PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTERY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY OF MOHAMED BOUDIAF - M'SILA

FACULTY: Mathematics and Informatics
DEPARTMENT of Computer Science
N°:……

DOMAIN: Mathematics and Computer Science
FIELD: Computer Science
SUB-FIELD: Information Communication Technologies

A Dissertation in Fulfillment
For the Requirements of the Degree of Master

By: Zeyneb BOUREZG

SUBJECT

Multilingual information retrieval based on ontology

Defended publicly on: 07/06 /2017, to the jury:

Board of Examiners

Dr. Bourahla Mustapha	University of M'sila	Chairman
Dr. Kadri Said	University of M'sila	Supervisor
Dr. Mahdjoubi Roussafi	University of M'sila	Examiner

Academic year: 2016 /2017

PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
MINISTERY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY OF MOHAMED BOUDIAF - M'SILA

FACULTY: Mathematics and Informatics
DEPARTMENT of Computer Science
N°:……

DOMAIN: Mathematics and Computer Science
FIELD: Computer Science
SUB-FIELD: Information Communication Technologies

A Dissertation in Fulfillment
For the Requirements of the Degree of Master

By: Zeyneb BOUREZG

SUBJECT

Multilingual information retrieval based on ontology

Defended publicly on: 07/06 /2017, to the jury:

Board of Examiners

Dr. Bourahla Mustapha	University of M'sila	Chairman
Dr. Kadri Said	University of M'sila	Supervisor
Dr. Mahdjoubi Roussafi	University of M'sila	Examiner

Academic year: 2016 /2017

Dedication

To my dear parents;

To my dear brothers and sisters

To my dear teachers of the university of Mohamed Boudiaf

To all my friends near and far

This work is dedicated.

Acknowledgement

I cannot complete this project if it is advice, aid, and consolidation of the share of people who are now forever etched in my thesis.

I express my greatest thanks to my dear family for their moral support and encouragement.

May my gratitude also go to Dr. Kadri Said, despite its many duties, has agreed to provide leadership for this work, his counsel, his dedication, his availability, his comments and corrections are relevant led to the culmination of this work.

I thank all the teachers of our university who have formed us during this cycle of license and master. The success of this project is due, mainly, to the knowledge that I have been taught in previous years.

I especially thank the teachers who advised me and towards the right path when necessary from the first year to the last year.

All those who contributed in one way or another, to make this work, can be found here, the crowning of their efforts.

Table of content

General introduction 1

Chapter 01: The development of Web and information retrieval systems

1.1. Introduction 4
1.2. The evolution of web 4
1.2.1. The internet 4
1.2.2. World Wide Web 4
1.2.3. The web 1.0 5
1.2.4. The web 2.0 5
1.2.5. The Web 3.0 5
1.3. Information retrieval 6
1.3.1. Definitions 6
1.3.2. Basic Concepts of IR 7
1.3.2.1. Document 7
1.3.2.2. Collection of documents 7
1.3.2.3. Need for information 8
1.3.2.4. Query 8
1.3.2.5. Representation 8
1.3.2.6. Research model 9
1.4. Information retrieval models 9
1.4.1. The Boolean Model 9
1.4.2. The vector model 10
1.4.3. The probabilistic model 11
1.5. Information retrieval system 12
1.5.1. Definition 12
1.5.2. Information retrieval process 12
1.5.3. Main phases of the information retrieval process 12
1.5.3.1. Indexing 13
1.5.3.2. Weighting function 14
1.5.3.3. The matching query-document 14
1.6. Search for information on the web 15

1.7. Conclusion 21

Chapter 02: Semantic web and ontology

2.1. Introduction 23
2.2. Semantic Web 23
2.2.1. history of semantic Web 24
2.2.2. Semantic web Architecture 25
2.2.3. Objectives of the semantic web 27
2.2.4. Standards apply to the semantic web 28
2.3. Ontologies 28
2.3.1. Definitions 28
2.3.2. History 29
2.3.3. Example of ontology 31
2.3.4. Components 31
2.3.5. Types of ontologies 32
2.3.6. Ontology construction methodologies 33
2.3.6.1. Method of Uschold and King 1995 33
2.3.6.2. Method of Uschold and King 1996 33
2.3.6.3. Method of Bernaras and al 1996 33
2.3.6.4. Swartout and al. 'SENSUS' method 34
2.3.6.5. Method of Assenac-Grilles and al 2000 34
2.3.6.6. Method of Bachimont 2000 34
2.3.6.7. Kassel's 2002 OntoSpec Method 34
2.4. Conclusion 34

Chapter 03: Conception and realization of multilingual ontology

3.1. Introduction: 37
3.2. A look for some statistics of users and languages on the web: 37
.3.3 Our system proposed: 40
3.4. Architecture of the proposed system: 40
3.5. Detailed architecture 40
3.5.1. Module 01 : web content database 41

3.5.2. Module 02 : Semantic Knowledge Base 41
3.6. Steps of conception of “OntoSam” Ontology 43
3.6.1. Definition of domain and objectives of “OntoSam” Ontology 43
3.6.2. Definition of classes , their properties and their hierarchy 43
3.6.3. Ontology classes hierarchy 49
3.6.4. Definition of relations between classes of “OntoSam” ontology 51
3.6.5. creation of instances (individuals): 51
3.6.6. Edition of ontology 52
3.6.7. Using RDF annotations for a multilingual ontology 55
3.7. Development environment: 55
3.7.1. Software environments 55
3.7.2. languages used: 56
3.7.3. Server used to create and manage database 57
3.7.4. Tools used: 57
3.7.4.1. Protégé 5.2: 57
3.7.4.2. Sublime Text3 57
3.7.5. Frameworks used 58
3.8. The first interface of our IRS “OntoSam”: 58
3.9. Conclusion: 59
General conclusion 60
Bibliography 61

General introduction

Since years, the Internet has become an essential medium for the dissemination of multilingual resources. There are currently around 6,900 living languages around the world. Although it is difficult to estimate the exact number of used languages among them, due mainly to the lack of reliable and available sources of information. It can be assumed that many languages will eventually generate documents in textual format or otherwise. The Web is a vast universe of diverse human knowledge and culture, enabling the sharing of ideas and information without borders. However, the performance of the various retrieval systems varies considerably when locating documents written in one or more languages different of the one used in the application. Linguistic differences are often a major obstacle to exchange scientific, cultural, educational and commercial documents. In addition, the search for information on the Internet is faced with the problem of over-abundance of results. This problem, far from dwindling, is gaining momentum with the growth of the Web and the emergence of a wide variety of languages in the last years. Access to this multiplicity of information has become a major challenge.

In documentary literature search, the indexing of documents is relatively more or less structured terminological resources. These resources provide descriptors (or preferred terms), used to unambiguously represent a contained notion in a document during the indexing process.

The evolution of the paradigms of knowledge representation in computer systems based on artificial intelligence led to the notion of ontology in the 1980s. Ontologies, in the computer sense of the term, are "explicit and formal specifications of a domain of knowledge" used in information systems. These structures have a fundamental role in the creation of the semantic Web, a tool that allows computers to access the meaning of data available on the Web to assist users in a more "intelligent" way.

Outline:

Our manuscript is organized as follows:

An introduction,introduce the work generally.

The first chapter, we shall set out the essential points of the subject: we will introduce the issues of information related to the development of the Web and retrieval information systems, its definitions, process and some of information retrieval tools.

The second chapter is dedicated to the notions of semantic web and ontologies. We start talking about the semantic web by giving a short history and the architecture proposed by Tim berners lee. Then we propose a set of definitions, including a notion of describing an ontology, its components, types and the methodologies used for the construction.

Finally in chapter three, we will focus on the realization of our IRS "OntoSam" starting with a small look of some statistics about the use of languages on the web (the problem of multilingual information) .Then we will talk about the conception of our realized system notably: conception of the used ontology and the methodology proposed by Noy and McGuiness .after that we present the tools used during the development of our IRS.

Finally, we will conclude this thesis by a general conclusion and perspectives.

Chapter 01:

The development of Web and information retrieval systems

1.1. Introduction:

Information retrieval (IR) can be defined as an activity whose purpose is to locate and deliver a set of documents to a user based on his need for information. The challenge is to be able, from the large volume of documents available, to find those that best match the expectations of the user. The operationalization of IR is carried out by computerized tools called Information Retrieval Systems (IRS), the purpose of which is to match a representation of the user's need with a representation of the contents of the documents .The rise of the web has put the IR in the face of new challenges of access to information, this time to find relevant information in a diverse and large space. These difficulties gave rise to a new discipline called Web Information Search.

This chapter contains three main parts: the first part introduces the evolution of web and presents the basic concepts of information retrieval and the various models that have been proposed. The second part describes the IR process. The third part will be devoted to IR on the web by presenting the tools of retrieval of information on the web, how it works and evolution to semantic web.

1.2. The evolution of web:

- **The internet:**

The internet is the global system of interconnected computer networks that use the Internet protocol suite (TCP/IP) to link devices on World Wide Web. It is a network of networks that consists of private, public, academic, business, and government networks of local to global scope, linked by a broad array of electronic, wireless, and optical networking technologies. The Internet carries an extensive range of information resources and services, such as the inter-linked hypertext documents and applications of the World Wide Web (WWW), electronic mail, telephony, and peer-to-peer networks for file sharing. **[42]**

- **World Wide Web:[41]**

The World Wide Web is an information space where documents and other web resources are identified, interlinked, and can be accessed via the internet **[52]**. English scientist Tim Berners-Lee invented the World Wide Web in 1989. He wrote the first web browser in 1990.

The World Wide Web has been central to the development of the information age and is the primary tool that billions of people use to interact on the Internet. **[48][54][44]**Web pages are primarily text documents formatted and annotated with HTML. In addition to formatted text, web pages may contain audio, video, images and software components that are rendered in the user's web browser as coherent pages of multimedia content.

The web is undoubtedly a major technology of the 21st century. And while its nature, structure and use have evolved over time, it is clear that this has also profoundly changed our business and social practices.

- **The web 1.0:**

Still called traditional web, is above all a static web, centered on the distribution of information. It is characterized by product-oriented sites, which require little intervention from users. The first e-commerce sites date from this area. The cost of proprietary programs and software is enormous and the explosion of the dot.com bubble, in 2000, challenges this approach to the web. **[37]**

- **The web 2.0:**

Or social web, totally changes perspective. It favors the dimension of sharing and exchanging information and content (texts, videos, images or others). He sees the emergence of social networks, smartphones and blogs. The web is becoming more democratic and dynamic. The opinion of the consumer is solicited constantly and he takes a taste for this virtual socialization. However, the proliferation of content of unequal quality generates infobesity that is difficult to control **[37]**.

- **The Web 3.0:**

Also called semantic web, aims to organize the mass of information available according to the context and the needs of each user, taking into account its location, preferences, etc. It is a web that tries to give meaning to the data. It is also a more portable web and that makes more and more the link between real world and virtual world. It meets the needs of mobile users, always connected through a multitude of media and malicious or playful applications **[37]**.

- **The web 4.0:**

Evoked by some as the intelligent web, scares as much as fascinates, since it aims to immerse the individual in a more and more prevalent (web) environment. It pushes to the paroxisic the way of the personalization opened by the Web 3.0 but it raises at the same time many questions regarding the protection of the privacy, the control of the data, and so on. It is a field of experimentation where not everyone is ready to venture **[37]**.

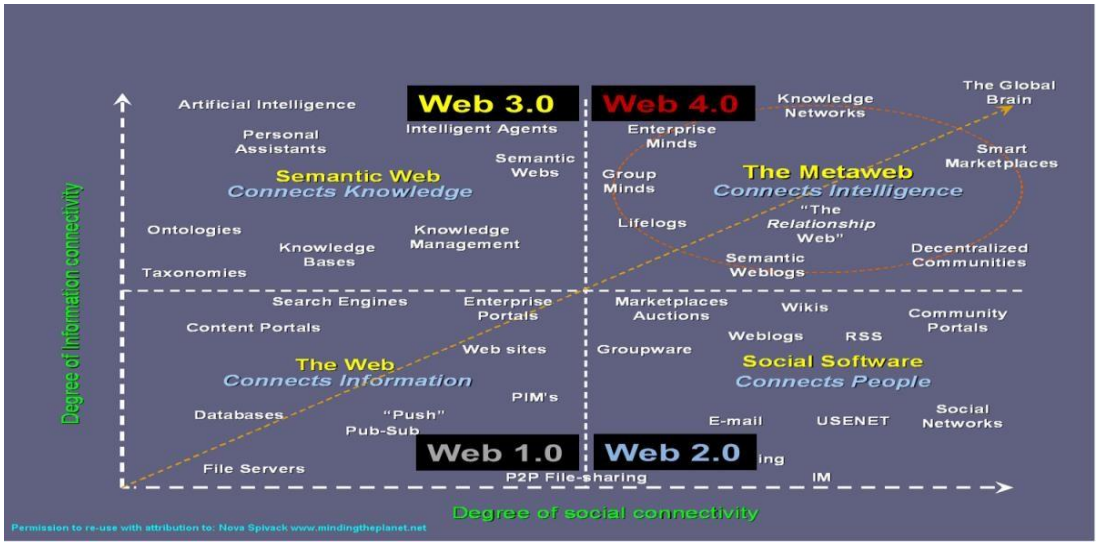

Figure 1.1 The evolution of Web and its characteristics **[37]**.

1.3. Information retrieval:

Information retrieval is an area linked to the information sciences and library science, which have always been concerned with document representations in order to retrieve information through the construction of indexes. Information technology has enabled the development of tools to process information and establish the representation of documents at the time of their indexing, as well as to search for information. It can now be said that the search for information is a transdisciplinary field which can be studied by several disciplines using approaches that should lead to finding solutions to improve its effectiveness **[1]**.

- **Definitions:**

Many definitions have been emerged for the retrieval of information in recent years; we cite in this chapter some ones:

- **Definition 1:** information retrieval is an activity whose purpose is to locate and to supply documentary granules to a user according to his / her need in information **[27]**.

- **Definition 2:** information retrieval is a branch of computing that is concerned with the acquisition, organization, storage, and research and selection information **[5]**.

- **Definition 3:** Information retrieval is a research discipline that integrates models and techniques to facilitate access to relevant information for a user with a need for information **[18]**.

All these definitions share the idea that IR is to extract from a document or a set of documents relevant information that reflects a need of information.

- **Basic Concepts of IR:**

The retrieval of information is traditionally considered as the technique of selecting from a collection of documents those that are likely to meet the needs of the user. The management of this information implies storage, retrieval and exploration of relevant documents **[17] [28]**. From this observation, several key concepts can be defined, so we have found it useful to clarify them. We can identify the following concepts:

- Document:

The document is the basic information of a collection of documents. The elementary information, also known as a document granule, can represent all or part of a document.

- Collection of documents:

The collection of documents constitutes all the information that can be exploited and accessible. It consists of a set of documents. In the general case and for the sake of optimality,

the database constitutes simplified but sufficient representations for these documents. These representations are studied of such sorts that the management or the query of the base are done in the best conditions of cost.

- Need for information:

The need for information is often assimilated to the need of the user. Three types of user requirements have been defined by: **[30]**

- **Known thematic need:** the user seeks to clarify, review or find new information in a known topic and domain. A need of this type can be stable or variable; it is very possible indeed that the need of the user refines during the research. The need can also be expressed in an incomplete way, that is to say that the user does not necessarily state everything he knows in his query but only a subset.
- **Verification need:** the user tries to verify the text with the known data it already has. It therefore searches for a particular datum, and often knows how to access it. The search for an article on the Internet from a known address would be an example of such a need. Another example would be to look for the date of publication of a book whose reference is known. A verifying type requirement is said to be stable.
- **Unknown thematic need:** this time, the user seeks new concepts or new relationships outside the subjects or domains that are familiar to him. The need is intrinsically variable and is always expressed in an incomplete way.

- Query:

The query is an expression of the user's need for information. It represents the interface between the IRS and the user**[1]**. Various types of query languages are proposed in the literature. A query is a set of keywords, but it can be expressed in natural, Boolean, or graphic language.

- Representation model:

A representation model is a process allowing extracting from a document or a query a parametrized representation that best covers its semantic content. This conversion process is called indexing. The result of indexing is the descriptor of the document or query, which is a list of terms or groups of terms (concepts), significant for the corresponding textual unit, to

which weights are generally associated, to differentiate their degrees of Representativeness of the semantic content of the unit in question. The set of terms recognized by the ISS is stored in a structure called the dictionary constituting the indexing language. This type of language guarantees the recall of documents when the query uses the dictionary terms to a large extent. On the other hand, there is a significant risk of loss information when the request moves away from that vocabulary **[1]**.

- Research model:

It represents the core model of an ISS. It includes the fundamental decision function which allows associating with a request all the relevant documents to be returned. It is used for the actual information retrieval and is closely related to the representation model of documents and queries.

1.4.Information retrieval models :

An IR model is to provide a formalization of the IR process and a theoretical framework for modeling the relevance measure. There are a large number of textual IR models developed in the literature. These models have in common the vocabulary of indexing based on keyword formalism and differ mainly by the query-document pairing model.

The model plays a central role in IR. It determines the key behavior of an IRS.

Many models exist. In the following we will first present the Boolean model which is historically one of the first models studied and which served as a starting point for the research of the domain then the vectorial model (algebraic approach) and finally, the probabilistic model.

- **The Boolean Model:**

The Boolean model is the first model of the IR. It is based on set theory. A document is represented by the set of terms that compose it. The Boolean model can be explained by considering a query consisting of a term as an unambiguous definition of a set of documents.

Thus the retrieval query simply defines the set of all indexed documents with the term retrieval; Queries can consist of several terms linked together by operators of Boolean logic. Georges Boole defined three basic operators: the logical product AND, the logical sum OR, the logical difference NOT.

A query combining two terms connected by a AND will find a set of two), the result is the union of the two sets.

This model has many advantages: it is easy to implement and is fully functional. It allows users to express structural and conceptual constraints. Users find that the use of synonyms (using the OR clause) and word groups (using the AND clause) is useful for the formulation of the query. The Boolean approach has a great power of expressiveness: it is perfectly adapted to queries that require an exhaustive and unambiguous selection. Finally the Boolean approach can be quite useful in the end of the search process because of the clarity and accuracy with which the concepts are represented.

Nevertheless, this model has received several criticisms: the first is that it is difficult for a non-expert user to formulate adequate queries using Boolean expressions .The AND and OR operators cause problems: the logical AND does not differentiate between two different cases: if no term satisfies the query or if all but one satisfy the query the operator does not find a document (Null Output Problem).

Symmetrically, too frequent use of OR brings too many documents (Overload Output problem).

The major problem with the Boolean approach is that documents that respond to the query are returned in any order, and are all identical to the query **[1]**.

- **The vector model:**

The vector model is an algebraic model in which documents and queries are represented by vectors in a multidimensional space whose dimensions are the terms derived from the indexing **[1]**. The creation of the index involves the search for the relevant terms, the lexical

treatment of the terms used, and finally the statistical analysis of the distribution of these terms in the documents and in the collection to give them a weight. Thus, documents and query are represented as vectors in the benchmark of terms. The comparison of the query to the document is made by comparing their respective vectors.

Let ***R*** be the vector space defined by the set of terms: ***<t1, t2, ...,tn>*** a document *d* and a query ***q*** can be represented by weight vectors as follow:

$$d \rightarrow <w_{d1}, w_{d2}, \dots, w_{dn}>$$

$$q \rightarrow <w_{q1}, w_{q2}, \dots, w_{qn}>$$

W_{di} and W_{qi} correspond to the weights of the term t_i in the document di and in the query ***q*** and n corresponds to the number of terms in the space **[7]**.

- **The probabilistic model:**

Several approaches **[32] [7] [19]** attempted to define weighting more formally, often based on probability theory.

The notion of probability of occurrence of an event, for example the probability of relevance P(R) is formalized through the concept of experimentation which is the process by which the observation is made. The set of values that a fact can take is the starting space.

For P(R) the starting space is, {relevant, irrelevant}. The probabilistic model considers that the indexing terms are independent that is to say that their probability of occurrence is the same with or without the presence of the other terms. Under this assumption, an attempt is made to estimate the probability that a document is relevant to a request.

PERT and NPERT respectively represent relevance and irrelevance (or equivalent, the set of relevant documents and the set of irrelevant documents).

The probabilistic model tries to estimate the probability P (PERT | D) (resp. P (NPERT / D)) that a document d belongs to the class of the relevant documents (irrelevant). In other words, relevance or irrelevance is observed in document D. Only the presence and absence of

terms in the documents and in the queries are regarded as observable characteristics. In other words, the terms are not weighted, but take only the values 0 (absent) or 1 (present).

1.5.Information retrieval system

- **Definition**

According to Alan Smeaton **[2]** "The purpose of an information retrieval system is to retrieve documents in response to a request from users, so that the contents of the documents are relevant to the initial need for information of the user".

An information retrieval system is defined by a document representation language (which can be applied to different corpus of documents) and queries that express a user's need (e.g., keywords), And a function of mapping the user's need and the corpus of documents in order to provide, as results, documents relevant to the user, that is to say, responding to his need for information.

- **Information retrieval process**

An information retrieval system manipulates a corpus of documents that it transposes by means of an indexing function into an indexed corpus. This corpus allows it to resolve queries translated from user requirements. Such a system relies on the definition of an information retrieval model that performs these two transpositions and matches the documents to the requests. Transposing a document into an indexed document is based on a document template. Similarly, the transformation of the user need into a query is based on a query model. Finally, the correspondence between a request and documents is established by a relation of relevance **[14]**.

- **Main phases of the information retrieval process:**

The fundamental purpose of an IR process is to select the documents closer "to the user's need for information described by a request. This two main phases in the process: indexing and matching request/documents.

- Indexing:

An IRS manages the various collections of documents by organizing them an intermediate representation to reflect as faithfully as possible their semantic content. The interrogation of this documentary background by means of a query requires the representation of the latter in a form compatible with that of the documents. This conversion process is called indexing (also known as for the request).Indexing is a very important step in the IR process. It consists of determining and extracting terms representative of the contents of a document or a request, which best covers their semantic content. The quality of the research depends greatly on part of the quality of indexing.

The result of indexing constitutes the so-called descriptor of the document or of the request. The latter is often a list of terms or group of terms the corresponding textual unit, usually with weight representing their degree of representativeness of the semantic content of the unit they describe.

The descriptors of the documents (words, group of words) are arranged in a structure called dictionary constituting the indexing language.

A group of words is a priori semantically richer than the words that composed separately. This argument leads us not to consider simply the words simple as basic units in the indexing language but also groups of words. This group of words forms what is called a thesaurus. The latter includes linguistic relations (equivalence, association, hierarchy) and statistics (Weighting) **[28]**.

Indexing can be characterized by its mode and weighting function.

Indexing modes: indexing can be manual, automatic or semiautomatic:

- **Manual indexing:** each document is analyzed by a specialist in the domain or by a documentalist.
- **Automatic indexing:** each document is analyzed using a process fully automated.

- **Semi-automatic indexing:** the final choice remains with the domain specialist correspondent or documentalist, who often intervenes to establish relationships semantics between keywords and choose significant terms.

- Weighting function:

[20]The weighting makes it possible to assign to each indexing term a value that measures its importance in the document where it appears. The power to discriminate terms to describe the content of documents is not identical for all terms. To find the terms of the document that best represent its semantic content,defined the weighting function of a term in a document known as Tf, Idf, which is used in different versions by the majority of IRS. There are:

- **TF (term frequency):** this measure is proportional to the frequency of the term in the document. The underlying idea is that the longer a term is used in a document, the more important it is in the description of this document.
- **Idf (Inverse of Document Frequency):** measures the importance of a term throughout the collection. The underlying idea is that the terms that appear in few documents in the collection are more representative of the content of these documents than those that appear in all the documents in the collection.

- The matching query-document:

The IRSs integrate a search / decision process that the information deemed relevant to the user. For this purpose, a measure of similarity (correspondence) between the indexed query and the descriptors of the documents in the collection is computed. Only documents whose similarity exceeds a predefined threshold are selected by the IRS.

The correspondence function is a key element of an SRI because the quality of the results depends on the ability of the system to calculate a relevance of the documents as close as possible to the judgment of relevance of the user **[20]**.

There are two types of matching:

- **Exact Match**

The result is a list of documents that exactly match the specified query with specific criteria. Returned documents are not sorted.

- **Approved matching**

The result is a list of documents that are supposed to be relevant to the query. The returned documents are sorted according to a measurement order. This order reflects the document / request.

1.6.Search for information on the web:

- **Information retrieval tools:**

There are many tools for searching information on the Web, these tools that specialize according to the services used and the type of information they identify. We often describe any search and query interface of the search engine anywhere, regardless of the source queried and the computer system used **[3]**. Different types of search tools should be distinguished on the Internet.

A first criterion for the classification of IR tools is the search mode offers. It distinguishes between tools by tree navigation (such as directories) or hypertext (such as bookmarks lists), and tools by query (such as engines, based on the use of keywords). This distinction is no longer relevant today, as the interweaving is strong between the same tools **[10]**.A second criterion remains valid, despite appearances: that of the mode indexing of resources. According to this criterion, one distinguishes the thematic directories, which carry out a SEO of websites and search engines, which operate by automated collection and indexing of web pages (not sites). This distinction, which is 'historical', is less clear today, because of the interweaving of directories and engines: Google uses the directory of the open directory; Yahoo has its own engine, and so on.

As part of our thesis, we distinguish three categories of tools for the search for information on the web: search engines, directories and meta-engines. This distinction, which also relies on indexation, remains essential, since it induces very different uses and

technologies. So a thematic directory will refer websites, where an engine will index all the pages of a site? Indeed, the directory will facilitate the clearing, the first identification of resources in a domain or a sector defined by the proposed tree organization, while a search engine will find a very precise document. Finally, the meta-engines make it possible to interrogate at the same time various search tools, whether directory or motor type. In the following, we present these three categories of IR tools.

- Search engine:

A search engine is a web application allowing finding resources from a query in the form of words. Resources can be web pages, articles from UseNet forums, images, videos, files, etc. Some websites offer a search engine as their main feature; then the site itself is called the search engine **[41]**.

Search engines are inspired by documentary search tools (based on inverted files, alias index files) used on mainframes since the 1970s, such as the STAIRS software on IBM. The mode of filling their databases is however different, because network oriented. Moreover, the distinction between formatted data ("fields") and free text no longer exists, although starting from 2010 to reintroduce itself via the semantic web.

Historical engines were Lycos (1994), Altavista (1995, first 64-bit engine) and Backrub (1997), ancestor of Google.

These are web-based research tools without human intervention, which distinguishes them from directories. They are based on "robots", also known as bots, spiders, crawlers or agents who regularly scan the sites and automatically discover new URLs. They follow the hyperlinks that link the pages to each other, one after the other. Each identified page is then indexed in a database, accessible by keywords.

It is by abuse of language that we also call search engines of websites offering directories of websites: in this case, they are research tools elaborated by persons who index and classify

websites deemed worthy of a website, Interest, not indexing robots - such as DMOZ and

formerly Yahoo.

Search engines do not only apply to the Internet: some engines are software installed on a personal computer. These are so-called desktop engines that combine search among files stored on the PC and search among websites - Exalead Desktop, Google Desktop and Copernic Desktop Search, Windex Server, etc...

There are also meta-engines, that is to say, websites where a single search is launched simultaneously on several search engines, the results then being merged to be presented to the user. We can cite in this category Ixquick, Mamma, Kartoo, Framabee or Lilo.

- How it works:**[41]**

The functioning of a search engine as any research tool is broken down into three main processes:

- **Exploration or crawl:** the web is systematically explored by a robot of indexing recursively all the hyperlinks that it finds and recovering the resources deemed interesting. The crawl is launched from a pivot resource, such as a web directory page. A search engine is first of all an indexing tool, that is to say it has a technology for collecting documents remotely on Web sites, via a tool called robot or Bot. An indexing robot has its own signature (like every web browser). Googlebot is the user agent (signature) of Google crawler.

The indexing of the recovered resources consists in extracting the words considered as significant from the corpus to be explored. The extracted words are recorded in a database organized as a gigantic reverse dictionary or, more precisely, as the terminological index of a book, which makes it possible to quickly find in which chapter of the work a given significant term is situated. Insignificant terms are called empty words. Significant terms are associated with a weight. This reflects both the probability of the word appearing in a document and the "discriminating power of that word" in a language, in accordance with the principle of the formula TF-IDF.

The search matches the query part of the engine, which returns the results. An algorithm is applied to identify in the document corpus (using the index) the documents that best correspond to the words contained in the query, in order to present the results of the searches in order of supposed relevance. Research algorithms are the subject of numerous scientific investigations. The simplest search engines are satisfied with Boolean queries to compare the words of a query with those of the documents. But this method quickly reaches its limits on voluminous corpora. The more advanced engines are based on the vector model paradigm: they use the TF-IDF formula to relate the weight of the words in a query to those contained in the documents. This formula is used to construct word vectors, compared in a vector space, by a cosine similarity. To further improve the performance of an engine, there are many techniques, the best known being that of the Google PageRank which makes it possible to weight a measurement of cosines using a index of notoriety of pages. The most recent searches use the latent semantic analysis method which attempts to introduce the idea of co-occurrences in the search for results (the term "car" is automatically associated with its close words such as "garage" or a name In the search criterion).

Similarly, an article on the harvesting of wheat in France will be considered relevant as a candidate for a reply on a question concerning the cultivation of cereals in Europe.

Complementary modules are often used in combination with the three basic bricks of the search engine. The best known are:

- **The spell checker:** it allows to correct the errors introduced in the words of the request, and to make sure that the relevance of a word will be taken into account in its canonical form.
- **The lemmatizer:** it allows to reduce the sought-after words to their lemma and thus to extend their research range.
- **The anti-dictionary:** Used to delete both "empty" words (such as "from", "the") both in the index and in the queries that are non-discriminating and disrupt the score Research by introducing noise.

- Evolution to Semantic Web:

More and more content producers, following the recommendations of the W3C on the semantic web, index their bases with metadata or taxonomies (ontologies), in order to allow the search engines to adapt to the semantic analyzes **[46] [41]**.

These forms of research and analysis of corpus of information by computer are still only potentialities.

By comparison with full-text searches, searches made on the semantic web must be more user-friendly:

The user of a semantic system must be able to directly ask his question in natural language.

The semantic search engine provides the precise answer to a question rather than a list of answering pages.

There is not yet a semantic search engine that allows us to understand a question in natural language and to adapt a response according to the results found.

There are, however, some attempts to try to answer intermediate forms to this problematic of meaning in the search for information:

- Powerset bought by Microsoft and partially integrated with Bing;
- NLGbAse which allows to query an ontology extracted from Wikipedia;
- The INRIA Edelweiss research project, which develops tools exploiting the RDF triplets;
- KartOO, of the company Kartoo, which displayed semantic graphs within the framework of its research maps (closed in 2010);
- SYNOMIA offers an intelligent search engine for Saas websites. Large semantic analysis capabilities (EDF, Minefi, AXA ...)
- Antidot Finder Suite of the French publisher Antidot, which performs semantic search

from ontologies in RDF, for example for the ISIDORE project of the CNRS;

- Sinequa CS of Sinequa, which was one of the first drivers of responses implemented in real-life situations on the institutional site of Gaz de France;
- WolframAlpha, a search engine that answers natural language questions from a database.
- Yatedo, people search engine uses semantics to extract information about a person on a web page.
- Verticrawl, a Saas search solution, integrates semantic functions dedicated to indexed space to provide consistent answers to the ontology of semantically unambiguous content.

- Directories: **[41]**

A directory is a website offering a classified list of websites.The ranking is typically in a category tree, meant to cover all or part of the visitors' interests. Each category contains: sub-categories for more specific aspects of a given subject; Hyperlinks to sites with a description.

- Types of directories:

A directory can be generalist, specialized (thematic) or geographical: The generalists do not exclude, a priori, any center of interest; The specialized and thematic directories focus exclusively on websites or web pages dealing with a certain topic or intended for a certain audience; Finally, geographical directories can be both generalist and specialized; In both cases, they relate to a country, a region, a locality.

Unlike search engines, the classification in the directories is performed by humans. Three models compete:

- The "enterprise" model: a company adopts this activity1 in order to provide this service, often free of charge for generalist directories. Its revenues are based on advertising and / or the provision of ancillary services.
- The "community" or "collaborative" model: volunteers take responsibility for a part of

the directory tree, depending on their interests, skills and availability. They select the sites proposed in the directory section, depending on the project policy. Open Directory Project is a sample directory using this template.

- The "robot-pre-filled" community-wiki model: a robot aspires published public data contents to pre-populate the indexed data directory and the community takes over to classify and re-index information and business content. As with Wikipedia, the vision is to rely on the knowledge of each visitor. The AboutUs.org template is an example of a directory using this template to list existing domain names around the world.

- Meta-engine:

A meta-engine or a meta-researcher is a search engine that draws its information through several general search engines. More precisely, the meta-motor sends its queries to several search engines and returns the results of each of them. The Meta engine allows users to enter the subject of their search only once while accessing the responses of several different search engines **[41]**.

A meta-motor eliminates similar results; For example, if Google and Yahoo! refer to the same two links, the Meta-motor will only display it once in the results list. Some indicate which search engines come from each result, such as Iquick (with stars and a tooltip) and Seeks (with respective logos and tooltips). Finally, a meta-motor sorts the results to provide first the pages provided by several engines. Some meta-motors also allow to mix a directory function (the results are classified by themes) and a motor function. This allows a double view on the results.

1.7.Conclusion:

In this chapter we have presented the main concepts of information retrieval, information retrieval systems and web search tools. Through the various sections we have presented, we conclude that the search for information focuses on defining models and systems in order to facilitate access to a set of documents found in documentary databases or on the Web. The goal is to allow users to locate documents whose content meets their information needs.

Chapter 02 :

Semantic web and ontology

2.1. Introduction:

Information retrieval is the set of procedures and techniques for selecting from a set of documents the relevant information in response to a need for information expressed by the user through a request. Recent advances in web technologies and the generalization of search engines have made finding information on the web an increasingly difficult task. The root causes of these difficulties stem mainly from the disparity and the quantity of documents to be managed on the one hand and the multiplicity of requests from users on the other.

Given these limitations, IR approaches have shifted to a new generation of research systems based on semantic access to information. The domain of semantic IR has emerged recently; its objective is to exploit knowledge related to the query in order to better meet the user's information needs. Similarly, the emergence of the semantic Web has favored research in the field of IR semantics. The aim is to explain the knowledge contained in the websites and to formalize it so that the information retrieval agents can exploit it via inference mechanisms and provide better answers to the user's needs.

This chapter is organized in two parts: in the first, we present the semantic web: the history, the components ...In the second part, we concentrate on the notion of ontology by giving its definition, its different types…etc.

2.2. Semantic Web:

The Semantic Web (more technically known as a "Web of data") allows the machines to understand the semantics, the significance of the information on the Web. It extends the network of links between Web pages classic by a network of link between structured data thus allowing agents automated to more intelligently access different sources of data on the Web and, in this way, tasks (research, learning, etc.) more accurate for users. The term was coined by Tim Berners-Lee, co-inventor of the Web and W3C Director, who oversees the development of Semantic Web standards proposals **[38][52]**.

Most of the time, when we pronounce Semantic Web, we're talking about different technologies that are hidden behind. The well-known examples are RDF (Resource Description Framework) which corresponds to a model of information, and the exchange of data in RDF formats for communicating between different applications (RDF/XML,

RDF/JSON, N3, Turtle, N-Triples and others). In the field of the Semantic Web, the semantics of the data is described by ontologies .it will be defined further in section with languages intended to provide a formal description of concepts, terms, or relationship to a field any. These languages are RDFS (Resource Description Framework Schema) and OWL (Web Ontology Language). There are also languages description structured data in XHTML so that tools do an automatic processing of these data. These languages are Microformat and RDFa and newly arrived with HTML 5 Microdata also. Then, to finish with the list of technologies, there is a query language, as well as SQL for relational databases, SPARQL, performing queries, but on RDF triplets. There are others (BBA and RDQL), but they are much less used.

2.2.1. history of semantic Web:[33]

In 1994, the first Conference WWW in Geneva, specifically at CERN, takes place the announcement of the creation of the W3C. It is also at this time that Tim Berners-Lee gives the objectives of the W3C and shows needs to add semantics to the future Web. It then shows what hypertext links or, more precisely, how it connects the documents on the Web is too limited to enable machines to automatically link the data on the Web to reality. Given the ambition of such a project, this idea raises a few resistors and controversies that are classically encountered as soon as it addresses issues related to the field of artificial intelligence.

After this conference, apart from implementing the necessary recommendations to the construction of documents, the newly created W3C began the first reflections on the implementation of the Semantic Web. These considerations lead to the publication of a first draft of recommendations on the semantic web in October 1997 and a second in April 1998. That same year, Tim Berners-Lee publishes a paper on the first versions of which will be later called the Semantic Web. These versions are to implement the different technologies of the Semantic Web. In this document, he presents the Semantic Web as a kind of extension of the Web of documents, which is a database on a global scale, so that machines can better bind the data to the Web. This roadmap is materialized by a graphical representation, the "layer cake", which shows the layout of the different technological building blocks of the Semantic Web.

Furthermore, in 1999, Tim Berners-Lee publishes the book *Weaving the Web* in which he draws up a portrait of the Web and the tracks for his future. The ideas of the Semantic Web are obviously not absent. It is in this same year he stated his famous quote:

I have a dream for the Web in which computer become capable of analyzing all the data on the Web - the content, links, and transactions between people and computers. A 'Semantic Web', which should make this possible, has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The 'intelligent agents' people have touted for ages will finally materialize.

2.2.2. Semantic web Architecture:

The current vision of the semantic Web proposed by Berners-Lee can be represented in a multi-layered architecture (Figure 2.1) **[36][35]**.

The lowest layers provide syntactic interoperability: Uniform Resource Identifier (URI) provides universal standard addressing to identify resources while Unicode is a universal text encoding for symbol exchange. Recall that the URL (Uniform Resource Locator), like URI, is a short string of characters that is also used to identify (physical) resources by their location.

XML (eXtensible Markup Language) provides a syntax to describe the structure of the document created and manipulated instances of documents. It uses the namespace to identify the names of the tags used in XML documents. The XML schema allows defining vocabularies for valid XML documents. However, XML imposes no semantic constraint on the meaning of these documents, syntactic interoperability is not sufficient for a software program to "understand" the content of the data and manipulate it in a meaningful way.

RDF M&S (RDF Model and Syntax) and RDF Schema are considered to be the first foundations of semantic interoperability. They make it possible to describe the taxonomies of concepts and properties (with their signatures). RDF provides a way to insert semantics into a document. The information is kept mainly in the form of RDF declarations. The RDFS schema describes the hierarchies of concepts and relationships between concepts, properties and domain / co-domain restrictions for properties.

The next layer Ontology describes heterogeneous distributed and semi structured information sources by defining the consensus of the common domain shared by many people

and communities. Ontologies help the machine and the human to communicate concisely using the semantic exchange rather than syntax only.

Rules are also a key element of the semantic Web vision.

The rules layer offers the possibility and the means of integration, derivation, and transformation of data from multiple sources, etc.

The Logical layer is located above the ontology layer. Some consider these two layers to be on the same level, as ontologies based on logic and allowing logical axioms. By applying logical deduction, new knowledge can be inferred from information explicitly represented.

The Proof and Trust layers are the remaining layers that provide the capacity to verify statements made in the Semantic Web. We are moving towards a reliable and secure semantic Web environment in which we can perform complex tasks in safety.

On the other hand, the origin of knowledge, data, ontologies or deductions is authenticated and ensured by digital signatures, in the case where security is important where the secret is necessary, and then encryption is used.

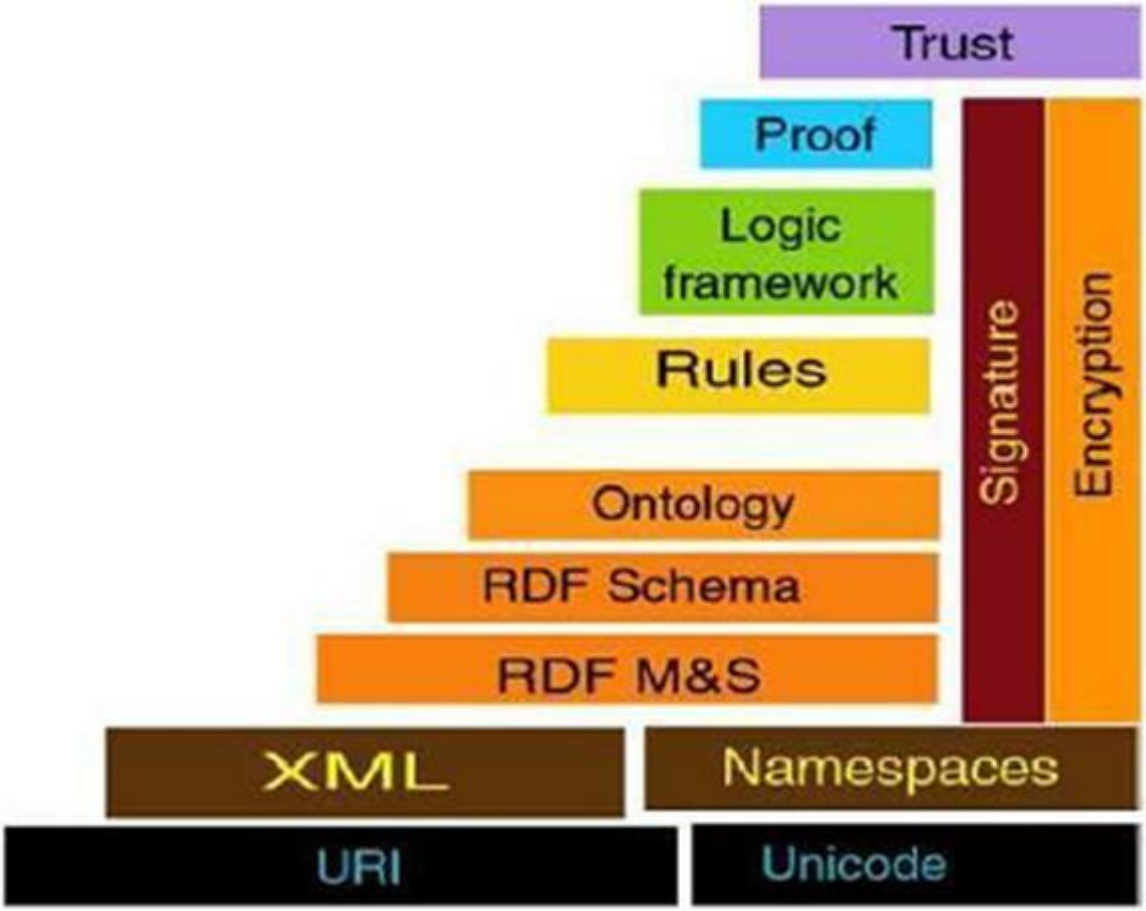

Figure 2.1 Semantic web vision (Layer cake architecture).**[50]**

2.2.3. Objectives of the semantic web:

The community of the Web as a whole tends to say that the two terms 'Semantic Web' and 'Web 3.0' roughly the same concept, if it is not completely interchangeable. The definition continues to vary based on the people you speak with. The general view is that Web 3.0 is certainly the next big revolution, but it is that, for the moment, is only speculation as to what it could be. There will be big improvements, but keeping most of the properties of Web 2.0. There are some who claim that Web 3.0 will be more application and will focus its efforts towards more graphical environments, others who claim that it will be more focused on the research of geographic information based on geolocation, or even use the many advances in artificial intelligence **[15]**.

2.2.4. Standards apply to the semantic web:

From a technical point of view, the Semantic Web consists primarily of three technical standards:

- **RDF (Resource Description Framework)**: The data modeling language for the Semantic Web. All Semantic Web information is stored and represented in the RDF **[4]**.
- **SPARQL (SPARQL Protocol and RDF Query Language)**: The query language of the Semantic Web. It is specifically designed to query data across various systems.
- **OWL (Web Ontology Language):** The schema language, or knowledge representation (KR) language, of the Semantic Web. OWL enables you to define concepts composably so that these concepts can be reused as much and as often as possible. *Composability* means that each concept is carefully defined so that it can be selected and assembled in various combinations with other concepts as needed for many different applications and purposes.

Though there are other standards sometimes referenced by Semantic Web literature, these are the foundational three **[22].**

2.3. Ontologies:

The semantic web approach consists in adding metadata to Web resources that describe their contents and their functionalities, which must be based on ontologies in order to be able to be shared. Ontologies constitute one of the most important bases of the semantic Web approach **[12]**.

2.3.1. Definitions:

Several definitions of ontologies have been proposed. The first was proposed by Neches **[31]** "An ontology defines the basic terms and relationships of the vocabulary of a domain and the rules that allow to combine terms and relationships in order to extend the vocabulary."

Two years later, Gruber **[34]** gives the definition that has become the most widely used in the literature: "An ontology is an explicit specification of a conceptualization."

In this same logic ,Guarino **[26]** propose their definition: "An ontology is a logical theory offering an explicit and partial view of a conceptualization".

Since then, many definitions, both complementary and precise, have emerged. Aussenac-

Gilles**[25]** emphasize the dependence between the formalization of the ontology and the application in which it will be used: "An ontology organizes in a network concepts representing a domain. Its content and degree of formalization are chosen according to an application ".

Thus ontology expresses an explicit consensus on the formalization of the knowledge of a domain in order to facilitate the sharing and the re-use of this knowledge by the members of a community or by software agents.

It is in this perspective that ontologies present themselves as a pillar of the semantic web because they allow people and machines to communicate using the semantics shared by the different actors of the web and describing its resources **[16][21]**.

2.3.2. History:[39]

The word ontology is built from the Greek roots: “ontos” which means "what exists", "the existent", and logos for "discourse", "study". In other words, Ontology means the study of what exists, the science of being.

The term "ontology" comes from the field of philosophy that is concerned with the study of being or existence. In philosophy, one can talk about ontology as a theory of the nature of existence (e.g., Aristotle's ontology offers primitive categories, such as substance and quality, which were presumed to account for All That Is). In computer and information science, ontology is a technical term denoting an artifact that is designed for a purpose, which is to enable the modeling of knowledge about some domain, real or imagined.

The term had been adopted by early Artificial Intelligence (AI) researchers, who recognized the applicability of the work from mathematical logic and argued that AI researchers could create new ontologies as computational models that enable certain kinds of automated reasoning. In the 1980's the AI community came to use the term ontology to refer to both a theory of a modeled world (e.g., a Naïve Physics) and a component of knowledge systems. Some researchers, drawing inspiration from philosophical ontologies, viewed computational ontology as a kind of applied philosophy.

In the early 1990's, an effort to create interoperability standards identified a technology stack that called out the ontology layer as a standard component of knowledge systems. A widely cited web page and paper associated with that effort is credited with a deliberate definition of ontology as a technical term in computer science. The paper defines ontology as an "explicit specification of a conceptualization," which is, in turn, "the objects, concepts, and other entities that are presumed to exist in some area of interest and the relationships that hold among them." While the terms specification and conceptualization have caused much debate, the essential points of this definition of ontology are:

- An ontology defines (specifies) the concepts, relationships, and other distinctions that are relevant for modeling a domain.
- The specification takes the form of the definitions of representational vocabulary (classes, relations, and so forth), which provide meanings for the vocabulary and formal constraints on its coherent use.

One objection to this definition is that it is overly broad, allowing for a range of specifications from simple glossaries to logical theories couched in predicate calculus. But this holds true for data models of any complexity; for example, a relational database of a single table and column is still an instance of the relational data model. Taking a more pragmatic view, one can say that ontology is a tool and product of engineering and thereby defined by its use. From this perspective, what matters is the use of ontologies to provide the representational machinery with which to instantiate domain models in knowledge bases, make queries to knowledge-based services, and represent the results of calling such services. For example, an API to a search service might offer no more than a textual glossary of terms with which to formulate queries, and this would act as an ontology. On the other hand, today's W3C Semantic Web standard suggests a specific formalism for encoding ontologies (OWL), in several variants that vary in expressive power. This reflects the intent that an ontology is a specification of an abstract data model (the domain conceptualization) that is independent of its particular form.

2.3.3. Example of ontology:

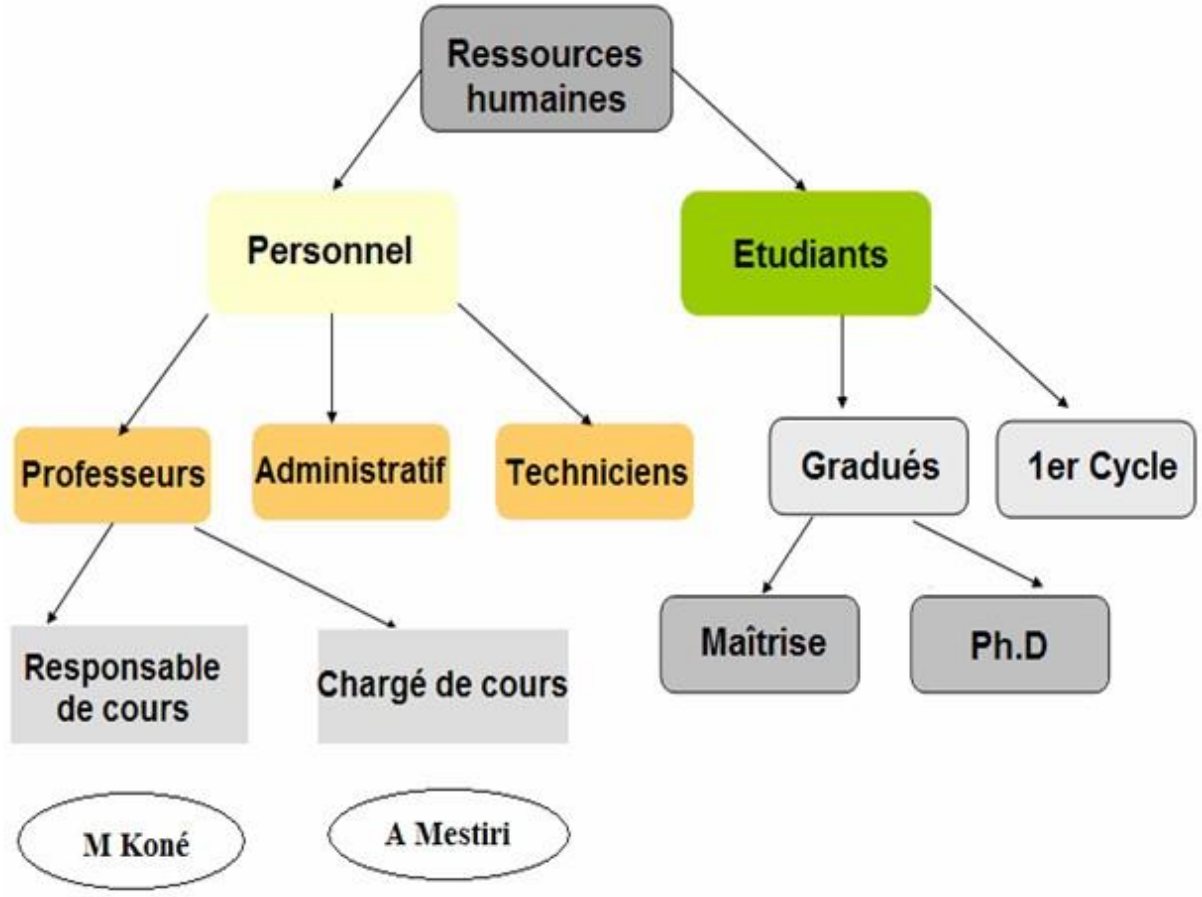

Figure2.2 An ontology of human ressources **[51].**

2.3.4. Components: [29]

Contemporary ontologies share many structural similarities, regardless of the language in which they are expressed. As mentioned above, most ontologies describe individuals (instances), classes (concepts), attributes, and relations. In this section each of these components is discussed in turn.

Common components of ontologies include:

- **Individuals:** instances or objects (the basic or "ground level" objects)
- **Classes:** sets, collections, concepts, classes in programming, types of objects, or kinds of things.
- **Attributes:** aspects, properties, features, characteristics, or parameters that objects (and classes) can have.
- **Relations:** ways in which classes and individuals can be related to one another

- **Function terms:** complex structures formed from certain relations that can be used in place of an individual term in a statement.
- **Restrictions:** formally stated descriptions of what must be true in order for some assertion to be accepted as input.
- **Rules:** statements in the form of an if-then (antecedent-consequent) sentence that describe the logical inferences that can be drawn from an assertion in a particular form.
- **Axioms:** assertions (including rules) in a logical form that together comprise the overall theory that the ontology describes in its domain of application. This definition differs from that of "axioms" in generative grammar and formal logic. In those disciplines, axioms include only statements asserted as a priori knowledge. As used here, "axioms" also include the theory derived from axiomatic statements.
- **Events:** the changing of attributes or relations.

2.3.5. Types of ontologies: [9]

Van Heijst define two main typologies of ontologies: a typology based on the structure of conceptualization and the other based on the subject of conceptualization.

In the first typology, they distinguish three categories:

- Terminological ontologies (lexicons, glossaries ...);
- Information ontologies .
- The ontologies of the knowledge models.

In the second typology, which is the most cited, they distinguish four categories:

- Application ontologies: they contain all the necessary information to model knowledge for a particular application.
- Domain ontologies: they provide a set of concepts and relationships describing the knowledge of a specific domain.
- Generic ontologies (also called high-level ontologies): they are similar to ontologies, but the concepts defined therein are more generic and describe knowledge such as state, action, space, and components.

Generally, the concepts of domain ontology are specializations of concepts of a high-level ontology.

- Ontologies of representation (meta-ontologies): they provide primitives formalization for the representation of knowledge. They are generally used to write domain ontologies and high level ontologies.

2.3.6. Ontology construction methodologies:

2.3.6.1. Method of Uschold and King 1995: **[23]**

They proposed the first "general" engineering method, the result of their ontology construction work in the field of business management. Initially, this method was based on four steps:

- Identify the purpose and scope of the ontology.
- Building ontology: capturing knowledge, coding, reusing and integrating existing ontologies.
- Evaluate the ontology.
- Document ontology.

2.3.6.2. Method of Uschold and King 1996: **[24]**

Distinguish three possibilities to identify the concepts that will be present in the ontology:

- We start from the most generic concepts that will be declined in concepts more and more specific. It is a top-down approach.
- On the contrary, we start with specific concepts that we organize with more generic concepts. It's a bottom-up approach.
- Identify the most important concepts (not necessarily specific or generic) and start from these to find the most generic and specific concepts that will be needed. This approach moves from the middle to the end (or MIDDLE OUT).

In practice, there is no purely "TOP DOWN" or "BOTTOM UP" approach especially when an already existing ontology is reused.

2.3.6.3. Method of Bernaras and al 1996:**[11]**

It is conditional on the development of an application. It is based on three points:

- Specify the ontology-based application, in particular the terms to be collected and the tasks to be performed using this ontology.
- Organize terms using meta categories: concepts, relationships, attributes, etc.

- Refine the ontology and structure it according to principles of hierarchical modularization and organization.

2.3.6.4. Swartout and al. 'SENSUS' method: [11]

Starting with the reuse of a vast common ontology in which the relevant concepts are identified in order to extract the initial skeleton of the future ontology. The initial ontology behaves like a hinge between the different ontologies developed.

2.3.6.5. Method of Assenac-Grilles and al 2000:[13]

The methodology of constructing an ontology based on text proposed by Aussenac-Gilles insists on the stage of conceptualization.

2.3.6.6. Method of Bachimont 2000:[11]

Proposes to determine the meaning of a concept (node) in the ontological tree (taxonomy). This method is based on four principles:

- The principle of community with the father.
- The principle of difference with the father.
- The principle of difference with the brothers.
- The principle of community with the brothers.

2.3.6.7. Kassel's 2002 OntoSpec Method:[11]

Developed by the IC team of LARIA of Amiens is based on the notion of semantic axis grouping the sub-concepts of a concept according to the characteristics involved in the definition of their differentiation.

Despite the large number of methods and approaches proposed, at present there are about thirty of them, none have been able to impose themselves. These methodologies can deal with the whole process and guide the ontologist on all the stages of ontology construction.

2.4. Conclusion:

In this chapter we have presented the two main notions that we use as a support for the modeling of our propositions. This is the notion of semantics and ontology.

In semantic IR, ontologies aim to represent knowledge by being simultaneously interpretable by man and machine. The purpose of this knowledge is to facilitate the modeling of the research process. First, knowledge can be useful for understanding the content of

information granules by providing a semantics helping to interpret the words that compose it. They can also help

Taking into account the task of the user, relying in particular on the exploitation of the metadata which are associated with the granules. Finally, they can also be useful in understanding the need of the user both by the user himself and by the system. The consideration of the semantics in IR thus gives a general view on the knowledge available in a corpus and can help to specify the need for information of a user.

Chapter 3 :

Conception and realization of multilingual IRS “OntoSam”

3.1. Introduction:

The conceptual study is the most important stage of a computer project. Its goal is to determine the choices of information and processing to be handled in the information system.

After giving a general overview of IRS and ontologies, we will devote this chapter to the conception and realization of our application . Its general architecture, its operation and its modules will be described briefly describing its implementation, the implementation tools used and the interfaces of the application.

3.2. A look for some statistics of users and languages on the web:

Internet usage has seen tremendous growth. From 2000 to 2009, the number of Internet users globally rose from 394 million to 1.858 billion. By 2010, 22 percent of the world's population had access to computers with 1 billion Google searches every day, 300 million Internet users reading blogs, and 2 billion videos viewed daily on YouTube. In 2014 the world's Internet users surpassed 3 billion or 43.6 percent of world population, but two- thirds of the users came from richest countries, with 78.0 percent of Europe countries population using the Internet, followed by 57.4 percent of the Americas. **[41][55]**

The prevalent language for communication on the Internet has been English. This may be a result of the origin of the Internet, as well as the language's role as a lingua franca. Early computer systems were limited to the characters in the American Standard Code for Information Interchange (ASCII), a subset of the Latin alphabet.

After English (27%), the most requested languages on the World Wide Web are Chinese (25%), Spanish (8%), Japanese (5%), Portuguese and German (4% each), Arabic, French and Russian (3% each), and Korean (2%). By region, 42% of the world's Internet users are based in Asia, 24% in Europe, 14% in North America, 10% in Latin America and the Caribbean taken together, 6% in Africa, 3% in the Middle East and 1% in Australia/Oceania. The Internet's technologies have developed enough in recent years, especially in the use of Unicode, that good facilities are available for development and communication in the world's widely used languages. However, some glitches such as *mojibake* (incorrect display of some languages' characters) still remain.

According to forecasts by Euromonitor International, 44% of the world's population will be users of the Internet by 2020. Splitting by country, in 2012 Iceland, Norway, Sweden, the

Netherlands, and Denmark had the highest Internet penetration by the number of users, with 93% or more of the population with access.

This figures show us some statistics about users and languages on the web:

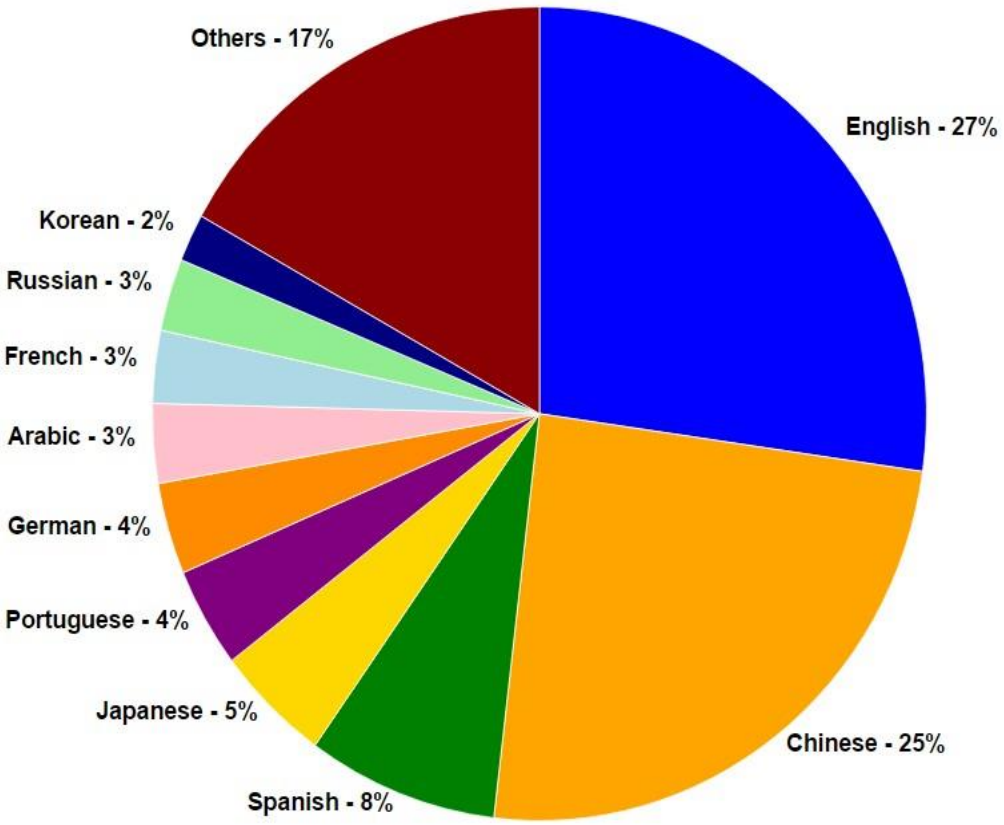

Figure 3.1. Percentage of internet users by language**[45]**

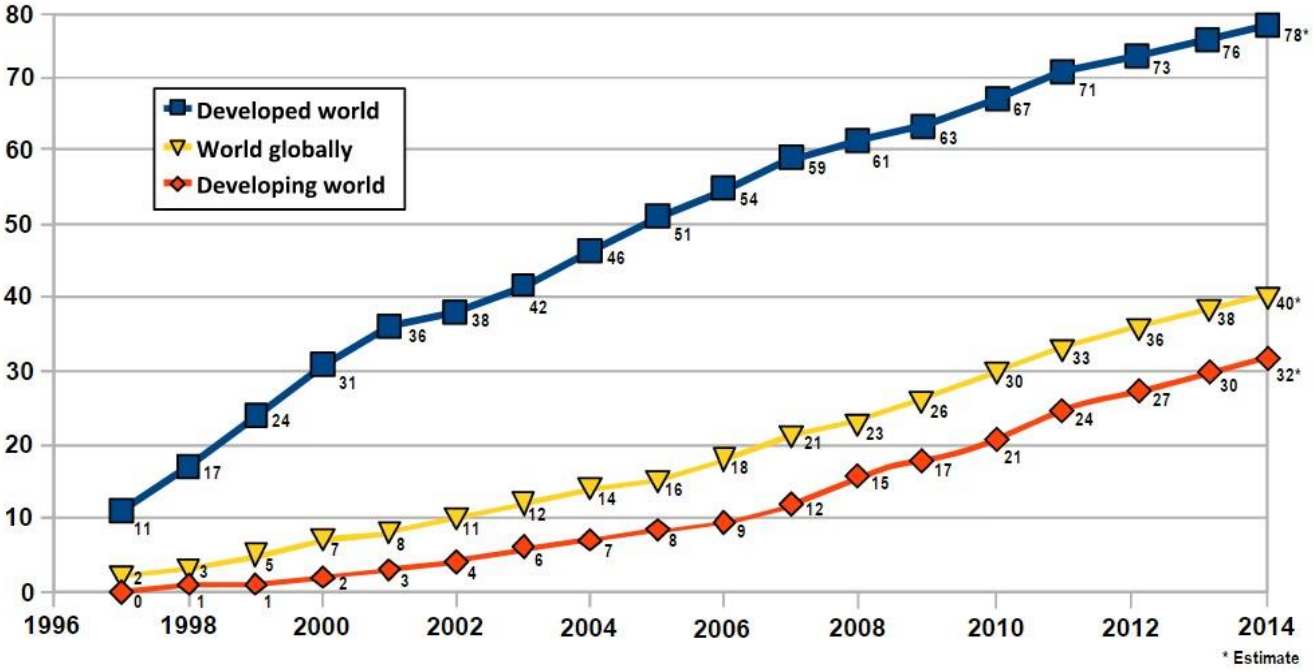

Figure 3.2 Internet users per 100 inhabitants.**[55]**.

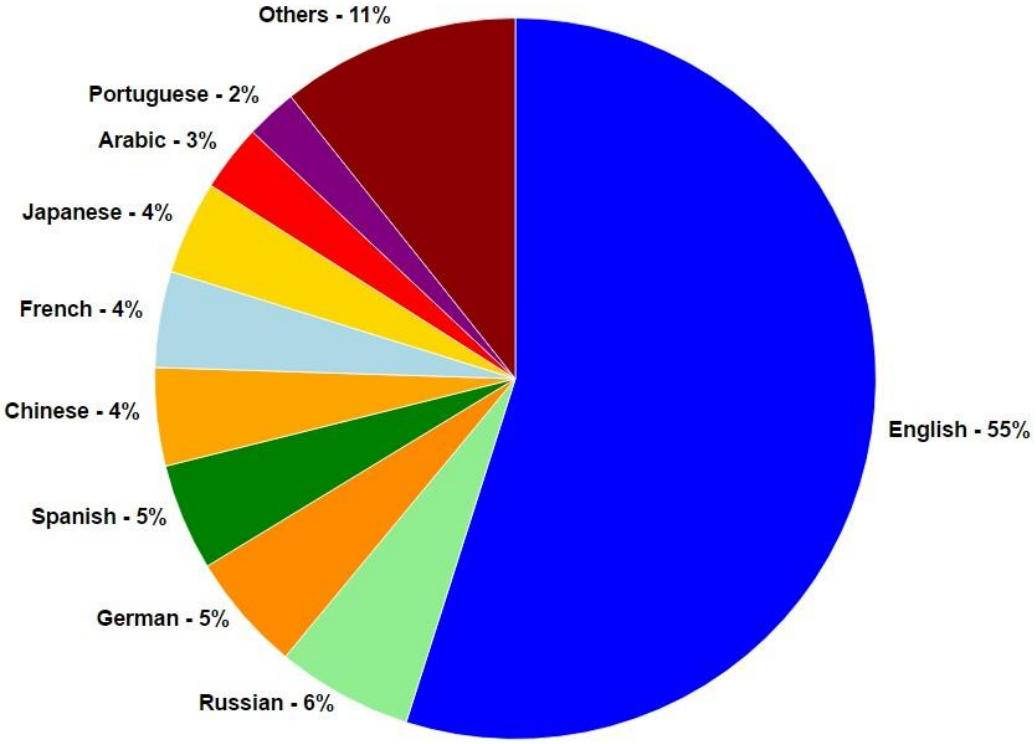

Figure 3.3 Content languages for websites (April 2013) **[53]**.

3.3. Our system proposed:

3.4. Architecture of the proposed system:

The system has been carried out in three modules:

- Web Content database (Module 1)
- Semantic Knowledge Base (Module 2)
- Search (Module 3)

The architecture of the system, in accordance to the modules, is shown below

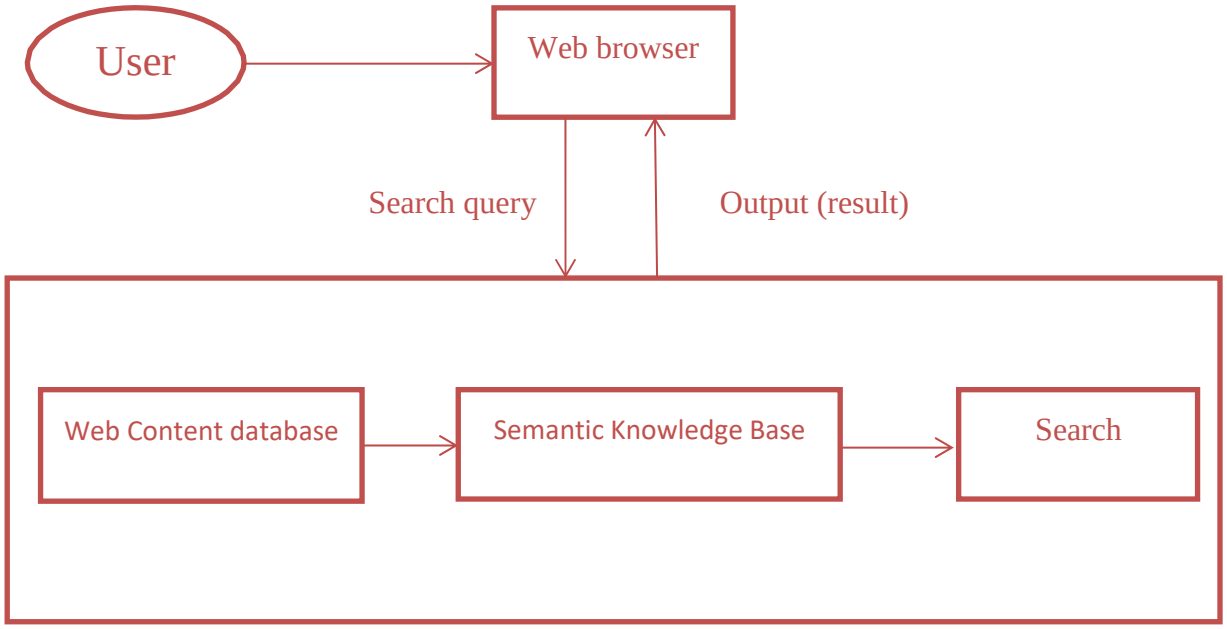

Figure3.4. Architecture of the proposed system.

Here, the user enters the search query into the semantic application. The semantic application itself consists of the three modules: the database that contains the list of URLs (module 1), the ontology created in accordance to the database (module 2) and module 3 parsed through the created ontology to return the result. This result was then displayed to user through the semantic application on the browser.

3.5. Detailed architecture:

In this part we will concentrate and detailed on the three modules or parts shown in figure 3.4.

3.5.1. Module 01 : web content database :

This module consists of the created database, this database contains 4 parts:

- ID : ID number of URLs returned in result , primary key , auto increment
- Title : the title of URL
- URL : returned in the result
- Description : a text description of the URL

This figure show us the database created in PHP MY ADMIN (MySQL)

Figure3.5 The database of the proposed system.

3.5.2. Module 02 : Semantic Knowledge Base:

In this module we will present the ontology that was used in this system (the methodology of construction of ontology)

➢ Choice of methodology used in the construction of the ontology:

A methodology is considered as a set of systematically connected construction principles, applied successfully by an author in ontology construction. Then we tried to build a multilingual ontology in the field of products of the Samsung brand according to proposed measures By Noy and McGuiness in **[6]**, we named it "OntoSam Ontology". We will then follow the following steps, inspired from the **[6]** guide with additions, mergers and eliminations of steps considered as unnecessary:

- **Step 1. Determination of the domain and the scope of the ontology**

We suggest starting the development of ontology by defining its domain and scope. That

is, answer several basic questions :

- What is the domain that the ontology will cover?
- For what we are going to use the ontology?
- For what types of questions the information in the ontology should provide answers?
- Who will use and maintain the ontology?

The answers to these questions may be changed during the ontology-design process, but at any given time they help limit the scope of the model.

- **Step 2. Enumerate important terms in the ontology:**

It is useful to write down a list of all terms we would like either to make statements about or to explain to a user. What are the terms we would like to talk about? What properties do those terms have? What would we like to say about those terms?

- **Step 3. Define classes and their hierarchy:**

There are several possible approaches in developing a class hierarchy (Uschold and Gruninger 1996)**[24]**:

- A **top-down** development process starts with the definition of the most general concepts in the domain and subsequent specialization of the concepts.
- A **bottom-up** development process starts with the definition of the most specific classes, the leaves of the hierarchy, with subsequent grouping of these classes into more general concepts.
- A **combination** development process is a combination of the top-down and bottomup approaches: We define the more salient concepts first and then generalize and specialize them appropriately.

- **Step 4. Define the properties of classes—slots**

The classes alone will not provide enough information to answer the competency questions from Step 1. Once we have defined some of the classes, we must describe the internal structure of concepts.

We have already selected classes from the list of terms we created in Step 3. Most of the remaining terms are likely to be properties of these classes. These terms include, for example, a name, age, color …etc

For each property in the list, we must determine which class it describes. These properties become slots attached to classes.

- **Step 5. Define the facets of the slots**

Slots can have different facets describing the value type, allowed values, the number of the values (cardinality), and other features of the values the slot can take. For example, the value of a name slot is one string. That is, name is a slot with value type string. A slot produces can have multiple values and the values are instances of the main class.

- **Step 6. Create instances**

The last step is creating individual instances of classes in the hierarchy. Defining an individual instance of a class requires (1) choosing a class, (2) creating an individual instance of that class, and (3) filling in the slot values.

3.6. Steps of conception of "OntoSam" Ontology:

3.6.1. Definition of domain and objectives of "OntoSam" Ontology:

like we mentioned before, we have constructed an ontology with the domain of products of Samsung brand . It is built to achieve certain objectives:

- the first and the main objective of construction a multilingual ontology because of lack of this kind of ontologies , we find many ontologies but with one language.
- The second objective is to obtain a web application of retrieval information more efficient with a wide range of results for a given query.

3.6.2. Definition of classes , their properties and their hierarchy:

the classes of our ontology " OntoSam " ontology, their properties and their sub class are grouped in this table below :

Class	Properties of class	Sub class
Products	Code Name Type	Samsung_company
Accessories	Name Type Compatible device	Products
Mobiles	Code Name Type	Products
TV/AV	Code of tv/av	Products

	Type	
Mobile	Name of mobile Type of mobile	Mobiles
Mobile_apps	Name of apps Type of apps Type of OS Version	Mobiles
Windows_tablets	Code Name Type	Mobiles
Audio_and_video	Code Name Type	TV/AV
TV	Code tv Name tv Type tv	TV/AV
TV_apps	Name of apps Type of apps Type of OS Version	TV/AV
Smartphones	Name Code Type	Mobile
Tablets	Name Type Type of OS	Mobile
Wearables	Category Carrier Color OS Size Battery	Mobile

	Memory Connectivity Processor Display Audio Service and apps	
Mobiles_apps	Code Name Type Version Type of OS	Mobiles
Galaxy_a	Code Name Processor Size Camera resolution Color Memory Connectivity OS Physical specification Battery Audio and video Services and applications	smartphones
Galaxy_j	Code Name Processor Size Camera resolution Color Memory	Smartphones

	Connectivity OS Physical specification Battery Audio and video Services and applications	
Galaxy_s	Code Name Processor Size Camera resolution Color Memory Connectivity OS Physical specification Battery Audio and video Services and applications	Smartphones
Other_phones	Code Name Processor Size Camera resolution Color Memory Connectivity OS Physical specification Battery Audio and video	smartphones

	Services and applications	
A_series	OS Processor Display Camera Memory Network/bearer Connectivity Color Battery	Tablets
Other_series	OS Processor Display Camera Memory Network/bearer Connectivity Color Battery	Tablets
S_series	OS Processor Display Camera Memory Network/bearer Connectivity Color Battery	Tablets
tabPro_s	OS	Tablets

	Processor Display Camera Memory Network/bearer Connectivity Color Battery	
Smart_fitness band	Connectivity OS Display Color Sensors	Wearables
smartWatches	Code Name Processor Size Camera resolution Color Memory Connectivity OS Physical specification Battery Audio and video Services and applications Phones comapatibles	Wearables
Galaxy_apps	Name of apps Type of apps Type of OS	Mobiles_apps

	Version	
Smart_home	Name of apps Type of apps Type of OS Version	Mobiles_apps
Smart_switch	Name of apps Type of apps Type of OS Version	Mobiles_apps

Table 3.1 Definition of classes , their properties and their sub classes.

3.6.3. Ontology classes hierarchy:

we used top-down strategy for the construction of hierarchy of concepts starting with the root classe " Samsung_company"

the relations between classes is : **is a**

the figure below show us the hierarchy of classes of "OntoSam" ontology

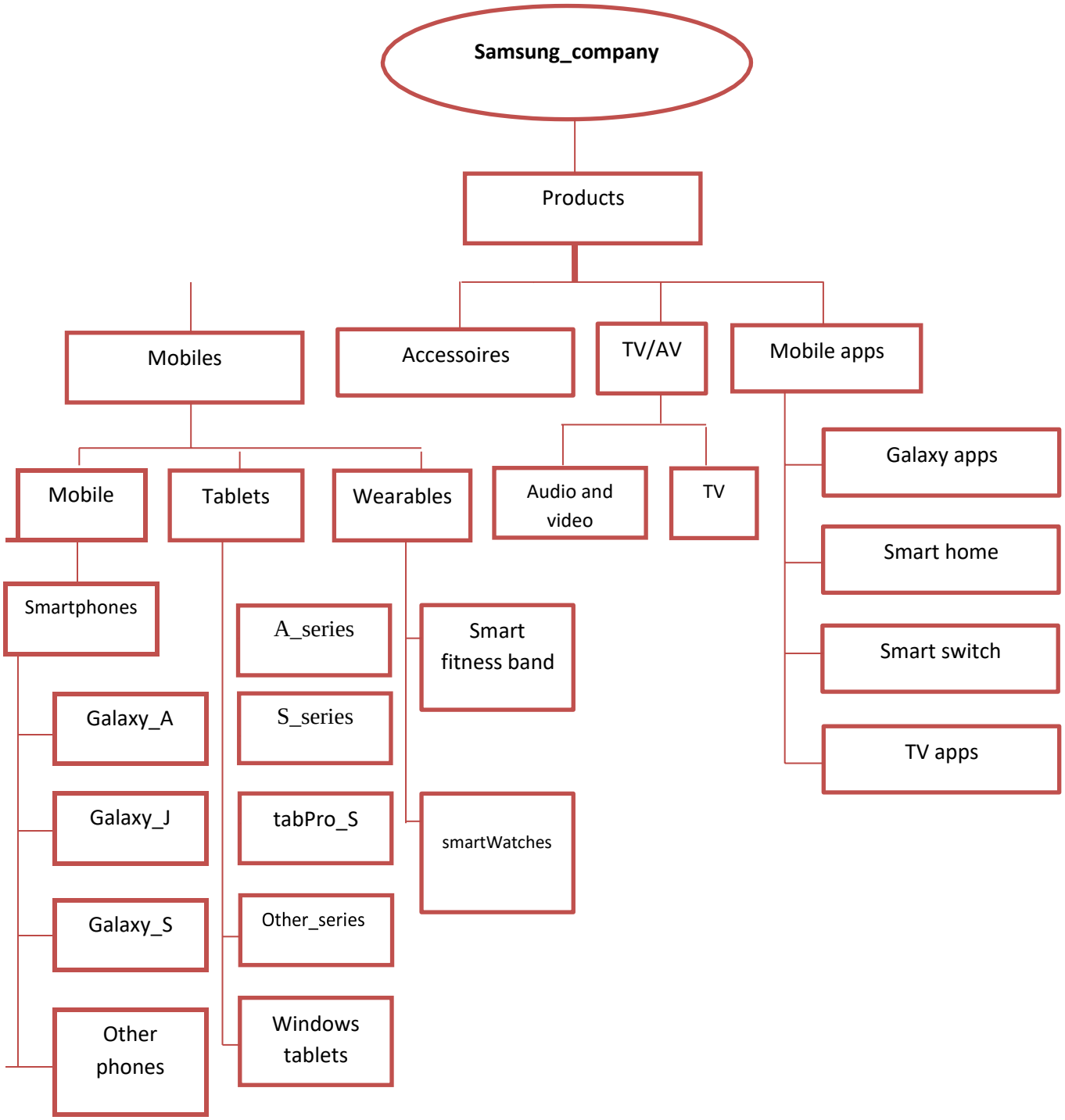

Figure 3.6 The hierarchy of concepts of "OntoSam" ontology.

3.6.4. Definition of relations between classes of "OntoSam" ontology:

we will present this step in an organized table which contains relations between classes of "OntoSam" ontology

Name of relation	Concept source	Target concept	Cardinality source	Target cardinality
Are from	Products	Samsung company	(1,n)	(1,1)
Has	Smartphones	Smart switch	(0,n)	(1.n)
Contains	Mobile	Mobiles apps	(0,n)	(1,n)
Has many	Samsung company	products	(1,1)	(1,n)
May has	Mobiles	Accessories	(1,n)	(0,n)
Compatible with	Mobiles	Mobiles apps	(1,n)	(1,n)

Table 3.2 Relations between classes .

3.6.5. creation of instances (individuals):

It was mentioned previously in the last step of methodology of construction an ontology creation of instances starting with defining an individual instance of a class requires (1) choosing a class, (2) creating an individual instance of that class, and (3) filling in the slot values.

In this step we will organize it by a table that contains some of instances created in "OntoSam" ontology.

Sub class	Instance	Attributes	values
Galaxy_ A	Galaxy_A5	Code	SM-A520FZBAXEF
		Name	Samsung galaxy A5 2017
		Processor	1.9 GHZ
		Size	5.2" (132.2mm)
		Camera resolution	16 MP
		Color	Gold
		Memory	32 GO

		Connectivity OS Physical specification Battery Audio and video Services and applications	Wifi Android 159g 3000mAh MP4 Gear
Galaxy_S	Galaxy S6 Edge	Code Name Processor Size Camera resolution Color Memory Connectivity OS Physical specification Battery Audio and video Services and applications	SM-G925FZDAXFE Samsung galaxy s6 edge 2.1 GHZ 5.1" (129.2mm) 16 MP Black 24.9 GB Wifi Android 132g 2600 mAh MP4 Gear circle

Table 3.3 Some of instances created in "OntoSam" ontology.

3.6.6. Edition of ontology:

The ontology edition is a primordial step to move to a specified operational ontology using a representation language. We chose in the context of our work, Protégé ontology editor.

The figure below shows us the hierarchy of the concepts in protégé

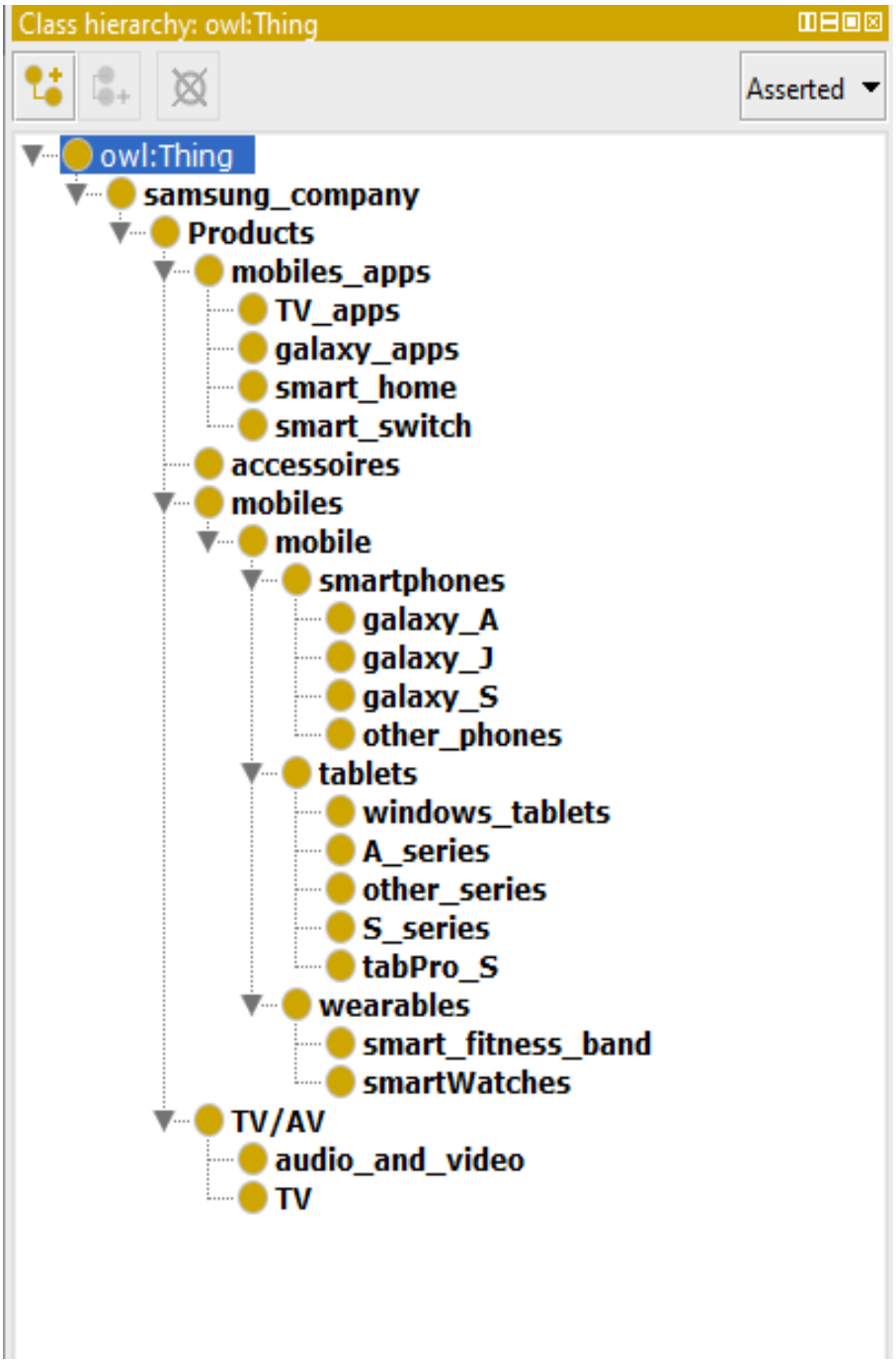

Figure 3.7 presentation of "OntoSam" ontology in protégé.

This figure also show all the relations of "OntoSam" ontology using protégé

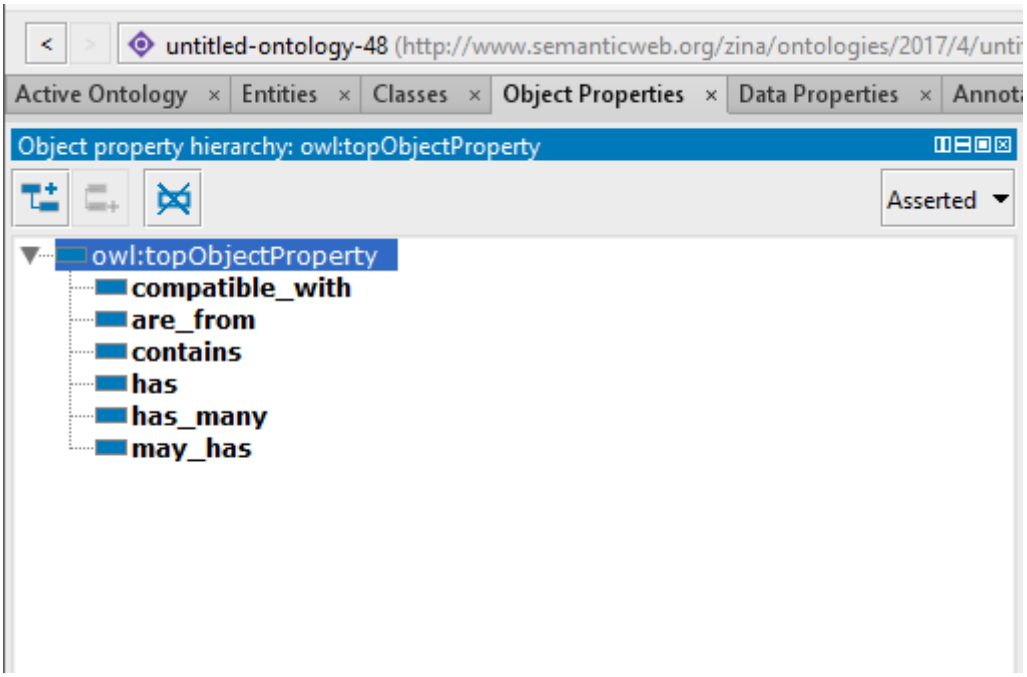

Figure 3.8 relations between classes.

When we constructed an ontology, we obtain an OWL document . Here we present some of declarations:

```
<!-- http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#TV_apps -->

<owl:Class rdf:about="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#TV_apps">
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#mobiles_apps"/>
</owl:Class>



<!-- http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#accessoires -->

<owl:Class rdf:about="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#accessoires">
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#Products"/>
</owl:Class>



<!-- http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#audio_and_video -->

<owl:Class rdf:about="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#audio_and_video">
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#TV/AV"/>
</owl:Class>



<!-- http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#galaxy_A -->

<owl:Class rdf:about="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#galaxy_A">
    <rdfs:subClassOf rdf:resource="http://www.semanticweb.org/zina/ontologies/2017/4/untitled-ontology-48#smartphones"/>
</owl:Class>
```

Figure 3.9 Some declarations of OWL document.

3.6.7. Using RDF annotations for a multilingual ontology:

The simplest way to store additional information in an ontology is to use the rdfs: label and rdfs: comment annotations. In fact, they allow to associate strings of characters with concepts, relations and instances. The rdf: lang annotation also allows multilingualism to be taken into account. It is important to note that we mention annotations (owl: AnnotationProperty) and no attributes (owl: DatatypeProperty) because attributes could only be applied to instances.

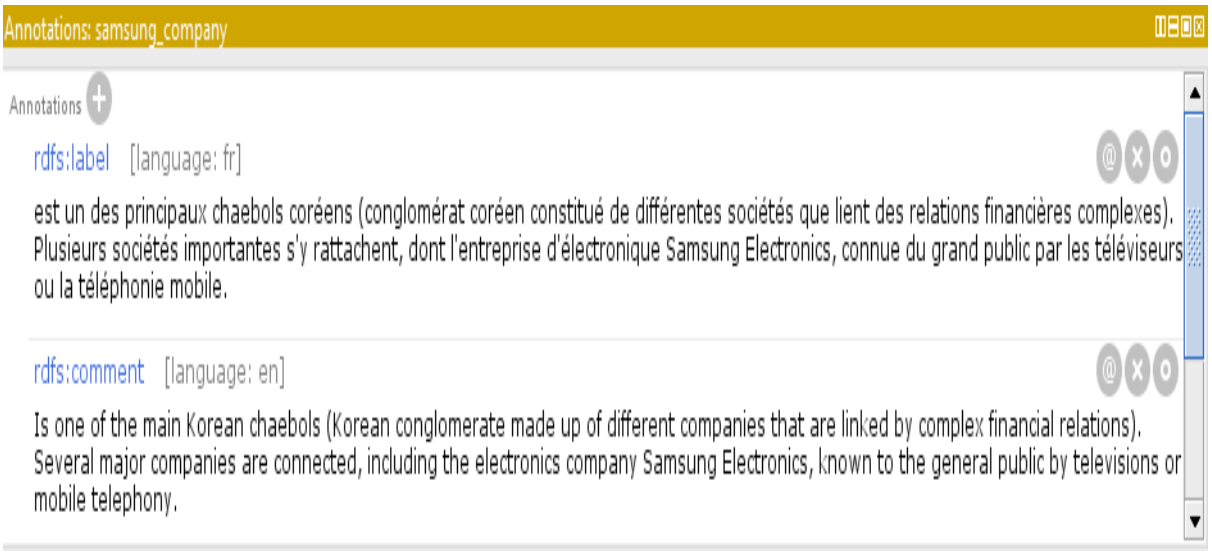

Figure 3.10 An example of using rdfs:labes, rdfs:comment and language on Samsung company class.

3.7. Development environment:

3.7.1. Software environments:

Operating system: Microsoft Windows10

Tools of development: Protégé 5.2 , Sublime Text 3

Application Server: Wamp Server

Data Base Server: MySQL database

3.7.2. languages used:

- **for Web Presentation:**

CSS : "Css (Cascading Style Sheets) is the language for describing the presentation of Web pages, including colors, layout, and fonts. It allows one to adapt the presentation to different types of devices, such as large screens, small screens, or printers. CSS is independent of HTML and can be used with any XML-based markup language".**[40]**

Html5: "HTML5 is the specification that defines the 5th major revision of the core language of the World Wide Web: the Hypertext Markup Language (HTML). HTML5 is the cornerstone of the Open Web Platform, a full programming environment for cross platform applications with access to device capabilities; video and animations; graphics; style, typography, and other tools for digital publishing; extensive network capabilities; and more" .**[41]**

- **Programming languages:**

PHP:[ ] is a server-side scripting language designed primarily for web development but also used as a general-purpose programming language. Originally created by Rasmus Lerdorf in 1994,the PHP reference implementation is now produced by The PHP Development Team. PHP originally stood for *Personal Home Page*,but it now stands for the recursive acronym *PHP: Hypertext Preprocessor*.**[41]**

PHP code may be embedded into HTML or HTML5 markup, or it can be used in combination with various web template systems, web content management systems and web frameworks. PHP code is usually processed by a PHP interpreter implemented as a module in the web server or as a Common Gateway Interface (CGI) executable. The web server software combines the results of the interpreted and executed PHP code, which may be any type of data, including images, with the generated web page. PHP code may also be executed with a command-line interface (CLI) and can be used to implement standalone graphical applications.

The standard PHP interpreter, powered by the Zend Engine, is free software released under the PHP License. PHP has been widely ported and can be deployed on most web servers on almost every operating system and platform, free of charge.

The PHP language evolved without a written formal specification or standard until 2014, leaving the canonical PHP interpreter as a *de facto* standard. Since 2014 work has gone on to create a formal PHP specification.

3.7.3. Server used to create and manage database:

MySQL [54] is an open-source relational database management system (RDBMS).Its name is a combination of "My", the name of co-founder Michael Widenius' daughter,[and "SQL", the abbreviation for Structured Query Language. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL was owned and sponsored by a single for-profit firm, the Swedish company MySQL AB, now owned by Oracle Corporation.For proprietary use, several paid editions are available, and offer additional functionality.

MySQL is a central component of the LAMP open-source web application software stack (and other "AMP" stacks). LAMP is an acronym for "Linux, Apache, MySQL, Perl/PHP/Python". Applications that use the MySQL database include: TYPO3, MODx, Joomla, WordPress, phpBB, MyBB, and Drupal. MySQL is also used in many high-profile, large-scale websites, including Google(though not for searches).

3.7.4. Tools used:

3.7.4.1. Protégé 5.2:

[55] Protégé is a free, open source ontology editor and a knowledge management system. Protégé provides a graphic user interface to define ontologies. It also includes deductive classifiers to validate that models are consistent and to infer new information based on the analysis of an ontology. Like Eclipse, Protégé is a framework for which various other projects suggest plugins. This application is written in Java and heavily uses Swing to create the user interface. Protégé recently has over 300,000 registered users. According to a 2009 book it is "the leading ontological engineering tool".

Protégé is being developed at Stanford University and is made available under the BSD 2-clause license. Earlier versions of the tool were developed in collaboration with the University of Manchester.

3.7.4.2. Sublime Text3:

[2] Sublime Text: is a proprietary cross-platform source code editor with a Python application programming interface (API). It natively supports many programming languages and markup languages, and its functionality can be extended by users with plugins, typically community-built and maintained under free-software licenses.

3.7.5. Frameworks used :

ARC2 is a PHP 5 framework for the semantic Web, specifically for the RDF and some related standards: it has a series of parsers for various formats of triplets representation, can serialize them, store them in a MySQL database and Offer a SPARQL access point, in addition to extractors of HTML formats (microformats, RDFa, etc.). It also supports extensions.**[43]**

The ARC2 library available in PHP was used for parsing through the OWL file created in the Semantic Knowledge Base module. ARC2 is a flexible RDF system for semantic web and PHP. It provides SPARQL and easy RDF parsing for LAMP systems. The ARC2 library was made available in the PHP module, which was then used to query the ontology using SPARQL.

3.8. The first interface of our IRS "OntoSam":

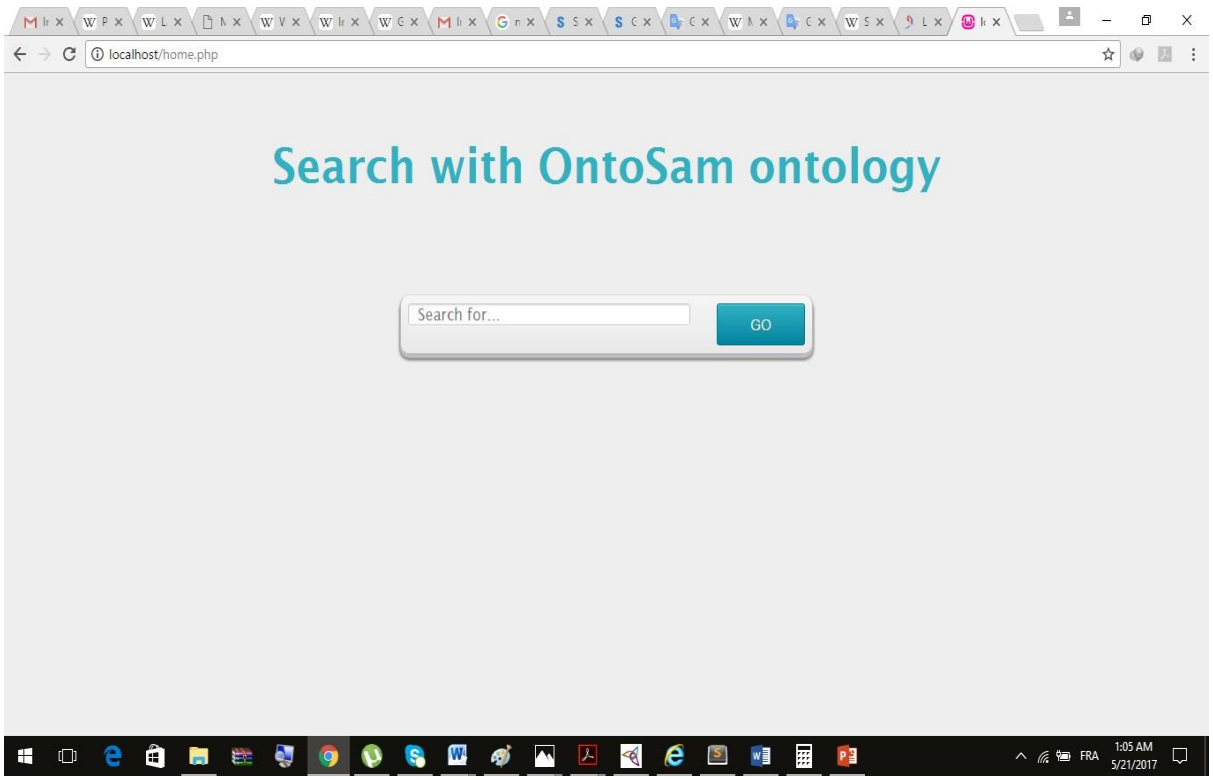

Figure3.11 First page of "OntoSam" IRS.

```php
<html>
<head>
    <title> result</title>
</head>
<body>
<?php
$con_bdd= @mysql_connect("localhost", "root", "");
$db=mysql_select_db("onto",$con_bdd);
$squery=isset($_POST['search']) ? $_POST['search'] : NULL;
include_once('C:\wamp64\www\semsol-arc2-1cba048\ARC2.php');
$config= array(
    'db_host' =>'localhost' ,
    'db_name' => 'onto',
    'db_user' => 'root',
    'db_pwd' => '',
    'store_name' => 'onto',

     );
$store= ARC2::getStore($config);
if(!$store -> isSetUp()){
    $store -> setUp();
}
$store-> query('LOAD <file:///www/ontosam.owl');
$q= ' PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
      SELECT ?url ?title ?description WHERE {
      ?Query rdfs:Label ?Label.
      ?url a ?Query.
      ?url rdfs:comment ?description.
      ?url rdfs:label ?title.
      FILTER regex(?label,"' . $squery . '" , "i")
      }
      ';
if ($rows= $store->query($q, 'rows')){
    foreach ($rows as $row ) {
    echo '<a href="'.$row['url'].'">' . $row['title'] . '</a><br/>' . $row['description'] . '<br/>' ;
```

Figure 3.13 Querying the ontology using ARC2 and SPARQL.

So, the ontology, saved as an OWL file, was accessed and queried in the PHP application, where the query entered by the user was passed as the query to the file. Here, SPARQL is used to query the OWL file, which returns the results from the created ontology. These results are then displayed to the user in the output page. HTML5 and CSS are used for designing the application to make it look more user-friendly. This user interface is displayed in Figure 3.11. Also, to provide more semantically relevant results, the information related to the entered query can also be displayed along with the URLs.

3.9. Conclusion:

In this chapter, we have presented the different steps we have followed for the implementation of our IRS. The conceptual study enabled us to follow the required steps to implement the system. This study also allowed us to make the different modules of the system.

We also described the global architecture of the application by giving the language and tools used for implementation as well as the main interfaces of the application.

General conclusion

The work presented in this thesis is part of the general context of the information retrieval. It aims the realization of a system of multilingual information retrieval based on ontology.

With the development of the internet and the exchange of documents between countries. There are more and more documents written in different languages. The search then becomes multilingual: it is necessary to find all the documents concerned by a need for information, whatever their language.

There is therefore a real need to provide effective tracking mechanisms, ie tools allowing easy, fast and reliable access to multilingual collections, so that individuals can benefit from them. Access to multilingual information must be seen as a crucial issue for both the individual and the business community. However, the process of locating information in multilingual collections is often very complex. Multilingual information retrieval systems are therefore trying to find a solution to this problem of linguistic diversity of information.

Today, in information retrieval systems, research is no longer based solely on the matching of keywords, rather than on the basis of a mere lack of meaning. It is here that the notion of ontology intervenes, organizing it in the form of a set of concepts by semantic relations. This is a technical way to simulate knowledge and take the initiative to increase search results.

This work, like any other work, makes several ideas and perspectives emerge. Among these ones we can add other languages on the ontology to make the retrieval better, also we can rich our ontology with new concepts, new domains …etc.

Bibliography

[1] A.bouramoul, «Recherche d'information contextuelle et semantique sur le web », thèse doctorat, université de MENTOURI de Constantine, 2011.

[2] A.F. Smeaton. "Information retrieval and natural language processing",In proceedings of a conference jointly sponsored by ASLIB, University of York, page 2, march 1989.

[3] A.Largouet, "La recherche d'informations sur Internet. Rapport de recherche",Service Commun de Documentation - Université Michel de Montaigne - Bordeaux3,2005.

[4] Baget, J.F., Canaud, E., Euzenat, J., Hacid, M.S."Les langages du web sémantique. Information – Interaction-Intelligence" Revue en Sciences du Traitement de l'Information. Toulouse. 2000.

[5] F. Boubekeur, "Contribution à la définition de modèles de recherche d'information flexibles basés sur les CP-Nets", thèse de doctorat en informatique, Université Paul Sabatier. 2008.

[6] F. Natalya Noy and Deborah L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology", Stanford University, Stanford, CA, 94305.

[7] G. Salton and M. McGill. "Introduction to Modern Information Retrieval". McGraw-Hill, New York, 1983.

[8] G. Salton. "A comparison between manual and automatic indexing methods. Journal of American Documentation", 20(1):61–71, 1971.

[9] G. Van Heijst, A. Schreiber, B. Wielinga. "Using explicit ontologies in KBS development". Int. J. of Human-Computer Studies, 46(2/3):183–292, 1997.

[10] G. Vignaux. "La recherche d'information : Panorama des questions et des recherches". Rapport de synthèse de recherche CNRS-MSH Paris Nord.

[11] Gargouri. Faiez, "Ontology Theory, Management and Design: Advanced Tools and Models", IGI Global, 30 avr. 2010.

[12] K. Ottens. « Un système multi-agent adaptatif pour la construction d'ontologies à partir de textes ». page 14, Octobre 2007.

[13] K.Weller, **"**Knowledge Representation in the Social Semantic Web**"**, Walter de Gruyter, 29 oct. 2010.

[14] L. Maisonnasse. "Les supports de vocabulaires pour les systèmes de recherche d'information orientés précision : application aux graphes pour la recherche

d'information médicale". thèse de doctorat en informatique, Université Joseph Fourier- Grenoble I, France, 2008.

[15] L.Feigenbaum, « The Semantic Web in Action », Scientific American, 1er mai 2007.

[16] L.K. Khelif "Un modèle général de recherche d'information : Application à la recherche de documents techniques par des professionnels". Thèse de doctorat en informatique, université Joseph Fourier-Grenoble I, 2006.

[17] M. Baziz. "Indexation conceptuelle guidée par ontologie pour la recherche d'information, Thèse de doctorat en informatique", Université Paul Sabatier de Toulouse, 2005.

[18] M. Daoud. "Accès personnalisé à l'information : approche basée sur l'utilisation d'un profil utilisateur sémantique dérivé d'une ontologie de domaines à travers l'historique des sessions de recherche", thèse de doctorat en informatique, Université Paul Sabatier,2009.

[19] M. E. Maron and J. L. Kuhns. "On relevance, probabilistic indexing and information retrieval". J. ACM, 7(3) :216–244, 1960.

[20] M.abassi,la recherché d'information(chapitre 1) , mémoire de master, université de ouregla.

[21] M.K.Khelif. "Web sémantique et mémoire d'expériences pour l'analyse du transcriptome". Thèse de doctorat en informatique, Université de Nice-Sophia Antipolis-UFR sciences, pages 7-16, Avril 2006.

[22] M.K.Smith, C.Welty, D.L.MCGuinness. "OWL Web Ontology Language", 2004.

[23] M.Uschold and M. King,"Towards a Methodology for Building Ontologies",AIAI_TR,July 1995.

[24] M.Uschold and M.Gruninger,"Ontologies :Principles,Methods and Applications,AIAI_TR, February1996.

[25] N. Aussenac-Gilles, B. Biébow, N. Szulman. "Revisiting Ontology Design: a method based on corpus analysis". Proc of KAW'2000. Juan-Les-Pins (F). Oct 2000. Lecture Notes in Artificial Intelligence Vol 1937. Springer Verlag. pp. 172-188, 2000.

[26] N. Guarino, P. Giaretti . "Ontologies and knowledge bases: Towards a terminological clarification". In Towards Very Large Knowledge Bases. N. J. I. Mars, Ed., IOS Press: 25-32, 1995.

[27] N. Hernandez, "Ontologie de domaine pour la modélisation du contexte en recherche d'information", thèse de doctorat en informatique, Université Paul Sabatier, 2006.

[28] N.D.Y. Kompaoré. « Fusion de systèmes et analyse des caractéristiques linguistiques des requêtes: vers un processus de RI adaptatif ». Thèse de doctorat en informatique, Université Paul Sabatier de Toulouse, 2008.

[29] O. Corcho, Fernandez-Lopez, M., & Gómez-Pérez, A. "Ontological Engineering-Principles, Methods,Tools and Languages". http://oa.upm.es/5457/. 2006.

[30] P. Ingwersen. "Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction". In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval., pages 101-110, 1994.

[31] R. Neches, R.E. Fikes, T. Finin T, T.R. Ruber, R. Patil, T. Senator, W.R. Wartout. "Enabling technology for knowledge sharing". AI Magazine, 12(3), 16- 36, 1991.

[32] S. Robertson and K. Sparck Jones. "Relevance weighting for search terms". Journal of The American Society for Information Science, 27(3) :129–146, 1976.

[33] T. Berners-Lee, Fischetti, Mark, Weaving the Web, HarperSanFrancisco, 1999 (ISBN 978-0-06-251587-2).

[34] T. Gruber. "A translation approach to portable ontology specifications. Knowledge Acquisition". 5(2):199–220, 1993.

[35] T.Berners-Lee, James Hendler, Lassila. Ora, « The semantic web », Scientific American 284, no.5. 2001.

[36] T.Berners-Lee, James Hendler. and Ora Lassila. "The semantic web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities". Scientific American, (285), 2001.

Websites:

[37] C marketing, www.c-marketing.eu, consulted on march 2017

[38] Crazyegg, **www.crazyegg.com,** consulted on april 2017

[39] Data versity, **www.dataversity.net,** consulted on February 2017

[40] developer Mozilla, **www.developer.mozilla.org,** consulted on may 2017

[41] English Wikipedia, **www.en.wikipedia.org,** consulted on February 2017

[42] Futur sciences, **www.futura-sciences.com**, consulted on February 2017

[43] Github, **www.github.com,** consulted on march 2017

[44] Internet live stats, **www.internetlivestats.com/,** consulted on april 2017

[45] Internet world stats, **www.internetworldstats.com,** consulted on april 2017

[46] Journal du net, **www.journaldunet.com,** consulted on march 2017

[47] Openclassrooms, **www.openclassrooms.com,** consulted on may 2017

[48] Pew internet, **www.pewinternet.org**, consulted on march 2017

[49] Protégé université de Stanford, **www.protege.stanford.edu,** consulted on February 2017

[50] Semantic scholar, **www.semanticscholar.org,** consulted on February 2017

[51] Theses ulaval, **www.theses.ulaval.ca,** consulted on may 2017

[52] w3, **www.w3.org**, consulted on march 2017

[53] W3techs, **www.w3techs.com,** consulted on February 2017

[54] Washington post, **www.washingtonpost.com**, consulted on april 2017

ملخص :

أصبحت شبكة الانترنت مصدرا شائعا جدا للمعلومات المكتوبة بلغات مختلفة. أدى تطور نماذج تمثيل المعارف في الذكاء الاصطناعي في سنوات الثمانينيات الى ظهور ما يعرف بالأنطولوجيا حيث أنها تلعب دورا في خلق الويب الدلالي ,اتجاه جديد من الويب التقليدية التي تسمح للمستخدم الوصول الى البيانات المتاحة على شبكة الانترنت بطريقة أكثر ذكاءا.

العمل الحالي هو تصميم و تنفيذ أنطولوجيا متعدد اللغات الذي يهدف الى توجيه البحث الدلالي متعدد اللغات, و بالتالي تحسين دقة النتائج المتحصل عليها.

الكلمات المفتاحية: استرجاع المعلومات، الأنطولوجيا، متعدد اللغات، الويب الدلالي

Abstract :

The web has become a very frequent source for information described in different languages. The evolution of models and paradigms of knowledge representation in AI led to the notion of ontology in the 1980s. It plays a primordial role in creating the semantic web a new trend of the classic web that allows the user to access within the data available on the web in a more intelligent way.

The present work consists in designing and implementing a multilingual ontology whose main task is to guide the multilingual semantic retrieval of information and consequently to improve the accuracy of the results obtained by the retrieval.

Keywords: Information retrieval, ontology, multilingual, semantic web

Résumé :

Le web est devenu une source très fréquente pour l'information décrite dans des langues différentes .L'évolution des modèles et de paradigmes de représentation de connaissances en IA a débouché dans les années 80 sur la notion d'ontologie. Elle joue un rôle primordial dans la création du web sémantique, une nouvelle tendance du web classique qui permet à l'utilisateur d'accéder à des données disponibles sur le web de façon plus intelligente.

Le présent travail consiste à concevoir et mettre en place une ontologie multilingue ayant comme mission principale de guider la recherche sémantique multilingue d'information et par conséquent d'améliorer la précision des résultats obtenus par la recherche.

Mots-clés : Recherche d'information, ontologie , multilingue , web sémantique.