SCIREA Journal of Computer

http://www.scirea.org/journal/Computer

November 16, 2021

Volume 6, Issue 1, February 2021

How to extract knowledge of Qualitative Data from Big Textual Data

Jouis Christophe [1]*and Orús-Lacort Mercedes [2, **]

[1] Centre d'Analyse et de Mathématiques Sociales - CAMS; cjouis45@gmail.com

[2] Online teachers at College Mathematics; mercedes.orus@gmail.com

* Centre de Recherche en Épistémologie Appliquée (CREA), Under the direction of Dr. Jean Petitot

** Correspondence: cjouis45@gmail.com; mercedes.orus@gmail.com

Abstract

In this article, we will analyze how to obtain pertinent Information in the form of Qualitative Data graphically represented from unstructured Big Textual data. Unstructured data refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner (80-90% of all information). Obviously, it is not useful to accumulate large amounts of information if we cannot find a particular piece of information. The current methods prove to be expensive and the results are too often inappropriate. The goal of the research described here is to present an approach for automating the detection and the extraction of meaning from unstructured data using its normalized part: Web of data & Linked Open data (LOD). On the other hand, in structured indexes, classification systems, thesauri, conceptual structures or semantic networks, relationships are too often vague. One possible approach to this problem consists in organizing the relationships in a typology based on logical properties. For instance, we typically use only the general relation “Is-a” (too vague). We propose an original method: Contextual Exploration. This is implemented in the EC3 software. EC3 does not need syntactic analysis, statistical analysis nor a "general" ontology. EC3 uses only small ontologies called "linguistic ontologies" which depend on the knowledge of the language.

Keywords: Big Textual Data, Extraction of Pertinent Information, Data Mining for Information Retrieval, Big Data Analysis, Web Semantics, Computational linguistics, Ontologies, Text Mining, Web Mining, (Big, Linked, Smart) Data, Semantic relations, Contextual Exploration, Qualitative Research, Relationships, Hierarchies, Concept, Semantics, Logical Properties, EC3.

1. Introduction

Nowadays when technology is available to everyone, those of us who investigate how to build new ways to help and improve any professional sector, we do not cease in the search of those ones that can be the best solution.

The amount of information that exists today is of such magnitude that the most common is to find it unstructured. It is what we call unstructured Big Data. And it is precisely from unstructured Big Data, where arises the need to obtain consistent Qualitative Data providing the information that we need.

Many tools exist on the market. Not all of them low cost, and many requiring properly computer equipment (not low cost either).

In this article, we show our advanced Contextual Exploration technique, on which our EC3 software is based. And for this, we introduce the reader from the most basic, zstarting from the following subsections.

1.1. Towards an enriched Google?

It seems quite reasonable to assume that the reader of this article begins by reading the title, the abstract, and the keywords. Then, the reader has probably searched the bibliography to see if it contains enough references. In addition, the reader expects to read something typical for an article book. These include:

1. An introduction: this is what the reader is reading. An introduction is indispensable, because it must at least in particular announce what the authors will say, and a plan;
2. A state of the art, indispensable, which makes it possible to pinpoint the article;
3. The central part of the article, that is, what the authors bring back;
4. And finally, a conclusion with perspectives, all with a substantial bibliography.

The title is not easy to write, because this article is inserted into an organized set of other articles that talk about rather similar subjects, but according to different approaches, methods, objectives, and points of view.

The summary is also difficult to write, because it is the authors themselves who wrote the summary, after having written the body of the article. But, a priori, it is not easy to describe exactly how we authors build it.

The key words are also placed by the authors. But here again, it is not easy to know exactly how the authors found them. In addition, it is not specified how and for what purpose these will be used later. All these keywords must still respect at least some common-sense principles. Among these principles, we state below two of them:

1. Obviously, the keywords must be related to the book themes: it is not too difficult since it is enough to look at the presentation of the objectives of the book, and then to integrate terms in adequacy with the themes of the book;
2. At the same time, for the article to be read, the keywords must emerge from the many other articles in the article. In other words, they must allow this article to be discerning, to distinguish itself from others, because the reader does not know where to start, or which

article to read in depth, facing a significant number of articles. But there is a problem for the authors: they do not know in advance the other articles. So, there is a tactic: to see the content of the previous book texts. But this tactic has a major disadvantage, it always refers to... the PAST.

All these preliminaries are only some of the problems that the reader may have noticed. These problems are obviously not new, and they reveal the following question: How to find relevant information from a very important set of information organized?

Nowadays, this issue is becoming more and more important because of the computerization of information of all types in a more and more massive way, especially on the Web, or in large libraries. Increasingly, companies, organizations, and countries collect information, and make cards structured or organized. Mainly, they digitize the information. Very quickly, we realize that then, for this massive data to be used for something, we must give ourselves the means to quickly find accurate information, and only that one. The difficulties in finding relevant information are not only related to the size of the data, but also to other extremely important factors:

1. Information is dynamic, especially on the Web (databases constantly change content and localization, increase, others are outdated, disappear, etc.). As a result, at any given moment, the data that we consider only reflect, for a large part, the PAST;
2. The data are heterogeneous, both on the form (computer format, computer encoding, reading direction, etc.), and on the nature and style of the content (scientific, technical, literary, religious, philosophical, political, socio-cultural, etc.);
3. The data are distributed in different sites (WEB) or geolocated in libraries, with different classification systems of documents: it is therefore necessary to consider contexts (the author for example linked to the period associated with the well thinking majority of the moment), because the Information is linked to viewpoints;

4. The data are mainly textual and at the same time multilingual and multicultural: it is therefore necessary to add analytical tools which do not depend neither on a given language, nor on socio-cultural environment;

5. The data are of different sizes: relevant information can be found in an "SMS" (a SMS is not necessarily syntactically correct) or in a book of several hundred pages.

To answer these problems, very briefly, (since we develop in the first part a "state of the art"), the generally adopted approach is to index information / documents. As rightly pointed out in 2012 A. Das & J. Jain [1], in application to the World Wide Web: "(…) The last two decade have witnessed many significant attempts to make this knowledge discoverable. These attempts broadly categories:

1. Classification of web pages in hierarchical categories (directory structures), championed by the likes of Yahoo! And Open Directory Project;

2. Full-text index search engines such as Excite, Alta Vista, and Google".

We must position ourselves from the point of view of the user. The user "(...) extracts and tries to distinguish" essential "information:

1. By a superficial reading;

2. By a fast reading;

3. As part of an increasingly important "noise" of information (media: TV, radio, newspapers, internet, etc.);

4. Without necessarily having knowledge on the subject.

What most systems return too often is:

1. Too much information;

2. With poor ergonomic results, usually in the form of lists.

In addition, these systems have the disadvantage of being expensive in the sense that it is necessary to have large servers that store indexations of information / documents. "[2].

Also, in addition to these methods, and especially to the Google method, we propose an added value to Google's strategy: Contextual Exploration (CE), whose main fundamental principles have been proposed by Professor J.- P. Desclés, already more than twenty years ago. Let's mention them among others [3], that we will detail in this article.

Contextual exploration has been implemented and tested in a large number of software packages, the first of which, really tested and validated in an industrial context, implemented in 1993 [4] (SEEK). Many other CE systems were subsequently tested and validated, including (EXCOM-2) and [5], among others.

Also, the idea that we propose is complementary to the traditional indexing and is the continuation of works such as [6] and now EC3 software [7], [8] & [9] to apply the method of Contextual Exploration for Data Mining for Information Retrieval.

As announced, in the following, we present a state of the art, then we present in detail the principles of our approach with examples, and finally we conclude.

1.2. Big Data: a survey

No one can argue that the use and management of Big Data (BD) is one of the most important issues of the beginning of this century. The huge amount of data produced every day [10] from various sources is increasing considerably due to the massive expansion and use of communication tools everywhere in our daily life [14]. The necessity to explore and to find a universal methodology to deal with BD is crucial to both Academics and Industrials and can affect all the sphere of modern societies [13].

Finding and extracting relevant information from big contextual and multilingual data (BCMD) is very valuable for a lot of people from different background and activity sectors [11]. The literature presumes that BCMD is very important for a very large amount of organizations and very central for Facebook, Twitter, Google and other social media as well as the governmental entities, and communication companies. Collecting, storing and analyzing BCMD can be very efficient to improve decision making, competitiveness, innovation, and efficiencies in a very large scale of institutions. For Scientific Research, BCMD is an asset and challenge to improve and understand the quality of life.

A big number of software and methods to deal with this problem are in use but most of these methods are very expensive, inappropriate, and owned by big firms. Millions are spent every year to improve contextual data management to advance decision making and efficiency in little and big organizations. Academics can propose new low cost analytics, methods, and models in order to find relevant information in a big amount of contextual data. This work will advance our understanding of multilingual texts from different cultures and communities.

The main question here is "could we really find these low cost methodologies to extract relevant information from big multilingual contextual data (BCMD)? Could this methodology be effectively multilingual in regard to the different linguistic systems? How can we implement and integrate EC3 software and methodology to extract relevant information from BCMD? Undoubtedly these questions are vital to NLP Professionals and Academics and could lead to a new and meaningful methodology to enhance research and innovation in this field [12].

1.3. Data Mining for Information Retrieval using Contextual Exploration

Data Mining is understood as a process of automatically extracting meaningful, useful, previously unknown, and ultimately comprehensible information from large databases. Data mining (the analysis step of the Knowledge Discovery in Databases process), a relatively young and interdisciplinary field of computer science, is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. With recent technical advances in processing power, storage capacity, and interconnectivity of computer technology, data mining is seen as an increasingly important tool by modern business to transform unprecedented quantities of digital data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. The growing consensus that data mining can bring real value has led to an outburst in demand for novel data mining technologies.

In addition to traditional methods, we propose to add the contextual exploration method. In this section, we begin with a presentation of the contextual exploration in an intuitive way, and then we give a more formal description. Next, we show how contextual exploration is implemented as a computer system. Finally, we describe our EC3 system and give some examples of results.

1.4. An intuitive presentation of Contextual Exploration

We begin by giving a real example, which is not chosen at random, because it makes it possible to highlight the method: “Il ne m’est Paris que d’Elsa”, Louis Aragon (A reference in French poetry, taught in French high school), in French:

" Il était une fois la fin qui commence
Un Paris de remparts de tramways et de pigeons
Où à Saint-Michel descendant l'Arpajon
De faux moulins faisaient semblant de tourner sur la Butte

Il était une fois un Paris sans raison
Qui n'avait d'autre plan que celui des trams ?
Dispersés les rois déchus de leurs chars
Sur les bancs d'asphalte et le seuil des maisons

Les signaux clignotaient vainement par endroits
Et en pleine lune ou en plein midi

La chaussée ressemblait à une main mendiante
Le vent soufflait les chaises des terrasses

Plus personne ne s'asseyait dans les cafés
Il pourrait bien faire aussi bien beau que la pluie
Un ciel de zinc ainsi qu'un drapeau de lavoir
Immobile sur ce conte de fées..."

The least the reader can say is that this text is unusual. Indeed, the syntax is strange. Already, the title is difficult to translate. Then there is no punctuation, not really a "sentence". It's in the surrealist style. The first line is illogical "(...) commence fin". However, even for a reader without knowledge about Paris, ″thing″ come out: "mental images, ″messages", and many other things.

How to explain that this text, which was written by a well-known author, can still deliver information to the reader? Our hypothesis is as follows: the reader finds, only thanks to his knowledge of the language, linguistic markers that enable him to find information. Here are some markers that can extract information:

- "Il était une fois: this is an imaginary story, at the limit of a tale,
- (...) commence fin ": the notion of time is not considered: would it be a description?

″où (…)": this is a place, etc.

In the following, we present a state of the art, then we present in detail the principles of our approach with examples, and finally we conclude with perspectives.

2. Related Works

The interest of our work in unstructured data, and the lack of the semantically annotated data (required for data analysis), locates naturally our work in the context of big data (consisting mainly of unstructured data) and Linked Open data (LOD) (consisting mainly of semantically annotated data). In what follows, we will present some works and a comparison of latest tools in the field of semantic analysis of unstructured data in the context of the (Big, Linked, Smart) Data. These motivate our work and our approach.

2.1. Linked Open Data (LOD) context

The issues addressed in this context are of two types: those that exploit the LOD for the analysis of unstructured data (such as disambiguation) and those that enrich the LOD by connecting new data. In the latter case, the inclusion of a word in the LOD first requires its disambiguation beforehand, i.e. associate required relations and attributes to determine its real sense in order to link it to the right nodes. Many studies have been done in this area. A very interesting work included in LODifier: Generating Linked Data from Unstructured Text [15]. The LODifier approach combines several technologies to perform semantic analysis of texts and their integration into the LOD. It is based on the named entity recognition and disambiguation of words based on controlled vocabularies. Its purpose is to extract entities and relationships from text; convert them into RDF, and then link them to DBpedia or WordNet RDF. The process consists of the tokenization of a text to extract lexical units, then the named entity recognition, then the generation of URIs, using the dedicated module (wikified) and then "mapping" with DBpedia. The relationships between entities are detected by additional operations modules (C & C) and (Boxer) (Curran et al.) [16].

The first module uses the techniques of determination of lexical categories (POS) and grammatical groups (chunking) to provide a well annotated parse tree. The second uses the results of the first to produce the representation of parts of speech (Discourse Representation Structures, DRS) introduced by (Hans Kamp) in the Theory of Speech Performances (Discourse Representation Theory DRT) [17] to formally represent speech text.

To disambiguate words (Word Sense Disambiguation WSD), it operates a Lemmatization module to search for the lemmas of words and relationships obtained by Boxer for "mapping" with the relationships of the term if it exists in DBpedia (remember, relationships and properties of terms in DBpedia are obtained from the info-boxes of Wikipedia pages, used in this case for the disambiguation). If relations do not exist in DBpedia, we proceed in the same way with WordNet relations to determine the synced and do the "mapping" with RDF WordNet: LOD version of WordNet. In both cases, the resulting graph is transformed into RDF, integrated into the LOD via DBpedia and / or RDF WordNet.

A second work, also interesting in that category, is the one in Semantator (Cui Tao et al.): Semantic annotator for converting biomedical text to linked data [18], around the conversion of biomedical text data in the LOD [19] [20] [15]. Semantator is a solution based on two components: an application to automatically generate and supply semi-automatically an ontology from a corpus, i.e., in this case, a medical corpus; and a web service that can take as input a text and recognize there all the ontology elements. The user can intervene manually to refine the suggested elements. We obtain an annotated document following the RDF standard, allowing the linking of each element to an element of the LOD. The user interventions are used, in addition to enriching the local ontology, to define its own ontology for targeted and specific uses. There may be mentioned, also, other works directly using the LOD for semantically annotate documents (Delia Rusu et al.), in Automatically Annotating Text with Linked Open Data (2011) [19].

2.2. Big data context

Big Data is characterized by large volumes of varied data, generated and shared quickly: 3V (Volume, Velocity, and Variety). The variety concerns data formats: structured, semi-structured, but mostly unstructured which constitute the majority of data Big Data, about 80%. To cope with this large amount of varied data, in order to optimize their processing, it is necessary to implement intelligent and automated processes that can analyze and convert data from Big Data to their structured and semantically annotated format. New research in the domain of semantic analysis and exploitation of unstructured Big Data, are of two types: those that directly address the analysis of texts as raw materials and those based on a middle layer like NOSQL.

For NOSQL, systems and storage schemas are simplified as independent aggregates, easily distributable physically on clusters processing. This architecture is based only on the concept of key / value, where key is the identifier of the aggregate, and value can be a basic data, an unstructured document, or semi-structured documents such as XML, RDF, ... its main feature is to simplify the problem of managing schemas of relational data models, by moving the issue to the processing of unstructured data.

That is, instead of spending energy to define a conceptual / relational data model; the data is stored, as is, and then we proceed with the implementation of solutions for analyzing contents and deducing the concepts, relationships and meaning. The relations considered in this case are semantic enriched types that go beyond the classical relationship of relational models. The intelligence of NOSQL systems is in the processing of data level and not in terms of their structuring. The NOSQL technology began to emerge in the fields of storage and processing of data Big Data because, firstly, the text data cannot be supported by a fixed data model: each use can be associated with a specific data model. Moreover, the dynamic of textual data representation models cannot be supported by conventional relational models (Minelli 2013) [21]. Much work and achievements have been made in this domain. Include for example, the solutions, Bitable implementation by Google, Dynamo by Amazon, Cassandra by Facebook, etc. In all cases (NOSQL or not), all the solutions implementing and exploiting Big Data converge to support at least a common requirement: the se-mantic analysis of unstructured data.

Many works have also been undertaken in this area, there are for example those conducted by (aC Boury-Brisset) around the extraction of semantic data from Big Data [22]; the work of Dimitrov in Ontotext project: From Big Data to Smart Data [23]; or the work of (E. Khan) on semantics and Big Data Addressing Big Data Problems using Semantics and Natural Language Understanding [24] and the development of the tool called Semantic Engine using Brain-Like approach (SEBLA). This tool is able to paraphrase a text, translate it into another language, to answer questions on the text, and to deduce inferences about the text. As applications in the Big Data, this approach has been used in several areas: research and information extraction, question & answer, summary generation, converting data into information, knowledge and artificial intelligence (E. Khan) Processing Big Data with Natural Semantics and Natural Language Understanding using Brain-Like Approach [25].

- From our side, and to demonstrate the utility of the use of technologies related to the management of semantics in Big Data, we started a work (H. Fadili) "Towards a new approach of an automatic and contextual detection of meaning in text, Based on lexical-semantic relationships and the concept of the context" [26]. The aim is, firstly, to improve the detection of meaning in the text, then use the results to optimize three use cases processes: semantic indexing, linguistic research and extraction of information from textual data. These processes can help unclog the Big Data; allowing, for example, the reduce of the size of an index by dividing it by a large number, and to have a positive impact in terms of elimination of noise and silence, on linguistic search and information extraction.

- In the same direction and goal, the second author (C. Jouis) began in 1990 an approach similar to the previous one: the detection of semantic relationships and concepts/Named Entities in technical texts [4]: SEEK software, developed in an industrial context (CR2A/IBM) and within the University of Paris Sorbonne and the CAMS (Center for Analysis and Social Mathematics – UMR CNRS/EHESS), laboratory (LALIC team: http://lalic.paris-sorbonne.fr). SEEK is based on a general computational and linguistic model to analyze languages [27]: ACG (Applicative and Cognitive Grammar), first developed by the Professor Jean-Pierre Desclés, University Paris-Sorbonne, director of the LALIC team. Derived from this general model, the idea of the Contextual Exploration strategy (CE) [28] begins in 1983, thanks to Professor Jean-Pierre Desclés.

Initially, SEEK is a text analysis software for natural language modeling an area of expertise. SEEK was developed to help build the static descriptive level of an ontology. This level models the domain objects and relationships that these objects have with each other.

SEEK provides interactive linguistic assistance by searching in texts relations between domain objects (class hierarchy / subclass, attribute / value class instance, relation between an object and its parts identification, comparison, incompatibility, etc.). It produces a visual representation of conceptual graphs (neighboring to the model of Sowa [29]). Hyperlinks allow finding, for each identified relationship, the part of the text used in its construction. SEEK is based on the contextual exploration. The contextual exploration is a linguistic and computer model of understanding of linguistic units inserted into context.

The EC hypothesis can be formulated as follows: "Some independent knowledge of knowledge of the external world - the language skills - can be directly used to extract information from text (words, concepts, relationships, etc.), especially verbalizations an expert in a field skill".

In other words, we consider the automatic processing of texts for building semantic representations:

1. Defining the semantic value systems;
2. Building computer systems that seek these semantic values in texts**.**

These systems are in the form of systems knowledge base. They consist of declarative rules, the rules for expressing the observer (the linguist) a decision-making skill of a human reader who have to interpret a text. These rules are in the form: IF condition THEN conclusion. Rule conditions express the presence or absence of certain linguistic clues in the context. The findings help to gradually build semantic representations.

SEEK is therefore to search the records of textual clues that identify a particular relationship. SEEK is an autonomous semantic analysis module that requires no comprehensive dictionary of the language, or full parsing for work.

The tool is divided into three parts:

1. Publisher hypertext editing a report of expertise, hyperlinks;
2. Module contextual exploration: research relationships;
3. Display module: graph visualization / sub-graphs obtained and hyperlinks to return to parts of the text represented by the graphs.

After SEEK development and validation on real industrial projects in 1993, many of CE ameliorations and applications inside LALIC where conduced, as we briefly cited before. The latest ones are EXCOM-2 [30] [31] [32] [33] [34], SeekBio [35] (in cooperation with ACASA team) and BioExcom [36].

2.3. Smart data context

Smart data are interpreted data, unambiguous, usable and useful, outcome generally from the results of semantic analysis of texts. ″…Smart Data is data which has useful semantics associated with it...Wikipedia″.

The generation of Smart Data can be a good exploitation of big data. It allows, ultimately, in the process of analyzing unstructured data, the extraction of the interpretations as relevant information, such as the generation of indicators, summaries, etc. "…Smart Data means information that actually makes sense. It is the difference between seeing a long list of numbers referring to weekly sales vs. identifying the peaks and troughs in sales volume over time…Wikipedia".

2.4. Test and comparison of some popular platforms

This section summarizes some estimated relevant information, according to a study and tests of some tools and approaches in the field of semantic analysis of textual data. Thi11s enabled us to deduce two basic categories: those related to the first levels of basic linguistic processing, and those related to the following layers interested in meaning and semantics processing. We chose to focus on tools below, for the first category (NLP layer): GATE, Weka, AUTOMAP and OpenNLP; and two sub-categories of the second category related to the integration of semantic data in the context of Linked Data and Big Data.

2.5. Semantic analysis of textual data

Data Preparation phase (data filtering). This phase corresponds to the linguistic pre-processing of text data, necessary, to the upper layers including automatic semantic processing. The process consists generally in the analysis and the generation of a network of words and relations, called a morpho-syntactic network, meaningless, representing only the elements constituting the text. The generation of such network follows "in general" the following process:

- Segmentation: Tokenization splits a text into individual tokens and segmentation splits a text into sentences;
- Morphological analysis: Stemming reduces inflected words to their word root and Lemmatization assembles different inflected forms of a word;

- Parsing: Part-of-speech tagging (POS): determines the lexical category, grammatical categories annotation, grammatical groups analysis (chunking): determines the grammatical groups - nominal, verbal, etc.

Semantic analysis phase. This phase uses the results of the previous phase to implement sophisticated processing to infer semantics in the context. This completes the general process of text analysis:

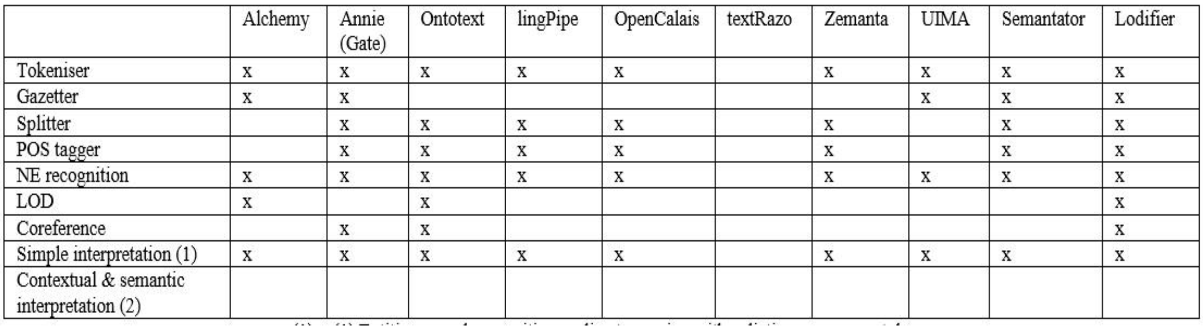
- Semantic Analysis: Capitalize on the previous phase and symbolic and / or mathematical statistics models for meaning analysis.

This phase will be described in detail through our approach in the section dedicated to our contribution. In general the results of analysis of the two phases are stored in structured formats like XML, CSV, etc. exploitable by third party applications.

2.6. Test and Comparison

These tests have allowed us to deduct a technical comparison of modules used, and the level of support in each application; from segmentations a text, to the semantic interpretation of each word and the text itself.

[1] Table 1.: Comparison chart: (1) Entities named recognition or direct mapping with a dictionary or ontology, (2) Contextual interpretation of the meaning augmented by semantic relationships



	Alchemy	Annie (Gate)	Ontotext	lingPipe	OpenCalais	textRazo	Zemanta	UIMA	Semantator	Lodifier
Tokeniser	x	x	x	x	x		x	x	x	x
Gazetter	x	x						x	x	x
Splitter		x	x	x	x		x		x	x
POS tagger		x	x	x	x		x		x	x
NE recognition	x	x	x	x	x		x	x	x	x
LOD	x		x							x
Coreference		x	x							x
Simple interpretation (1)	x	x	x	x	x		x	x	x	x
Contextual & semantic interpretation (2)										

These results show that the various solutions support almost NLP basic techniques in their early stages of analysis. This is in order to prepare texts to additional treatments on the semantic analysis or other specific processing. The concepts of words segmentation, sentences segmentation, POS tagging and named entity recognition are almost supported by all applications early in their process. But beyond that, each solution implements its own technology to improve the management of semantic analysis and its integration in various fields (indexing, extracting, and searching), in big data exploitation, enrichment and / or exploitation of Linked open data, etc. For the management of semantics, most of the tools support the named entity recognition; some go a little beyond, but without the full and effective management of semantics. The quality of services offered by these solutions depends on the quality of the semantic analysis of the proposed texts. In other words, more semantic management is better; the exploitation of the texts is better.

So there is a real need for development of new technology improving analysis of the semantics of texts for a better exploitation of unstructured data in the context of (Linked, Big, Smart) data.

2.7. Our contribution

We propose a new version of Contextual Exploration application in order to extract pertinent information: EC3, based on ACG (see below) for use in heterogeneous textual parts in WEB 3.0. EC3 aims to search for semantic values in texts. Completely rewritten in Python language, it prolongs and generalizes the all the previous CE applications. EC3 provides a more general environment for the extraction of information from texts by using an original method: contextual exploration. We process texts automatically in order to build semantic representations:

1. By defining semantic systems of values;
2. Then by building information processing systems that seek these semantic values in texts. These se antic systems are knowledge-based systems. They are composed of declarative rules, these rules that model for the observer (the linguist) a decisional expertise of a reader having to interpret a text. The conditions of the rules express the presence or the absence of certain linguistic indices in the context. The conclusions of the rules enable to build semantic representations.

3. Foundation of EC3: Contextual Exploration (CE)

Contextual exploration originates from a global model of language processing: Applicative and Cognitive Grammar (ACG) an extension of Shaumyan's Applicative Grammar, and more recently GRACE [37].

3.1. Linguistic Basis of CE

ACG articulates three levels of representations: (i) The level of the morpho-syntactic structures: each language is apprehended in the diversity of linguistic expressions, which are directly observable. With such a level, a morpho-syntactical analysis of a text builds a representation where all the morphological and syntactical information are present and narrowly dependent on characteristics of the language (order of words, presence or absence of cases for instance) ; (ii) The level of the applicative or predicative structures: it is about the syntactic-semantic structures of the language, represented by means of a descriptive metalanguage. Descriptions are presented in the form of applicative expressions (operators applied to operands from different types) and (iii) Cognitive level: with this level, the meanings of linguistic units may be analyzed under the form of layouts (semantic-cognitive representations) in order to constitute the knowledge representations associated to a given text.

Compared to a typical architecture of automatic treatment of the natural language, a system of contextual exploration fits simultaneously on several levels: at the morphological level, at the lexical level and at the semantic level, but does not use pragmatic information.

3.2. CE principles

General Contextual Exploration principle can be quickly defined as « A linguistic and Computational method which states that semantic information associated to textual segments can be identified by linguistic primary marks (called indicators) and a set of clues that would handle their polysemy and interpretation”, as J.-P. Desclés writes [3] (translated).

In detail, it means: [3], [4] “(…) The Contextual Exploration is based on two principles:

- A first principle is to exploit only the linguistic knowledge present in the texts, which implies that complementary markers and indices are general indices, totally independent of ontologies, conceptual classifications and semantic networks. This principle constitutes one of the cognitive hypotheses formulated by J.-P. Desclés and this can be illustrated by the following observation: a non-specialist reader of a domain is quite capable of identifying in a text certain organizational relation of the knowledge as well as the textual organizations put in place by the author.
- The second principle consists in asserting that the computer processing of natural language does not need beforehand very complex analysis and representations. In the traditional approach of automatic language processing the lexical and morphosyntactic processing steps, parsing and semantic analysis are chained together. Unlike this approach J.-P. Desclés proposes to first determine the semantic values that correspond to the task to be performed. Language work involves identifying and collecting explicit discourse markers that potentially express these values. In some works, the identification of semantic values has led to the production of a semantic map.

In general, the decision procedure of the CE relies on two types of components: (1) a set of linguistic markers; (2) a set of contextual exploration rules.

3.3. Linguistic Markers and CE Rules

The linguistic markers are subdivided into linguistic indicators and complementary indices. Language indicators are linguistic markers of Semantic values (grammatical or discursive) deemed relevant for task resolution. They play both a prominent branding role - it is from this mark that the context needs to be explored - and a role of linguistic knowledge - it is from this mark that the linguist organizes this knowledge. Complementary indices are linguistic indices that make it possible to confirm or reject the semantic hypotheses proposed by the presence of an indicator. The complementary indices are not necessarily lexical: by way of example, the place of the textual element which contains the indicator in the text or a property of this textual element such as its length can act like complementary indices to raise a semantic indeterminacy attached to an indicator.

The rules are indexed by the indicators; that is, the presence of an indicator triggers the application of a set of rules associated with it. The context specifies the textual space in which to look for the indices. This context is explicitly stated in the rules.″

From these components the decision procedure of the contextual exploration method for accomplishing a task t on a text T is as follows: (1) detect in T the indicators u1..., a relevant for the resolution of t; (2) for each indicator ui identified in the text examine the set of rules associated with the unit ui. The general form of a rule is: "if the indicator ui is identified in T and if one finds the presence of the indices Ip in the contexts Cp then take the decision Dj".

This is one example of CE rule:

```
RULE ing48
LET x1, x2, x3, x4 of linguistic units; P a proposition
IF x1 is an occurrence of the verb "to be" or one of the commas,
2point, or hyphen punctuation symbols
AND x2 is a past participle of one of the three LIN3, LIN4 or
LIN5 lists
AND x3 is a marker of the list (with / from / by / ...)
AND x4 is the lexical unit "of"
AND x1 x2 x3 x4 follow each other in the same
proposition P
THEN
create a link with the part-of relation in proposition P.
```

Where the markers concerned are:

LIN3 = {built, built, created, manufactured, done, given, provided, produced, ...};

LIN4 = {constituted, composed, formed, ...};

LIN5 = {derived, obtained, born; ...}.

This rule will trigger on the statement: "(...) Each system [Airbag] is composed of:

- An [airbag] and its gas generator mounted on the steering wheel for the driver and in the dashboard for the passenger;
- An [electronic box] (...)".

The following figure (see Fig. 1) is a visual representation of the result:

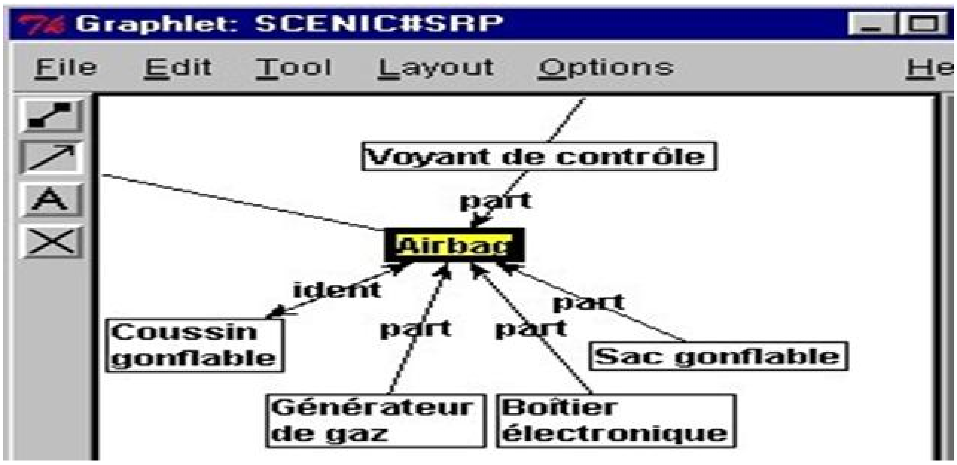


Figure 1. Visual representation of the result

Afterwards, when the user selects a node or a link of the graph, there is a hypertext link which refers to the part of texts that allowed the CE rules to be triggered. For details, refer to [7].

4. Application: EC3 Software

We will now zoom in on a principal aspect of our approach, i.e. automatic detection of meaning in unstructured data. Currently, we exploit mainly two important concepts: the concept of context and the concept of semantic relationships. The various components of this process will then be encapsulated in specific modules, by using, in addition, the machine learning technology, either to exploit LOD data for data analysis, or to use the Big data analysis results to feed and connect new data to LOD.

To do that, we propose a new version of Contextual Exploration application in order to extract pertinent information: EC3 software, based on ACG (see before) for use in heterogeneous textual parts, see [38].

EC3 is an advanced and tested project, which now has many applications in the field of intelligence and cyber security in Europe, and in the field of very large libraries (documentary information technology), but the most important EC3 application is an assistance against a tax escape.

EC3 aims to extract pertinent information from heterogeneous textual parts in Semantic Web or big Documentary Databases.

EC3 aims to search for semantic values in texts. Completely rewritten in Python language, it prolongs and generalizes the all the previous CE applications. EC3 provides a more general environment for the extraction of information from texts by using an original method: contextual exploration.

EC3 does not need syntactic analysis, statistical analysis nor a Big and General Ontology. However, EC3 only uses little ontologies called "Linguistic Ontologies", that express the linguistic knowledge of a user who must find and focus on pertinent information from a point of view. This is why EC3 runs very quickly on big corpora (many texts: from de "part of speech" such as SMS to books).

As output, EC3 proposes a Visual representation of information in the form of labeled graph of nodes (Named Entities) linked by semantic relations (depending of the point of view). The graph is represented to the user on a computer screen using an original approach: "Memory Islands".

A new application of the project with high potential is currently envisaged in the legal field. Lawyers (lawyers, magistrates, unions) would use it, to defend textual databases of jurisprudence. However, with the new French and European laws, it is urgently necessary to update the indexation of jurisprudence.

In short, with EC3, we process texts automatically in order to build semantic representations:

(1) By defining organized semantic values (linguistic ontology), and then

(2) By building information processing systems that seek these semantic values in texts.

These semantic systems are knowledge-based systems. They are composed of declarative rules. These rules model for the observer (the linguist) a decisional expertise of a reader having to interpret a text. The conditions of the rules express the presence or the absence of certain linguistic indices in the context. The conclusions of the rules enable to build semantic representations. The EC3 improvements are:

1. EC3 is not written in Java because the newer versions of Java are "uncertain". Instead, it is written in Python, which is faster.
2. EC3 is not built according to the architecture client / server: it operates directly on the target machine, and has been tested on Windows, Mac OS, Linux and Chrome OS.
3. The most important improvement concerns the linguistic marker lists. Indeed, some lists are not disjoint, which caused problems in previous versions of the Contextual Exploration.
4. Finally, we use graph viewers to visualize the results of very large sizes, in turn using efficient algorithms. Currently, we test GEPHI [39].

5. Conclusions and Perspectives

5.1. A logical system for semantic relationships

With a view to better designing the knowledge structures underlying the concepts of an organization, and more specifically, the indexing of documents and/or information retrieval, we use structured set of relationships, based on a linguistic model.

We indeed use an organization of terminologies founded on a semantic and logic model. This model proposes a semantic and logical concept organization using linguistic links: an enriched terminology [40], [41]. The semantic and logical organization of the terminologies is founded on a semantic model of language processing: Applicative and Cognitive Grammar (ACG, Desclés 90) [42], [43], and an extension of this model to Terminology and Information Retrieval [44], [45], [46].

The ACG postulates three levels of representation of languages especially the Cognitive Level. At this level, the meanings of lexical predicates are represented by semantic cognitive schemes. In this perspective we propose a set of semantic concepts, which defines an organized system of meanings.

Relations are a part of a specification network built on a general terminological relation schema (i.e. a coherent system of meanings of relations). The general schema of relation (″an entity X is in relation with an entity Y″) is further specified according to the algebraic properties in more precise relations that are axiomatically attributed to them. In this system, a relation may be specified in other more precise relations in terms of its properties:

1. Its functional type (the semantic type of arguments of the relation).
2. Its algebraic properties (reflexivity, symmetry, transitivity, etc.).
3. Its combinatorial relations with other entities in a same context (the part of the text where a concept is defined for instance).

Within the cognitive level, we distinguish four categories of primitives: elementary semantic types of entities, formation operators, fundamental static Relations between terminological units, fundamental dynamic relations between terminological units.

5.2. Elementary semantic types of entities

We distinguish several elementary types of entities. For instance:

1. Boolean entities (noted H) are objects whose value is either true or false;
2. Individualizable entities are entities that can be designated and shown by pointing. They may be counted individually or regrouped by distributive classes. Entities such as John, table, chair, man, child are distinctive. Individualizable entities are noted J: [J: table];
3. Distributive classes regroup individual entities with one identical property. They are noted D. For instance: [D: to-be-a-square].

Collective classes are distinguished from Distributive classes in that they represent objects that form a "whole" from more elementary objects. They are noted C. Thus, [C: geographical entities], [C: molecule] represent collective classes.

The "whole" is seen as the "accumulation" of elements that constitute it, disjoint or not. Lesniewsky (1886-1939) proposed a general theory of wholes and parts (mereology), in response to the problem of set theory (Cantor, 1932, 1962). A detailed analysis of mereology was carried out by Miéville (1984). Lesniewsky arrives at the conclusion that the notion of class contains two features: the distributive one and the collective one.

The following example, borrowed from Grize, 1973, [47] gives an idea of the difference: "A distributive class is, to be strictly correct, the extension of a concept. If p is the concept planet, the statement that Jupiter is a planet is either to pose p(Jupiter) or Jupiter ∈ {x / p(x)}, and the transmitted information is the same one in the two writing. Thus p = {Mercury, Venus, Earth, Mars, Jupiter, Saturn, Neptune, Pluto} is a distributive class. It contains nine elements and nothing else. The polar caps of Mars, the red Jupiter spot, the rings of Saturn do not belong to p. Yet, all that and a thousand other things deal with the concept planet. The notion of collective class must mitigate this gap" (see Fig. 2).

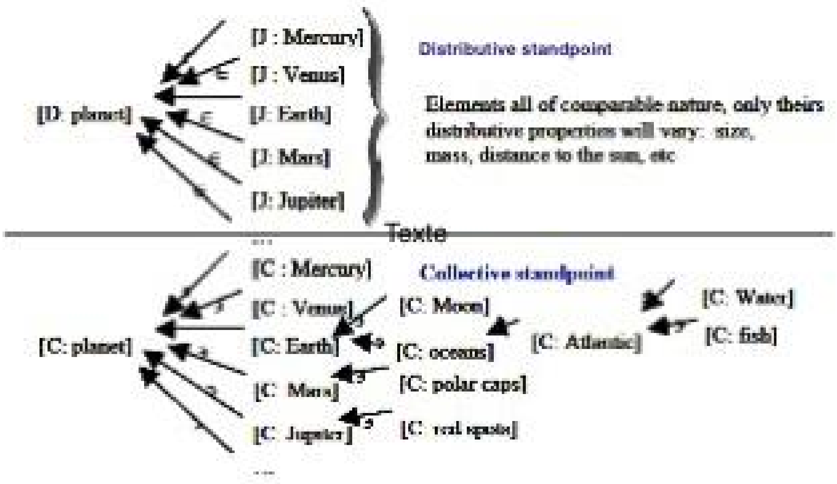


Figure 2. Distributive vs Collective (part-of) classes: different but logical and coherent standpoint.

Formation operators. Formation operators create more complex types from elementary types (lists, arrays, functional types, etc.). For instance, it is possible to define functional types.

From the set of elementary types S = {H, J, D, C, …} we define a system of more complicated types in a recursive way starting from the following rules:

1. The elements of S are elementary types;
2. If x and y are types, then Fxy is a functional type.

Then, an entity E of type Fxy (noted [Fxy: E]) is a unary operator which takes for its argument an object of the type x to provide a result of the type y. If we consider an entity A of type x, the application of E to A will build a certain entity B of the type y: ([Fxy: E] [x: A]) >[y: B].

For example, type FJH is that of an operator which, when applied to an individualizable entity (J) returns a value of truth H (set of individuals, or ″concept″ such as [FJH: ″to-be-a-square″]). A relation between an individual entity and a distributive class will have type FJFDH. Because this relation is a binary operator, the application is done in two steps.

For example, we have the following types: [J: Jean], [D: Human] and [FJFDH: inclusion]. The inclusion applies initially to Jean to return an operator of the type FDH: ([FJFDH: inclusion] [J: Jean]) > [FDH: inclusion_Jean].

The result is an operator of type FDH that applies to the distributive class Human to return a value of truth of type H: ([FDH: inclusion_Jean] [D: Human] > [H: True].

All representations of the cognitive level are typified in this manner.

Fundamental static Relations between terminological units. They are binary relations. Static relations permit the description of some states (static situations) related to an area of knowledge. We distinguish more than twenty relations, especially: identifications (or equivalence between two entities), incompatibility among entities, measures, cardinality, comparisons, inclusions (among distributive classes), belonging of one individualizable entity to another distributive class, localizations of one entity in one place (interiority, exteriority, boundary and closure of a location, boundary of a locality, orientations, etc.), relations part/whole among collective classes (direct, necessary, atomic, typical, unique, quantifiable, etc.).

Fundamental dynamic relations between terminological units. Dynamic relations permit descriptions of processes or events related to an area of knowledge (terminological unit denoting dynamical situations): movements, changes of state, conservation of a movement, iterations, intensity, variation, constraints, causes, etc. From the set of elementary types, it is possible to define a system of more complicated types with the help of formation operator of functional types.

Specification of "Is-a" relationships. In our system, a relation may be specified in more precise relations in terms of its properties. We assume that general relation "Is-a" is characterized by asymmetry. This asymmetry is specified in:

1. The belonging of one individualizable entity to a distributive class (noted ∈). Of type FJFDH, this relation is NEVER-reflexive, asymmetric and NEVER transitive. It is expressed in statement such as: π is a real;
2. Inclusion among distributive classes, noted ⊂, (e.g. Bacteria are microorganisms), which is of type FDFDH, is NEVER-reflexive, asymmetric and transitive. It should be noted that, in many thesauri or semantic network models, we typically use only the general relation "Is-a" without distinguishing belonging from inclusion. However, there is a fundamental difference, since the first is NEVER transitive while the second is transitive and allows inheritance of properties;
3. The relation part of (or "composition"), noted ∋, is reflexive but (generally) nontransitive. It is expressed in statement like "The hand forms part of the arm". Its type is FCFxH, where x is of type J or of type C. Part of is specified in several relations. Indeed, there are a great number of properties describing the relationships between the composing object and the total object (collective class), for example:

- Atomic composition versus non-atomic composition (The smallest component of a program is the bit versus A book breaks up into articles, which themselves break up into paragraphs);

- Atomic composition does not admit transitivity, but non-atomic composition authorizes it;
- Direct composition versus non direct composition (Opium appears among the primary component of Lamaline versus A molecule consist of neutrons, protons and electrons, which are part of atoms). An Object-Part OP is a direct component of the Object-Whole OW, if there is no object OP1 (different from OP) such that object OP is a component of object OP1 and object OP1 is a component of object OW Otherwise, OP is a non-direct component. (See Fig. 3). Non direct composition is transitive, while direct composition is nontransitive;

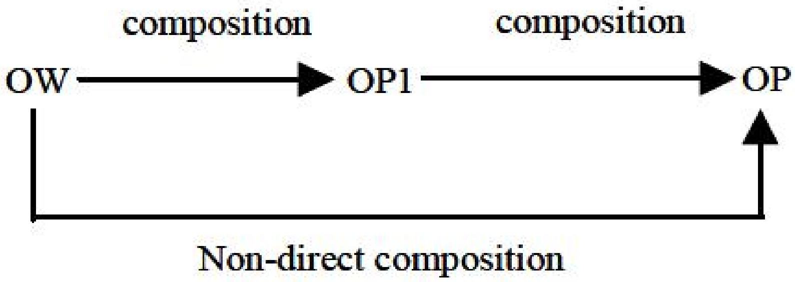


Figure 3. Direct vs. Non-directive composition

- Necessary composition versus no necessary composition (The processor is one of the essential components of a computer versus A CD-ROM drive is an accessory component of a computer). The characteristics necessary and non-necessary are transitive within the relation of composition;
- Single composition versus non single composition (A young star is made up exclusively of atoms of hydrogen versus the atmosphere is a mixture of several gases, whose principal one are oxygen and nitrogen);
- Quantifiable composition versus nonquantifiable composition (The hand is made up of five fingers; Each human cell contains 46 chromosomes versus Water consist of atoms of oxygen and atoms of hydrogen).

Localizations Relationships. The type of places (noted L) represents areas or groupings of positions of the same entity (individualizable, collective or distributive): Paris, garden, house.

We notice that the same linguistic unit can be assigned several types depending on the point of view. For example, Paris can be seen as a place (Jean is in Paris), as a collective class (Paris is composed of 20 arrondissements), as an individual entity (Paris is a capital).

Localization relations in relation to a place are expressed for example in the following statements: Paris is in France, A garden surrounds the house, the book is on the table, etc. An X (localized) is located relative to Y (the locator). The localization relations are of type FxFLH where x is of type J or of type L, according to the context of the localized one (each occurrence of an object, in a particular pragmatic environment, determines a place, or neighborhood in the topological sense).

Position primitives can be defined by using some rudimentary concepts of the general topology. A place L is then visualized either in its interiority, in its exteriority (excluding its interiority and its borders), or in its globality (limits and interiority). We introduce the topological determination operators of a place x: in (x), ex (x), fr (x) and fe (x), respectively determining the inside, the outside, the border, and the closure of x. The properties of these four operators allow us to specify the following four bit location relationships:

1. loc-in: "to be-in";
2. loc-ex: "to be out of";
3. loc-fr: "to be-at-the-border-of";
4. loc-fe: "be-at-the-closing-of".

Let's mention some properties of these relationships. The loc-in relation (example: The case contains the electronic board) is transitive, antisymmetric and non-reflective. The loc-ex relation (example: The cooling tank is outside the radioactive zone) is irreflexive. The relation loc-fr (example: the cell is delimited by its membrane) is incompatible (in the same context) with the outside and the inside and more precise than the closure. The relation loc-fe (example: Jean is in Paris) is incompatible (in the same context) with the outside and redundant with the border and the interior.

5.3. Contextual Exploration, an uncontainable way complementary classical method?

To conclude, we hypothesize that the most published and diffuse work for over a thousand years cannot be indexed by conventional documentary methods. It seems that our hypothesis is correct, in at least one ancient text: the Koran. According to several collaborators familiar with classical Arabic, the keyword "God" would not be a good candidate to index this book, among a large set of books, whose central themes are not related to "God", such as in large libraries:

1. The Koran is divided into 114 suras (or articles). Most of them are titled "(...) Merciful God". But, within the very text of the sura, the term "God" would be implicit;
2. One would write, wrongly, that in the Koran there would be 99 names of "God". In fact, in Arabic, there is not as in English or French the verb "to be": We must therefore speak of "qualifiers" and it is the context (that is to say the words on the left and to the right of a term that indicate the qualifiers);
3. It is only in Sura 59/114 that we are talking about the 99 qualifiers of "God".

So, the keyword "God" would not be a good candidate to index the Koran, using classical indexation?

However, for Christians, there is the Bible, which is divided into 2 parts:

1. The Old Testament (which partly recapitulates the Jewish Bible, but not all and not in the same order, and after translated in old Greek);
2. The New Testament constituted essentially by the Gospels (but not only).

In the Old Testament, in its French version ("The Bible of Jerusalem"), the term "God" appears under several denominations: "Yahve", "Jehovah", "Elohim", etc. In the New Testament, there is the notion of Trinity, which corresponds to three points of view on the same entity, God, namely, to go quickly and horribly simplify:

1. The Father (God-Creator);
2. Jesus (the son of God, who is God made man);
3. The Holy Spirit (God-always-present).

Also, to know from what point of view God is spoken of in the New Testament, we must refer to the context, that is to say again words that are left and right, and what is more, of the whole paragraph (verse) sometimes. Is the keyword "God" a good candidate for indexing the Christian Bible?

But more generally, can we say that the keyword "God" is insertable in a general classification, an ontology, an index? Is "God" a universal concept? Is it therefore necessary to consider contexts (the author for example linked to the period associated with most of the moment, a socio-cultural environment, etc.), because a concept is linked to viewpoints? Finally, is that not what we said at the beginning of this article: Contextual Exploration? equation never contains an equation number to its right, and this unique property distinguishes it from a numbered equation.

References

[1] A. Das & A., Indexing the World Wide Web: The Journey So Far, In Next Generation Search Engine, Advanced Models for Information Retrieval, pp. 1-28, C. Jouis, I. Biskri, J.-G. Ganascia, M. Roux (Eds): IGI Global, PA, USA (2012).

[2] C. Jouis, Contextual Exploration (EC3): A strategy for the detection, extraction and visualization of target data, 4th International Conference on Big Data Analysis and Data Mining, conferenceseries.com, September 07-08, 2017, Paris, France, DOI: 10.4172/2324-9307-C1-014, Paris, France (2017).

[3] Desclés, J.-P : Système d’Exploration Contextuelle, In C. Guimier (Ed.), In Cotexte et calcul du sens (pp. 215-232), Caen, France, Presses Universitaires de Caen, France (1997).

[4] Jouis, C. : Contributions à la conceptualisation et à la Modélisation des connaissances à partir d'une analyse linguistique de textes : réalisation d'un prototype : le système SEEK, PhD. Thesis, Paris, Under the direction of J.P. Desclés, EHESS & Centre d'Analyse et de Mathématiques sociales (Paris), en convention CIFRE: EDIAT/CR2A/IBM.

[5] Alrahabi, M.: Plateforme d'annotation automatique de catégories sémantiques: conception, modélisation et réalisation informatique: applications à la catégorisation des citations en arabe et en français, 2010, Under the direction of Jean-Pierre Descle΄s, Paris, Université Paris-Sorbonne -Paris IV, France (1993).

[6] Makkaoui O.: PhD. Thesis, Construction de fiches de synthèse par annotation sémantique automatique des publications scientifiques: application aux articles en biologie, Under the direction of Jean-Pierre Descle΄s et Christophe Jouis, Paris, Université Paris-Sorbonne -Paris IV, France (2014).

[7] Djioua B., Descle΄s, J.-P., Alrahabi, M., Searching and Mining with Semantic Categories, pp. 115-137, In Next Generation Search Engine, Advanced Models for Information Retrieval, pp. 1-28, C. Jouis, I. Biskri, J.-G. Ganascia, M. Roux (Eds): IGI Global, PA, USA (2012).

[8] Fadili, H., Jouis, C.: towards an automatic analyze and standardization of unstructured data in the context of big and linked data, Proceedings of the 8th ACM International Conference on Management of Digital Ecosystems, November 2016, Henday, France (2016).

[9] Fadili, H., Jouis, C. : Exploration Contextuelle (EC3) : une stratégie de détection, d'extraction et de visualisation des données cibles, Séminaire TIM 2017 (DGA, Ecole Militaire), Paris, France (2017/07/5).

[10] Jouis C., EC3 project, on Researchgate, https://www.researchgate.net/project/EC3-project.

[11] Agrawal, R., & Dhar, V. Editorial – Big data, Data science, and analytics: the opportunity and challenge is research. Information System Research, 25(3), 443-448, (2014).

[12] Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., and Zhou, X. Big data challenge: a data management perspective. Frontiers of computer Science, 275, 314-347, (2013).

[13] Gandomi, A., & Haider, M. Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management, 35(2), 137-144, (2015).

[14] VanDijck, J. Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. Surveillance & Society, 12(2), 197-208, (2014).

[15] Liu, Y & Jin, H. Building a network highway for big data: architecture and challenge. IEEE Network, 28(4), 5-13.FNM Surname (2018). Article Title. Journal Title, 10(3), 1–10, (2014).

[16] Augenstein, I.: Lodifier: Generating Linked Data from Unstructured Tex". ESWC (2012).

[17] Curran J. R., Clark S., and Bos J.: Linguistically Motivated Large-Scale NLP with C&C and Boxer. Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo), pp.33-36, (2007).

[18] Kamp H.: A Theory of Truth and Semantic Representation. In P. Portner & B. H. Partee (eds.), Formal Semantics - the Essential Readings. Blackwell. 189-222, (1981).

[19] Tao C., Song. Sharma, & Chute, G., Semantator: Semantic annotator for converting biomedical text to linked data. Journal of Biomedical Informatics, Volume 46, Issue 5, Pages 882-893. DOI: 10.1016/j.jbi.2013.07.003, (2013)

[20] Rusu D., Fortuna B., M., Dunja: Automatically Annotating Text with Linked Open Data. Venue: In 4th Linked Data on the Web Workshop (LDOW 2011), 20th World Wide Web Conference, (2011)

[21] Gupta A., Viswanathan K., Joshi, Finin, T. & Kumaraguru, P.: Integrating Linked Open Data with Unstructured Text for Intelligence Gathering Tasks. Proceedings of the Eighth International Workshop on Information Integration on the Web, 28/03/2011.

[22] Chan, J. O. "An Architecture for Big Data Analytics.", Communications of the IIMA 13.2: 1-13. ProQuest Central. Web. 6 May 2014, (2013).

[23] Boury-Brisset, A.-C. Managing Semantic Big Data for Intelligence., in Kathryn Blackmond Laskey; Ian Emmons & Paulo Cesar G. da Costa, ed., 'STIDS', CEUR-WS.org, pp. 41-47, (2013).

[24] Dimitrov, M.: From Big Data to Smart Data. Semantic Days, (2013).

[25] Khan, E.: "Addressing Big Data Problems using Semantics and Natural Language Under-standing," 12th International Conference on Telecommunications and Informatics (Tele-Info '13), Baltimore, September 17-19, (2013).

[26] Khan, E.: "Processing Big Data with Natural Semantics and Natural Language Understanding using Brain-Like Approach", (2014).

[27] Fadili., H.: Towards a new approach of an automatic and contextual detection of meaning in text, Based on lexicosemantic relations and the concept of the context., IEEE-AICCSA, (2013).

[28] Jouis, C. "Contextual Approach: SEEK, a linguistic and computational tool for use in knowledge acquisition", in Proceeding of the First European Conference "Cognitive Science in Industry", 28th -30th September 1994, Luxembourg, pp. 259-274, Luxembourg (1994)

[29] Desclés, J.-P. Contextual exploration processing for discourse and automatic annotations of texts. In FLAIRS Conference, 281–284, Florida, USA (2006).

[30] Sowa, J.-F., Conceptual structures, Information Processing in mind and machine, Addison-Wesley, (1984).

[31] Alrahabi, M. EXCOM-2 : plate-forme d'annotation automatique de catégories sémantiques : Applications à la catégorisation des citations en français et en arabe. PhD. Dissertation, Université Paris-Sorbonne, France (2010).

[32] Atanassova, I. 2012. Exploitation informatique des annotations sémantiques automatiques d'Excom pour la recherche d'informations et la navigation. PhD. Dissertation, Université Paris-Sorbonne, France (2012).

[33] Bertin, M. Biblio sémantique : une technique linguistique et informatique par exploration contextuelle. PhD. Dissertation, Université Paris-Sorbonne, France (2011).

[34] Djioua, B. ; Flores, J. J. G. ; Blais, A. ; Desclés, J.-P. ; Guibert, G. ; Jackiewicz, A. ; Le Priol, F. ; Nait-Baha, L. ; and Sauzay, B. Excom: An automatic annotation engine for semantic information. In FLAIRS Conference, 285–290, (2006).

[35] Desclés, J.; Alrahabi, M.; and Desclés, J.-P. BioExcom: Detection and categorization of speculative sentences in biomedical literature. In Human Language Technology. Challenges for Computer Science and Linguistics. Springer. 478–489, (2011).

[36] Makkaoui O. : Construction de fiches de synthèse par annotation sémantique automatique des publications scientifiques : Application aux articles en biologie, PhD. Dissertation, Université Paris-Sorbonne, France (2014).

[37] Descle ́s, J.P. & Faiz R. Méthode automatique d'annotations sémantiques et indexation de documents textuels pour l'extraction d'objets pédagogiques. Boutheina Ben Ali, France (2014).

[38] Descles J.-P. : Langages applicatifs, langues naturelles et cognition, Hermès, Paris, France (1990).

[39] Jouis C. & Shafei B.: Big textual data: how to find relevant information (with low cost)? Invited Paper. In Proceedings of the 10th International Conference on Management of Emergent Digital Ecosystems (MEDES'18). ACM, Tokyo, Japan (2018).

[40] Heymann S., GEPHI, Encyclopedia of Social Network Analysis and Mining, pp.612-625, (2014).

[41] Jouis, C, (1995). «SEEK, un logiciel d'acquisition des connaissances utilisant un savoir linguistique sans employer de connaissances sur le monde externe». In Actes des 6ème Journées Acquisition, Validation, (JAVA 95), INRIA, pp. 159--172, Grenoble, France (1995).

[42] Mustafa, W. & Jouis, C. "Terminology Extraction and acquisition from textual data: criteria for evaluating tools and methods" In Proceedings of the First International Conference on Language Resources and Evaluation, Granada (Spain): 28- 30 May 1998, organized by ELRA (European Language Resources Association). Granada : ELRA, Vol. 2, pp. 1175-1180, Spain (1998).

[43] Descles, J.-P., Langages applicatifs, langues naturelles et cognition, Hermès, (1990).

[44] Descles, J.-P. & Guibert, G. La fonction première du langage, Champion, Paris, France (2011).

[45] Mustafa, W., Jouis C. "Natural Language Processing-based Techniques and their Use in Data Modelling and Information Retrieval", In Proceedings of 6th International Study Conference on Classification Research, Knowledge Organization for Information

Retrieval, 16-19 June 1997, University College of London, London, FID/CR, & ISKO. The Hague: FID, pp. 157-161, UK (1997).

[46] Mustafa, W. & Jouis, C. "Natural Language Processing-based Systems for Terminological Construction and their Contribution to Information Retrieval", in Proceedings of the Fourth International Congress on Terminology and Knowledge Engineering (TKE'96), Vienna, Austria, INDEX Verlag, Frankfurt/Main. 118- 130, Austria (1996).

[47] Jouis, C. & Ferru, J.-M. (2004). « Intranet Try To Find Project (ITTF): An approach for the searching of relevant information inside an organization », LREC 2004: Language Resources and Technology Evaluation within Human Language Technologies, pp. 1325-1329, ELRA – European Language Resources Association, Lisbon, Portugal (2004).

[48] Grize, J.-B. Logique Moderne. (Fascicule II). Paris : Mouton/Gauthier-Villars, France (1973).