to Patent Processing

We have evaluated the interest and the relevance of the GRISP terminology with a information retrieval system for patent documents system called PATATRAS (PATent and Article Tracking, Retrieval and AnalysiS) described in (Lopez and Romary, 2009) and developed for the CLEF IP 2009 track (Roda et al., 2009). The collection consists of all patent publications from the European Patent Office until 2000, approx. 1,9 million documents in English, French and German (more than 3 billion words). The goal of the CLEF IP track was to realize a prior art search for a total number of 10.000 patents, referred to as *patent topics*. The automatic evaluation was based on the documents cited by patent examiners in the official search reports and examination procedures, with an average of approx. 6 relevant documents per patent topic. PATATRAS has been ranked first for all subtasks of the evaluation track among 14 participants (Roda et al., 2009).

As one of our goal was to perform a complete conceptual indexing of this collection, a terminology covering a large spectrum of technical and scientific domains in three languages was needed. We performed a complete conceptual indexing of this collection based on GRISP. The terms of the GRISP terminological database have been used for annotating the textual data of the whole multilingual collection. A term annotator able to deal with such a large volume of data has been developed specifically for this track. After a POS tagging and a lemmatisation of the whole collection, this annotator matched the term variants following morphological variations. The concept disambiguation was realized on the basis of the IPC classes of the processed patent which indicates one or several basic domains where the patent document belongs. Approx. 1.1 million different terms present in GRISP have been identified at least one time in the collection resulting in more than 176 million annotations.

In addition to this large scale conceptual index, we created three additional indexes of word forms (one per language) and a phrase index for English. We report in Table 5 the results obtained with the KL divergence model for each index in term of MAP (Mean Average Precision). A preprocessing based on patent specific metadata and classification information was first realized to prune the search space. The queries were based on the whole content of the 10.000 patents for which the prior art was realized.

index	lang	MAP
word form	en	0,1589
word form	fr	0,1234
word form	de	0,1218
phrase	en	0,1344
concept	-	0,1476

Table 5: Retrieval accuracy following different indexing models for the CLEF IP 2009 track.

Table 5 shows that the retrieval based on English word form surpasses the conceptual based retrieval. The fact that a control terminology covers usually only a part of a text, implies an information loss as compared to a word form indexation and retrieval (Zhou et al., 2007). However, as shown by table 6, the conceptual model presents a very strong complementarity with the word-form models. By combining retrieval models on the basis of confidence estimations, it is possible to exploit the different retrieval models, in particular conceptual results, for refining the overall accuracy.

Model	# better than baseline	# best overall
word form en (baseline)	-	1341
word form fr	3480	839
word form de	3392	781
phrase en	3559	869
concept	**4832**	**1692**

Table 6: Complementarity between results sets for the XL patent topic set (10.000 documents). The concept index based on GRISP provided the highest number of results better than the baseline and the highest number of best results compared to the other index models.

In the present case, the combination of models was based on a linear interpolation of ranked results sets, the coefficients being computed by SVM regression models using query-specific features and existing search reports present in the patent collection, see (Lopez and Romary, 2009) for details. The accuracy after the merging of the retrieval model and after a final post-ranking based on specific patent metadata and statistics, are presented Table 7. The combined multilingual result set which integrate the concept results shows a MAP 43.5% higher than the one of the best monolingual individual result set.

Measures	Combined models	After Post-Ranking
MAP	0.2281	0.2802
Prec. At 5	-	0.2768
Prec. At 10	-	0.1776

Table 7: Final Results for the CLEF IP 2009 track.

In bioinformatics, it is known that information retrieval based on well curated resources as MeSH or UMLS can be more effective than word-based retrieval models (Zhou et al., 2006). The combination of word-based language model and concept-based language model for Information Retrieval in Genomics results in significant performance improvements (Zhou et al., 2007). The present work shows that, even with lower standard and less complete terminological resources, a model combination can improve baseline retrieval results. A conceptual terminological model such as GRISP provides specialized and precise representations which are complementary to the word-based models.

To our knowledge, this was the first time that a controlled conceptual indexing was realized on such a large scale for multiple scientific and technical domains in a realistic multilingual task.

8. Conclusion and Future Works

We have proposed a method for creating a massive multilingual terminological database for multiple scientific and technical domains based on various existing free terminologies and knowledge bases. The accuracy of the concept merging between several resources have been evaluated following several methods.

The resulting resource has been used successfully for improving the accuracy of an information retrieval system applied to a collection of 1.9 million of patent documents covering multiple technical fields in the context of the CLEF IP 2009 track.

Within this framework, any new specialized terminologies, not specifically created for text processing applications, can be aggregated and merged to GRISP, providing new vocabulary and complementary semantic descriptions, with minimal manual efforts.

We plan to release a free version of GRISP corresponding to the merging of the subset of resources which are free and permits the distribution of derived versions for non-commercial use.

Future works include the experiments of GRISP for more applications, in particular the automatic classification of scientific publications and patent documents following different classification schemes. We also foreseen the integration of more languages such as Japanese and Chinese which are essential for scientific and technical information.

9. References

K. Ahmad. 1996. Pointer project final report. Technical report.

L. Bentivogli, P. Forner, B. Magnini, and E. Pianta. 2004. Revising the wordnet domains hierarchy: semantics, coverage and balancing. In *Proceedings of COLING Workshop on Multilingual Linguistic Resources*, Geneva, Switzerland.

M. T. Cabré, Sager J. C., and J. A. DeCesaris. 1999. *Terminology: theory, methods, and applications*. John Benjamins Publishing Company.

C.-C. Chang and C.-J. Lin. 2001. Libsvm: a library for support vector machines. Technical report.

K. Degtyarenko and al. 2008. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.*, 36:344–350.

A.H. Doan, P. Domingos, and A. Halevy. 2001. Reconciling schemas of disparate data sources: a machine-learning approach. In *Proc ACM SIGMOD Conf.*

E. Gabrilovich and S. Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*.

A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. 2005. A framework for modeling and evaluating automatic semantic reconciliation. *VLDB Journal*, 14(1):50–67.

M. Khayari, S. Schneider, I. Kramer, and L. Romary. 2006. Unification of multi-lingual scientific terminological resources using the iso 16642 standard. the termsciences initiative. In *International Workshop Acquiring and representing multilingual, specialized lexicons: the case of biomedicine*, Genoa, Italy.

P. Lopez and L. Romary. 2009. Multiple retrieval models and regression models for prior art search. In *CLEF 2009 Workshop*, Corfu, Greece.

J. Madhavan, P.A. Bernstein, and E. Rahm. 2001. Generic schema matching with cupid. In *Proc. 27th Intl. Conference on Very Large Databases (VLDB)*, Rome, Italy.

B. Magnini and G. Cavaglià. 2000. Integrating subject field codes into wordnet. In *Proceedings of LREC-2000, International Conference on Language Resources and Evaluation*, Athens, Greece.

D.L. McGuinness, R. Fikes, J. Rice, and S. Wilder. 2000. An environment for merging and testing large ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning*, Breckenridge, Colorado.

G. Roda, J. Tait, F. Piroi, and V. Zenz. 2009. Clef-ip 2009: Retrieval experiments in the intellectual property domain. In *CLEF 2009 Workshop*, Corfu, Greece.

L. Romary. 2001. An abstract model for the representation of multilingual terminological data: Tmf - terminological markup framework. In *TAMA (Terminology in Advanced Microcomputer Applications)*, Antwerp, Belgium.

B Sagot and D. Fišer. 2008. Building a free french wordnet from multilingual resources. In *In Ontolex*, Marrakech, Maroc.

Temmerman, R. 1997. Questioning the univocity ideal. the difference between socio-cognitive terminology and traditional terminology. *Hermes, Journal of linguistics*, 18:51–91.

The Gene Ontology Consortium. 2000. Gene ontology: tool for the unification of biology. *Nature Genet.*, 25:25–29.

S. Wang, A. Issac, Schopman B., S. Schlobach, and L. van der Meij. 2008. Matching multi-lingual subject vocabularies. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL)*, Corfu, Greece.

I.H. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition.

X. Zhou, X. Zhang, and X. Hu. 2006. Using Concept-based Indexing to Improve Language Modeling Approach to Genomic IR. *Lecture Notes in Computer Science*, 3936:444.

X. Zhou, X. Hu, and X. Zhang. 2007. Topic signature language models for ad hoc retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 19(9):1276.