# Extracting Relevant Learning Objects Using a Semantic Annotation Method

Boutheina Smine[1], Rim Faiz[2], Jean-Pierre Desclés[3]
[1]LaLIC, Paris Sorbonne University
28 Rue Serpente Paris 75006, France
[2]LARODEC, IHEC de Carthage
2016 Carthage Présidence, Tunisia
[3]LARODEC, Paris Sorbonne University
28 Rue Serpente Paris 75006, France
Boutheina.Smine@etudiants.univ-paris4.fr, Rim.Faiz@ihec.rnu.tn, Jean-Pierre.Desclés@paris4.sorbonne.fr

***ABSTRACT:*** *Information research refers, in our context, to information retrieval to obtain further learning information from documents. However, automatic tools for learning information retrieval from these documents based on semantic tags are not yet effective. We propose here a model which aims at automatically annotating texts with semantic metadata. These metadata will allow us to index and extract learning objects from texts. This model is composed of two parts: the first part consists of a semantic annotation of learning objects according to their semantic categories (definition, example, exercise, etc.). The second part uses automatic semantic annotation which is generated by the first part to create a semantic inverted index able to find relevant learning objects for queries associated with semantic categories. We have implemented a system called SRIDOP, on the basis of the proposed model and we have verified its effectiveness.*

**Keywords:** Semantic Annotation, Learning Objects, Contextual Exploration

**Received:** 12 August 2012, Revised 2 October 2012, Accepted 8 October 2012



## 1. Introduction

Searching learning information has become requested owing to the rapid development of the e-learning concept within the web technologies. In general, users usually enter keywords into search engines, and the returned results list all web pages containing the same character string as the chosen keywords.

These search engines are based on terms indexation without taking into account neither the semantics of pedagogical content nor the context. Besides, e-learning platforms using traditional informational retrieval technology are not useful for learning object retrieval. In fact, a key-word based approach may result in retrieving information appearing in the list of results but not relating to the subject of learning.

A better alternative is to retrieve learning information basing on the semantic annotation of learning objects. In this way, the learning information presented by the author of a document is captured and the learning or the teaching process for the student or the instructor is respectively facilitated. Extracting learning objects enables a person to combine multiple objects and

compose personal lessons for an individual learner. This paper explains how a new kind of learning objects retrieval system is implemented by using semantic and discourse automatic annotation of learning objects according to their types (Definition, Example, Exercise, etc.). We note that the automatic annotation of learning objects is not a simple task because the pedagogically related information depends to a great extent on context. Add to that it cannot be expressed at a generic level.

The rest of this paper is organized as follows: In Section 2, we present the learning object categories. section 3 deals with the presentation of related works on learning objects annotation and retrieval. Our model for learning objects retrieval is detailed in section 4. And before concluding, we illustrate the evaluation results of the different parts of our model in the fifth section.

**2. Learning objects categories**

When working with a computer, learners will manipulate digital artifacts to perform the learning activity they have been assigned to. With the spread of pedagogical resources on the web, the idea has emerged to capitalize these artifacts by learning objects [1].

Wiley [2] defines a learning object as any digital entity which we can use, re-use or refer to during a learning process. Learning objects are supposed to be small parts of courses that may be assembled together. In reality, a complete course is "*sliced* " to

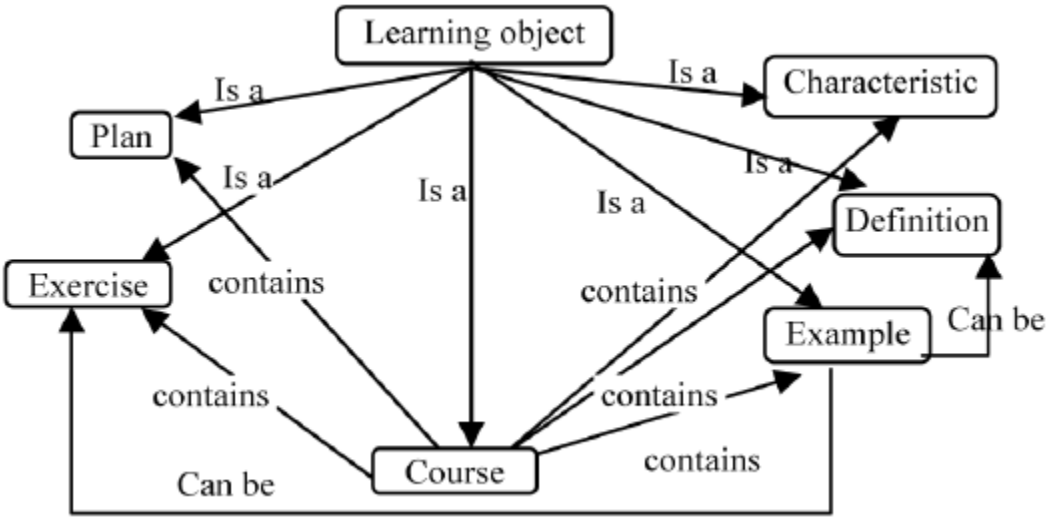


Figure 1. Relations binding learning objects categories

create several learning objects that can be composed together later on. For the purpose of this paper, we use a rather functional definition of a learning object as a textual segment (sentence, paragraph, and document) used in a learning process. This learning object is assigned to one of the types presented in Figure 1.

The first interest of learning objects is the creation of opportunities for institutions and instructors in their lesson planning and its execution. Learning objects are considered as cost and time efficient by emphasizing annotation, retrieval and reuse over individual creation. We propose to categorize learning objects according to 6 categories (Plan, Exercise, Example, Course, Characteristic and Definition). Figure 1 represents learning object categories and relations binding these objects. In our research, we are guided by the assumption that is: "*A user who searches relevant learning information proceeds by guided readings giving preferential processing to certain textual segments* (*sentences or paragraphs*)". The aim of this assumption is to reproduce: "*What does a human reader do naturally; in particular a learner who underlines certain segments relating to a particular learning object attracting his attention*".

Indeed, such a learner could be interested in a definition by formulating a request such as: find documents which contain "*The definition of the SQL Language*". Another user will look for "*Examples*" that can be applied on a certain concept (for instance "*social events*" in sociology, "*SQL requests*" in computer science, "*bacteria*" in biology, ..) by exploring many texts (specialized encyclopaedias, handbooks, articles). So these examples will be included to the user's pedagogical resources. While some users

may be interested in applying exercises to a concept, others require a course support for learning or teaching. The aim of the study dealing with learning object categories is a possible annotation of the textual learning objects. These annotations which correspond to a guided research enable to extract learning objects from texts.

Each learning category, as we mentioned above, is explicitly indicated by identifiable linguistic markers in the texts. Our hypothesis is that semantic learning objects leave some discursive traces in textual document. The learning object categories are described as follows: (1) On the one hand, a complex relation between different object categories (see Figure 1) and on the other hand a set of classes and subclasses of linguistic units (indicators and indices) structured inside the learning objects categories. (2) A set of rules: each rule connects a class of indicators with different clues.

**3. Related works**

We noticed, in the last decade, two search orientations in learning information retrieval. The first one is the Berners-Lee "*Web Semantic*" dealing with manual or semi-automatic annotations based on domain ontology. According to [3] the second one is qualified by the traditional information retrieval technology as keyword-based vector space model [4] and Decision trees.

Within the first orientation presented above, several works provide infrastructure and services for learning information annotation, indexing, and retrieval from documents. Among these works, we can mention:

QBLS [5] is a learning system for instructors and students. It proposes annotations using an RDF description. The course is structured referring to a pedagogical ontology constituted of cards (definition, example, procedure, solution, etc), then the pedagogical resources are created (course, topic, concept, and question). These resources deduced from the initial course are stored with their respective annotations in "*a database of pedagogical knowledge*". Students can thereafter practice how to resolve some questions, or learn more details about a definition, etc. When the user formulates a request, the search engine *Corese* is activated to search the pedagogical cards as response to the user's query.

We also denote the SYFAX system [6] which presents several annotations indicating: (1) correspondence of the document with the user profile (Yes/Not), (2) the user point of view on the document (interesting, average, not very interesting), (3) the type of documents (TD, TP, etc) based on the ontology "*Type of documents*" which was created manually and (4) concepts of the domain treated by the document referring to an ontology of the informatics domain built automatically from a dictionary named FOLDOC.

In order to index pedagogical documents, the various systems mentioned stored the generated annotations in knowledge databases from which relevant results are extracted. A refinement process of the request based on two ontology is suggested; one dealing with educational material types and the other one with the computer science domain. Thus, relevant documents have the same type and the same concept of the user's request.

Within the second orientation (traditional information retrieval technology), we can mention the work of [7] who explore the task of automatically identifying educational materials by classifying documents with respect to their educative value. The following features are associated with each document in the dataset: Educativeness (a four point scale ranging from non educative to strongly educative), Relevance (a four point scale ranging from non relevant to very relevant), Content categories (Definition, Example, Question & Answers, etc.), Resource type (Blog, Online book, forums, Presentation, etc.), Expertise (The expertise of the annotater in each of the selected topics on a four point scale). Authors experiment with automatic classifiers (Naïve Bayes and Salton Vector Machine) to annotate the educativeness of a given document.

We also denote the SOAF system [8] which proposes architecture to extract semantic descriptions of multimedia learning resources automatically. It is based on Latent Semantic Indexing using the representation of the resources in a vector space through their visual features. SOAF considers three types of metadata that might describe a learning object : (1) low-level features which generate automatic semantic indexing (2) High level descriptors provided by authors (title, date of creation, etc.), (3) collaborative annotations that are given by users.

The authors in [9] target the problem of finding educational resources on the web. They suggest providing metadata for educational web pages, considering first text classification, and then information extraction.

To sum up, we can say that, in the context of learning information retrieval, there exist several systems which offer manual annotation to retrieve pedagogical information. Yet, producing interesting semantic metadata manually is not interesting. In fact, providing a group of fields for users to fill in is one possible solution. However, this solution produces an inflexible system and learners still need to know various types of metadata which depend on the use of language, glossaries, expert opinions, personal experiences and so on. Automatic procedures exist but they can only fill in "*simple*" and "*low added value*" fields (e.g. a set of properties: Relevance, title, instructor, year, etc.). Therefore, their methods don't enable to reach the contents of the documents to analyse their textual segments. We support the task of automatically annotating discursive textual segments with semantic metadata relative to their learning categories.

In this paper, we propose a model which aims at automatically annotating learning objects according to their semantic categories (Definition, Example, Exercise, etc.) (cf. Section 4) in order to index and extract learning objects as response to the user's query.

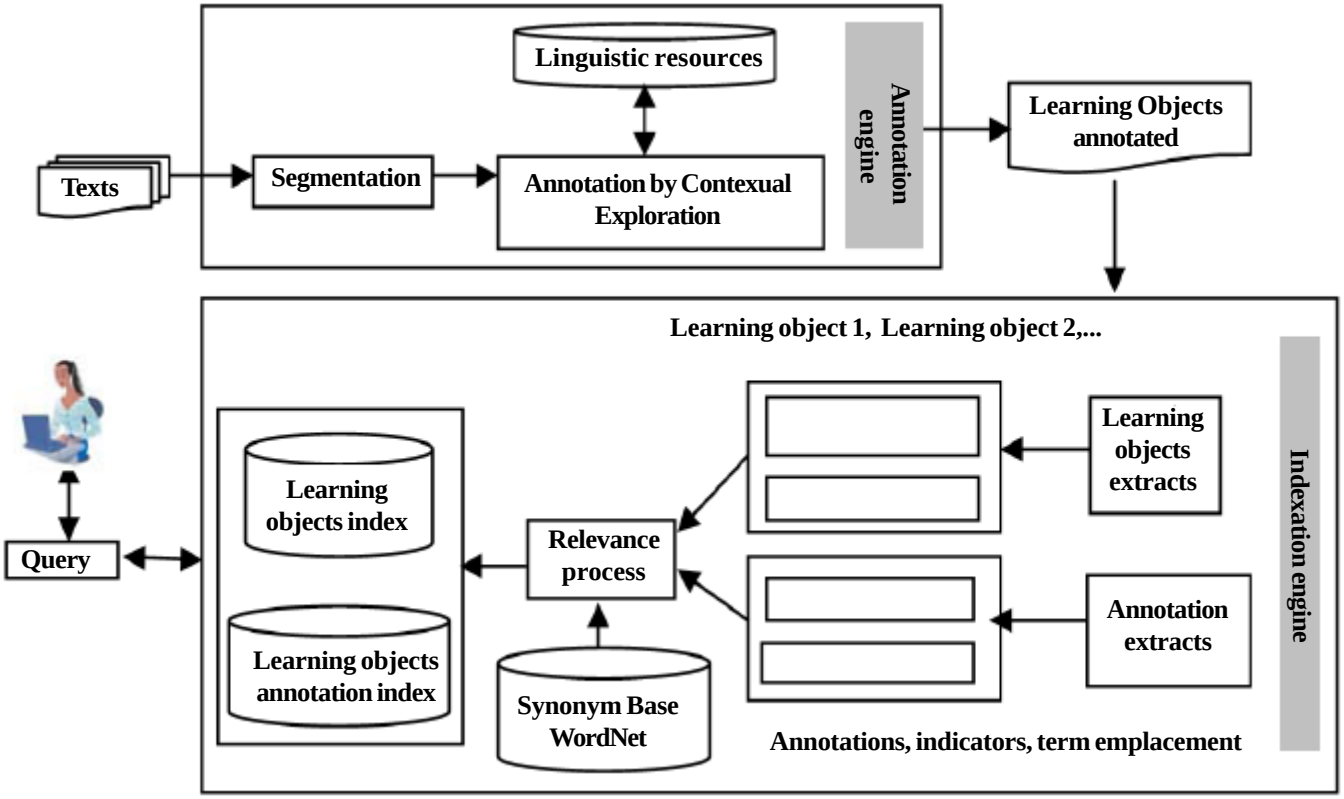


Figure 2. Our Learning Information Retrieval Model

**4. Our learning information retrieval model**

Our model is built up from two parts (see Figure 2): The first part consists of an automatic annotation of pedagogical texts according to learning object categories [10], [11]. The second part uses automatic semantic annotation which is generated by the first part to create a semantic inverted index which is able to find relevant objects for queries associated with learning categories such as *Definition, Exercise, Example*, etc.

**4.1 Learning Objects Annotation**

Before applying the annotation task, the content of the considered document has to undergo a segmentation action which lies in determining the unit's borders. Our plain text documents are transformed into XML structured documents (titles, sentences, paragraphs, etc.).

For the annotation process, we unfold the *Contextual Exploration technique 'EC'* [12], [13] which call upon knowledge exclusively linguistic and present in the texts. This linguistic knowledge is structured in form of lists and is capitalized in a knowledge base. There

are two kinds of lists: indicator lists on the one hand, contextual clue lists on the other hand. Indicators are specific to a given information learning category (i.e.: to recognize a *Definition*, to locate an *Example*, etc.). Each indicator is seen as associating a set of heuristic rules of Contextual Exploration. The application of a rule called by an indicator, amounts seeking explicitly, in the indicator context, the linguistic clues complementary to the indicator, in order to be able to solve the task (see Figure 3). In addition, it doesn't need a morpho-syntactic analysis which reduces considerably the execution time of the method [14], [15].

We focus on the learning object categories to construct our contextual exploration rules. We go through each document in order to extract linguistic structures that define the learning object categories, i.e. the category "*Definition*" can be expressed by several structures : "*...is defined as…*", "*The definition of ….is….*", "To define …, *we say that…*". These linguistic structures are expressed by discursive markers (indicators and clues) which are represented in a list of verbs, prepositions, nouns, etc. Relations binding indicators and clues are defined within Contextual Exploration rules. The rule is triggered when one of its indicators is detected within the textual segments. These rules must identify an indicator (Ii) then locate linguistic clues to the left (CLi) and/or to the right (CRi) context of the indicator, which involves the confirmation or not of the semantic value carried by the indicator.

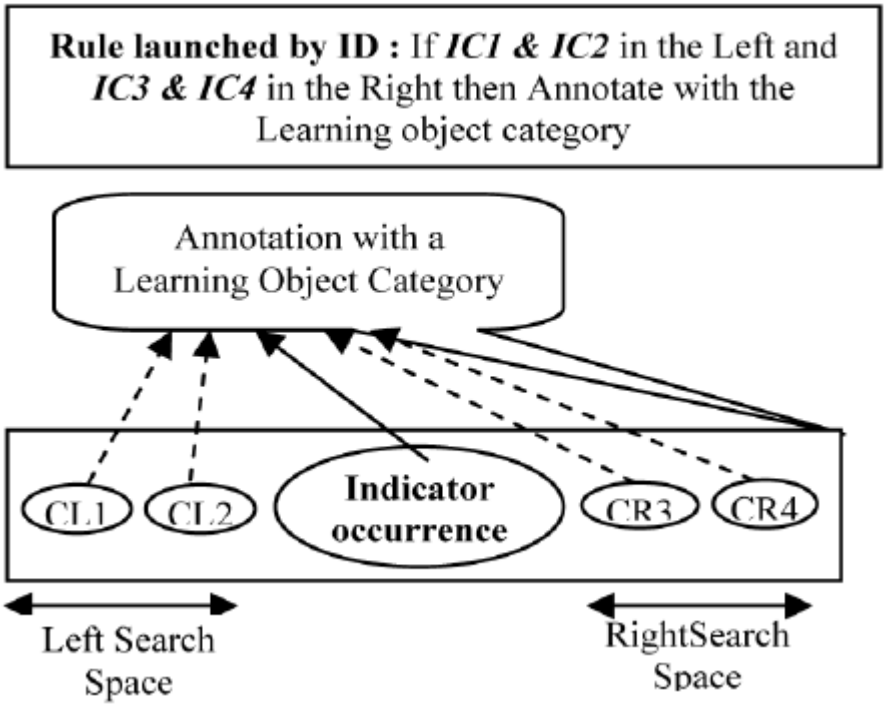


Figure 2. The annotation process

For each learning object category, we defined the set of rules which covers all the possible linguistic form of learning object. We have developed about 180 rules. We start from a textual example to generalize all linguistic structures. This method permits to define incrementally a solid base of rules. Indeed, we give the permission to the user to manage the EC rule base (adding, updating, deleting rules) through the MySQL Database system. The Table 1 shows some examples of rules. In this table, IdR denotes the identifier of the rule; $CL_1$, CL2 denote the left clues and $CR_1$, $CR_2$ denote the right clues.

The whole rules relative to the various categories and their respective indicators and clues constitute the linguistic resources that we employed to annotate learning objects. The result of the annotation process is a set of the learning objects existing in the documents and annotated with learning categories. We realize a learning object annotation engine for French language. Therefore, it can be easily extended to other languages such as English by adapting linguistic resources to English language.

The annotation process, as described above, may fail in the relevant learning object extraction due to the fact that documents unrelated to the subject of learning can be retrieved and shown in the list of results. So, we introduced another parameter to the EC rule which is the emplacement of the query's term. This parameter permits to perform the indexation process. The introduction of this parameter

is argued by the fact that the place of the term expressed in the user's query varies according to the rule applied to annotate the learning objects. We illustrate an example to detail this parameter:

For the category Definition: the term "*SQL Language*" can exist in the beginning of the sentence "*SQL Language is defined as the .....*", or in the middle of the sentence "*The person X has defined the SQL Language as ….…..…*".

We have designed this emplacement with a set of values, relatively to the indicator, and the clues of the rule (left/right of the indicator, left/right of the clues, Title, Section Title, etc.). For example, *LIND* indicates that the term emplacement is "*On the Left of the Indicator*".

| IdR | $CL_1$ | $CL_2$ | Indicator | $CR_1$ | $CR_2$ |
|---|---|---|---|---|---|
| *RD*1 | is \| are | | defined | as | |
| *RD*2 | | | is \| are | a \| an \| the | |
| *RC*1 | The \| A | | Characteristic\| Characteristics | of | is \| are |
| *RE*1 | This is | an \| the | example \| examples | of | |

Table 1. Examples of Contextual Exploration Rules

**4.2 Indexing Annotated Objects**

The aim of this step is to build up a multiple index composed of textual learning objects (sentences, paragraphs, etc), their learning categories (Definition, Example, Plan, etc.). Djioua and Desclés [16] proposed an index that would allow the user to search and extract semantic information, from texts, about "*Causality*", "*relations between concepts*", "*quotations*", etc. Inspired from this work, we have developed a model which deals with the indexation of learning objects.

Our index is realized using the Lucene platform. One of the main reasons for the use of Lucene is its scalability and performance. It works well with small collections of a few thousand documents and large collections with millions of documents. Additionally, the index files are organized to provide quick access to documents for queries. Besides, a large index can be distributed across multiple directories and searched concurrently. Our index consists of more segments. Each segment is a standalone index, holding a subset of annotated and indexed learning objects. There are separate files to hold the different files of the index (terms of the learning object, stored fields as (Document URI, Document Title, etc.), inverted index, etc. The SRIDOP inverted index is made up of textual learning objects and their learning categories. Each learning object is associated with several important pieces of information such as:

- Its learning category (Definition, Exercise, Plan, …) and subcategory (explanation significance, etc.) according to the semantic categories (Figure 1) used in the annotation process
- Document URI (Uniform Resource Identifier) for the identification of the document path
- The full-text content of the learning object for a relevant answer to users (the terms composing the learning object as used by Salton [24]).
- The emplacement of the term enounced in the user's query

We have implemented a learning information retrieval system, called SRIDOP (**S**ystème de **R**echerche d'**I**nformations à partir de **Do**cuments **P**édagogiques), using the Java language, on the basis of the proposed model.

The SRIDOP search user interface is built on top of the index that we constructed. So, the interface uses a query language which is based at the same time on both linguistic terms and semantic learning categories (definition, example, exercise …). Let us see some queries for the "*Exercise*" category. The answer to the query "*Exercises on SQL Language*", in French "*Exercices sur le langage SQL*" gives a set of learning objects (textual segments) grouped through a document URI (the annotated document by the annotation process). Each learning object presents a semantic learning category ("*Exercise*" for this example). A query is first submitted from the SRIDOP search interface to find learning objects. Then, the index is searched for matching documents and the associated metadata for each matching document (learning object) is extracted from a database and returned to the user.

The search engine proceeds as follows:

• The query, in French, has two important functions: a learning object category ("*Exercise*") and the term "*SQL Language*".

• SRIDoP extracts all learning objects found in the index associated with the annotation "*Exercise*".

• For each object extracted, SRIDoP searches the term "*SQL Language*" and its synonyms in the emplacement enounced in the index. For the term synonyms, we used a component of the synonyms dictionary WOLF (a French version of WordNet) to replace the query term by its synonyms. For example, if the term emplacement is *RIND*, the system looks for the term "*SQL Language*" in the right of the indicator.

• Selection from these learning objects, all objects within an occurrence of the term "*SQL Language*" or its synonyms in the well emplacement.

• Display all present information in the index related to each learning object selected.

**5. Experimentation and results**

We have implemented the SRIDoP system using the language *Java* and the Platform *Lucene* to annotate, index and sort the learning objects. To constitute the learning corpus for all the steps, we collect a data set covering the fifteen topics used in the step of "*Creation of learning card-index*"(i.e.Local Networks, Job-shop Scheduling, Programming language, Database, Maintenance, and so on.). Starting with each of these topics, a query is constructed and run against the Google search engine, and the top 20 ranked search results are collected. Note that the meaning of some terms can be ambiguous, e.g.., "*Base*" or "*Record"* and thus we explicitly disambiguate the query by adding the word "*data*". By performing this explicit disambiguation, we can focus on the learning property of the documents returned by the search, rather than on the differences that could arise from ambiguities of meaning.

Our testing corpus is composed of 1000 documents in French, mainly of learning nature: Support of Courses, Assignments, PowerPoint presentations, Syllabus, and documents of different nature. These documents are files in different formats (DOC, PDF, PPT, HTML, TXT, etc.) and have an average length of 53.6 pages.

**5.1 First step: Learning objects annotation**

To evaluate this step, our testing corpus was annotated by two experts: for each learning object spotted, they affect to it a category. The results of the SRIDoP annotation process are illustrated in the table below where **NOA**: Total number of annotated objects, **NOAC**: Number of objects annotated correctly, **NOMAC**: Number of objects annotated by the experts:

| Learning Object Category | NOA | NOAC | NOMAC | Precision (%) | Recall(%) | F-score (%) |
|---|---|---|---|---|---|---|
| **Plan** | 88 | 85 | 98 | 96,59 | 86,73 | 91,40 |
| **Course** | 72 | 60 | 85 | 83,33 | 70,59 | 76,43 |
| **Definition** | 228 | 140 | 266 | 61,40 | 52,63 | 56,68 |
| **Characteristic** | 139 | 124 | 156 | 89,21 | 79,49 | 84,07 |
| **Example** | 357 | 349 | 376 | 97,76 | 92,82 | 95,23 |
| **Exercise** | 760 | 705 | 776 | 92,76 | 90,85 | 91,80 |

Table 2. Experimentation results of the Annotation step

According to the experimentations presented above, the annotation results are promising. Indeed, the precision of the annotation exceeds the 85% for most learning categories (Example, Exercise, Plan, etc). But, concerning the "*Definition*" category, the corresponding precision is average. This derives owing to the fact that certain rules can annotate at the same time objects reflecting or not a "*definition*". Such the case of a "*Definition*" category rule which has as an indicator the occurrence "*is is | are*" and as clue "*a | an | the*". These indicators and clues may exist within a textual segment of a defining nature or not. During the experimental phase, we

could also note that the effectiveness of the annotation is closely related to the document segmentation effectiveness.

### 5.2 Second step: Indexing annotated objects

To test this module, we formulated 25 queries for each learning category. These queries deal with the fifteen topics of the learning and testing corpus. For each learning category, we illustrated the number of the returned results and the number of the relevant results given the whole set of the entered queries. The results are presented in the table below (Table III), where **NR**: Total number of results, **NRP**: Number of relevant results, **NRRU**: Number of relevant objects existing in the index.

| Learning Object Category | NR | NRP | NRRU | Precision (%) | Recall(%) | F-score (%) |
|---|---|---|---|---|---|---|
| **Plan** | 72 | 66 | 77 | 91,67 | 85,71 | 88,59 |
| **Course** | 43 | 35 | 54 | 81,40 | 64,81 | 72,16 |
| **Definition** | 156 | 112 | 193 | 71,79 | 58,03 | 64,18 |
| **Characteristic** | 94 | 86 | 112 | 91,49 | 76,79 | 83,50 |
| **Example** | 213 | 198 | 230 | 92,96 | 86,09 | 89,39 |
| **Exercise** | 517 | 465 | 520 | 89,94 | 89,42 | 89,68 |

Table 3. Experimentation results of the Learning Objects indexing

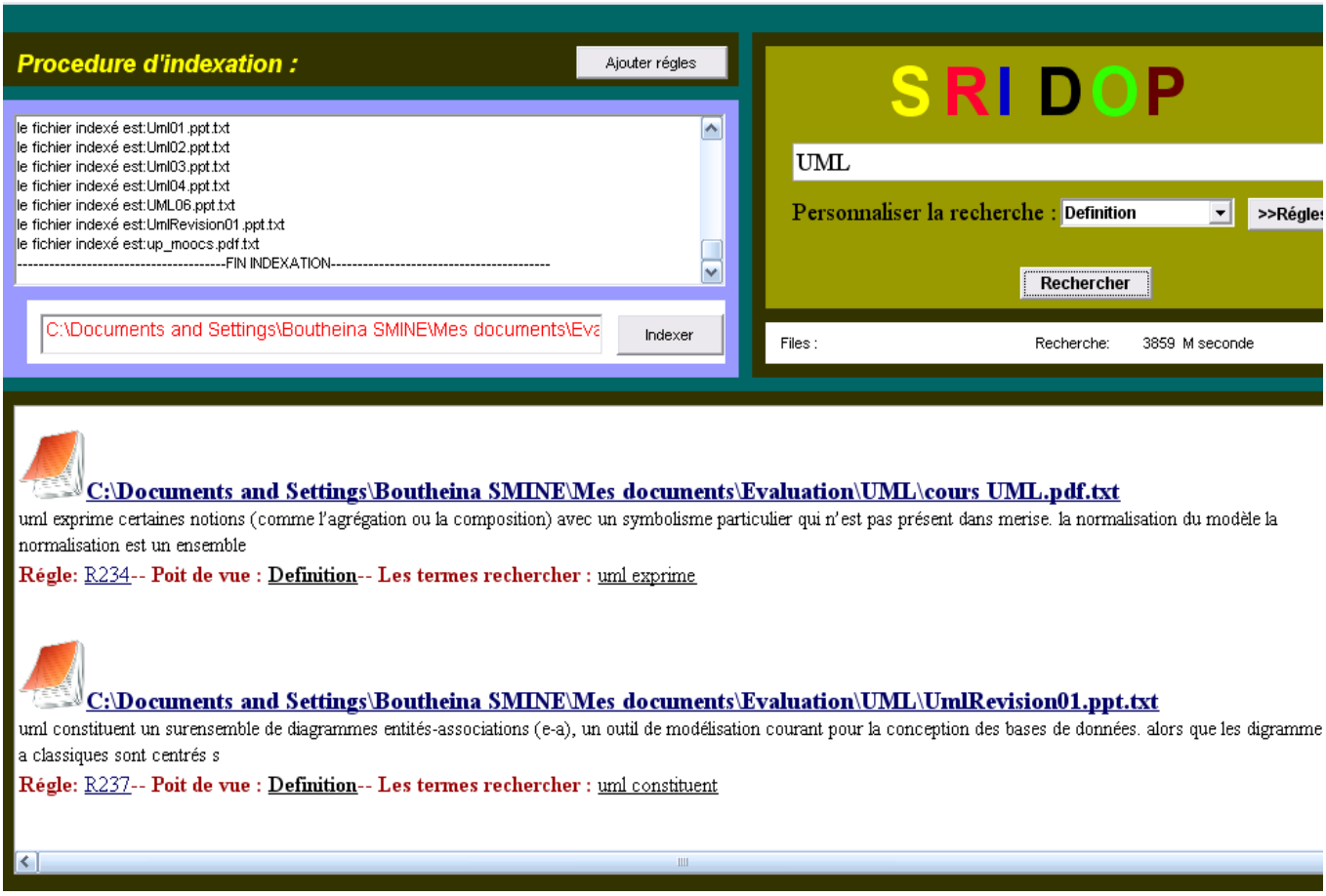


Figure 4. A screnshot of the SRIDOP system

The SRIDOP screenshot presented below (Figure 4) is divided into three parts (Left Higher part, Right Higher part and the Lower part). In the left higher part, the user chooses the set of documents to index and launch the indexation process. In the right higher part, the user enter his query which is composed at the same time of linguistic terms (UML in this case) and a learning object category (*Definition in this case*). In the lower part, the SRIDOP system displays a set of hits representing the document URI containing the relevant learning object as response to the user's query.

The Exploration Contextual rules are managed using an interface presented in the screenshot (Figure 5). The user can manage its linguistic resources (semantic map, rules, indicators and clues).

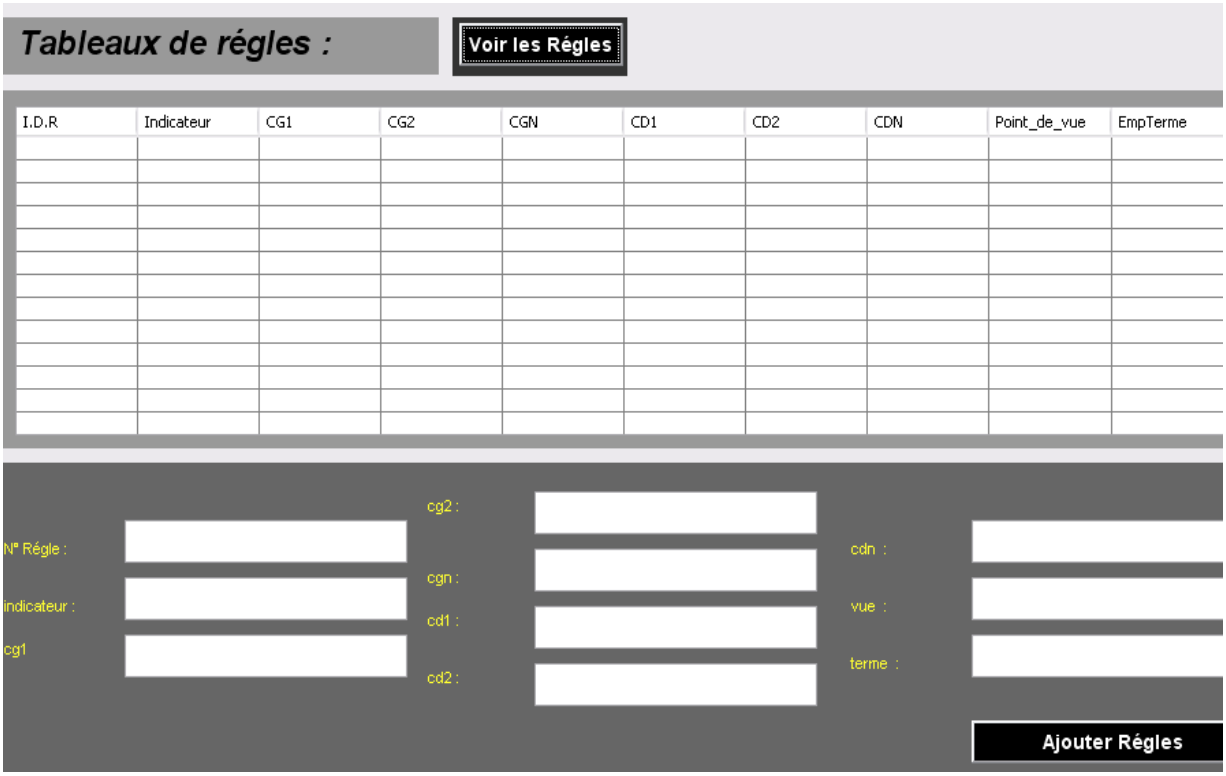


Figure 5. A screnshot of the EC Rules Management module

## 6. Conclusion and Future Works

In this article, we proposed a model for learning objects retrieval from documents. To develop it, we proceed by a semantic annotation of learning objects, then an indexation of these objects to find relevant learning objects for queries associated with semantic categories. Through the evaluation results, we observe the originality of a learning object indexation based on a semantic annotation relatively to a key-words searching system. This work comes within the context of learning objects processing and retrieval. Actually, it constitutes a considerable target in many application domains as the e-learning domain, training courses domain, data management systems, etc. One of the future works that we propose is to extend the semantic map of the pedagogical objects categories by other categories as Method, Author, etc. We also look forward to construct pedagogical card-index.

## References

[1] Christiansen, J-A., Anderson, T. (2004). Feasibility of course development based on learning objects: Research analysis of three case studies, *International Journal of Instructional Technology and Distance Learning*, 1 (3).

[2] Wiley, D. A. (2000). Connecting learning objects to Instructional design theory: a definition, a metaphor, and a taxonomy, *In*: Wiley (eds.), The Instructional Use of Learning Objects.

[3] Lee, M., Tsai, K., Wang, T. (2008). A practical ontology query expansion algorithm for semantic-aware learning objects retrieval. *Computer & Education*, 50, 1240-1257.

[4] Salton, G. (1991). Developments in automatic text retrieval. *Science*, 253 (5023) 974-980.

[5] Dehors, S., Faron-Zucker, C., Kuntz, R. (2006). Reusing Learning Resources based on Semantic Web Technologies, *In*: Proc. International Conference on Advanced Learning Technologies, Kerkrade.

[6] Smei, H., Ben Hamadou, A. (2005). Un système à base de métadonnées pour la création d'un cache communautaire-Cas de la communauté pédagogique, *In*: Proc. International E-Business Conference, Tunisia.

[7] Hassan, S., Mihalcea, R. (2009). Learning to identify educational materials, *In*: Proc. Conference on Recent Advances in Natural Language Processing (RANLP), Bulgaria.

[8] Cernea, D., Moral, E., Gayo, J. E. (2008). SOAF : Semantic indexing system based on collaborative tagging, *In*: Interdisciplinary *Journal of E-learning and Learning Objects*, 4.

[9] Thompson, C., Smarr, J., Nguyen, H., Manning, C. (2003). Finding educational resources on the web : Exploiting automatic extraction of metadata, *In*: Proc. ECML Workshop on Adaptive Text Extraction and Mining.

[10] Smine, B., Raiz, R., Desclés, J. P. (2010). Analyse de documents pédagogiques en vue de leur annotation, *Journal of New Information Technologies* (RNTI), E-19, Ed. Cépaduès, p. 429-434.

[11] Smine, B., Raiz, R., Desclés, J. P. (2011). Extraction d'Informations pédagogiques pertinentes à partir de Documents Textuels, Proceedings of the *Conference Traitement Automatique des Langues Naturelles* (*TALN* 2011), 27 June-01 July, Montpellier, France.

[12] Desclés, J. P. (1997). Systèmes d'exploration contextuelle, *In:* C. Guimier (ed.) Cotexte et calcul du sens, Presses Universitaires de Caen.

[13] Desclés, J. P. Contextual Exploration Processing for Discourse Automatic Annotations of Texts, in FLAIRS, invited speaker, Melbourne, Florida, p. 281-284.

[14] Djioua, B., Garcia-Flores, J., Blais, A., Desclés, J. P., Guibert, G., Jackiewe, A., Le Priol, F., Nait-Baha, L., Sauzay, B. (2006). EXCOM : an automatic annotation engine for semantic information, *In*: Proc. The Florida Artificial Intelligence Research Society (FLAIRS), AAAI Press, Florida, p. 285-290.

[15] Elkhlifi, A., Faiz, R. (2010). French-Written Event Extraction Based on Contextual Exploration, *In*: Proc. The Florida Artificial Intelligence Research Society (FLAIRS), AAAI Press, Florida.

[16] Djioua, B., Desclés, J. P. (2007). Indexing documents by Discourse and semantic contents from automatic Annotations of Texts, *In*: The Florida Artificial Intelligence Research Society (FLAIRS), AAAI Press, Florida.