

Création d'un  
système de  
recherche  
d'information

Bureau d'étude  
de recherche  
d'information

## Page de situation du document

*Table de révision*

Date	N° de version	Chapitre concerné	Cause de la modification
24/03/14	1.0	Version initiale du rapport	
24/03/14	2.0	V- Démarche de développement	Revue, réorganisation du rapport
25/03/14	3.0	V- Organisation et développement	Désaccord sur le graphique rappel/précision
26/03/14	4.0	/	Changement de la mise en page

## Remerciements

Nous tenions à remercier Mme Bahsoun pour son cours sur la gestion de projet qui nous a permis de mener à bien ce projet.

Nous remercions également Mme Lechani pour ses cours de recherche d'information, indispensable pour ce projet.

Et nous tenions également à remercier Mme Soulier pour nous avoir guidé durant les séances de TP et qui s'est montré disponible en permanence et nous a toujours apporté des solutions aux problèmes que nous avons rencontrés.

# Table des matières

I- Objet et but du document.....	5
1.1 Objet.....	5
1.2 But .....	5
II- Présentation de la problématique .....	5
Contexte .....	5
III- Documents applicables et documents de référence .....	7
3.1 Cahier des charges.....	7
3.2 Documents de références .....	12
3.3 Lexique de notation.....	13
3.4 Présentation du document.....	13
IV- Organisation et développement .....	14
4.1 Ressources Humaines.....	14
4.2 Ressources matérielles : méthodes et outils utilisés .....	15
4.3 Planning prévisionnel .....	16
V- Démarche de développement .....	17
5.1 Cycle de vie.....	17
5.2 Architecture d'un système de recherche d'information .....	18
5.3 SADT .....	19
5.4 Préparation de la collection .....	22
5.4.1Aspiration des pages web.....	22
5.4.2 Extraction du contenu et des métadonnées .....	23
5.4.3 Indexer la collection .....	23
5.4.4 Construction de tuples pour la base de données.....	24
5.5 Traitement des requêtes .....	25
5.5.1 Indexer des requêtes exemples .....	25
5.5.2 Calculer le score RSV .....	26
5.5.3 Trier les résultats .....	26
5.6 Évaluation des performances de recherche.....	27
5.7 Mise en œuvre d'une interface de recherche.....	30
5.7.1 Schéma de la base de données .....	30
5.7.2 Enrichir la base de donnée .....	31
5.7.3 Génération des interfaces .....	31

VI- Assurance et contrôle qualité .....	37
Revues .....	37
VII- Bilans.....	39
7.1 Bilan Fournisseur .....	39
7.2 Bilan Personnel.....	39
7.3 Conclusion .....	39
Annexes .....	40

## I- Objet et but du document

### 1.1 Objet

Ce document est destiné à l'Université Paul Sabatier. Il est le fruit du travail de deux élèves de L3 Statistiques et Informatique Décisionnelle et est une trace d'un projet mené durant le second semestre de l'année Universitaire 2013-2014.

L'objectif de ce projet réalisé dans l'UE *Recherche d'information : concepts et langages* permet de mettre en pratique les connaissances acquises durant la formation SID.

### 1.2 But

Ce document a pour objectif de montrer notre démarche de développement et notre aptitude à utiliser des logiciels qui nous sont fournis. Il permet ainsi de comprendre quels ont été les mécanismes de développement de l'outil et la gestion de projet appliquée pour mener à bien la mission.

## II- Présentation de la problématique

### Contexte

Aujourd'hui quand nous utilisons internet, nous sommes forcément confronté à l'utilisation d'un moteur de recherche. Cet outil est très utilisé mais sa conception est souvent inconnue.

Le premier moteur de recherche, « Archie », est né en 1990, il permettait de rechercher des documents sur internet. Développé par un étudiant de l'université de McGill au Québec, il était peu optimisé mais on peut dire aujourd'hui que tous les moteurs de recherche descendent d'Archie.

Le moteur de recherche le plus utilisé dans le monde aujourd'hui est Google. En février 2014, 92,4% des recherches en France se font sur le moteur de recherche Google. Ce

géant d'internet répertorie 30 trillions de documents en 2012. Il y a 100 milliards de requêtes faites chaque mois sur Google.

En somme, les moteurs de recherche ont énormément progressé en 20 ans. Et sont très utilisés mêmes au sein de site web contenant de nombreuses informations. Cela facilite la navigation et la recherche. Et apporte un côté "attractif" au site web.

La compréhension et l'apprentissage de la création d'un site web s'inclut donc parfaitement dans les formations informatiques.

Dans le cadre de la formation L3 SID nous devons donc analyser, concevoir, évaluer et implémenter un système de recherche d'information.

Ce projet s'est donc déroulé à l'Université Paul Sabatier sur une durée de 11 semaines, du 31 janvier 2014 à la mi-mars 2014.

## III- Documents applicables et documents de référence

### 3.1 Cahier des charges

**CMI SID**

**Bureau d'études**

**2013-2014**

<b>ANALYSE, CONCEPTION, IMPLEMENTATION ET EVALUATION D'UN SYSTEME DE RECHERCHE D'INFORMATION</b>
--

#### **I. Introduction**

---

##### **I.1. Objectifs**

L'objectif fondamental de ce bureau d'études est de mettre en œuvre de façon compilée, et à des degrés divers, les connaissances acquises dans le cadre de différentes unités d'enseignements. En particulier, les enseignements suivants :

- Concepts de recherche d'information
- Langage de développement Perl
- Bases de données : conception et interrogation
- Génie-Logiciel

En pratique, cet objectif sera poursuivi à travers l'analyse, la conception, l'évaluation et l'implémentation d'un système de recherche d'information. Les résultats attendus sont :

- Un module logiciel capable de sélectionner, à partir d'une collection de documents, une liste de documents pertinents en réponse à une requête utilisateur.
- Un document qui décrit le contexte du bureau d'études, les solutions techniques retenues ainsi que le déroulement du projet

D'un point de vue personnel, la réalisation du bureau d'études conduira à une meilleure maîtrise (1) des techniques de recherche d'information, (2) des langages de traitement de texte (Perl) et interrogation de bases de données (SQL), (3) les techniques de gestion d'un projet.



## **I.2. Organisation et déroulement**

a- Le travail sera réalisé en groupes formés de 4 étudiants au maximum. Il est recommandé de désigner un animateur (chef de projet) pour le groupe.

b- Le déroulement sera encadré durant des séances de travaux dirigés et travaux pratiques.

Des revues d'étapes seront réalisées régulièrement avec les intervenants.

c- Chaque étudiant doit être impliqué dans une tâche durant chaque séance pour optimiser au mieux le temps de travail.

d- Chaque groupe doit fournir, à la fin du projet, un support écrit qui résume le travail réalisé durant le bureau d'études.

## **I.3 Evaluation**

Le travail réalisé durant le bureau d'études sera noté. La note sera basée sur critères suivants :

a- Qualité du travail réalisé

b- Qualité du support écrit

c- Qualité de la soutenance : supports, temps de parole, réponses aux questions

**Cette note compte pour 40% de l'évaluation finale de l'enseignement « Concepts de RI »**

## **I.4 Dates importantes**

- Semaine du 13/01 : démarrage des séances de C/TD bureau d'études

- Semaine du 27/01 : démarrage des séances de TP

- in mars : soutenance du projet bureau d'études

## II- Travail à faire

---

Les étapes de mise en œuvre du système de recherche d'information sont les suivantes :

**Etape 1 :** Préparation de la collection

**Etape 2 :** Traitement des requêtes

**Etape 3 :** Evaluation des performances de recherche

**Etape 4 :** Mise en œuvre d'une interface de recherche

### II.1 Préparation de la collection

L'objectif de cette étape est de produire à partir d'une collection de documents WEB bruts téléchargés à partir de sources comme Wiki news (<http://wikipedia.fr>), ou Journal du Geek (<http://www.journaldugeek.fr>) : (1) des index et index inverses (*posting*) structurant le contenu et (2) une base de données comprenant des métadonnées. Le travail sera réalisé à l'aide d'un robot WEB, du langage Perl, et du SGBD Oracle.

Etape 1 : Préparation de la collection	
Travail à effectuer	Outil / Technologie/source
1- Collecter des documents WEB à partir d'un ensemble de sites ciblés	Robot Htrack ou Wget. Sources Wikinews ( <a href="http://wikipedia.fr">http://wikipedia.fr</a> ), Journal du Geek ( <a href="http://www.journaldugeek.fr">http://www.journaldugeek.fr</a> )
2- Analyser les documents pour extraire les données et informations <ul style="list-style-type: none"><li>• <i>Données</i> : auteurs, date de publication, source, date, avis, tags etc.</li><li>• <i>Informations</i> : contenu principal du document</li></ul>	HTML, PERL, informations de référencement, contenus des pages, métadonnées
3- Structurer les données et informations <ul style="list-style-type: none"><li>• <i>Données descriptives</i></li><li>• <i>Contenu indexable</i></li></ul>	PERL, SGBD oracle, SQL
4- Indexer la collection <ul style="list-style-type: none"><li>• <i>Construire le dictionnaire</i></li><li>• <i>Construire l'index des documents</i></li><li>• <i>Construire l'index inversé de la collection</i></li></ul>	PERL

## II.2 Traitement des requêtes

L'objectif de cette étape est d'implémenter et tester différentes stratégies d'interrogation de la collection à l'aide de requête en utilisant : 1) différentes fonctions de pondération ou d'appariement, 2) différents types de requêtes : booléenne, un seul terme, plusieurs termes, combinaison du contenu et métadonnées etc.

Etape 2 : Traitement des requêtes	
Travail à effectuer	Thème / Technologie
1- Préparer et indexer un ensemble de requêtes exemples	Perl
2- Calculer des scores RSV (requête, document)	Perl, Bibliothèque Perl/DBI, SGBD oracle, SQL
• <i>Tester des requêtes de contenu</i> <i>Tester des requêtes combinées (contenu, métadonnées)</i>	
• <i>Tester différentes fonctions de pondération, appariement</i>	

## II.3 Evaluation des performances de recherche

L'objectif de cette étape est d'évaluer l'efficacité du système de recherche d'information développé. Pour cela, un ensemble de requêtes test doit être préparé avec, pour chaque requête, la liste des documents pertinents attendus, identifiés à partir de la collection. L'évaluation sera basée sur des mesures de rappel/précision calculées sur les N premiers documents retournés à l'issue de l'évaluation de chaque requête

Etape 3 : Evaluation des performances de recherche	
Travail à effectuer	Thème / Technologie
1- Construire une base de test locale : collection de requêtes et jugements associés (documents pertinents associés à chaque requête)	Perl
2- Interroger le système développé avec les requêtes de test	P@N, MAP, SGBD oracle, SQL, Excel
3- Evaluer les précisions de recherche	
• <i>Calculer des précisions moyennes, précision à N documents</i> • <i>Tracer des courbes rappel-précision</i>	

#### **II.4 Mise en œuvre d'une interface de recherche**

*L'objectif de cette étape est de développer un module qui permet à des utilisateurs de soumettre une requête, lancer une recherche et récupérer les résultats de la recherche sous forme de liens, résumés, etc.*

<b>Etape 4 : Mise en œuvre d'une interface de recherche</b>	
<b>Travail à effectuer</b>	<b>Thème / Technologie</b>
4- Définir une syntaxe de requête et une fonction qui permet la saisie/vérification d'une requête	Perl, PHP, MySql, Oracle, SQL
5- Afficher les résultats triés de l'évaluation de requête	Perl, PHP, MySql, Oracle, SQL
6- Accéder aux métadonnées et contenus des résultats par navigation	Perl, PHP, MySql, Oracle, SQL

### 3.2 Documents de références

Pour mener à bien ce projet nous nous sommes appuyés sur plusieurs documents, tous issus de la formation L3SID.

En voici la liste :

UE	Semestre	Enseignant	Justification
Recherche d'information (TD + Cours)	2	Mme Lechani	Ce projet, s'intégrant dans l'UE Recherche d'Information, nous nous sommes forcément appuyés sur le cours. Il a été le fil conducteur de notre projet, de la création de la collection de documents, en passant par le traitement des requêtes et l'évaluation des performances de recherches, jusqu'à la mise en œuvre de l'interface de recherche (pour certains modules).
Recherche d'information (TP)	2	Mme Soulier	Durant ces TP nous avons 'amorcé' le projet, ce qui nous a fait gagner un temps considérable sur le projet en lui-même. Quand nous avons commencé le projet, nous avons ainsi pu nous inspirer des TP que l'on a faits.
Génie Logiciel	1 et 2	Mme Bahsoun	Tout ce qui concerne la gestion de projet réalisé dans cette mission a été vu dans ce cours.
Langage Perl	1	M Farinas	Ce langage nous a été indispensable lors de ce projet. Tout est codé en perl dans ce projet.
Conception de base de données	1	M Morvan, Mme Yin, M Mokadem	Ce cours nous a permis de construire la base de données associée au projet.
Langage de Requête	1	Mme Sauvagnat, Mme Soulier	Ce langage nous a permis d'administrer la base de données.

Nous avons également utilisé les informations contenues sur le site [github.com/](https://github.com/) pour réaliser les graphiques en Java Script.

### 3.3 Lexique de notation

Cette partie contient l'ensemble des notations récurrentes au rapport qui nécessite d'être défini :

RI : Recherche d'information. La recherche d'information peut se définir comme l'ensemble des opérations effectuées pour retrouver une information répondant à une question précise.

RSV : Relevance Status Value est le score de pertinence.

BD : Base de données

RP : Rappel précision

SQL : Langage permettant d'administrer une base de données

Perl : langage informatique adapté au traitement et à la manipulation de fichiers texte.

E/A : Entité association

Mot vide : mot très commun comme le, la, les qu'il est inutile de prendre en compte dans des traitements de l'information.

Index : utilisé pour représenter le contenu d'un document.

### 3.4 Présentation du document

La prochaine partie est consacrée à l'organisation et au développement du projet qui nous a été confié. Nous expliquerons notre gestion de projet et les personnes qui ont travaillé sur cette mission, ainsi que tous les outils utilisés.

Dans un second temps nous expliquerons notre démarche de développement et nous y détaillerons chacune des étapes.

Puis, pour assurer la qualité du travail fourni, nous détaillerons l'ensemble des revues qui ont eu lieu durant ce projet.

## IV- Organisation et développement

### 4.1 Ressources Humaines

Deux personnes ont participé à ce projet :

Nom	Rôle
Romain ROBERT	Responsable des livrables
David JEAUNEAU	Responsable la programmation

On a défini quatre grandes étapes dans ce projet. On a évalué la proportion de chaque tâche dans le projet. Il y a eu en tout 24h de TD consacré à ce projet, ce qui donne :

Étape	Heures	Proportion	Charge de travail
Préparation de la collection	8	33%	2,8
Traitement des requêtes	4	17%	2
Évaluation des performances de recherche	4	17%	2
Mise en œuvre d'une interface de recherche	8	33%	2,8

De ce tableau, nous avons pu déduire la courbe de Gauss. On peut constater que le 'pic' de travail se situe environ à la première et à la dernière étape du projet.

C'est donc là où nous devons être le plus présents, attentifs et réactif aux éventuels problèmes qui se déclareraient.

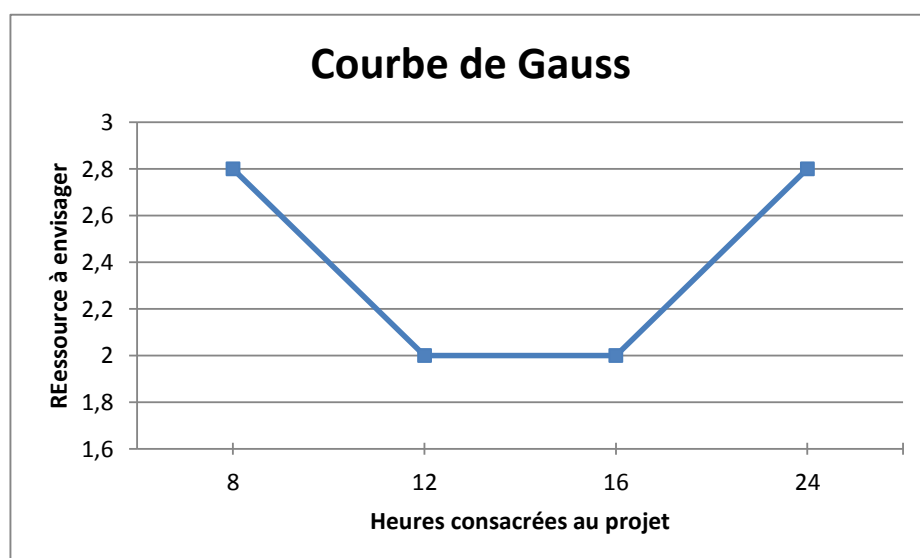


FIGURE 4.1 – Courbe de Gauss

## 4.2 Ressources matérielles : méthodes et outils utilisés

Pour mener à bien ce projet, nous nous sommes appuyés, sur plusieurs méthodes et outils.

- SQL

SQL est un langage informatique permettant d'administrer des bases de données

- PERL

Perl est un langage informatique adapté au traitement et à la manipulation de fichiers texte. Aujourd'hui, nous utilisons principalement Perl pour générer, mettre à jour, analyser des fichiers HTML et pour la conversion de formats de fichiers.

- HTML

HTML est un langage informatique utilisé pour créer des pages web.

- CSS

Le CSS est un langage informatique qui permet de gérer la présentation d'une page web.

- Word

Word est un logiciel qui permet de traiter du texte, nous l'avons donc utilisé pour rédiger ce rapport.

- Excel

Excel est un logiciel tableur, qui nous a permis de réaliser la majorité des graphes présents dans ce rapport et le planning prévisionnel.

- Powerpoint

Powerpoint est un logiciel de présentation qui nous a permis de réaliser les schémas E/A et le modèle SADT.

- HTTracks

Le logiciel HTTracks est un aspirateur web qui nous a permis d'aspirer les pages web du journal du geek.



### 4.3 Planning prévisionnel

Pour simplifier la gestion de projet, nous avons décidé de choisir en fil conducteur les quatre étapes définies dans le cahier des charges, auxquelles nous avons attribués des sous-étapes. Il y avait 24h de TD à répartir. On peut par ailleurs constater que le planning a très bien été respecté car nous avons eu seulement une séance de retard sur la totalité du projet.

Étapes	Sous-étapes	Séance 1	Séance 2	Séance 3	Séance 4	Séance 5	Séance 6	Séance 7	Séance 8	Séance 9	Séance 10	Séance 11	Séance 12
		JANVIER	FÉVRIER										MARS
		31	07	07	14	14	17	21	24	28	28	13	13
Préparation de la collection	Collecter des documents WEB à partir d'un ensemble de sites ciblés												
	Analyser les documents pour extraire les données et informations												
	Structurer les données et informations												
	Indexer la collection												
Traitement des requêtes	Préparer et indexer un ensemble de requêtes exemples												
	Calculer des scores RSV												
Évaluation des performances de recherche	Construire une base de test locale												
	Interroger le système développé avec les requêtes des test												
	Évaluer les précision de recherche												
Mise en œuvre d'une interface de recherche	Définir une syntaxe de requête et une fonction qui permet la saisie / vérification d'une requête												
	Afficher les résultats triés de l'évaluation de requête												
	Accéder aux métadonnées et contenus des résultats par navigation												
	Construction de la partie Statistiques du moteur de recherche												



 Durée théorique des étapes  
 Retard sur les étapes

FIGURE 4.2 – Planning prévisionnel

## V- Démarche de développement

### 5.1 Cycle de vie

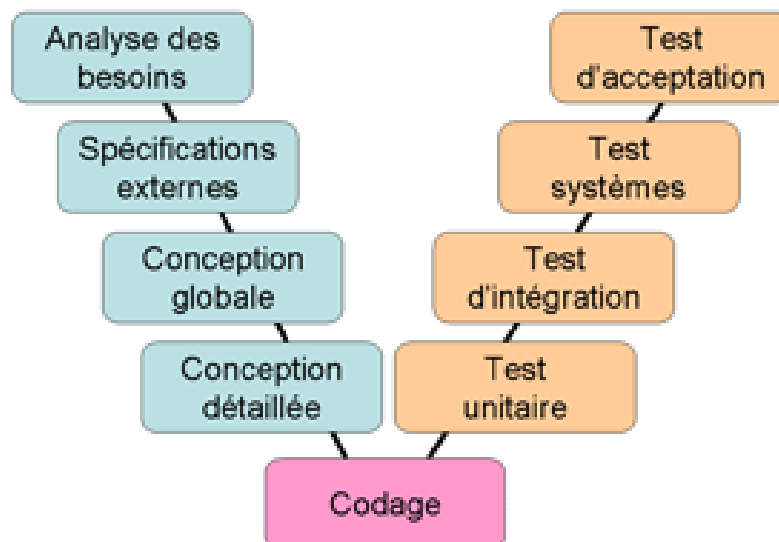


FIGURE 5.1 – Cycle de vie

Tout logiciel a un cycle de vie. Pour construire le cycle de vie du système de recherche que nous avons créé. Nous nous sommes appuyés sur la structure du cycle de vie en V.

Nous allons donc détailler succinctement chacune des étapes pour comprendre ce que nous avons fait au niveau du cycle de vie de notre moteur de recherche. Nous ne rentrerons pas dans les détails, car les étapes du développement sont détaillées plus tard dans le rapport.

Phase	Détails
Étude des besoins	Élaboration du cahier des charges
Spécification du logiciel	C'est à cette étape que nous avons décidé quels modules nous allions associer à notre site de moteur de recherche.
Conception globale	Nous avons formé 4 groupes de travail et construit une base de données commune.
Conception détaillée	Au sein de chaque groupe nous avons schématisé et conçu les algorithmes des éléments du module dont nous avons la charge.

Codage	Nous avons codé les algorithmes que nous avons mis en place.
Test unitaire	Nous avons fait des requêtes tests sur le moteur pour voir si nos modules marchaient.
Test d'intégration	Assemblage des différents modules créés par les quatre groupes.
Validation	Nous attendons la validation de notre site par les enseignants nous ayant confiés cette mission.
Exploitation, maintenance et évolution	/

## 5.2 Architecture d'un système de recherche d'information

L'architecture d'un système d'information répond à une structure précise, c'est à partir de ça que le plan du projet a été élaboré. Voici un schéma expliquant cette structure :

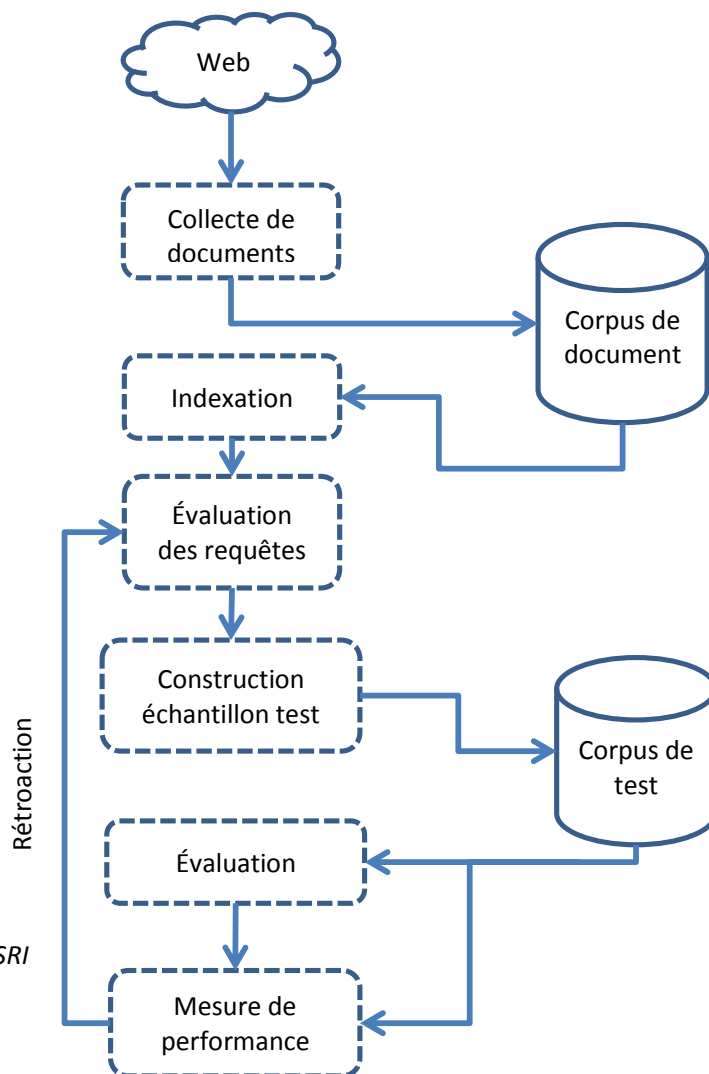
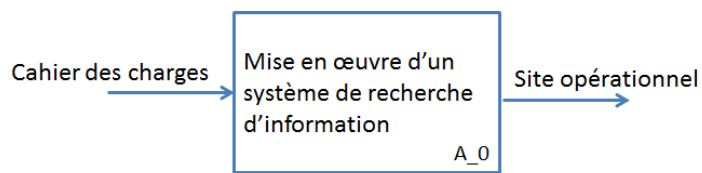


FIGURE 5.2 Architecture d'un SRI

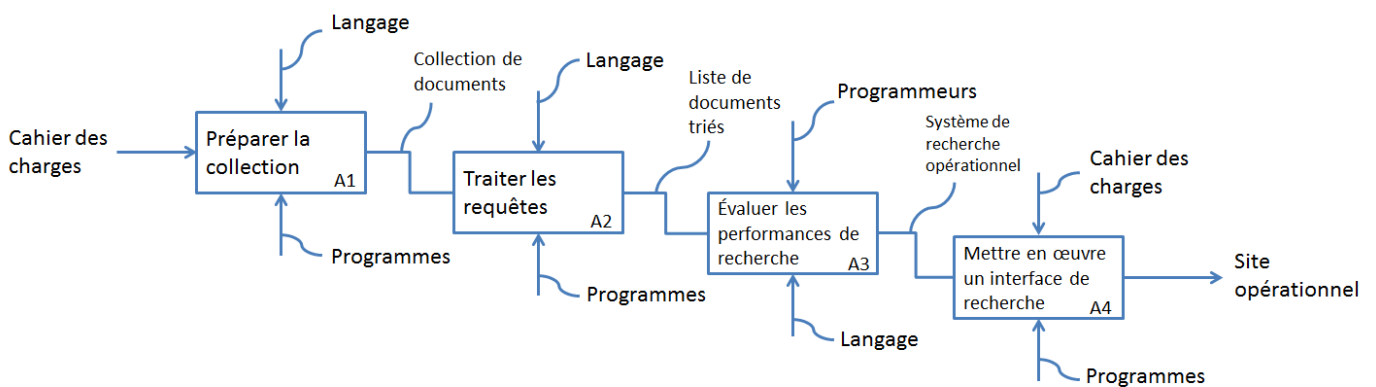
### 5.3 SADT

Nous avons utilisé la méthode SADT pour mener à bien ce projet.

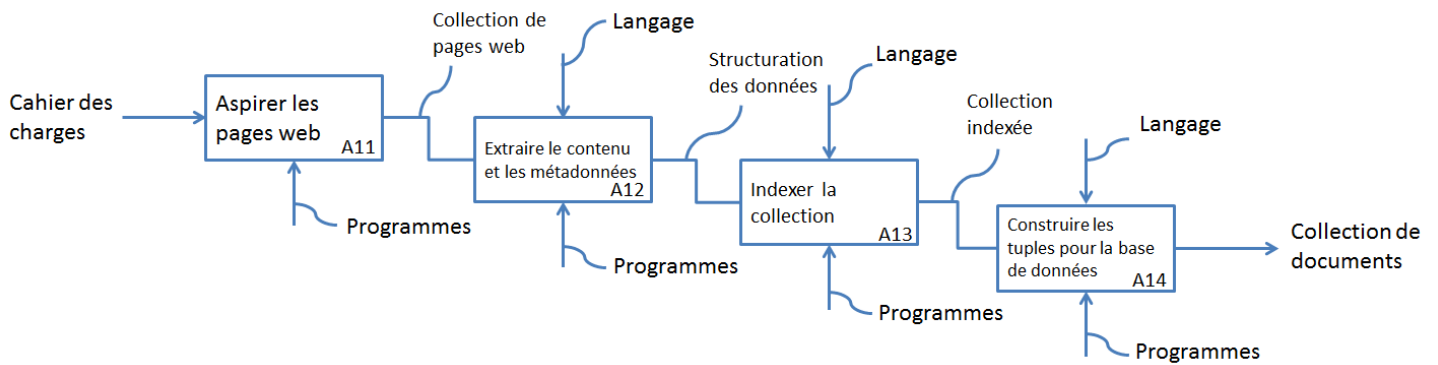
SADT permet non seulement de décrire les tâches du projet et leurs interactions, mais aussi de décrire le système que le projet vise à étudier.



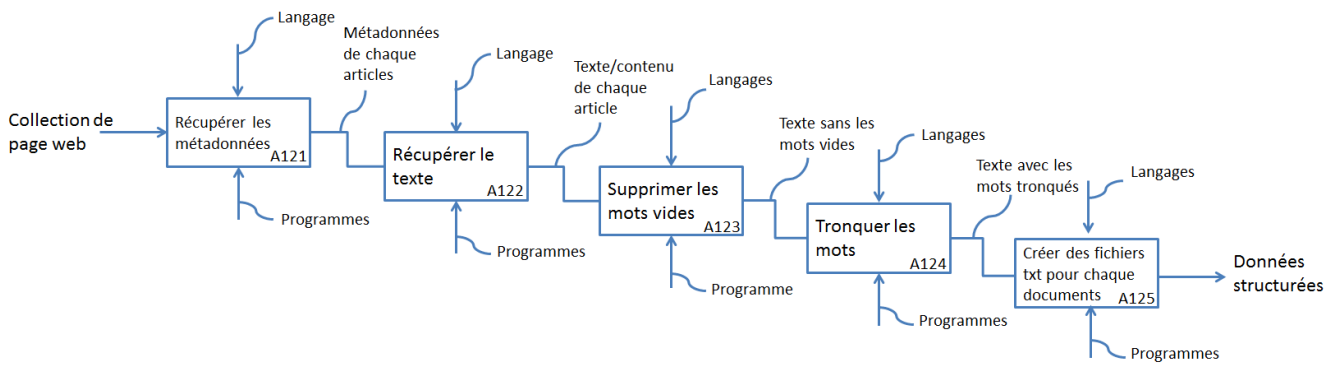
**Diagramme A\_0 : Mise en œuvre d'une système de recherche d'information**



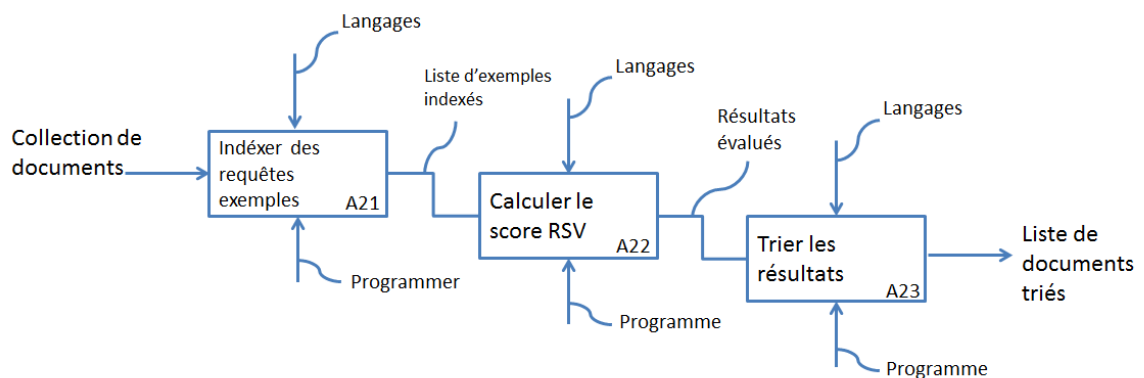
**Diagramme A0 : Mise en œuvre d'un système de recherche d'information**



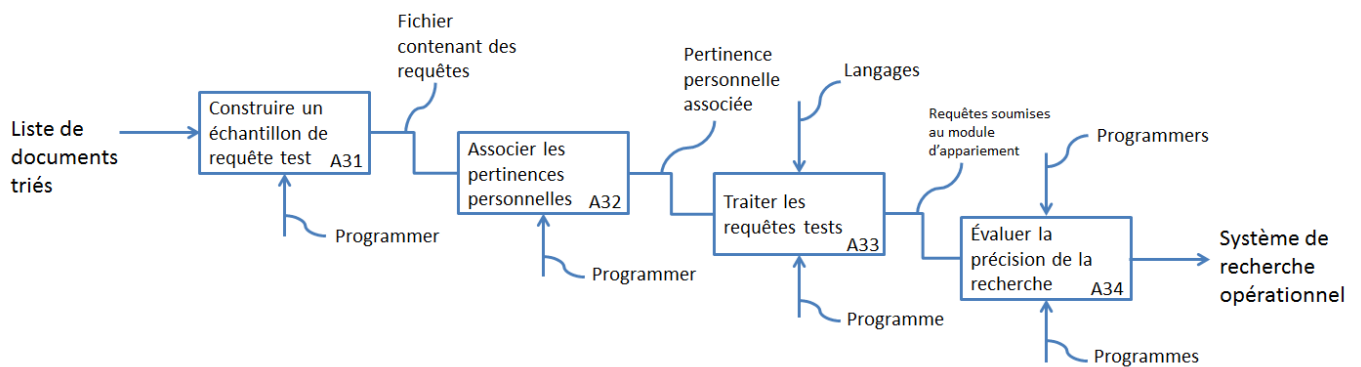
**Diagramme A1 : Préparer la collection**



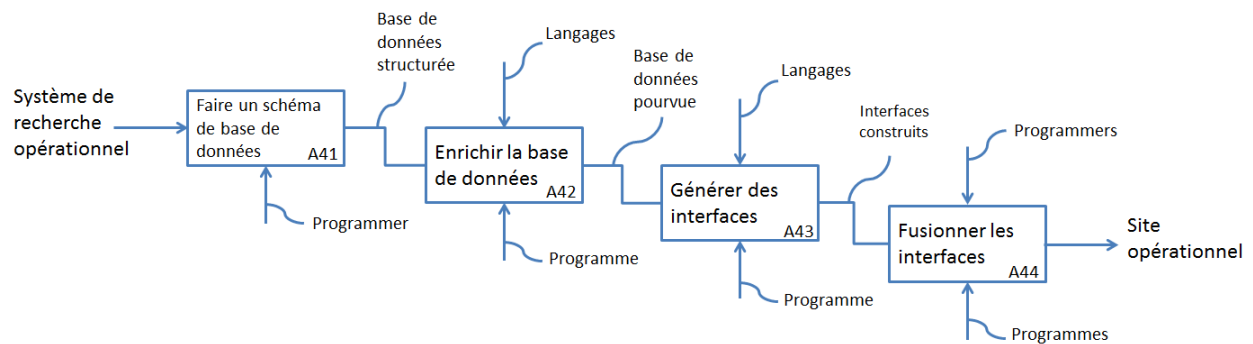
**Diagramme A12 : Extraire le contenu et métadonnées**



**Diagramme A2 : Traitements des requêtes**



**Diagramme A3 : Évaluer les performances de recherche**



**Diagramme A4 : Mise en œuvre d'une interface de recherche**

Notre projet se déroule donc en 4 grandes phases :

- Préparation de la collection
- Traitement des requêtes
- Évaluation des performances de recherches
- Mise en œuvre d'une interface de recherche

Et nous allons détailler chacune d'entre elles pour comprendre comment nous avons réussi à mener à bien le projet.

## 5.4 Préparation de la collection

### 5.4.1 Aspiration des pages web

Pour commencer le projet, nous avons dû constituer une collection de documents. Pour confectionner cette collection, nous nous sommes aidés du logiciel HTTrack.

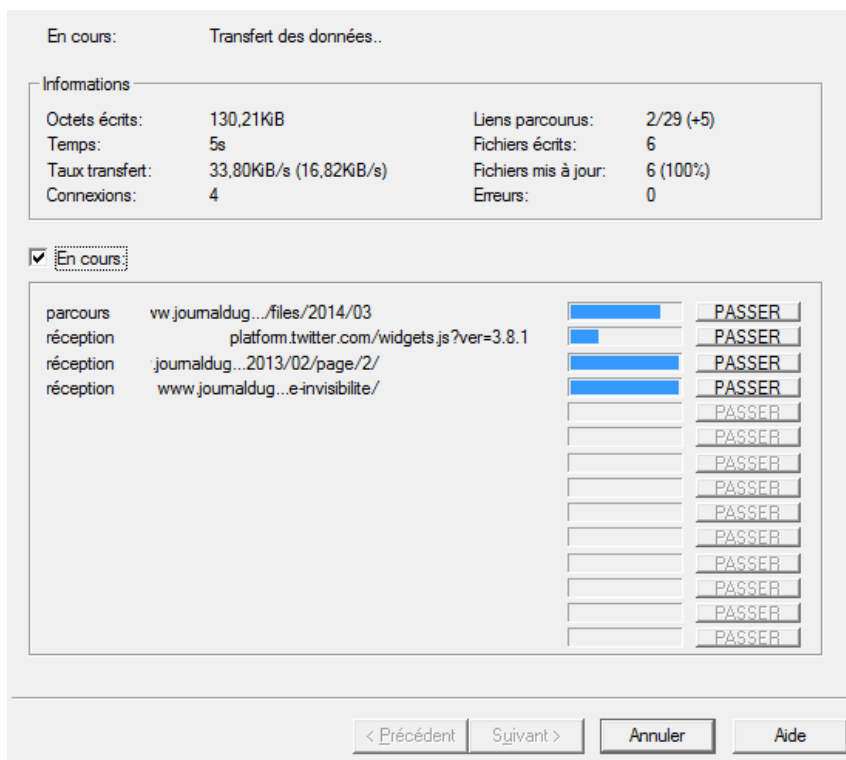


FIGURE 4.3 – Aperçu du logiciel HTTrack

Les quatre groupes de travail se sont occupés d'un des 4 mois de l'année parmi : Janvier, Février, Mars et Avril. Nous étions responsables du mois de février 2012. Nous avons le choix d'aspirer les articles du site Wiki news ou ceux du Journal du Geek. Nous avons décidé d'aspirer celui du journal du geek, car les articles nous paraissaient plus complets au niveau des métadonnées.

Une fois « l'aspiration » du mois de février réalisé par le logiciel HTTrack, nous nous sommes retrouvés avec 725 fichiers html à traiter.

Notre responsable de TP a renommé chacun des fichiers contenant les pages web aspirées pour chaque groupe pour nous faire gagner du temps, de sorte que l'on ait : 1.html, 2.html, ..., 725.html.

Une fois ceci fait, nous avons dû extraire le contenu et les métadonnées de chaque fichier.

#### 5.4.2 Extraction du contenu et des métadonnées

Nous avons donc créé un fichier Perl qui permet de :

- extraire les métadonnées (titre, auteur, date, catégories, tags, url), via des fonctions (getTitre, getAuteur,...).
- récupérer le contenu de l'article (contenu dans les balises 'body'), via la fonction getTexte qui supprime toutes les balises susceptibles de se trouver dans le 'body'.
- supprimer les mots vides (le, la, les, avec, sans,...)
- tronquer les mots à 7 caractères.
- passer tous les caractères en minuscules.
- remplacer la ponctuation par des espaces.
- enregistrer sous forme de fichier texte (1.txt, 2.txt, ..., 725.txt) le corps des articles « nettoyés » par les étapes précédentes

#### 5.4.3 Indexer la collection

Pour indexer la collection, nous avons décidé d'utiliser des tables de hachages.

Pour chaque article de la collection, nous avons donc généré du code SQL, via perl, permettant de créer une base de données.

Le script SQL regroupant les informations liées aux métadonnées de chaque article s'appuie sur le modèle Entité/Association suivant :

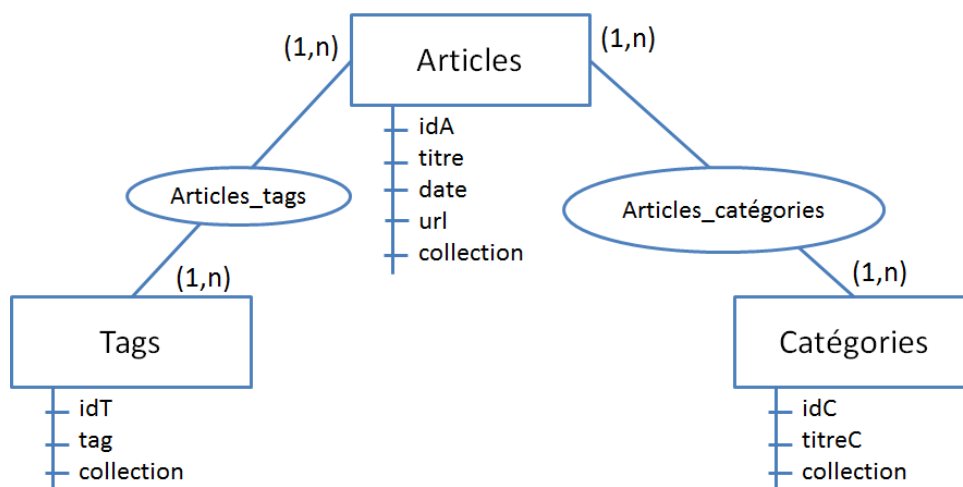


FIGURE 5.4 – Schéma E/A



Nous avons donc décidé de créer trois entités : Articles, Catégories et Tags, reliées par les associations Articles\_Tags et Articles\_Catégories.

Nous avons choisis cette configuration car cela nous permet d'associer à chaque articles, son/ses tags, sa/ses catégories.

Voici le schéma relationnel du modèle Entité/Association :

*Tags(idT, tag, collection)*

*Articles\_tags(idA, idT)*

*Articles(idA, titre, date, url, collection)*

*Articles\_catégories(idA, idC)*

*Catégories(idC, titreC, collection)*

#### 5.4.4 Construction de tuples pour la base de données

Notre fichier Perl, permet enfin de créer un fichier .txt ayant pour nom la relation qu'il doit pouvoir en tuples.

Voilà, par exemple, le fichier généré par le programme Perl, contenant les données relatives à la table Articles :

FIGURE 5.5 – Aperçu du  
fichier Articles.txt

```
322 INSERT INTO ARTICLES VALUES (334, pierre, "wikileaks devient un parti politique", "14-02-2013");
323 INSERT INTO ARTICLES VALUES (335, pierre, "xavier niel attaque un professeur d'économie en justice pour dénigrement",
324 INSERT INTO ARTICLES VALUES (336, pierre, "amazon accusé d'avoir engagé des gardes néonazis", "15-02-2013");
325 INSERT INTO ARTICLES VALUES (337, gogskable, "archos : des tablettes platinum aux écrans ips", "15-02-2013");
326 INSERT INTO ARTICLES VALUES (338, nathdahl21, "une météorite blesse 400 personnes en russie", "15-02-2013");
327 INSERT INTO ARTICLES VALUES (339, rédacteur invité, "conclusion sur le référencement sur mobile : l'outil d'appood.es"
328 INSERT INTO ARTICLES VALUES (340, pierre, "critique - a good day to die hard : la série à bout de souffle ?", "15-02-2
329 INSERT INTO ARTICLES VALUES (341, pierre, "explores le stade des colts d'indianapolis avec google street view", "15-0
340 INSERT INTO ARTICLES VALUES (342, pierre, "facebook a le droit d'imposer sa politique en allemagne", "15-02-2013");
341 INSERT INTO ARTICLES VALUES (343, nathdahl21, "le fisc réclame 52,5 millions à microsoft", "15-02-2013");
342 INSERT INTO ARTICLES VALUES (344, nathdahl21, "fujitsu présente sa tablette arrows tab ar70b", "15-02-2013");
343 INSERT INTO ARTICLES VALUES (345, kawaw, "gamestop : pas d'occasion, pas de nouvelle console", "15-02-2013");
344 INSERT INTO ARTICLES VALUES (346, pierre, "iron man 3 : iron patriot se dévoile en poster", "15-02-2013");
345 INSERT INTO ARTICLES VALUES (347, pierre, "la bataille de hoth à la maison", "15-02-2013");
346 INSERT INTO ARTICLES VALUES (348, nathdahl21, "la galaxy note 8.0 auto-leakée par samsung?", "15-02-2013");
347 INSERT INTO ARTICLES VALUES (349, pierre, "le blackberry z10 disponible chez orange", "15-02-2013");
348 INSERT INTO ARTICLES VALUES (350, pierre, "le htc one se dévoile", "15-02-2013");
349 INSERT INTO ARTICLES VALUES (351, kawaw, "les notebooks sega : attention déjà collector !", "15-02-2013");
350 INSERT INTO ARTICLES VALUES (352, nathdahl21, "microsoft n'a pas de plan b pour ses produits mobile", "15-02-2013");
351 INSERT INTO ARTICLES VALUES (353, pierre, "microsoft travaille sur une surface mini", "15-02-2013");
352 INSERT INTO ARTICLES VALUES (354, gogskable, "premières captures écran pour la montre samsung ?", "15-02-2013");
353 INSERT INTO ARTICLES VALUES (355, pierre, "ouya : 480 jeux seront compatibles à sa sortie", "15-02-2013");
354 INSERT INTO ARTICLES VALUES (356, pierre, "pal : etes-vous préparé à l'apocalypse zombie ?", "15-02-2013");
355 INSERT INTO ARTICLES VALUES (357, gogskable, "playstation 4 : la nouvelle manette en photo ?", "15-02-2013");
356 INSERT INTO ARTICLES VALUES (358, auré, "powerup 3.0 : contrôles vos avions en papier avec votre smartphone", "15-02-
357 INSERT INTO ARTICLES VALUES (359, nathdahl21, "samsung rex, des smartphones touchviss", "15-02-2013");
358 INSERT INTO ARTICLES VALUES (360, nathdahl21, "sfr : baisse des tarifs du roaming en europe, dom et etats-unis", "15-
359 INSERT INTO ARTICLES VALUES (361, fviou21, "sony vaio vpc-s23q9e/b - ultra portable 13,1", "15-02-2013");
360 INSERT INTO ARTICLES VALUES (362, pierre, "star wars episode vii : han solo de retour !", "15-02-2013");
361 INSERT INTO ARTICLES VALUES (363, gogskable, "steam pour linux : la version finale disponible", "15-02-2013");
362 INSERT INTO ARTICLES VALUES (364, pierre, "byron lamister dans x-men", "15-02-2013");
363 INSERT INTO ARTICLES VALUES (365, pierre, "un mystérieux nokia aperçu dans une publicité néerlandaise", "15-02-2013");
364 INSERT INTO ARTICLES VALUES (366, nathdahl21, "ventes de pc en baisse pour l'europe de l'ouest", "15-02-2013");
365 INSERT INTO ARTICLES VALUES (367, fabio, "[vidéo] un super mario qu'il casse des briques !", "15-02-2013");
```

A la fin de cette étape, nous avons constitué toute la collection et pouvions passer au traitement des requêtes.

## 5.5 Traitement des requêtes

### 5.5.1 Indexer des requêtes exemples

On a créé des requêtes 'tests' dans le but de faire marcher le système :

R0	voiture volante
R1	téléphones implantés
R10	Kim Dotcom mega
R11	android malware
R12	Google Glass
R13	Bon Plan smartphone
R14	smartphone vie privée
R15	Firefox OS
R16	star wars jedi
R17	hadopi streaming
R18	comparatif smartphones
R19	tablettes android
R2	mozilla vs explorer
R3	téléchargement illégal
R4	virus boîte mail
R5	free mobile
R6	Ecriture en chiffre romain
R7	humour
R8	smartphone samsung
R9	telechargement

FIGURE 5.6 – Échantillon de requête

Pour constituer cette liste de requête nous nous sommes évidemment inspiré des catégories et thèmes les plus récurrents au Journal du Geek, à savoir l'informatique et nouvelle technologie.

Puis nous avons codé le programme qui permet de calculer le score RSV grâce au modèle vectoriel et nous avons utilisé nos requêtes tests sur ce programme.

### 5.5.2 Calculer le score RSV

Pour calculer le score nous nous sommes basés sur l'inner product qui est une fonction de similitude vectorielle, dont le code est le suivant :

```
sub scoreInnerP {  
  my ($doc,@req)=@_;  
  my ($score)=0;  
  
  foreach my $mot (@req){  
    $N=N();  
    $Nt=Nt($mot);  
    $TF=TF($mot,$doc);  
    $TFIDF=TFIDF($TF,$Nt,@N);  
    $FT=FrequenceTerme($mot,@req);  
    $score+=$TFIDF*$FT;  
    # print "$mot\t$N, $Nt, $TF, $TFIDF, $FT\n";  
  }  
  # print "\n";  
  return $score;  
}
```

La fonction **N** donne le nombre de documents de la collection.

**Nt** donne le nombre de document qui contient le mot rentré en paramètre.

**TF** donne la fréquence du terme dans le document.

**TFIDF** est le poids associé au mot dans le document.

**FT** est la fréquence des mots dans la requête.

### 5.5.3 Trier les résultats

En sortie, nous avons calculé pour chaque couple requête/article retenu (20 maximum et si moins de 20 articles était pertinents nous avons ajouté aléatoirement le nombre d'article manquant), le score de RSV associé. Puis nous avons créé un fichier contenant ces résultats triés par requêtes et par score, du plus pertinent au moins pertinent.

R12 : 37	9.86058128319447
R12 : 557	9.86058128319447
R12 : 634	9.86058128319447
R12 : 513	9.1140228995326
R12 : 420	7.9353887166688
R12 : 668	7.9353887166688
R12 : 486	7.40569711380761
R12 : 211	5.88685626686129
R12 : 226	5.42412620793289
R12 : 402	5.38288874362268
R12 : 539	5.38288874362268
R12 : 203	4.80460022629893

R12 : 305	4.80460022629893
R12 : 31	4.80460022629893
R12 : 628	4.80460022629893
R12 : 235	4.39373949953754
R12 : 307	4.39373949953754
R12 : 382	4.39373949953754
R12 : 430	4.39373949953754
R12 : 705	4.39373949953754

## 5.6 Évaluation des performances de recherche

Pour cette étape nous avons repris le même échantillon de requête. (Voir figure 5.6).

Pour chaque couple requête/article du système de recherche, nous avons évalué (subjectivement), 'à la main' leur pertinence réelle.

Nous avons utilisé un système binaire pour cette évaluation :

- 0 si nous jugions le couple requête/article non pertinent
- 1 sinon.
- 

R12 : 37	1
R12 : 557	1
R12 : 634	1
R12 : 513	1
R12 : 420	1
R12 : 668	1
R12 : 486	0
R12 : 211	0
R12 : 226	0
R12 : 402	0
R12 : 539	1
R12 : 203	0
R12 : 305	0
R12 : 31	0
R12 : 628	0
R12 : 235	0
R12 : 307	0
R12 : 382	0
R12 : 430	0
R12 : 705	0

Nous avons ensuite sélectionné les 10 articles les plus pertinents parmi les 20 retournés par chaque requête, par le système de recherche.

Et nous avons calculé les valeurs de rappel et précision aux différents points de rappel.

Voici par exemple le calcul rappel/précision avec la requête n°12.

Id Doc	Pertinence	Rappel	Précision
37	x	0,14	1
557	x	0,29	1
634	x	0,43	1
513	x	0,57	1
420	x	0,71	1
668	x	0,86	1
486			
211			
226			
402			

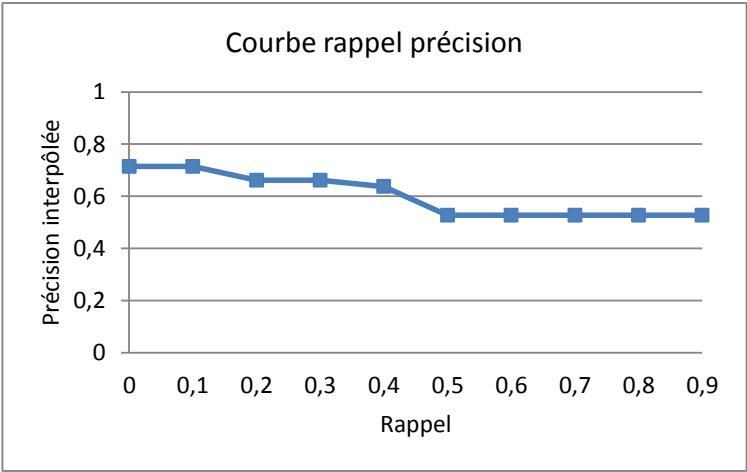
Rappel fixe	Précision interpolée
0	1
0,1	1
0,2	1
0,3	1
0,4	1
0,5	1
0,6	1
0,7	1
0,8	1
0,9	1
1	1

Nous avons fait la même chose avec les requêtes 13,18, 9, 8 et 15 et nous avons obtenus :

Rappel fixe	Précision interpolée					
	Requête n°12	Requête n°13	Requête n°18	Requête n°9	Requête n°8	Requête n°15
0	1	0,5	1	1	0,5	1
0,1	1	0,5	1	1	0,5	1
0,2	1	0,43	1	1	0,2	1
0,3	1	0,43	1	1	0,2	1
0,4	1	0,43	0,83	1	0,2	1
0,5	1	0,43	0,83	0,83	0,2	0,4
0,6	1	0,43	0,83	0,83	0,2	0,4
0,7	1	0,43	0,83	0,83	0,2	0,4
0,8	1	0,43	0,83	0,83	0,2	0,4
0,9	1	0,43	0,83	0,83	0,2	0,4

Nous avons pu en déduire la courbe rappel/précision suivante :

FIGURE 5.7 – Courbe rappel/précision



Nous n'avons pas pris les 20 requêtes par manque de temps, les 7 requêtes ont donc été choisies aléatoirement.

On peut constater que la précision est bien avec la méthode inner product que nous avons adopté. Nous sommes constamment au-dessus de 0,5 de précision, et ce pour un rappel variant de 0 à 0,9.

Pour information nous avons également calculé les P@X :

	P@5	P@10
R12	1,0	0,6
R13	0,2	0,3
R18	0,2	0,5
R9	0,29	0,57
R8	0,2	0,2
R15	0,4	0,2
Système sur ces requêtes	0,38	0,4

On constate que le système ramène environ 40% de documents pertinents parmi les 5 et 10 premiers documents.

## 5.7 Mise en œuvre d'une interface de recherche

### 5.7.1 Schéma de la base de données

Voici le schéma de notre base de données :

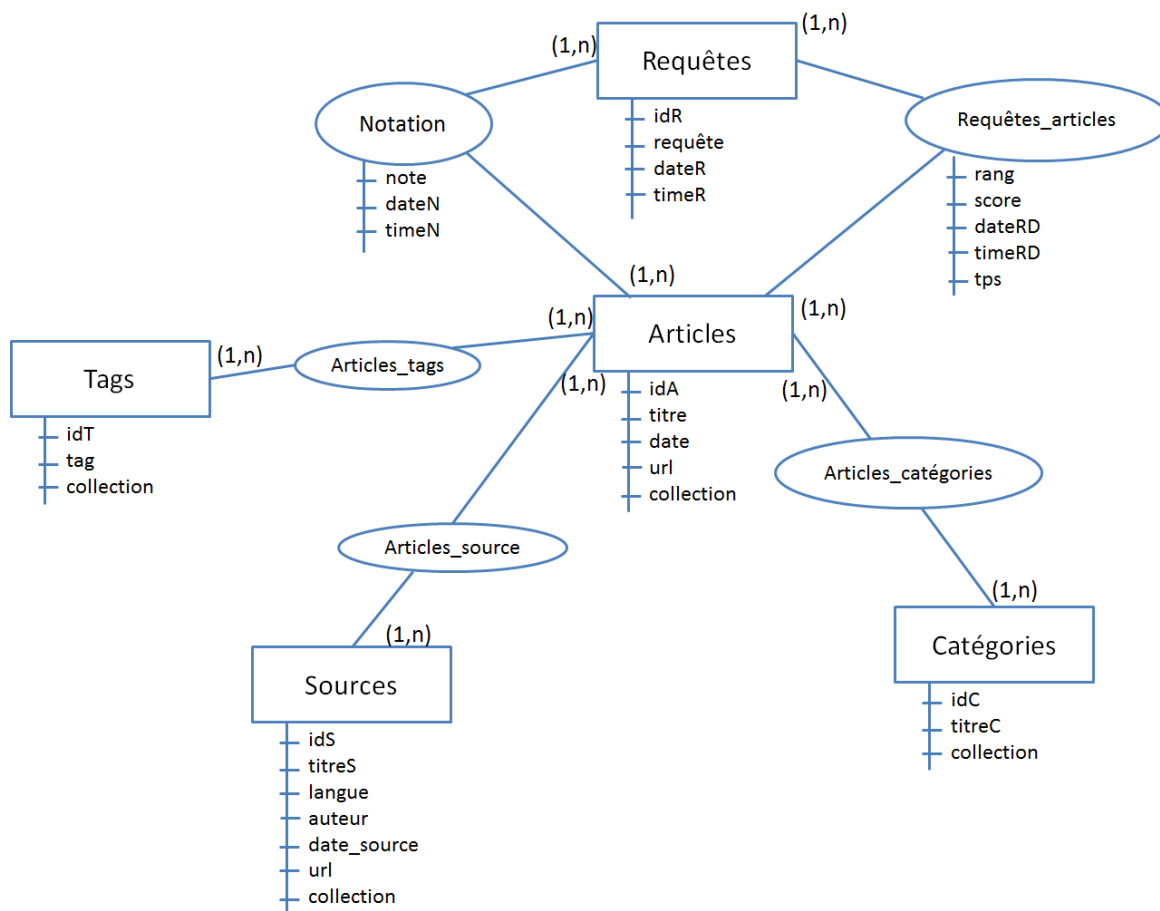


FIGURE 5.8– Schéma E/A

Pour le construire, nous nous sommes inspirés des bases de données que chacun des groupes avait préalablement construites pour leurs tests sur la collection de document (voir I- Préparer la collection).

On retrouve donc les relations Tags, Articles et Catégories de notre modèle Entité/Association initial, auquel a été ajouté les relations Sources et Requêtes. La relation

Requêtes permet donc de connaître le score de la requête en fonction de chaque article, mais nous permet également d'introduire la notion de « notation ». La notation a un impact du point de vue utilisateur. En effet, l'utilisateur a la possibilité de noter la pertinence des articles retournés (de 1 à 5 étoiles).

### **5.7.2 Enrichir la base de donnée**

Cette étape a été réalisée par Laure Soulier, à partir des données que nous avons recueillies lors de la préparation de la collection.

### **5.7.3 Génération des interfaces**

Le moteur de recherche que nous avons construit contient quatre interfaces :

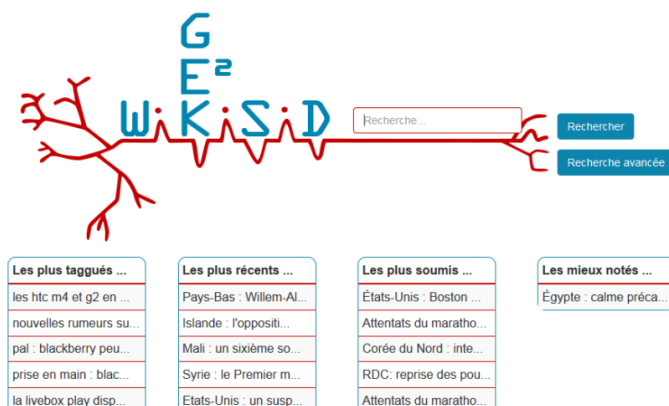
- la page d'accueil avec la recherche simple ou avancée
- la page d'affichage de résultat des requêtes
- la page de la gestion du moteur de recherche (côté administrateur) qui contient les informations concernant la recherche : pertinence des résultats, etc.
- une page de statistique de la recherche.

Les quatre groupes de projets ont créé une partie du moteur de recherche. Notre groupe s'est chargé de la page statistique.



Nous allons maintenant nous intéresser au moteur de recherche créé, et allons voir une de ses différentes interfaces.

### Page d'accueil



Le logo du moteur de recherche est inspiré de celui de la formation SID, c'est d'ailleurs pour cela que le moteur de recherche s'appelle «WikiSiD ».

La page d'accueil contient donc la barre de recherche, et quelques informations annexes à savoir : les articles possédant le plus de tags, les articles les plus récents, les articles revenant le plus dans les résultats et enfin, les articles les mieux notés.



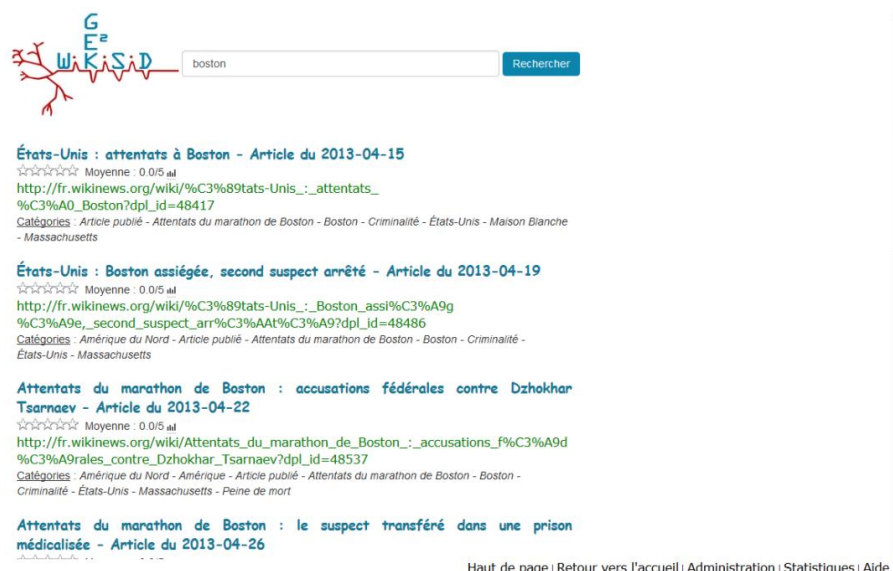
En plus de la recherche « classique », on propose également une recherche avancée qui amène à cette page :

### Recherche avancée

Requête	<input type="text" value="Requete..."/>
Titre	<input type="text" value="Titre..."/>
Catégorie	<input type="text" value="Indifférente"/> ▼
Date	<input type="text" value="AAAA-MM-JJ"/>
<input type="button" value="Rechercher"/>	

A la différence de la recherche classique, on peut spécifier vouloir des articles précis (en indiquant son titre), des résultats appartenant à une catégorie précise ou encore des articles parus à une date précise.

Une fois la recherche lancée, on tombe sur la page de résultat :



Pour chaque document, le résultat est affiché ainsi :

- Titre du document
- Possibilité de noter le document
- Moyenne de la note utilisateur
- Bouton Statistique du document
- Adresse url du document
- Catégories du document

On peut également constater qu'en bas de page, il y a des liens vers différentes interfaces du site. Les plus intéressants sont : « Administration » et « Statistiques ».

Comme son nom l'indique, le bouton « Administration » permet d'accéder à une interface de la gestion du système de recherche.

Page Administrateur

Statistiques et paramétrages

---

Statistiques

Répartition des scores doc/req

Distribution des termes

Précision / Rappel (histogrammes)

Paramétrage

Changement de collection

Changement de modèle

Quel modèle souhaitez-vous choisir ?

☒ Modèle 1 : InnerProduct

☐ Modèle 2 : Cosinus

Valider

Cette page se divise en deux parties : une partie Statistiques où l'on peut afficher des courbes rappel/précision, des graphiques contenant la distribution des termes en fonction de la collection à laquelle ils appartiennent.

Et une partie paramétrage qui permet de choisir, l'une ou l'autre des collections parmi celles disponible, à savoir : Wiki news ou Journal du Geek. Et on peut choisir également quelle mesure de similarité du modèle vectoriel que l'on souhaite appliqué à la requête.

Enfin, il y a le module statistique du moteur de recherche.

On peut accéder à cette page de deux façons différentes :

Soit en appuyant sur l'icône statistique en dessous d'un document, dans la page de résultat

Soit en cliquant sur le bouton Statistique, présenté tout à l'heure, situé en bas de la page résultat.

Voici un aperçu de la page statistique qui s'affiche quand on clique sur le bouton « Statistiques » en bas de la page de résultat.

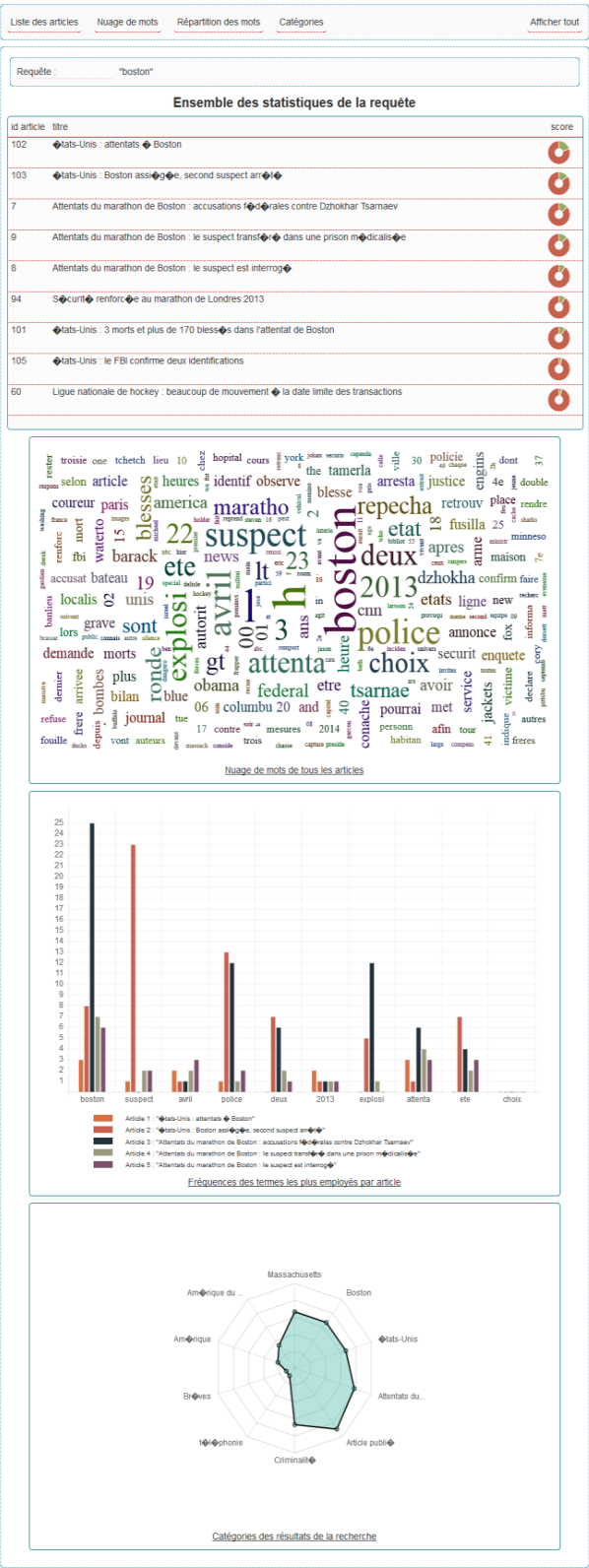
Cette page de statistiques porte sur l'ensemble des résultats de la requête formulée par l'utilisateur. Chaque graphique a été réalisé en Java Script.

On peut y trouver un tableau reprenant les documents retournés par la requête, associé à un graphique en anneau représentant la pertinence du document vis-à-vis de la requête (plus il est vers, plus le document est pertinent).

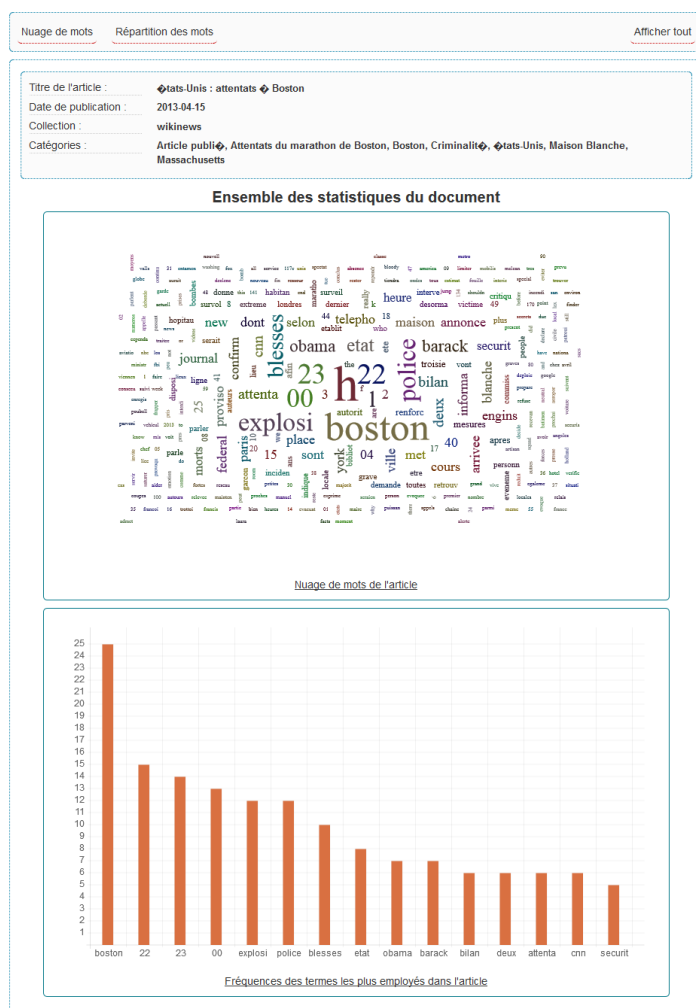
Nous avons ensuite un nuage de mot qui reprend maximum 400 mots revenant le plus souvent parmi les 4 premiers documents retournés par la requête.

Ce graphique affiche la fréquence des 10 termes les plus employés dans les 5 premiers articles retournés.

Enfin, on affiche les catégories qui reviennent le plus parmi les articles.



On accède à cette page en cliquant sur l'icône statistique en dessous d'un document.



En haut de cette page, il y a les métadonnées du document : titre, date de publication, collection et catégories.

Nous retrouvons ensuite le nuage de mots relatif au document.

Enfin, il y a un graphique récapitulant la fréquence des termes les plus employés dans l'article.

## VI- Assurance et contrôle qualité

### Revue

Revue 1	
Date: 7 février 2014	Lieu: bâtiment U1
Présent: Laure Soulier, David Jeuneau, Romain Robert	Absent:
Sujet: Validation de l'aspiration des documents avec HTTracks	
Problèmes évoqués:	Solutions envisagées:
Date de la prochaine revue: 17 février 2014	

Revue 2	
Date: 17 février 2014	Lieu: bâtiment U1
Présent: Laure Soulier, David Jeuneau, Romain Robert	Absent:
Sujet: Validation du traitement des requêtes	
Problèmes évoqués:	Solutions envisagées:
Date de la prochaine revue: 24 février 2014	

Revue 3	
Date: 24 février 2014	Lieu: bâtiment U1
Présent: Laure Soulier, David Jeuneau, Romain Robert	Absent:
Sujet: Validation de l'évaluation des performances de recherche	
Problèmes évoqués:	Solutions envisagées:
Date de la prochaine revue: 28 février 2014	

Revue 4	
Date: 28 février 2014	Lieu: bâtiment U1
Présent: Promotion entière + Laure Soulier	Absent:
Sujet: Répartition des modules au sein de la promotion	
Problèmes évoqués:	Solutions envisagées:
Date de la prochaine revue: 13 février 2014	

Revue 5	
Date: 13 mars février 2014	Lieu: bâtiment U1
Présent: Laure Soulier, David Jeuneau, Romain Robert	Absent:
Sujet: État de l'avancée de la construction du module	
Problèmes évoqués:	Solutions envisagées:
Date de la prochaine revue: non planifiée	

Revue 6	
Date: 14 mars février 2014	Lieu: bâtiment U1
Présent: Laure Soulier, David Jeuneau, Romain Robert	Absent:
Sujet: Module statistique	
Problèmes évoqués:	Solutions envisagées:
Difficulté à construire le nuage de mot en Java script	Recoder
Date de la prochaine revue: non planifiée	

## **VII- Bilans**

### **7.1 Bilan Fournisseur**

Du point de vue client, le système de recherche d'information mis en place est fonctionnel. En effet, les attentes du cahier des charges sont remplies, toutes les fonctionnalités attendues sont présentes. Il faut maintenant voir si le travail fourni sera à la hauteur des attentes du CMI.

### **7.2 Bilan Personnel**

Ce projet nous a permis de mettre en application toutes les connaissances acquises durant la formation SID. Nous regrettons d'être passé de trois à deux personnes dans le groupe ce qui a considérablement augmenté notre quantité de travail. Mais cela nous a appris à pallier toute éventualité.

Niveau programmation, ce projet a été très intéressant car nous avons dû associer plusieurs langages informatiques.

Et en ce qui concerne la gestion de projet, la méthode SADT et le cycle de vie du moteur de recherche ont été très instructifs.

### **7.3 Conclusion**

Ce projet a été mené à son terme, et est opérationnel. Ce que nous pouvons déplorer c'est le temps de recherche important quand les requêtes commencent à avoir plusieurs mots. Si nous avions eu plus de temps nous aurions pu peut-être atténuer ce problème.