

Application to refine a text corpus resulting from a web pages monitoring agent - Towards a modular software support of the monitoring process

Allan ZIMMERMANN, Xavier DELECROIX, Cyrille DUBOIS, Serge QUAZZOTTI

allan.zimmermann@tudor.lu, xavier.delecroix@tudor.lu, cyrille.dubois@tudor.lu,
serge.quazzotti@tudor.lu

[CRP Henri Tudor](#), 29 avenue John F. Kennedy L-1855 Luxembourg-Kirchberg
(Luxembourg)

Keywords:

Internet monitoring, monitoring agent, terminology, relevance, modular software system, lexical coloring, speed reading

Abstract

The Internet has become a major source of information for enterprises. It allows them access to a large number of information sources regarding their competencies, management or reputation. However, the professional use of information resulting from the Web can quickly become disappointing or counter-productive, as the various research means generate large volumes of information.

Even automated monitoring tools, usually used by professionals, do not include filtering means sufficiently adapted to precisely target the relevant information from the monitored pages. That is why we have developed a tool to refine the corpus resulting from a monitoring agent work.

This tool is based on the use of a precise terminology describing the subject, object of the surveillance, allowing a double selection, both automatic and human, of the newly detected information.

The developed tool is part of a general approach consisting of building modular software systems, based on existing interchangeable technical solutions, to develop monitoring products or services at an acceptable cost.

1. Introduction

The Technology Watch Center (Centre de Veille Technologique et Normative -CVTN-) of the Public Research Centre Henri Tudor supports innovation by integrating competitive and technological intelligence, industrial property, standards and regulatory information into products, services and specific trainings to companies.

Among its products, the CVTN offers the monitoring of Internet sources of information (Internet monitoring).

Due to the staggering amount of information it contains, Internet has become today's de facto source for information. Companies' web pages, weblogs, and news disseminated on the Internet can be of great interest for any company, which seeks to search for trends, or to identify the rumors related to subjects that are sensitive relative to the development of its business.

Because of the increasing amount of information sources on the Internet, the problems related to the capturing current and relevant information from the Internet are becoming more and more complex.

The monitoring of Internet pages is realized with the help of monitoring agents that detect the updates from a set of selected web pages.

However, the means of filtering used by the monitoring agents available at the CVTN are not well adapted to work with all the keywords used in multilingual terminology. Moreover, despite the keywords filtering, the detected updates still contain a high proportion of noise, making it essential to execute a complementary filtering, carried out by a human agent. Without dedicated tools, this level of complementary filtering, called "refining" in this paper, can be time consuming and expensive. This is why an experimentation and research work has been conducted at the CVTN in order to develop a tool to quickly and simply select relevant information gained from the surveillance process. Naturally this will also have the benefit of lowering the cost of Internet monitoring.

This article aims to present the work undertaken to realize a first version of a refining tool based on a lexical coloring process, and to show how this work is in line with the modular software approach developed for the Internet surveillance process. On one hand, the CVTN aims to reduce software constraints (Perbal, Dubois, & Schosseler, 2004) and, on the other hand to develop monitoring systems based on reusable elements proven to be reliable in the production phase.

Thus, we will first present the limits of the filtering means included in the monitoring agents, which implies the necessity of building a refining module. Then, we will expose the terminology structure to describe a subject to monitor, as well as the functioning of the refining module, which is based on this terminology. The software system which includes the monitoring agent with the refining module, will be briefly described. Then, the conclusion on the use of the agent in the techno-legal field will be drawn. Finally, the improvement perspectives will be discussed, particularly on carrying out analysis and visual representations on the relevant corpus.

2. Limits of the filtering means included in the monitoring agents

The main objective of a monitoring agent is to detect the updates in a set of web pages (or possibly in a set of documents such as pdf files for example). It also enables the creation of alarms (sound, email) when updates are detected on a targeted set of monitored web pages.

The updates detected by a monitoring agent are generally composed of hyperlinks and/or text. We will qualify as "linked document" any document one may reach by following the hyperlink generated by the update (see figure 1).

The objective of the Internet surveillance process we describe in this article is the creation of weekly alert bulletins, as well as the diffusion of monthly webographic reports on targeted subjects. An alert bulletin consists of a set of detected and relevant updates coming out from a set of monitored web pages. A webographic report is composed of a list of the webographic references of the relevant documents linked to the updates contained in the weekly bulletins.

The monitoring agents available at the CVTN make it possible to detect keywords within the monitored web pages, by associating lists of keywords and/or regular expressions to one or more monitored web pages. This mechanism enables us to filter updates on the basis of keywords, i.e. to automatically put aside the updates which do not contain any keyword. Then, these monitoring agents allow us to record the keywords which describe a surveillance subject and to keep only those updates likely to include relevant information.

Furthermore, the monitoring agents include a browser in which one can consult an updated version of the web pages whereby the new elements and the keywords are highlighted. This lexical coloring treatment of the keywords enables the user to locate possible relevant information more quickly in the updated web pages .

However, the gathering of keywords in the form of a list is not well adapted to manage a structured and relatively important terminology of keywords and/or regular expressions. For example, the keywords cannot be gathered by categories or connected explicitly to their various translations. Nevertheless, a relatively precise terminology is essential for guaranteeing a certain exhaustiveness of the information detected within the surveillance process, and for the best possible creation of lexical coloring in order to visualize information likely to be relevant as quickly as possible.

Moreover, keywords filters used within the monitoring agents do not prevent the updates from containing a high level of noise. For example, updates frequently comprise a list of documents extracts, from which only one part contains keywords (see figure 1). Sometimes, the majority of the excerpts are without keywords.

To achieve the monitoring goals mentioned previously at an acceptable cost, it is essential to develop a means to precisely target the relevant information comprised within the updates. This means, based on a terminology describing the surveillance subject, should also enable the removal of the noise contained in the updates, and the extraction of the webographic references from the relevant documents.

In the next paragraph, the construction principles of a descriptive terminology will be detailed.

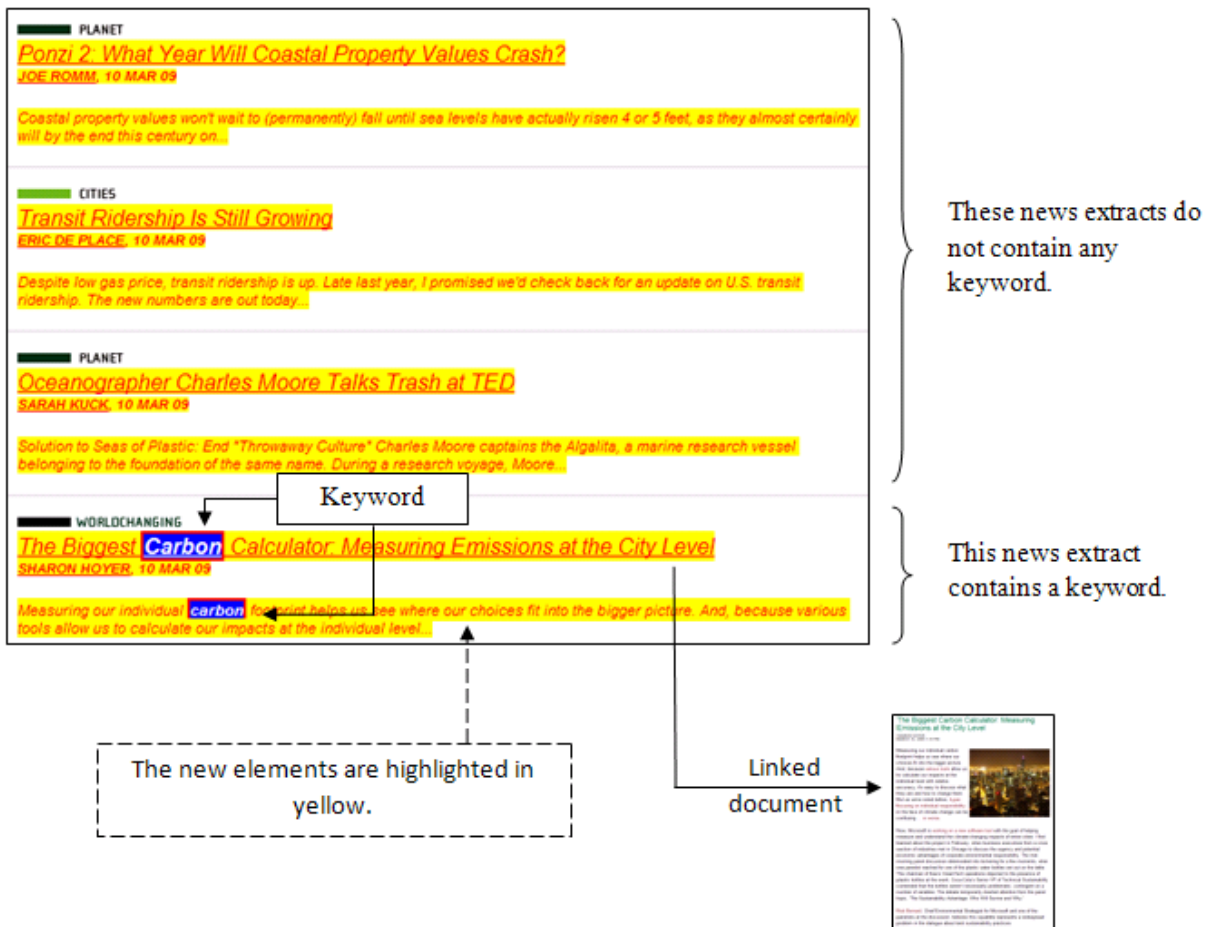


Figure 1: Extract of an updated webpage

3. Construction principles of a 3 levels terminology

Starting from simple and tested tools, like a database management system and a mechanism of regular expressions, a multilingual terminology, precise as well as simple to realize, has been built. To achieve this double objective of precision and simplicity, a model of terminology was designed with 3 levels: themes, keywords in the language of reference and regular expressions (see table 1).

3.1 Level 1: themes

Using sets of themes makes it possible to structure the subject of surveillance by gathering the terms considered to be equivalent, i.e. the terms which are close one to another (example: synonyms) or the terms which refer, directly or indirectly, to the same subject (example: the name of a working group on a standard and the designation of the standard itself).

There are two different sets of themes:

- Research themes.
- Secondary themes.

A research theme regroups the keywords (research keywords) which describe the main elements of the monitored subject. It enables the competitive intelligence analyst (CI analyst) to have a synthetic vision of

a monitoring topic. Such a vision is particularly useful when the CI analyst works on different monitoring subjects that he or she may not be familiar with. It also facilitates the transfer of the work to another collaborator. Moreover, these themes are treated with a lexical coloring allowing an identification of the monitored subject in a text, or only a part of it.

A secondary theme compiles keywords (secondary keywords) describing elements connected to the monitored subject or presenting a special type of information (example: it is possible to define a secondary theme "Event" to indicate the events, fairs, exhibitions...). The secondary theme aims to facilitate the updated reading via a treatment of lexical coloring. It may lead to the detection of new keywords. For example, within the framework of the techno-legal surveillance carried out previously, the theme "Event" can be used to identify, in the updates, the names of the events likely to provide documents related to the monitoring topic. These names can then be integrated into the terminology as research keywords.

3.2 Level 2: keywords in the language of reference

The language of reference is the main language used for the surveillance process (e.g. the mother tongue of the CI analyst or the working language of the customer); the others languages used are "only" translations. A keyword in the language of reference is a special grammatical form (e.g. a substantive) written in this chosen language.

The set of keywords comprises all the words the CI analyst has deduced from the customer request, culture and existing specialized terminology in order to describe the monitoring subject. The surveillance subject (and possibly the secondary themes) is described as precisely as possible, without taking into account, at this level, their translations or grammatical alternatives. However, it is at this level that the CI analyst should find the largest number of equivalent terms possible, to describe the monitoring subject in order to guarantee a certain level of exhaustiveness.

3.3 Level 3: regular expressions

In the language of reference, and possibly in other target languages, at least one regular expression is used to recognize one keyword or its radical. It is possible to use only one regular expression to define at the same time the verb, the adjective, the participles, plural and female in French or in English. The possible spelling mistakes can also be modeled using the regular expressions. Moreover, knowing the many Latin roots common to French and English and the many Germanic roots common to German and English, it is frequently possible to use a regular expression to indicate one or more grammatical alternatives of a keyword in at least 2 of these languages, sometimes even having 3 at the same time.

The technical means to carry out the keywords grammatical alternatives recognition (principle of lemmatization) in one of the target languages is composed by the regular expressions.

The keywords recognized by regular expressions associated with research subjects will be defined as research keywords. The same definition can be applied for secondary keywords and secondary subjects.

Table 1: Extract of a terminology

Set of subjects	Keywords	Regular expressions	Languages
Ecology	Ecology	[éeö][ck]olog	FR, EN, DE
	Sustainable development	développement durable	FR
		sustainable development	EN
		nachhaltige Entwicklung	DE
Fuel	Fuel	Fuel	EN
		carburant	FR
	Gasoline	essence	FR
		gasol[ei]n	EN
		benzin	DE

4. The refining module

The refining module (see figure 2) aims to identify relevant documents starting from the updates produced by a monitoring agent.

This module is composed of 4 phases:

- Phase 1: automatic selection of the updates produced by a monitoring agent.
- Phase 2: visual selection of the relevant information contained within the updates selected in phase 1.
- Phase 3: identification of the relevant documents linked to the updates selected in phase 2.
- Phase 4: processing of the updates without research keywords.

4.1 Phase 1: Automatic selection of the updates

When an update includes at least one research keyword recognized within a regular expression defined in the terminology, it is selected. This selection is divided into two sub corpuses: C1 containing updates with research keywords, C2 comprising updates without research keywords. As the common monitoring agents do not allow an easy management of a multi lingual and important terminology, this automatic selection is essential due to the choice of not using keyword filters or regular expression included within the usual monitoring agents.

4.2 Phase 2: Visual selection of relevant information

The updates included in C1 are then treated for a lexical coloring, as performed by the monitoring agents commonly used at the CVTN. A single code color is associated to each theme and allows for highlighting the related keywords, via regular expressions. It is thus feasible for an operator not only to quickly detect information which includes keywords within the updates, but also to quickly evaluate if this information is relevant, as it is possible to distinguish the various themes from the surveillance among the keywords highlighted (see figure 3).

During this phase, updates are also submitted to a double reduction process, both automatic and manual, in order to combine them into an alert bulletin. This reduction treatment consists of removing from an update all information without a keyword, and actually without relationship within the surveillance subject.

4.3 Phase 3: Identification of the relevant documents

From the set of information selected in phase 2, it is possible to access to the relevant documents by hyperlink using the Internet browser integrated within the refining module. The lexical coloring treatment may be integrated on the web pages containing the linked documents in order to facilitate the evaluation

of their relevance. The documents which are not web pages (PDF, DOC, RTF, XLS...) are not concerned with this processing of lexical coloring.

The documents considered to be relevant are then identified and documented by the way of their webographic references. These references are the metadata set of the documents: {source name, gathering date, document title, electronic format of the document, document type}. The objective of this simple and concise metadata set is to enable the customer to quickly evaluate the relative importance of a document compared to another.

The identified documents' metadata are partially and automatically edited using the electronic documents metadata and/or their URL. The title of a document included in a webpage can, for example, be taken from the <TITLE> tag of the HTML code of the webpage. However, manual editing is often necessary to correct the mistakes or the failures of the automatic editing.

4.4 Phase 4: Processing of the updates without research keyword

The updates contained within the complementary corpus C2 are gathered into a single document and are quickly reviewed, thanks to the lexical coloring, to identify possible new research keywords. In this case, only the secondary keywords are highlighted in the updates of C2. When a new keyword is detected, a regular expression is added to the terminology to ensure its future detection. When the reading of the complementary corpus is completed, the refining process is carried out another time to take into account the updates containing the newly detected keywords.

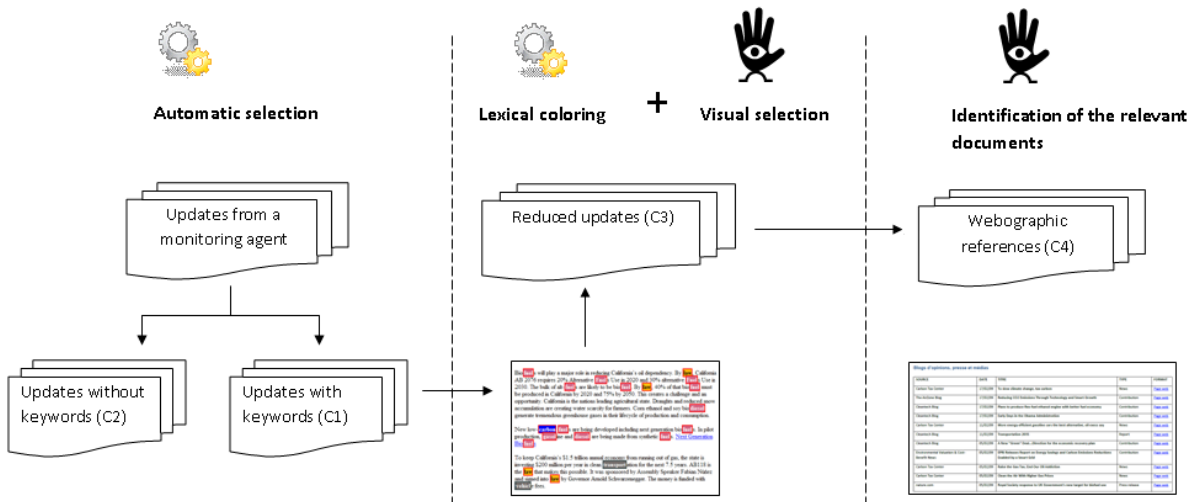
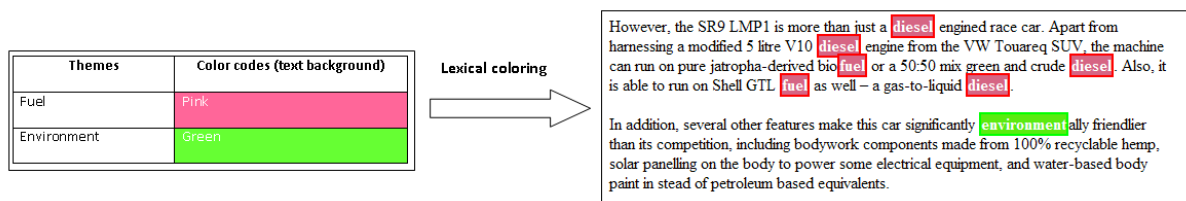


Figure 2: Extract of the refining module



Source : www.greencars.za.net

Figure 3: Example of lexical coloring

5. A specific software system

A software system to support a surveillance process in 2 phases which produces 3 types of deliverable has been set up (see figure 4):

- A set up report which comprises in particular a map of the supervised sources and the used terminology.
- A weekly alarm bulletin.
- A monthly report of the webographic references.

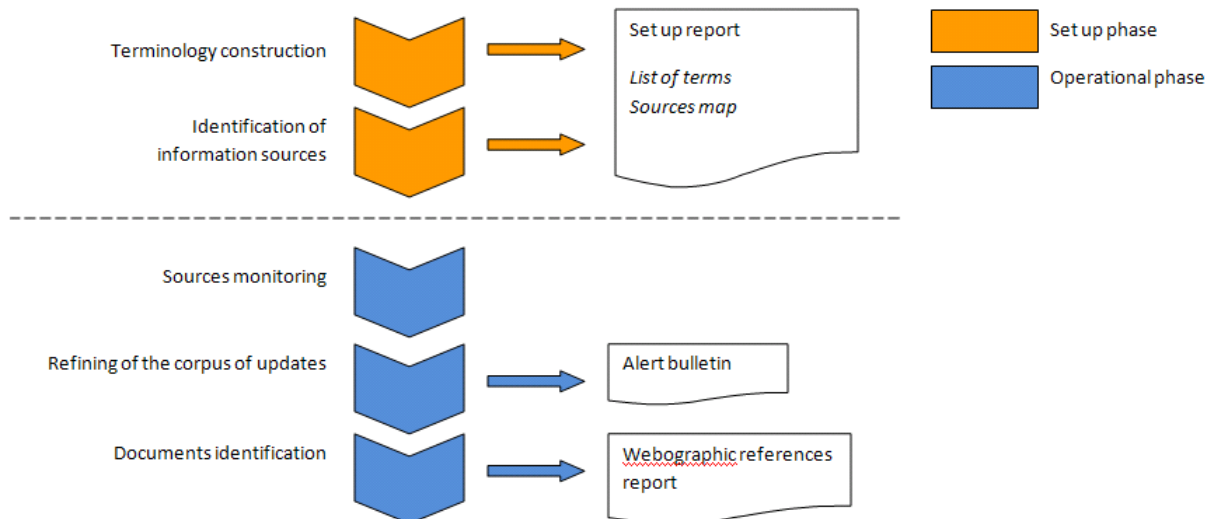


Figure 4: Surveillance process and associated deliverables

This system uses 4 different types of software (see figure 5):

- A monitoring agent.
- A conceptual mapping tool.
- A database management system (DBMS).
- A word processing tool used to build a bulletin of webographic references.

As previously mentioned, the monitoring agent makes it possible to record the monitored sources, the web pages, in order to detect the possible updates on these pages. The agent includes an export function to enable the download of its data, particularly the information source characteristics (URL, name...) as well as the text of the updates detected in these sources.

The conceptual mapping tool is used to give the customer a map of the sources monitored, classified according to the hierarchy used in the monitoring agent. The conceptual map gives the final user of the monitoring alerts and bulletins a synthetic and readable view of the monitored sources (linguistic cover, types of sources). This map is created by reformatting the XML file exported from the monitoring agent and containing the characteristics of the sources.

A procedure integrated into the DBMS makes it possible to import the updates enclosed in the updated XML files which were exported from the monitoring agent. The database integrates a terminology to describe the subject of surveillance. A refining module and a module for the automatic production of alert bulletins in HTML format were directly implemented into the DBMS. The refining module graphic user interface enables the operator to realize the manual and/or visual operations linked to the relevant information selection within the updates, as well as the identification of the linked relevant documents. The webographic references of the relevant documents are recorded in the database.

The word processing tool allows setting the layout of the webographic references comprised in the database via a mail merge function.

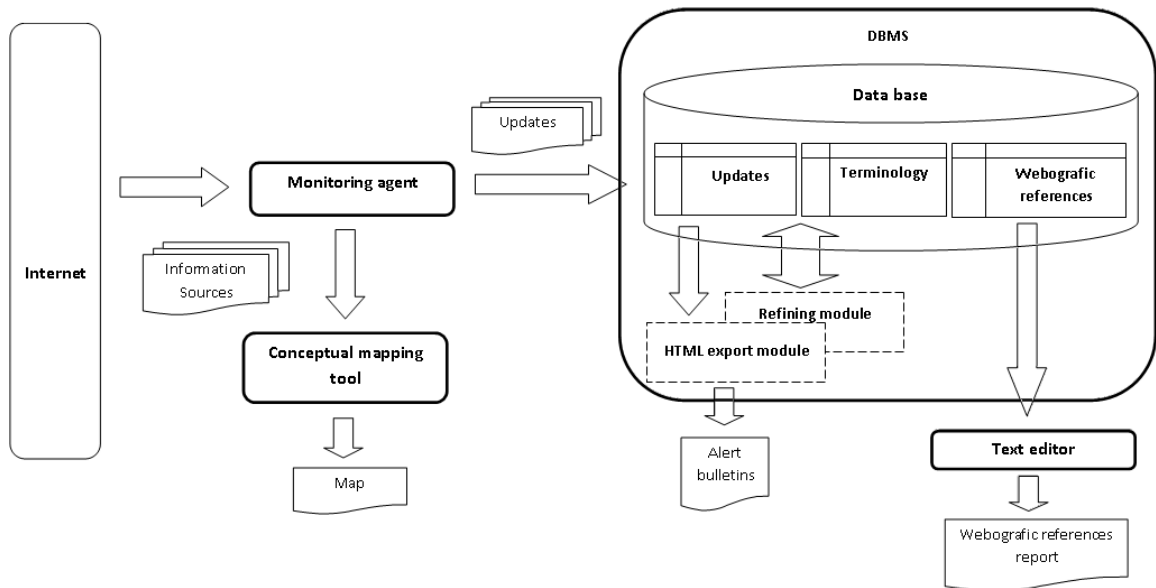


Figure 5: System composed of 4 types of software

6. Application test and limits

The tool was initially developed and successfully used to deliver webographic references bulletins in the framework of techno-legal surveillance.

This monitoring can be characterized by following dimensions:

Table 2: Surveillance dimensions

Terminology	<ul style="list-style-type: none"> • 3 target languages: English, French, German. • 9 themes. • 63 keywords (level 2). • 112 regular expressions.
Frequency of the deliverables	<ul style="list-style-type: none"> • Weekly alert bulletins. • Monthly webographic references bulletins.
Sources	<ul style="list-style-type: none"> • 400 web pages monitored.
Weekly updates	<ul style="list-style-type: none"> • On average, 203 updates collected weekly: <ul style="list-style-type: none"> ▪ On average, 125 automatically selected and submitted to a visual selection (see C1 in fig.3). ▪ On average, 30 manually selected and reduced (see C3 in fig.3).
Relevant documents	<ul style="list-style-type: none"> • On average, 70 documents submitted weekly to the visual selection. • On average, 39 documents selected weekly to appear in the monthly report (see C4 in fig.3).

The tool allows the user to carry out the surveillance for an average duration of 14 hours a month in which 9.5 hours are mainly devoted to the visual selection of the updates and of the documents.

However, the realization time still remains too long and dictates future improvements for the refining application (see § 7.1 Reducing the time to set-up a relevant corpus).

7. Future developments foreseen

In going beyond the general surveillance bulletins realized from the internet and with the goal of maximizing relevance while minimizing cost, the CVTN wishes to rationalize the realization of products with stronger added value. These products should be based on complementary analysis processing, whose results should be relatively easy to interpret by its end-users.

Thus, 2 axes of work are planned to achieve this goal.

7.1 Axis 1: Reducing the time to set-up a relevant corpus

The purpose of this axis is to improve the refining application in 2 ways:

- Direct filtering of documents rather than updates.
- Setting up a refining system by sequence of exclusive requests in order to allow for the automatic selection of one part of the filtered documents, without using a complementary visual selection phase.

To directly filter the documents, the work and/or techniques existing in online data retrieval will be investigated.

Querying a document with a sequence of exclusive requests (see figure 6) consists of defining a set of requests on the terminology, and then classifying them from the most precise to the least precise.

This order defines a sequence of execution within which each request is applied to the sub-corpus rejected by the preceding one in order to guarantee that no document is the object of more than one selection, whether this selection is automated or human. This system should also make it possible to define a threshold in the sequence of execution (at least during the first stage) from which all the sub-corpus are automatically selected (i.e. without human intervention). The relevance of the automatically selected documents will be guaranteed by the precision of the requests on one hand, and on the other hand by the few occurrences or even the absence of a polysemy of the keywords combined one with another.

Beyond the precision threshold, the documents will have to be selected visually by using the process of lexical coloring previously presented. Other means of rapid review or selection by sets will be possibly considered in addition to the lexical coloring, such as automatic summarizing tools that reduce the volume of information.

Finally, the corpus rejected by all requests (complementary corpus) will be processed in order to identify the relevant documents which could comprise new terms that would then be integrated into the research terminology.

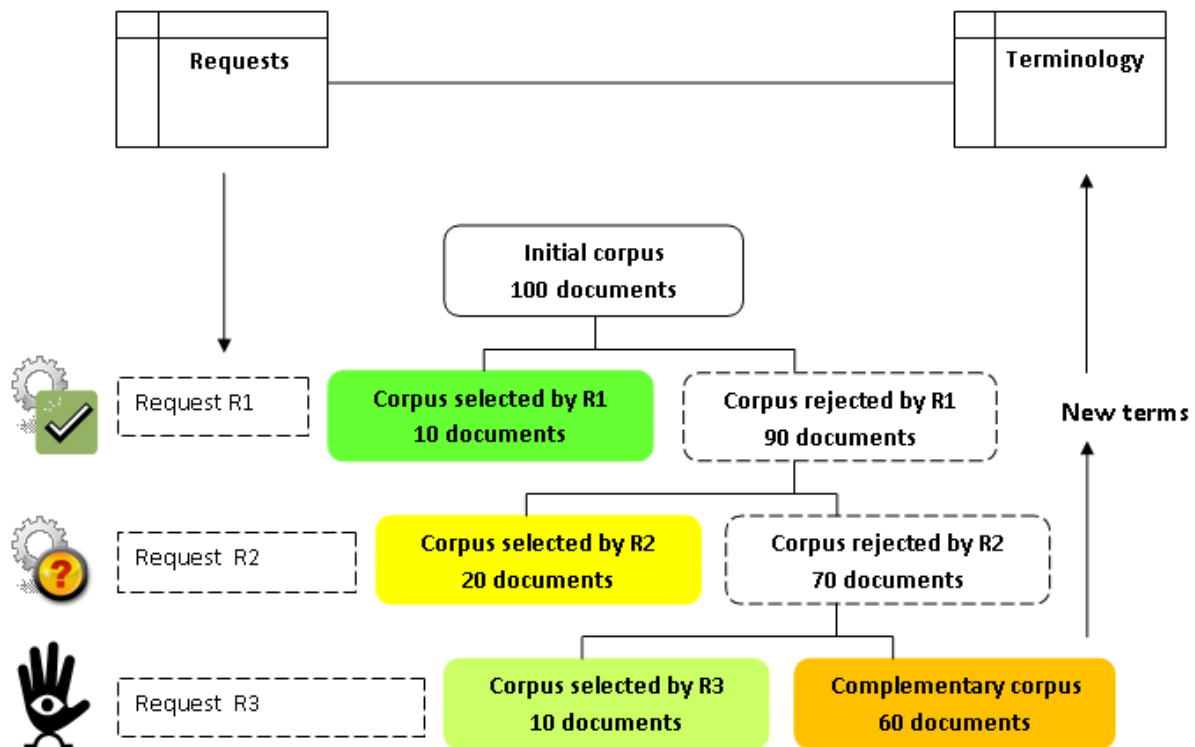


Figure 6: Example of refining by sequence of exclusive requests

This work will also give us the chance to have a closer look at ontologies (Hernandez & Mothe, 2007), in order to discuss the possible improvements that these tools could give to the refining module, particularly towards producing a terminology model and to extend it semi-automatically starting from any existing ontologies.

7.2 Axis 2: A modular approach to integrate new tools

This axis is in line with the already existing initiatives of building a surveillance system based on the combinations of distinct software modules (Denis, Simon, & Brunessaux, 2004) which can evolve according to the needs of the users.

The principle is on one hand to benefit from the possibility of obtaining relevant text corpuses from the Internet, and on the other hand to take advantage of the available software, freeware or shareware, and to articulate them in order to create products or monitoring services at an acceptable and predictable cost.

The integration of the text-mining tools into the system previously presented is also foreseen. This would allow for the visualization of new, or not easily accessible, relations through the documents of a corpus.

The visualization tools will also be analyzed in order to enhance the understanding of the results and to deliver to the customer visually attractive deliverables, using charts or visual representations of information.

Elements from online databases will also be integrated, with the objective of comparing the synthetic views achieved from different corpuses (scientific articles, standards, patents, articles of press...), targeted on the same subject.

However, in order to be integrated into a monitoring system, a tool for information processing, whatever its function is, must be able to receive and/or produce information in a format which is directly or indirectly usable by another tool. Thus arises the question of the tools compatibility for the information processing.

This question should be answered by operations of reformatting which will be necessary to transmit the information from one tool to another.

In doing this, a reference frame of compatible tools for information processing will be built. This reference frame will document in priority the data exchange capacities of each tool. It will also document the functions of each tool within the surveillance process and will provide an evaluation of its execution cost according to reference parameters. It will then allow for the improvement of the existing system by adding new tools or using the unexploited functions of particular software. Adapting the monitoring process on Internet could particularly be relevant for building renewable documents for state of the art or business plan types.

Within the framework of this axis, the benefit from the work completed in the field of "Business Intelligence" will be sought (Baumgartner et al., 2005). Those technical solutions, initially conceived to process internal management of company data are also aimed at the exploitation of data resulting from Internet.

References

- Baumgartner, R., Frölich, O., Gottlob, G., Harz, P., Herzog, M., & Lehmann, P. (2005). Web Data Extraction for Business Intelligence: the Lixto Approach. *Proceedings of BTW 2005*.
- Denis, X., Simon, G., & Brunessaux, S. (2004). Utilisation collaborative d'outils de text mining pour la veille sur Internet (pp. 235-243). *Proceedings of VSST'2004*.
- Hernandez, N., & Mothe, J. (2007). TtoO: Mining a thesaurus and texts to build and update a domain ontology. In H. O. Nigro, S. G. C. Císaro, & D. Xodo (Eds.), *Data Mining with Ontologies: Implementations, Findings, and Frameworks* (pp. 123-144). Buenos Aires, Argentina : IGI Global.
- Perbal, S., Dubois, C., & Schosseler, P. (2004). Exemple de mise en œuvre modulaire d'un processus de veille (pp. 540-546). *Proceedings of VSST'2004*.

Allan Zimmermann

M. Zimmermann joined Public Research Center Henri Tudor in 2007 as information broker. Graduated in Computer Science in 1999, he worked as computer engineer from 2001 to 2003. He gained a graduate in Competitive Intelligence in 2005 and worked as information broker from 2005 to 2007. He is currently involved in research to design a software prototype aiming at filtering and analyzing electronic documents. M. Zimmermann has a special interest in finding affordable ways to synthesize the knowledge available on the Internet.

Xavier Delecroix

M. Delecroix joined the Public Research Center Henri Tudor in 1998 after a University degree in Electronics and Electrotechnics and a Master degree in information sciences and competitive intelligence. He now works as products, services and trainings manager within the "Centre de Veille Technologique et Normative", department set up by a co-initiative of Public Research Center Henri Tudor and Intellectual Property Directorate of Luxembourg. He is mainly involved in the development and deployment of new added value services in Competitive Intelligence.

Cyrille Dubois

In 1995, after a postgraduation in Physical Chemistry and Strategic Information & Competitive Intelligence, M. Dubois joined the Public Research Center Henri Tudor in Luxembourg where he developed Technological and Competitive Intelligence products and activities for private companies. He also developed R&D projects in the Technological & Competitive Intelligence and Intellectual Property fields as well as Technological Intelligence activities for the research departments of the center.

Serge Quazzotti

PhD and post-doc in Organic Chemistry, M. Quazzotti is Director of “Centre de Veille Technologique et Normative”, department set up by a co-initiative of Public Research Center Henri Tudor and Intellectual Property Directorate of Luxembourg. He sets up activities regarding Technology Watch, Intellectual Property (IP) and Standardization. M. Quazzotti is also actively involved as coordinator or contractor in European projects in the field of Technology Watch and IP (FP 5 & 6, Leonardo da Vinci, Competitiveness and Innovation framework Programme).