# Language engineering techniques for web archiving

José Coch[1], Julien Masanès[2]

[1]Lingway, 33-35 rue Ledru-Rollin, 94200 Ivry-sur-Seine, France
Tel. +33 1 56 20 28 35, E-mail jose.coch@lingway.com

[2]Bibliothèque nationale de France, Quai François Mauriac, 75706 Paris, France
Tel. +33 1 53 79 81 17, E-mail julien.masanes@bnf.fr

**Abstract**: Advanced Information processing can enable automatic location of content on the Web, decisions making on its suitability for archiving and thus can ameliorate dramatically accuracy and efficiency for building of large scale Web Archive. This paper presents preliminary results from a research project (WATSON) aiming at adapting various Language Engineering technologies to facilitate large scale Web archiving as well as Web archives mining. The former includes pre-filtering and categorization of sites to define, based on criteria, a focus subset on the Web to be continuously crawled and site categorization to facilitate manual selection of important deep Web site. The result achieved in pre-filtering of commercial Web sites are 100% in precision for 70% of recall. A work station prototype aggregating useful information for professional is presented.

The latter encompasses collections mining with emphasis on content evolution study and analysis of political discourse. We present results applied to the 2002 French election collection made by BnF.

# 1. Information processing for large-scale Web archiving

### Information to be processed

In order to cope with the extremely large quantities of content on the Web an automated process must be put in place to collect it. Advanced Information processing is required to enable automatic location of content, decisions making on its suitability for archiving. This automatic process will be required to run continuously with the minimum of user intervention. It is based on the following types of information [1, 2]:

- Hypertext information (linking between content objects): this information is important both to discover new objects and to assess their position in the total information space. In the Web environment for instance, well-linked sites or pages are the most visible in the hypertext space and can then be considered as important (link weighting). Furthermore, clusters of [inter-]linked sites often indicate similarity of content (thematic clustering) or as originating from specific communities.
- Text information (the content of a document): this information is used to assess a document or sites' thematic classification. Analysis methods include rare word ranking or frequency distribution, key words and key phrases can help determine specific thematic classification or identify generic types, such as commercial, service, academic etc. Text can also be used to detect language and assess national origin or territoriality of sites.
- Technical information: this information consists in file format, embedded scripts, forms and other technical features relevant to assess the site technical profile and particularly possible deep part that crawler cannot access to.

Information is extracted and assessed at the page or document level but assessment for site selection is required at the site level, not the page level. Therefore all page or document level information have to be aggregated in such a way as to facilitate site level assessment, such as thematic classification and site level link weighting.

For large scale crawling, an automatic selection policy can be expressed as a combination of criteria like indexes, categorization schemes, and other parameters to enable decision making for 100k or millions of items (sites).

When this automatic selection is made in the first place, during the crawl, it can result in a significant reduction size in data discovered [3-6]. Automatic selection can also be done through post processing to focus updates out of a large snapshot [7] or domain crawl. By reducing number of sites to be followed up, this technique can allow a tuning of the crawling frequency associated to the site (the more is its importance, the more frequent is its crawling) and exploration depth. There is a "threshold" of importance, in other words, sites not reaching the threshold are just archived with the minimum frequency or not explored at all[1].

Even if most of Web archiving projects still tend to skirt the Deep Web issue, awareness of the necessity of finding solutions in this domain is growing. Site technical information can be used to determine the technical profile which can be very useful for tracking Deep Web sites [8] [9] and for triggering adapted procedures, like a specific harvesting [10, 11] or a deposit delivery.

The following diagram represent three processes (with information and output) of interest for large scale Web archiving projects.

On the top, the box encompasses analysis done at the global scale, during the domain snapshot for instance.

The bottom-left box groups processes applied to deep Web site and the bottom-right box include the focus crawl for updating a subset of Websites.

[1] This can be based on link weightening for instance. Linking can be interpreted as a vote for a page. Therefore at the global scale links represent a collective evaluation of pages. The idea is to use this for automatic selection. Millions of people actually do some sort of selection by their linking of pages. Sites that have more incoming links can be considered in a very broad sense as being the most ‘relevant’ or ‘important’. Using an iterative evaluation of importance (importance cast by link is weighted by the importance of the page where it comes from) provides even better results .

Comparing librarians’ manual evaluation of sites with the computed link weighting or page ranking for the site has tested this hypothesis; an important correlation has been found (Masanès, J., *Towards Continuous Web Archiving: First Results and an Agenda for the Future*. D-Lib Magazine, 2002. **8**(12).) .

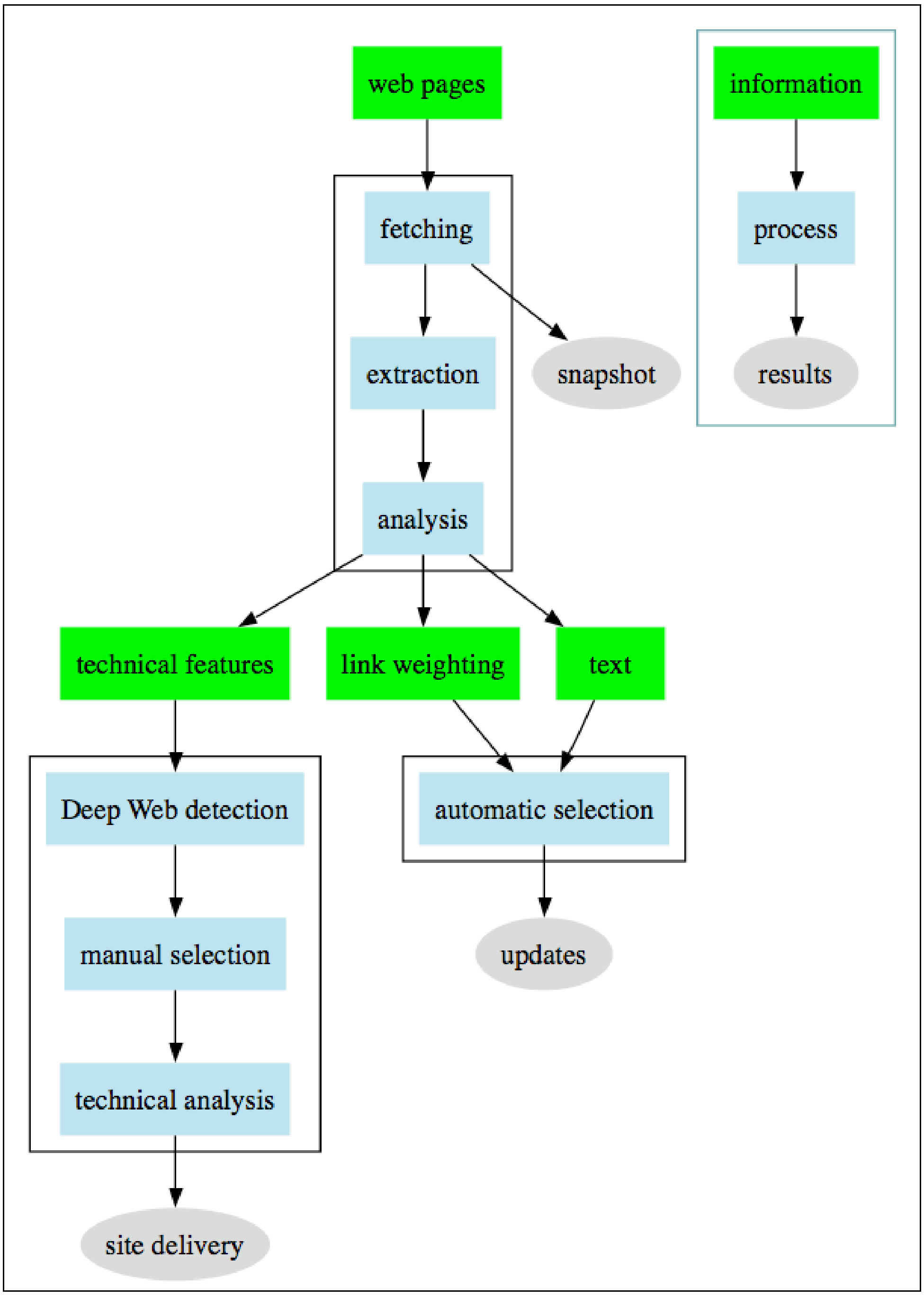


**Figure 1: Large scale Web archiving processes**

### Quantity

In the French context, based on experiments made by BnF [7, 12-15], the set of sites deemed of little relevancy according to BnF selection policy, is estimated to be more than 80% of the total publicly accessible surface Web. This encompasses among other service sites and 'commercial windows' sites that will be crawled only twice a year within the global snapshot. Note that site with no in-linking at all won't be part of this global snapshot. This excludes a lot of non-public Web sites, which is in line with the traditional mission of national libraries, based on published material.

More "important" sites will be followed and updated to fit more accurately to their actual change frequency (to avoid loss of information compare to crawling only twice a year newspaper sites for instance).

It is planed that BnF will be able to archive 1000 deep Web sites per year through site delivery [15]. A manual selection from a list of automatically detected deep Web sites will be done by reference librarians to assess importance and richness of hidden content in order to trigger a -relatively- heavy deposit procedure.

### Using language engineering

Several processes outlined below are based chiefly or exclusively on language engineering techniques.

These techniques can be used for:

- pre-filtering and categorization of sites to define, based on criteria, a focus subset that will be continuously crawled;
- site summarization and categorization to facilitate manual selection of important deep Web site.

Results of the WASTON project in this domain are presented in section 4.

In addition to being useful for professionals in the building of Web archives these techniques allows also the user (researchers doing researches these collections) to mine the collections, study evolution of content, analyze political discourse etc. A sample of potential analysis conducted during the WATSON project is provided in section 5.

## 2. Watson project: technical approach

The *Bibliothèque Nationale de France* (BnF) is involved in a two-year research project called WATSON ("*Web : Analyse des Textes, Sélection, et Outils Nouveaux*") which is funded by the French Ministry of Research in the framework of the Technolangue programme[2]. The project is coordinated by the Lingway company in association with two university laboratories. It aims at developing and adapting linguistic tools to Web sites content analysis. The present section describes the technical approach followed in the project. The next two sections describe the applications evaluated by the BnF.

### 2.1. Information extraction and Language engineering

Unfortunately, the vast majority of Web sites are not structured in order to quickly give a good idea of the salient facts about their content. Information extraction (IE) is the activity of automatically pulling out of relevant information from large volumes of text and notably from Web sites. This information is then usually entered automatically into a database or structured in an XML file, which may then be used to analyze the data for trends, to give a summary, for indexing in IR, etc.

The goal of Language Engineering (LE) is to produce software to process natural language.

Language engineering is essential for applications such as information retrieval, Web search, information extraction, data mining and text summarization. LE techniques for morphological analysis, named entity detection, part-of-speech tagging, word prediction, or term extraction are in use in real-world applications in these areas.

### 2.2. Watson project technical approach

The objective of our project is to develop, to generalize and to integrate LE tools for Web-mining, taggers, chunkers, named entities recognisers, XML taggers, co-reference solvers, extractors, categorizers, by focusing on performance and robustness for large collections of data (and notably large sets of Web sites).

Watson shows a very modular architecture and standard interfaces. This fact allows the use of the tools in several manners, either in an isolated way or integrated way.

The modules used in the Watson project are :

[2] The main objective of Technolangue is to set up a permanent infrastructure for producing and distributing linguistic resources, assessing written and oral language technology and participating in relevant national and international standardisation and information monitoring bodies 16.
*Portail technolangue.net..*

- pre-filtering

  *shallow term detection, weighted counting and computing of proportions of categories of terms (for example "our products", "our technical support" are associated to the* commercial *category, while "science laboratory", "research group" are associated to the* research *category) to quickly characterize a site*

- logical structuring of Web sites and pages,

  *splitting Web sites into parts, pages, sections, titles, paragraphs (not only based on HTML marks, but also on other aspects like fonts, blank lines, etc.)*

- segmentation into sentences,

  *splitting relevant parts of pages into sentences (not only based on "string starting by an Upper letter and ending by a dot", obviously)*

- named entity recognition,

  *recognizing and marking-up names of persons, organizations, geographical places and dates. For example for persons, based on huge lists of first names (*"Joseph Watson"*), regular expression rules describing the forms the name can have (*"Mr. J. Watson"*), and contextual rules (*"The meeting was adjourned by President Watson at 8:15 am. "*),*

- semantic mark-up,

  *recognizing and marking-up certain types of sentences (presentation, description, objective, citations, opinions, conclusions, etc.). For example* "Watson recalls that after 5 years of White Party government, there are 4 million poor people in this country." *is detected as being a citation. This detection is based on hundreds of* cue-phrases *as "recalls that", "declares", etc. The same applies for each type of sentence.*

- fact extraction,

  *extracting relations, for example between the author, the citation, the main object and the polarity (negative - positive). For example from "Watson thinks that it is not a good idea to enter in a new war" gives the relation :*

| **Who** | **Object** | **Polarity** |
|---|---|---|
| *Watson* | *to enter in a new war* | *negative* |

- morphological tagger,

  *marking-up simple and compound words and adding the morpho-syntactical tag (noun, verb, etc.). For example* "The last meeting was interesting" gives

```
<w tag="Det-Def" lemma="the" entry="the">
<w tag="Adj-Qual" lemma="last" entry="last">
```

```
  <w tag="Noun-Common" lemma="meeting"
entry="meeting">
  <w tag="Verb-Pt3p" lemma="be" entry="was">
  <w tag="Adj-Qual" lemma="interesting"
entry="interesting">
```

- chunker (shallow syntactic analysis),
  *marking-up simple syntactical groups (as* "the last meeting" *in the previous example)*
- term extraction,
  *marking-up terminological compounds (using patterns like adj+noun, noun+prep+noun, etc.)*
- co-reference solving,
  *marking-up the relation between a word or a term and its co-referent (typically a pronoun)*
- text summarisation,
  *producing a text summarising the salient facts on the Web site (see below)*
- categorization,
  *calculating a category for the Web site and adding this information to the metadata*

stressing on robustness and performance for large volumes of text.

### 2.3. Watson typical architecture

Figure 2 shows a typical integration of certain Watson modules (blue arrows show output directly used by the application).

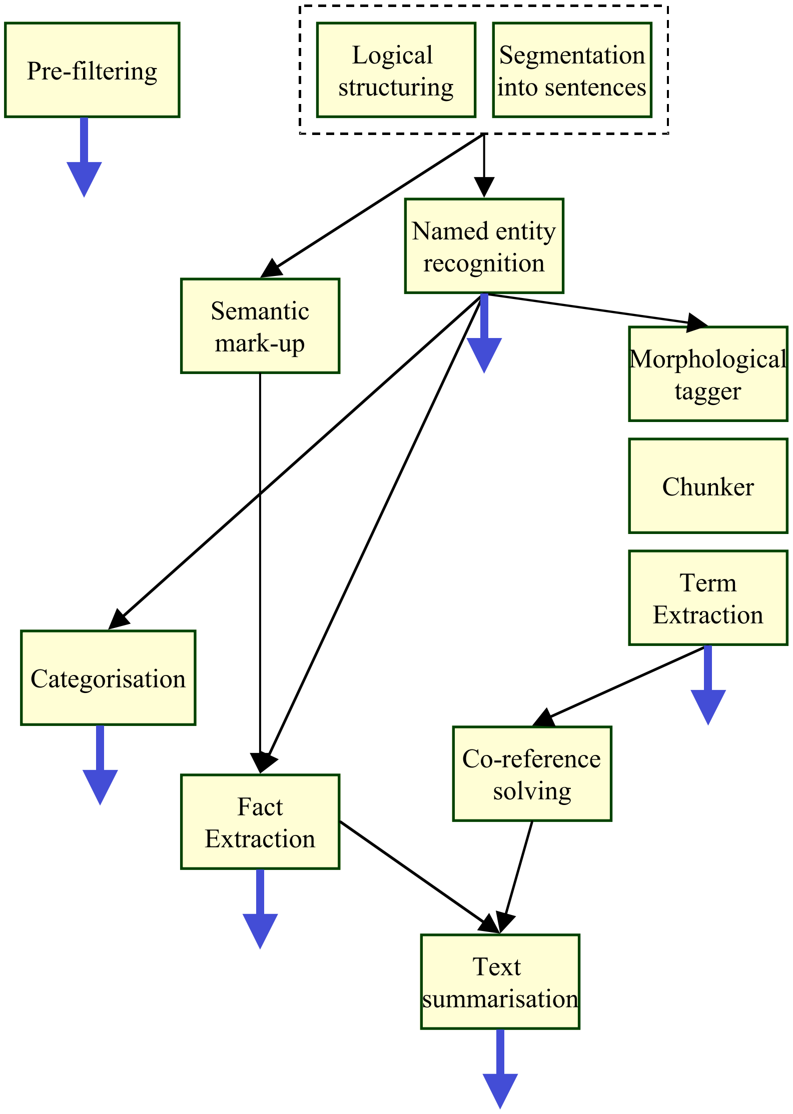


**Figure 2: WATSON modules**

## 3. Standards and Watson DTD

In order to facilitate integration of modules and interoperability with external tools, we have carefully selected a set of information exchange standards that are wrapped in an XML DTD.

### 3.1. Standards

The output of each module is in XML format. The following standards and models have been used to define relevant DTDs :

- TEI Lite for the general representation of text collections [17] ;
- Dublin Core for the metadata [18, 19];
- NIST / MUC format for the named entity representation [20, 21];
- Crossmarc format for the fact extraction [22].

### 3.2. Watson DTD

It would be too long here to describe all the aspects of the Watson DTD. We give in the next two sections the DTDs of the outputs of two crucial modules: the Logical structuring of the text, and the Named entity recognition.

### Output of the Logical structuring module

```
<!ELEMENT document (metadata,text) >

<!ELEMENT metadata
(title|description|keywords|subject|creator|
publisher|contributor|date|type|format|identifier|
source|language|relation|coverage|rights)* >

<!ELEMENT title (#PCDATA) >
<!ELEMENT creator (#PCDATA) >
<!ELEMENT keywords (#PCDATA) >
<!ELEMENT subject (#PCDATA) >
<!ELEMENT description (#PCDATA) >
<!ELEMENT publisher (#PCDATA) >
<!ELEMENT date (#PCDATA) >
<!ELEMENT type (#PCDATA) >
<!ELEMENT format (#PCDATA) >
<!ELEMENT identifier (#PCDATA) >
<!ELEMENT relation (#PCDATA) >
<!ELEMENT coverage (#PCDATA) >
<!ELEMENT rights (#PCDATA) >
<!ELEMENT source (#PCDATA) >
<!ELEMENT language (#PCDATA) >

<!ELEMENT text (group|section*) >

<!ELEMENT group (document*) >

<!ELEMENT section (sectionTitle?,(paragraph|section)*)
>
<!ATTLIST section
level (1|2|3|4|5|6|7|8|9) #REQUIRED
type CDATA #IMPLIED >

<!ELEMENT sectionTitle (#PCDATA) >
<!ATTLIST sectionTitle
type CDATA #IMPLIED >

<!ELEMENT paragraph (#PCDATA) >
```

One can recognize the influence of the Dublin Core’s metadata representation.

The text element can contain a group of documents, or the parts of a single document as in the TEI Lite. For example, a text element can contain a corpus, a set of messages, a Web site, or even a set of Web sites.

Output of the Named entity recognition module

```
<!ELEMENT section
(sectionTitle?,(paragraph|section)*)>
<!ATTLIST section
level (1|2|3|4|5|6|7|8|9) #REQUIRED
type CDATA #IMPLIED >

<!ELEMENT sectionTitle (#PCDATA|ENAMEX|TIMEX|NUMEX)* >
<!ATTLIST sectionTitle
type CDATA #IMPLIED >

<!ELEMENT paragraph (#PCDATA|sentence)* >

<!ELEMENT sentence (#PCDATA|ENAMEX|TIMEX|NUMEX)* >

<!ELEMENT ENAMEX (#PCDATA) >
<!ATTLIST ENAMEX
type (PERSON|ORGANIZATION|PLACE) #REQUIRED
subtype CDATA #IMPLIED >

<!ELEMENT TIMEX (#PCDATA) >
<!ATTLIST TIMEX
type (DATE) #REQUIRED >

<!ELEMENT NUMEX (#PCDATA) >
<!ATTLIST NUMEX
type CDATA #IMPLIED >
```

ENAMEX is the tag for the Named Entities, TIMEX for the Time expressions (typically dates) and NUMEX for the Numerical expressions. Named entities can be (at least) typed into PERSON, ORGANISATION, and PLACE. These entities can be subtyped into:

- person : function or role (president, doctor, etc.),
- organization : company, university, laboratory, library, etc.,
- place : country, region, town, etc.

Very simple example of document:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE document SYSTEM "watson3ne.dtd">
<document>
        <metadata>
             <title>Le Monde</title>
             <language>FR</language>
             <source>http://www.lemonde.fr</source>
             <identifier>Web/imprimer_article/232260.html</iden
        tifier>
             <date>2003-06-10</date>
        </metadata>
        <text>
             <section level="1">
                    <sectionTitle> Le Monde.fr : <ENAMEX
                    type="PERSON">Raffarin</ENAMEX>
                    s&#39;explique,
                    <ENAMEX type="PERSON">Sarkozy</ENAMEX>
               n&#233;gocie,
                    les syndicats persistent
                    </sectionTitle>
               <section level="2">
                    <sectionTitle>
                    <TIMEX type="DATE">mardi 10 juin
                    2003</TIMEX>
                    </sectionTitle>
                    <paragraph>
                         <sentence><ENAMEX
                         type="PERSON">Jacques Chirac</ENAMEX>
                         bla bla.
                         </sentence>
                         <sentence>Bla bla bla bis.</sentence>
                         <sentence>Blabla bla
                    blablater.</sentence>
                    </paragraph>
               </section>
             </section>
        </text>
</document>
```

# 4. Tools for professionals (librarians): results

## 4.1. Overview

As explained in section §1, the Watson project has prepared three different tools for the professionals of a institutions in charge of Web archiving (in our case, the BnF):

- pre-filtering of less important sites,
- working station for helping the professional to characterize a site,
- categorization of sites.

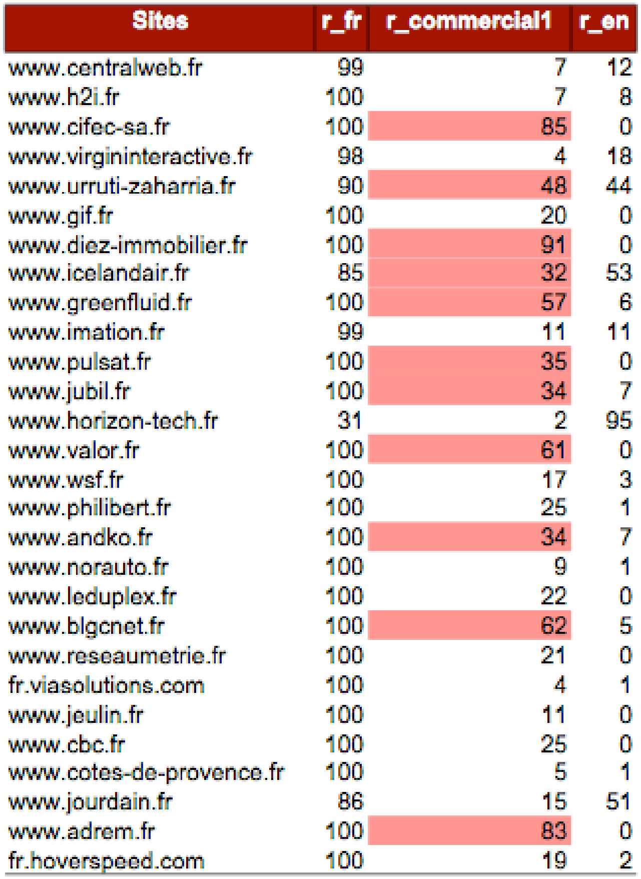
### 4.2. Pre-filtering: automatic characterizing Web sites

*Technical approach and description*

A previous experience described in [7] had been conducted before the beginning of Watson. The idea was to use an algorithm similar to the now famous *PageRank* [12, 23], used by Google; in other words to count link pointing to the site to have an idea of its "visibility" or "popularity", which are concepts close to "importance". That algorithm gives interesting and relevant results, but there is a type of sites for which the results of the algorithm are not so good: the very specialized sites, notably related to research or high-tech, because these sites are not so "popular" but they are considered as "important" by the BnF experts. Thus, the requirement for Watson pre-filtering was to propose an alternative or complementary way to balance the popularity ranking, in order to correctly assess this type of sites [7] .

After the described previous experience, we observe that less-important sites were very often purely commercial-windows style, sites. The Watson pre-filtering uses this idea to detect this type of sites. More concretely, the algorithm is based on a set of 120 weighted terms frequently used in commercial sites. The total weight of the commercial terms is compared to the total weight of all the terms. After tuning of thresholds, the algorithm allows to detect a considerable subset of the less-important sites (see §Evaluation subsection below), and of course, all the less-popular important sites wrongly processed by the previous algorithm are not classed as less-important by the Watson pre-filtering algorithm.

As a sub-product of this process, a secondary result of the Watson pre-filtering is an information about the proportion between texts in French (r-fr) and in English (r-en). Table 1 shows the result of Watson pre-filtering applied to 28 sites. For Watson, 11 (pink) sites are less-important, and one of the sites (horizon-tech.fr) has more text in English than in French.



| Sites | r_fr | r_commercial1 | r_en |
|---|---|---|---|
| www.centralweb.fr | 99 | 7 | 12 |
| www.h2i.fr | 100 | 7 | 8 |
| www.cifec-sa.fr | 100 | 85 | 0 |
| www.virgininteractive.fr | 98 | 4 | 18 |
| www.urruti-zaharria.fr | 90 | 48 | 44 |
| www.gif.fr | 100 | 20 | 0 |
| www.diez-immobilier.fr | 100 | 91 | 0 |
| www.icelandair.fr | 85 | 32 | 53 |
| www.greenfluid.fr | 100 | 57 | 6 |
| www.imation.fr | 99 | 11 | 11 |
| www.pulsat.fr | 100 | 35 | 0 |
| www.jubil.fr | 100 | 34 | 7 |
| www.horizon-tech.fr | 31 | 2 | 95 |
| www.valor.fr | 100 | 61 | 0 |
| www.wsf.fr | 100 | 17 | 3 |
| www.philibert.fr | 100 | 25 | 1 |
| www.andko.fr | 100 | 34 | 7 |
| www.norauto.fr | 100 | 9 | 1 |
| www.leduplex.fr | 100 | 22 | 0 |
| www.blgcnet.fr | 100 | 62 | 5 |
| www.reseaumetrie.fr | 100 | 21 | 0 |
| fr.viasolutions.com | 100 | 4 | 1 |
| www.jeulin.fr | 100 | 11 | 0 |
| www.cbc.fr | 100 | 25 | 0 |
| www.cotes-de-provence.fr | 100 | 5 | 1 |
| www.jourdain.fr | 86 | 15 | 51 |
| www.adrem.fr | 100 | 83 | 0 |
| fr.hoverspeed.com | 100 | 19 | 2 |

**Table 1: Result of Watson pre-filtering applied to 28 sites**

*Pre-filtering evaluation*

Concerning the speed, the pre-filtering is very rapid: less than 5 seconds per site in average on a standard PC. In other words, 20.000 sites in 24 hours.

Concerning the recall and precision measures, let us underline that the precision has priority, because the software must avoid to consider as less-important a site which has to be selected as important. The results of measures performed on 100 sites previously analyzed "by hand" by expert librarians are :

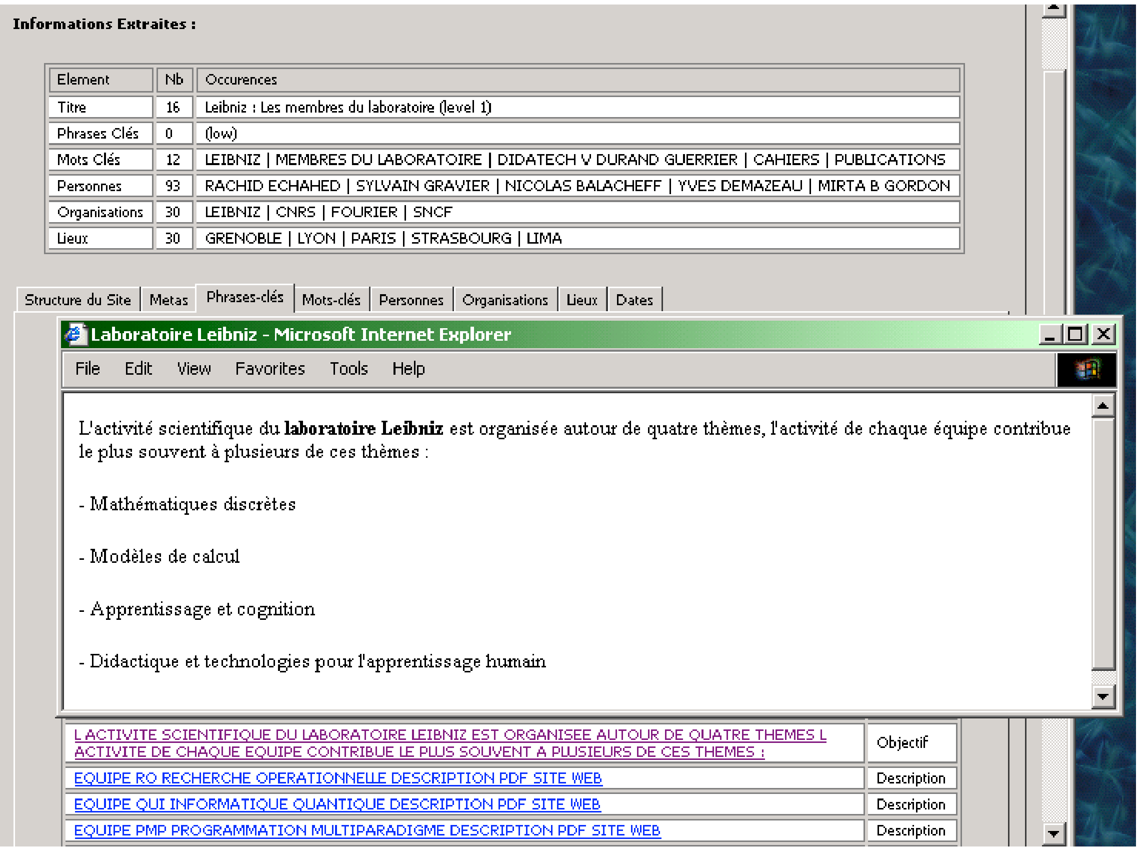
- Precision : 100%
- Recall : 70%

One can conclude that 70% of the less-important sites are rapidly identified by Watson. On the other hand, 30% of the less-important sites have still to be identified. That is the goal of the next step.

### 4.3. Helping the professional to characterize a Web site

Apart from the fact that complementary systems can be used to identify less-important sites, obviously it will always remain a set of sites difficult to identify automatically. That is why we conceived a working station for the librarian to rapidly have a good feedback on the content of a site.

By the moment, the Watson station shows information about the structure of the site, the metadata information, the key sentences, the key words and the named entities (persons, organizations, places and dates). All this information is organized in a single screen which is dynamic and "hypertextual". In other words, if the librarian is interested in something (for example in a key sentence) he/she can click on it and a pop-up window with the context appears. Figure 3 shows a screen copy with the Watson working station and a window pop-up once the librarian clicks on a key sentence.



**Informations Extraites :**

| Element | Nb | Occurences |
|---|---|---|
| Titre | 16 | Leibniz : Les membres du laboratoire (level 1) |
| Phrases Clés | 0 | (low) |
| Mots Clés | 12 | LEIBNIZ \| MEMBRES DU LABORATOIRE \| DIDATECH V DURAND GUERRIER \| CAHIERS \| PUBLICATIONS |
| Personnes | 93 | RACHID ECHAHED \| SYLVAIN GRAVIER \| NICOLAS BALACHEFF \| YVES DEMAZEAU \| MIRTA B GORDON |
| Organisations | 30 | LEIBNIZ \| CNRS \| FOURIER \| SNCF |
| Lieux | 30 | GRENOBLE \| LYON \| PARIS \| STRASBOURG \| LIMA |

L'activité scientifique du **laboratoire Leibniz** est organisée autour de quatre thèmes, l'activité de chaque équipe contribue le plus souvent à plusieurs de ces thèmes :

- Mathématiques discrètes

- Modèles de calcul

- Apprentissage et cognition

- Didactique et technologies pour l'apprentissage humain

| | |
|---|---|
| L ACTIVITE SCIENTIFIQUE DU LABORATOIRE LEIBNIZ EST ORGANISEE AUTOUR DE QUATRE THEMES L ACTIVITE DE CHAQUE EQUIPE CONTRIBUE LE PLUS SOUVENT A PLUSIEURS DE CES THEMES : | Objectif |
| EQUIPE RO RECHERCHE OPERATIONNELLE DESCRIPTION PDF SITE WEB | Description |
| EQUIPE QUI INFORMATIQUE QUANTIQUE DESCRIPTION PDF SITE WEB | Description |
| EQUIPE PMP PROGRAMMATION MULTIPARADIGME DESCRIPTION PDF SITE WEB | Description |

**Figure 3: Watson working station**

The relevance of the working station will be evaluated by the BnF during Q3-Q4 2004 (end of the project).

### 4.4. Categorization

The standard learning algorithm is used in the first version of Watson categorization. For each site, Watson uses a partial crawling of the site (max. 100 pages per site), extracts the text, calculates relevant *bigrammes* (two-word terms), drops empty words, and learns relations between text and categories. For the learning step, manually categorized sites are used. To categorize a new site, the same idea is followed: partial crawling, text extraction and finally, comparison between the extracted text and the learned corpus in order to identify the best category.

For evaluating the categorization module of Watson, the BnF's bookmarks or « *signets* » (http://signets.bnf.fr/) have been used. The learning corpus had 9500 pages crawled from categorized sites divided into 12 categories corresponding to the actual division of collections departments at BnF. The test has been made on 3800 others pages.

The difficulty comes from the non sheer thematic division of these categories (including the Reserve and Manuscript department for instance). Still results are quite encouraging.

Several methods (Rocchio, EM, Bayes) have been used and compared. The following table shows the results:

- Rocchio : 78,1 % of precision,
- EM: 77,2 %
- Bayes: 75,6 %

### 4.6. Future work

We are working now on the general improving of the functions delineated above (notably a new version of the Categorization more based on a linguistic extraction) also on adding a new function: the site summarization.

We are preparing a site summarization, which will be a text showing a short extraction on (most important) key sentences, typically describing the site or the main object of the site, and some extra information on special content found on the site, for example, the number of audio or video files, or of large images, which can give to the librarian interesting hints to identify a site as being important.

## 5. Tools for researchers: exploring Web sites

### 5.1. Overview

Watson proposes several tools to explore Web sites, notably:

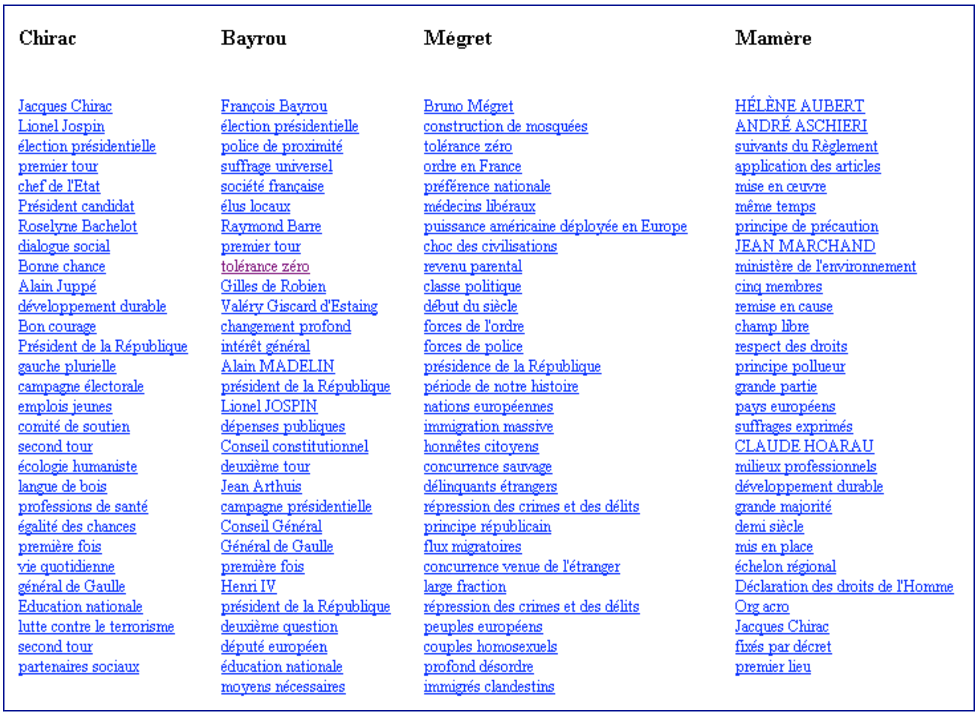
- extraction of most used terms and named entities,
- graphical representation of associations (co-occurrences),
- sentence extraction,
- monitoring of the site's content evolution.

A first experience has been done using the sites of the candidates to the French presidential elections (2002).

Evaluation of the tools for researchers is planned for Q3-Q4 2004.

### 5.2. Extraction of terms and named entities

A first idea on site’s content is given by the list of most frequent terms and named entities used in it. Figure 4 shows the most frequent items found in the April 2002 version of four candidates to the French presidential election's Web sites.



| Chirac | Bayrou | Mégret | Mamère |
|---|---|---|---|
| Jacques Chirac | François Bayrou | Bruno Mégret | HÉLÈNE AUBERT |
| Lionel Jospin | élection présidentielle | construction de mosquées | ANDRÉ ASCHIERI |
| élection présidentielle | police de proximité | tolérance zéro | suivants du Règlement |
| premier tour | suffrage universel | ordre en France | application des articles |
| chef de l'Etat | société française | préférence nationale | mise en œuvre |
| Président candidat | élus locaux | médecins libéraux | même temps |
| Roselyne Bachelot | Raymond Barre | puissance américaine déployée en Europe | principe de précaution |
| dialogue social | premier tour | choc des civilisations | JEAN MARCHAND |
| Bonne chance | tolérance zéro | revenu parental | ministère de l'environnement |
| Alain Juppé | Gilles de Robien | classe politique | cinq membres |
| développement durable | Valéry Giscard d'Estaing | début du siècle | remise en cause |
| Bon courage | changement profond | forces de l'ordre | champ libre |
| Président de la République | intérêt général | forces de police | respect des droits |
| gauche plurielle | Alain MADELIN | présidence de la République | principe pollueur |
| campagne électorale | président de la République | période de notre histoire | grande partie |
| emplois jeunes | Lionel JOSPIN | nations européennes | pays européens |
| comité de soutien | dépenses publiques | immigration massive | suffrages exprimés |
| second tour | Conseil constitutionnel | honnêtes citoyens | CLAUDE HOARAU |
| écologie humaniste | deuxième tour | concurrence sauvage | milieux professionnels |
| langue de bois | Jean Arthuis | délinquants étrangers | développement durable |
| professions de santé | campagne présidentielle | répression des crimes et des délits | grande majorité |
| égalité des chances | Conseil Général | principe républicain | demi siècle |
| première fois | Général de Gaulle | flux migratoires | mis en place |
| vie quotidienne | première fois | concurrence venue de l'étranger | échelon régional |
| général de Gaulle | Henri IV | large fraction | Déclaration des droits de l'Homme |
| Education nationale | président de la République | répression des crimes et des délits | Org acro |
| lutte contre le terrorisme | deuxième question | peuples européens | Jacques Chirac |
| second tour | député européen | couples homosexuels | fixés par décret |
| partenaires sociaux | éducation nationale | profond désordre | premier lieu |
|  | moyens nécessaires | immigrés clandestins |  |

**Figure 4: Most frequent items in 4 presidential Web sites**

## Technical approach

Named entity extraction is based on several types of rules :

- checking the first name and the name in a Knowledge base (for example, “Jacques” is a French first name and “Chirac” is known by the system),
- first name known and following word or words starting by a capitalized word (ex. “Bruno Mégret”),
- the structure of the sequence corresponds to a known entity (“Dr. XXX” => “XXX” is a person subtyped as a Doctor),
- the contexts contains the information (for example “ZZZ is a company which…”).

Term extraction is based on a linguistic analysis of the text, then a filtering step looking for patterns associated to terms (like adjective+noun, noun+preposition+noun, etc.), and finally a statistical step.

### 5.3. Graphical visualization of content association

The next step is to show the associations found in the sites. These associations are calculated by co-occurrences (in other words, a term is associated to an other one if they occur often in the same Web page). The more often a term is associated to an other term, the closer they appear in the graphical representation.

Figure 5 is a copy of the screen showing on the left the graphical representation of the terms and named entities most associated to "*tolérance zéro*" ("zero tolerance"). On the right side, the list of alternative choices of terms and named entities, which can be sorted by frequency, by score or alphabetically.

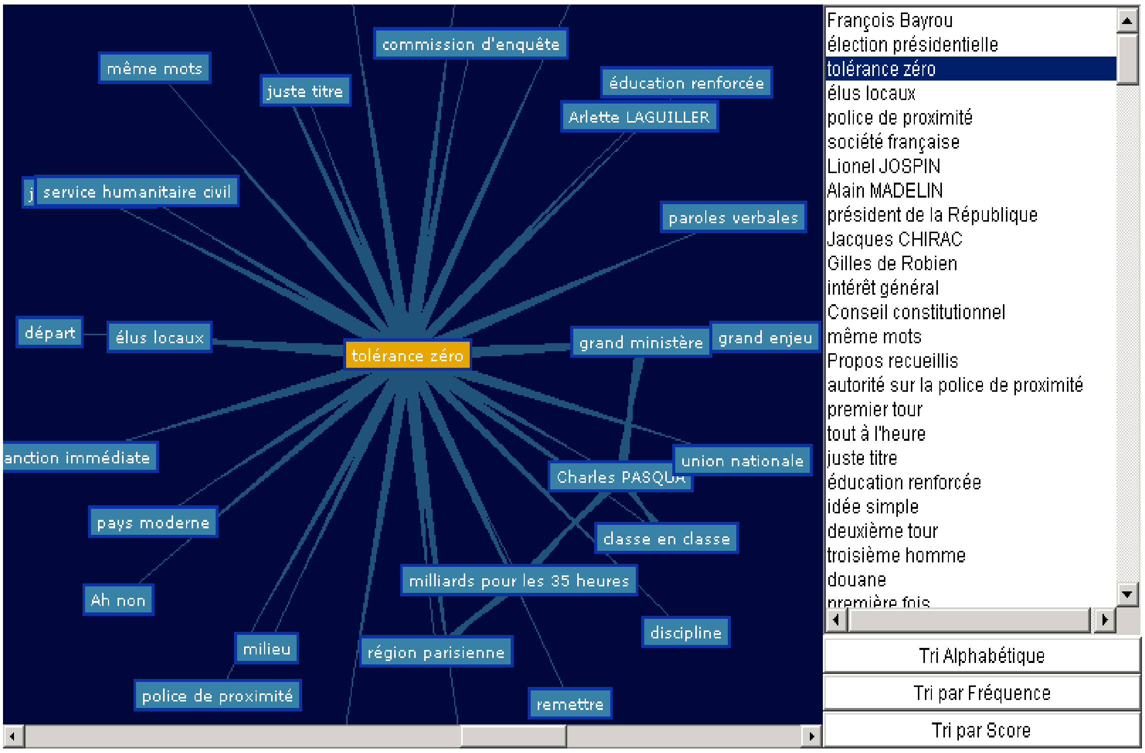


**Figure 5: terms and named entities most associated to "tolérance zéro"**

### 5.4. Monitoring the site's content evolution

Watson allows the researcher to have an idea of the evolution of the content of a site. The following example is taken again from the French presidential elections of 2002.

The idea is simply to represent in a graph the evolution of the frequency of each item (term or named entity).

The figure 6 shows the most frequent persons cited in the Web site of Jacques Chirac (except for Jacques Chirac himself). Let's recall that in France the election for president has two rounds. Only the two most voted candidates in the first round are allowed to continue in the second one. Lionel Jospin, Socialist Party candidate for president was the prime minister of Jacques Chirac, who was the president (and who belongs to the Right).

In the graph, it is possible to see that the frequency of Lionel Jospin reduces significantly between April and May (actually the first round was on the 21st of April and only Mr. Chirac and Mr. Le Pen continued in the second round). On the other hand, Mr. Le Pen, who was virtually inexistent on the Chirac's site in April, shows a considerable growing. Other growing people are Christian Blanc, François Bayrou, Jean-Pierre Raffarin, who were competitors before the 1st round but supported Mr. Chirac for the 2nd round.

Analogous analyses are being conducted on political campaign dominant themes.

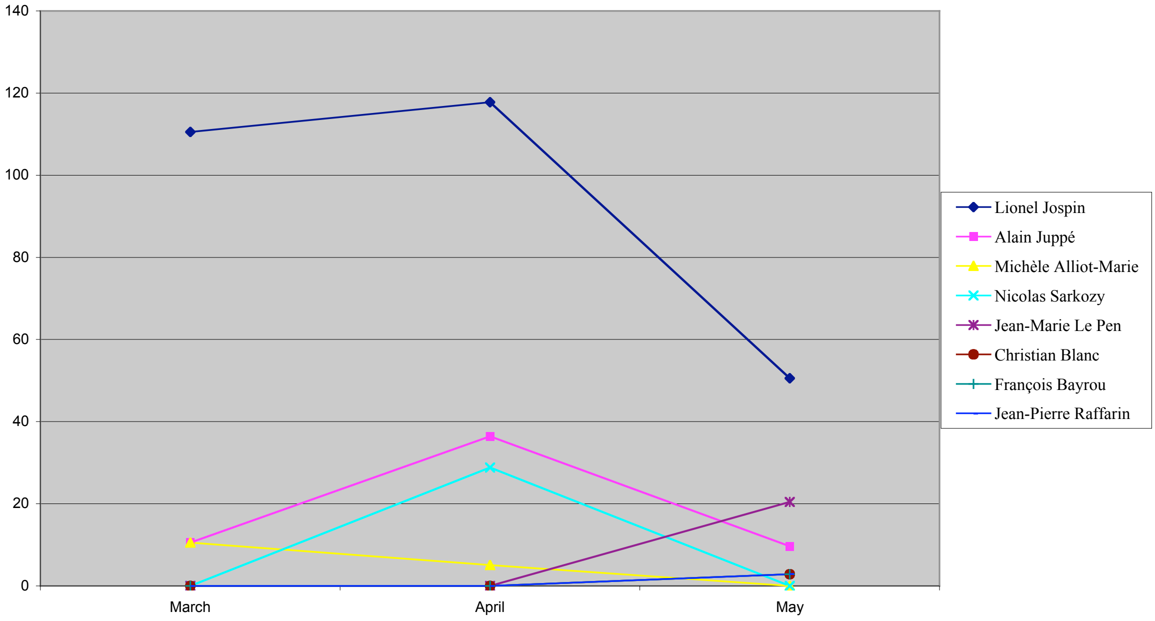


**Figure 6: most frequent persons cited in the Web site of Jacques Chirac**

## Related works

There is a large set of projects exploiting non structured information, thinking a web site as a « bag of words ». Their results have some interest, however they are rather limited in terms of knowledge extracted (less typed information, less relevant, noise, silence, etc. or even giving only a categorization of the Web site analysed). In this section, we present some work related to the Watson project, in terms both of technical approach and type and quality of results.

Bauer and Sharl [24, 25] conducted a set of analysis at the site level on textual content (what they call 'exploratory textual analysis'). Though limited to lematization and keyword extraction, it allows analysis based on site type (environemental organization, banking, infoTech and Tourism). They also proposed longitudinal study of web site textual content, correspondence analysis to improve the web development cycle as well as business analysis.

An example of some more structured information processing is the one given by Craven et al. [26]. The prototype they describe is using an ontology, "machine learning" algorithms, and a set of learning examples, and produces as its output structured knowledge bases. One interesting point is that the work unit is the Web site, not the Web page for which the results are less satisfactory. To compare with Watson, we can say that our objectives are larger than only a categorization in an ontology, and our approach is even more based on knowledge (not only semantic, but also linguistic, as for example the ones represented by dictionaries and grammars).

The TyPWeb project [27] associates several techniques to describe Web site typologies, notably to characterise commercial Web sites. It is interesting to note that the TyPWeb approach is largely based on counting some categories of words (like personal pronouns or subjective verbs) and on assessing a level of complexity of the site and the regularity of the hyperlinks. Some ideas we applied in Watson have been inspired by TyPWeb (notably for the pre-filtering). However, the counting of the personal pronouns does not appear as operational for characterizing a site as commercial, and nowadays, most of Web sites, even those which are not commercial, are build by specialists, they can be very complex and the links are regular very frequently.

Some other approaches specialize on one family of web sites, for example Academic sites in [28] or Job advertising pages [22]. In both cases, the use more detailed and hierarchical knowledge, which is very difficult to generalize to the whole World Wide Web.

## Acknowledgements

This project was funded, in part, by the French Ministry of Research under the Technolangue action.
We are very grateful to our partners from the Lallic Laboratory (University of Paris IV): Jean-Luc Minel, Ghassan Mourad and from the Lattice-Talana

Laboratory (University of Paris VII): Alexis Nasr, Alexa Volanschi, Mélodie Soufflard, Laurence Danlos.

We also like to acknowledge our colleagues from the Bibliothèque nationale de France: Sara Aubry, Thierry Cloarec, Catherine Lupovici, and from Lingway: Marie-Hélène Candito, Bernard Normier, Hugues de Mazancourt, Frédéric Touitou.

## References

1. Baldi, P., P. Frasconi, and P. Smyth, *Modeling the Internet and the Web - Probalilistic Methods and Algorithms*. 2003, New York;London;Sydney: Wiley.
2. Chakrabarti, S., *Mining the Web : discovering knowledge from hypertext data*. 2002, San Francisco, CA: Morgan Kaufmann Publishers. xviii, 345 p.
3. Abiteboul, S., M. Preda, and G. Cobena. *Adaptive on-line page importance computation*. in *Proceedings of the twelfth international conference on World Wide Web*. 2003: ACM Press.
4. Bergmark, D. *Collection synthesis*. in *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries*. 2002: ACM Press.
5. Bergmark, D., C. Lagoze, and A. Sbityakov. *Focused Crawls, Tunneling, and Digital Libraries*. in *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*. 2002: Springer-Verlag.
6. Chakrabarti, S., M.v.d. Berg, and B. Dom, *Focused crawling: a new approach to topic-specific Web resource discovery*. Computer Networks (Amsterdam, Netherlands: 1999), 1999. **31**: p. 1623--1640.
7. Masanès, J., *Towards Continuous Web Archiving: First Results and an Agenda for the Future*. D-Lib Magazine, 2002. **8**(12).
8. Cope, J., N. Craswell, and D. Hawking. *Automated discovery of search interfaces on the web*. in *Proceedings of the Fourteenth Australasian database conference on Database technologies 2003*. 2003: Australian Computer Society, Inc.
9. Zhang, Z., B. He, and K.C.-C. Chang. *Understanding Web query interfaces: best-effort parsing with hidden syntax*. in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. 2004: ACM Press.
10. Lage, J.P., et al. *Collecting hidden weeb pages for data extraction*. in *Proceedings of the fourth international workshop on Web information and data management*. 2002: ACM Press.

11. Raghavan, S. and H. Garcia-Molina. *Crawling the Hidden Web*. in *Proceedings of the 27th International Conference on Very Large Data Bases*. 2001: Morgan Kaufmann Publishers Inc.
12. Abiteboul, S., et al. *A First Experience in Archiving the French Web*. in *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*. 2002: Springer-Verlag.
13. Masanès, J., *Bilan des expérimentations sur le dépôt légal de l'Internet*. 2002, Bibliothèque nationale de France.
14. Masanès, J. *BnF's project for Web archiving*. in *IWAW'01*. 2001. Darmstadt.
15. Masanès, J. *Archiving the Deep Web*. in *IWAW'02*. 2002. Roma.
16. *Portail technolangue.net*.
17. Sperberg-McQueen, L.B.C.M., *TEI Lite: An Introduction to Text Encoding for Interchange*. 1995.
18. Powell, A., *Guidelines for implementing Dublin Core in XML*. 2003.
19. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*.
20. Chinchor, N. *MUC-7 Information Extraction Task Definition*. in *MUC-7*. 1998.
21. Chinchor, N., et al., *Named Entity Recognition Task Definition*. 1999.
22. Petasis, G., et al. *Adaptive, Multilingual Named Entity Recognition in Web Pages*. in *ECAI*. 2004.
23. Brin, S. and L. Page, *The anatomy of a large-scale hypertextual Web search engine*. Computer Networks and ISDN Systems, 1998. **30**: p. 107--117.
24. Bauer, C., D. Bauer, and A. Scharl. *Towards the Measurement of Public Web Sites: A Tool for Classification*. in *ISI Cutting Edge Conference on The Measurement of E-Commerce (MEC-99)*. 1999. Singapore: International Statistical Institute.
25. Scharl, A., *Evolutionary Web Development*. Applied Computing, ed. R. Paul, P. Thomas, and J. Kuljis. 2000, London: Springer. 314.
26. Craven, M., et al. *Learning to extract symbolic knowledge from the world wide web*. in *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*. 1998. Madison, Wisconsin, United States.
27. Beaudoin, V., et al. *Décrire la toile pour mieux comprendre les parcours*. in *CIUST'01. Colloque International sur les usages et services des Télécommunications*. 2001. Paris.
28. Rehm, G. *Towards Automatic Web Genre Identification A Corpus-Based Approach in the Domain of Academia by Example of the Academic's Personal Homepage*. in *HICSS-35*. 2002: IEEE.