# Les Technologies de l'Information et du réseau Internet au service de la Veille Technologique et de l'Intelligence Economique : WebProcess

Carine Dou [1,2], Bruno Mannina [1,2], Eric Giraud [2]

[1] *Intelligence Process, Bd Abbadie, Esp. Rossignol, Bât.2, 73730 Saint Victoret*
[2] *CRRM, Saint-Jérôme, Case 161, Av. Escadrille Normandie Niemen, 73397 Marseille Cedex 20*

carine@intelligence-process.com, bruno@intelligence-process.com, qiraud@crrm.u-3mrs.fr

**Résumé :**

Le volume et la variété des informations disponibles sur le réseau Internet alliés à la facilité d'accès **a** ces ressources, le plus souvent gratuites, contribuent à l'utilisation d'internet comme un outil privilégié du « Veilleur Technologique ».

Mais, le nombre exponentiel de ressources et d'informations disponibles sur Internet implique l'utilisation d'outils performants pour faire une Surveillance active de l'environnement informationnel de l'entreprise.

C'est dans ce contexte que WebProcess, agent intelligent de surveillance du réseau a été élaboré. Cet outil est le fruit d'une collaboration entre Intelligence Process et le CRRM, c'est-à-dire, une jeune start-up innovante et son laboratoire de rattachement.
Le principe de fonctionnement est simple et intuitif. L'utilisateur reçoit un E-mail hebdomadaire d'alerte sur les nouveautés dans le domaine étudié, et **il** aura à sa disposition sur un Internet sécurisé de l'ensemble des documents collectés.

WebProcess est en parfaite adéquation avec les évolutions que subit actuellement le réseau Internet. **Il** correspond à une utilisation professionnelle de l'information et avec sa version bridée **SearchProcess**[2] gratuite, **il** est en parfaite harmonie avec l'esprit Internet.

**Mot-Clé :**

Agent intelligent, traitement de l'information, Internet, WebProcess, SearchProcess, Bibliomètrie, Réseau Virtuel

[2] http://www.searchprocess.com

# Les Technologies de l'Information et du réseau Internet au service de la Veille Technologique et de l'Intelligence Economique : WebProcess

## 1 Introduction

De par ses qualités de communication et de diffusion de l'information, le réseau Internet est devenu la plus grande base de données d'information informelle. En outre, ses objectifs commerciaux contribuent a faire d'internet un média incontournable.

Le volume et la variété des informations disponibles alliés à la facilité d'accès aux ressources, le plus souvent gratuites, contribuent à l'utilisation d'internet comme un outil privilégié du « Veilleur Technologique ».

La figure 1 décrit à ce propos les différents types d'information auxquels un veilleur technologique est confronte.

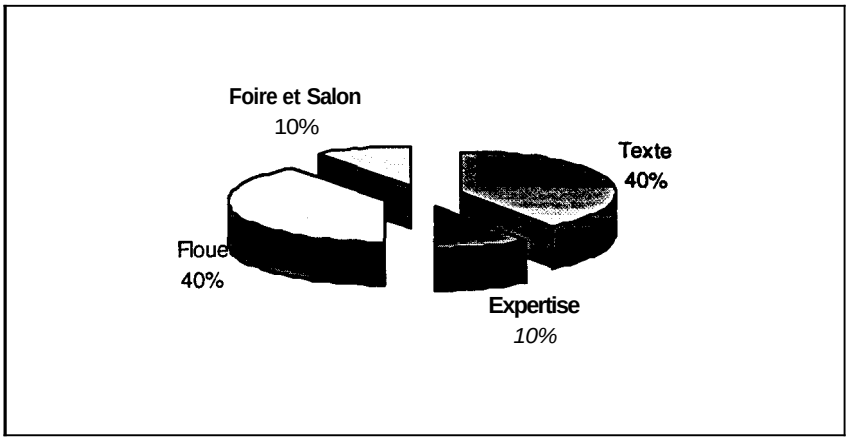


**Figure 1 : Stéréotype du marché de l'information**

L'observation de la Figure 1 permet de conclure que tous les types d'information « traditionnels » sont représentés sur Internet, seules les méthodologies de recherche et de collecte seront affectées par la nature des sources.

## II. Enjeux de l'information sur Internet

Le marché préfiguré par la transmission d'informations sur Internet revêt les caractéristiques suivantes :

- Concurrence exacerbée ;
- R&D omniprésente ;
- Information ;
- Désinformation.

Dans ce contexte, et de part sa diversité, le réseau des réseaux doit être considéré comme une source primordiale d'information. En effet quelque soit le domaine d'activité de l'entreprise, les informations que l'on peut extraire du réseau, une fois correctement valorisées, se révèlent très importantes pour

l'activité industrielle. Ainsi, une entreprise désireuse d'effectuer une partie de sa veille sur internet obtiendra de précieux renseignements sur :

- Les sites Internet des concurrents,
- Les sites Internet « non-officiel» (Rapports de stage, désinformation...)
- Les Banques et Bases de Données spécifiques au domaine,
- Les périodiques du domaine,
- La presse spécialisée ou non,
- Les sites spécialisés dans l'Intelligence Economique ou la Veille Technologique,
- Les groupes de News spécialisés,
- Les rapports de colloques et de congrès.

Cette diversité de sources en exponentielle croissance, font des ressources du réseau Internet des éléments indispensables à la stratégie de développement de l'entreprise.
Au fil de cette présentation, nous nous attacherons a décrire un outil d'exploitation de l'information a caractère scientifique, technique et économique qu'il est possible de trouver sur Internet

## III. Le besoin

Comme il a été exposé précedemment, l'Information Scientifique et Technique (IST) ainsi que de précieuses données sociales et économiques peuvent être recueillies sur Internet.
L'éceuil le plus évident est lié au nombre impressionnant de sources qu'une simple requête recense.
Une des principales préoccupations consistera, d'une part à optimiser les requêtes de l'utilisateur, et, d'autre par à filtrer judicieusement les réponses obtenues par différents moteurs de recherche.
L'autre enjeu d'un outil de collecte et de valorisation de l'information issue d'internet sera de manipuler tous les formats de documents vehiculés par les différents services du réseau.
La réalisation d'un tel outil de recherche, de collecte et d'analyse confronte le concepteur à un cahier des charges fonctionnel intégrant les contraintes suivantes :
L'outil doit être capable de :

C1 : Reformuler une requête établie par un non spécialiste
C2 : Acceder à des moteurs de recherches diférents
C3 : Garantir une actualisation des sources et des moteurs
C4 : Fournir un résultat de recherche directement exploitable
C5 : Proposer un mode d'exploitation des résultats à des fins de décision

Dans cette optique, nous proposons d'étudier les caractéristiques de l'outil WebProcess, issu de concepts élaborés par des équipes de recherche universitaires et élaboré par une société spécialisée dans le développement d'outils internet, la société Intelligence-process.

## Présentation ces s

Dans ce contexte de npétitivité accrue surveiller son environnement de façon performante et régulièr est un facteur clé de succès d'une repr inf englobe les côtés irre mi ix et ting de t WebI est un outil alliant les intelligents de nation avec les l liées à Internet.

Les reproches formules a l'Internet portent le plus souvent sur la perte de temps induite mettant en cause la lenteur du système et la surfréquentation de certains sites Internet, ainsi que la localisation des sites, la lecture et l'exploitation des ressources.

WebProcess permet d'outrepasser ces problèmes, puisque les recherches sont effectuées sur des lignes Internet spécialisées a haut débit et les résultats de recherche sont filtrés, stockés, réorganisés et mis à jour régulièrement. De plus, toutes les ressources essentielles sont détectées et intégrées directement dans les résultats des recherches.

De plus, WebProcess est un outil qui englobe toutes les étapes de la Veille Technologique :

- Une phase d'identification des sources
- Une phase de recueil de l'information
- Plusieurs étapes comprenant le traitement, la vérification, l'analyse de l'information...
- Une phase de diffusion de l'information

Les sources d'information de WebProcess sont multiples et englobent toutes les ressources disponibles sur Internet :

- Internet
- La recherche d'information sur Internet est effectuée à partir des **principaux moteurs de recherche** (Aitavista, Lycos Fr, Lycos Com, Hot Bot, Déjànews).

La restriction à ces principaux moteurs est volontaire dans la mesure où l'ajout d'autres moteurs fait perdre énormément de pertinence. Cette perte est dû au fait que les autres moteurs n'utilisent pas de syntaxe pointue de recherche comme la troncature, l'adjacence,et les recherches complexes avec parenthèses.

- Forums de discussions
- Bouquet de Bases de Données (évolutif) : Medline, INIST (Institut National de l'Information Scientifique et Technique)
- 500 revues de Presse Nationale dont la presse économique (le Monde, La Tribune.,), médicale (le Quotidien du Médecin ...), normative (le Journal Officiel Lois et Décrets), le Journal du Net...

[2] **WebProcess : outil développé par Intelligence Process : http://www.intelligence-process.com/webprocess/**

De plus, WebProcess, étant commercialisé en tant que service et non en tant que logiciel, est en constante amélioration et intègre régulièrement de nouvelles sources.

## V. PRINCIPE DE FONCTIONNEMENT

Le principe de fonctionnement est simple et intuitif. L'utilisateur reçoit un E-mail hebdomadaire d'alerte sur les nouveautés dans le domaine étudié, et il aura à sa disposition sur un Internet sécurisé de l'ensemble des documents collectés.
Le corpus global des informations ainsi collectées, est dénommé « Pôle d'information ». Ce pôle comprend les informations suivantes :

- Les différents protagonistes du domaine classés par thèmes (universités, concurrents, organismes...)
- Les experts du domaine et le contenu de leurs messages sur les groupes de discussion
- L'analyse chronologique des informations
- Les archives de Medline, de l'INIST et de la Presse
- Les archives des forums

De l'analyse de ces informations, il est possible de décrypter des signaux faibles, ainsi que des informations de la plus haute importance, à savoir :

- Les technologies émergentes
- Les réseaux de partenariats
- Les stratégies des concurrents et l'évolution de leurs images sur Internet
- La perception commerciale et marketing de l'entreprise
- etc…

Régulièrement (bi mensuellement ou mensuellement), l'utilisateur du service de WebProcess sera informé des :

- Nouveautés technologiques
- Nouveaux concurrents
- Nouveaux experts du domaine
- Nouvelles publications économiques
- Nouvelles références de Medline
- Nouvelles références de l'INIST
- Nouveaux messages des forums de discussion

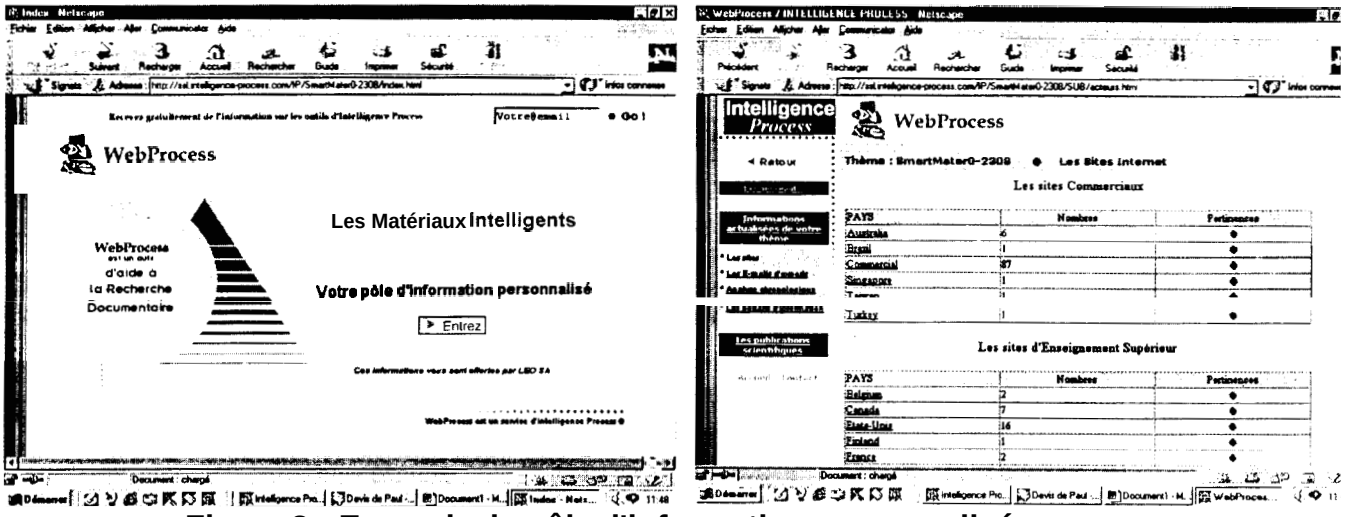


**Figure 2 : Exemple de pôle d'information personnalisé**

## VI.Caractéristiques des recherches d'information

### VI.1 Distinctions entre les différents types d'informatio

II est important de ne pas mélanger les différents types d'information présents dans le **pôle** d'information.

- Les nouveautés sont distinctes du corpus global d'informations

II est important de pouvoir extraire les nouveautés du corpus global d'information pour régulièrement se maintenir en alerte face aux nouvelles informations

- Distinction entre l'information formelle et informelle

Les informations issues des pages Internet sont distinctes des informations issues des bases de données scientifiques et de la presse.
WebProcess offre dans un même ensemble l'information informelle provenant d'Internet, très actuelle mais brute avec l'information formelle des bases de données, structurée et validée mais plus ancienne car elle est passée dans le filtre du domaine éditorial

### V1.2. les Traitements de l'information

- Possibilité de télédécharger la recherche en format bibliométrique (pour des traitements de Text-Mining)

Ainsi, il est possible d'intégrer toutes les références dans un logiciel spécifique pour analyser de façon très pointue les informations issues du réseau.
De nombreux champs sont extraits des pages HTML comme :
Titre, pertinence, nombre de mots dans la page, Méta Tags, Mots-Clés, domaine du serveur, date, date de modification de la page, environnement des mots-clés, taille du fichier, mails de la page, chemin d'accès...

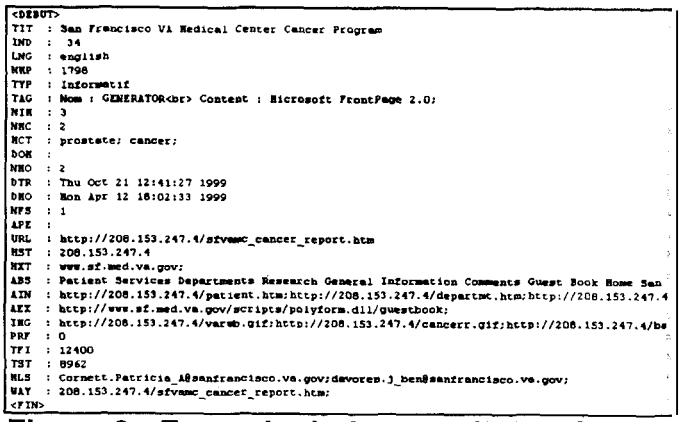


**Figure 3 : Exemple de format bibliométrique des pages internet**

- Cartographie des liens entre **les** sites

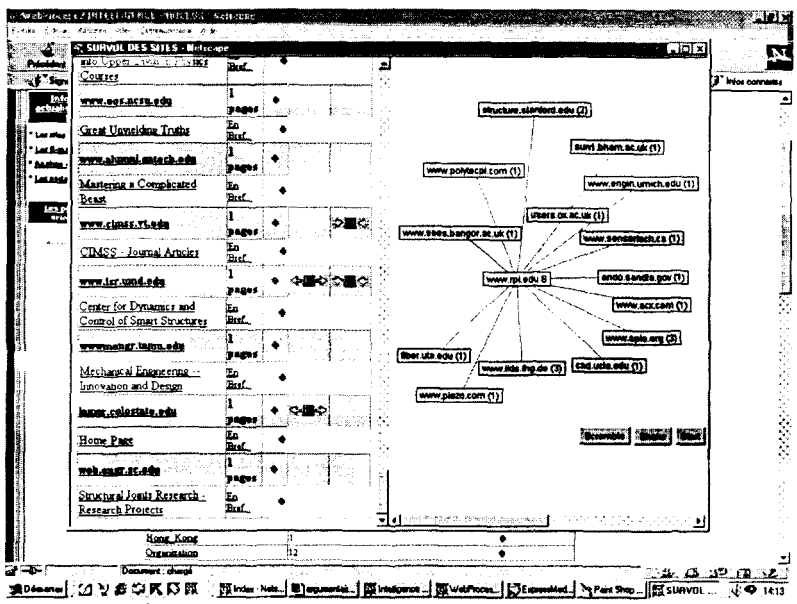


**Figure 4 : Exemple de réseau virtuel**

Ce type de schéma permet d'avoir **un** aperçu rapide des réseaux virtuels qui peuvent exister entre les différents protagonistes du domaine.

Il est maintenant possible de connaître le degré de notoriété d'un site (quand **il** y a beaucoup de pages Internet qui le citent en référence), ainsi que le réseau de travail d'un site (quand **il** fait référence à beaucoup de sites). Ainsi les sites portails sont rapidement détectés.

- **Récupération de toutes les adresses E-Mail présentes sur les pages et contenu de leurs messages dans les groupes de discussion**

Cette fonction permet de connaître tous les E-mails des pages qui sont dans le pôle et de les analyser dans les groupes de discussion. Il sera donc possible de détecter les réels experts du domaine en analysant leurs messages dans les Newsgroups, mais aussi de détecter les groupes de discussion qui sont réellement intéressants.

- **Analyse chronologique des informations : année, trimestre et mois**

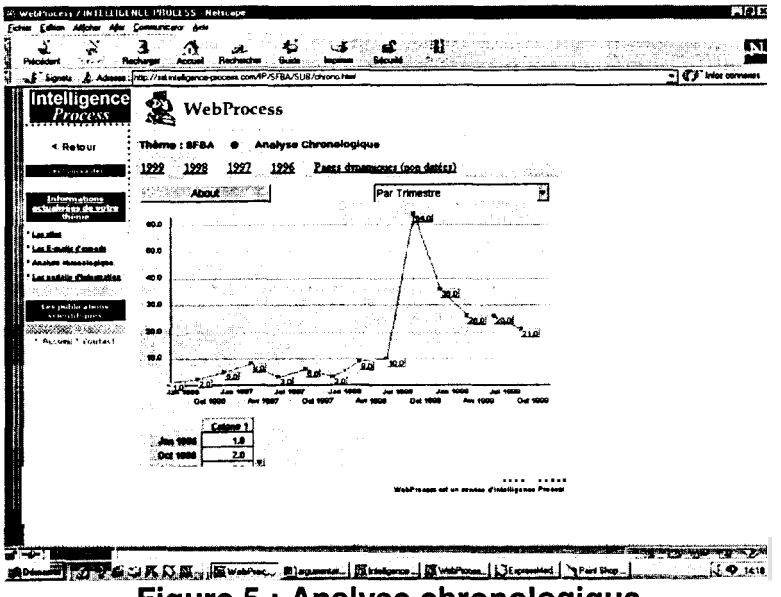


**Figure 5 : Analyse chronologique**

Cette analyse permet de cerner l'évolution de son domaine.

D'autres part les sites précurseurs du domaine seront facilement détectables, ainsi que ceux qui n'ont aucun dynamisme sur leurs sites Internet.

- **Extraction des sites portails de références (Gateways)**

Les sites portails sont en train de devenir des points d'entrée de l'internaute incontournable.
De plus, le temps d'indexation sur les moteurs de recherche et index est tellement long qu'il est souvent préférable d'utiliser ce genre de pages portails pour trouver l'information.

Sur cette page, le nombre de liens externes ainsi qu'un degré de pertinence est calculé.

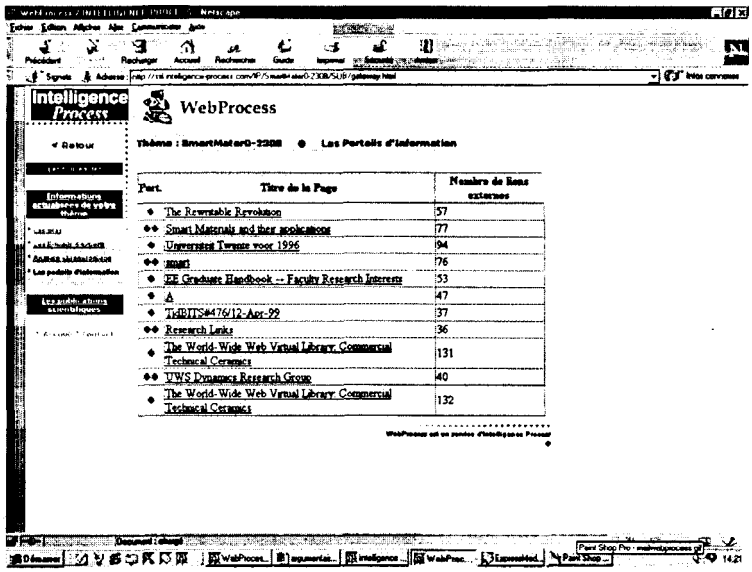


Figure 6 : Les portails i ti

## V1.3. é de recherche de l'information dans le pôle

Pour naviguer dans un corpus de document qui peut atteindre un nombre de pages infini, il est nécessaire de permettre à l'utilisateur de WebProcess de retrouver rapidement l'information spécifique qu'il recherche dans le pôle.

- Mise en sur brillance des mots-clés recherchés
- Moteur de recherche interne sur les pôles d'information
- Traduction des adresses internet des sites pour une meilleure interprétation de l'information
- Calcule un score de pertinence
- Elimine les doublons et les liens morts

## VI.4. L'interactivité de WebProcess (Phase de diffusion)

Cette phase de diffusion de l'information représente autant le push d'information que WebProcess envoie régulièrement **a** l'utilisateur (les nouveautés de son domaine), mais aussi, l'utilisateur a la possibilité de diffuser de l'information annotée de commentaires, c'est le rapport d'étonnemnt.

- Alerte périodique par E-mail sur les nouveautés du domaine

WebProcess envoie régulièrement toutes les nouveautés qu'il a détectées sur le réseau.

- Possibilité de créer automatiquement des rapports d'étonnement

Les rapports d'étonnement peuvent être sous forme de page HTML ou d'E-mail.

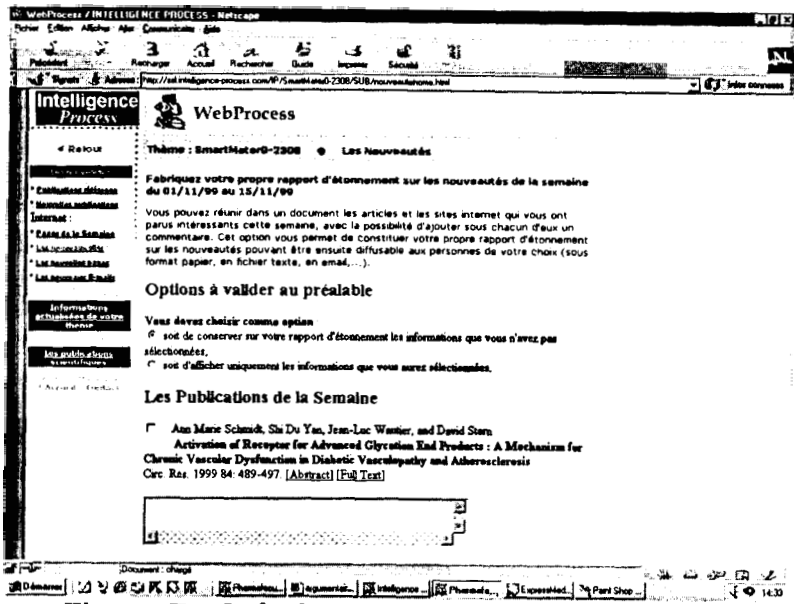


**Figure 7 : Création du rapport d'étonnement**

Les nouveautés détectées par nos outils peuvent être validées par des experts ou des veilleurs technologiques afin d'apporter une plus value supplémentaire.

Cette page ou cet e-mail, une fois validé, génèrent automatiquement un rapport d'étonnement qui sera envoyé en e-mail aux personnes concernées, et/ou apparaîtra en page HTML sur un serveur Internet ou Intranet.

En interne, ces rapports d'étonnement ainsi que les pôles d'information sur tous les sujets critiques (concurrence, domaines de R&D interne) constituent des bases de connaissances qu'il est important d'exploiter dans un système de gestion de l'information interne, de type Intranet, incluant des forums et outils de communication performants.

## VII. Les avantages supplémentaires de WebProcess

Par rapport aux outils existants sur le marché, WebProcess est un le seul permettant d'accéder directement au service clé en main. En effet, l'utilisateur de WebProcess n'aura aucune formation préalable à prévoir : Pas besoin de connaître la méthodologie pour les stratégies de recherche, ainsi que les opérateurs, pas besoin de se former à l'utilisation d'un logiciel complexe.
De plus, les recherches sur Internet se font sur des lignes spécialisées à haut débit, l'utilisateur n'a donc pas à être constamment connecté à Internet pour effectuer ses recherches. Ce gain d'argent et de temps est confirmé dans la mesure où les résultats des recherches peuvent être téléchargés pour être consultés « Hors Ligne ».

En ce qui concerne directement les résultats de recherche, WebProcess permet

de consulter sur une seule page l'intégralité des sites lnternet qui ont été détectés dans le domaine. Cet avantage donne a l'utilisateur de WebProcess la vision d'ensemble des données.

De plus, WebProcess se situe a la frontière des agents intelligents et des logiciels bibliomètriques. En effet, WebProcess est le fruit d'une active collaboration d'étudiants en doctorat et chercheurs au CRRM[3]. L'application au marché de concepts développés au sein de l'université a permis à AURESYS[4] de s'adapter aux réels besoins des entreprises.
Une permanente collaboration entre cette jeune start-up et le CRRM permet à Intelligence Process de rester activement à l'écoute des dernières nouveautés en matière de traitement de l'information et de Bibliomètrie.

En conclusion, WebProcess, de part ses Caractéristiques, est en parfaite adéquation avec les évolutions que subit actuellement le réseau Internet. En effet, le réseau lnternet tend vers trois caractéristiques principales : la division d'lnternet entre la partie ludique et la partie professionnelle, l'entrée des internautes par des sites portails et la gratuité sur Internet.
WebProcess correspond à une utilisation professionnelle de l'information et avec sa version bridée SearchProcess[5] gratuite, il est en parfaite adéquation avec la Netiquette.

## Bibliographie :


"A smart ltsy Bitsy Spider for the Web"
*Hsinchun, Yi-Ming Chung, Marshall Ramsey, C.C. Yang,* Journal of the american society for information science, May 15, 1998

"AURESYS 2.0 : Un agent Intelligent au service de l'information stratégique"
*Quoniam Luc, Bruno Mannina, Dou Henri,* CRRM, SFBA'97, Ile Rousse

"Construction automatique de réseaux : un outil pour mieux appréhender l'information provenant de I'lnternet"
*Eric Boutin, Bruno Mannina, Hervé Rostaing et Luc Quoniam ;* JADT, Février 1998

"Veille Technologique et Compétitivité", H. Dou, 1995, Dunod

"Organisation Intelligente et Système d'Information stratégique", **J-A.** Bartoli, **J-L.** Le Moigne, Economica, 1996.

"Systèmes d'Information *et* Management des Organisations." Robert **Reck.** Vuibert Gestion. 1995.


[3] **CRRM : Centre spécialisé en Veille Technologique et Intelligence Economique** : http://crrm.u-3mrs.fr
[4] **AURESYS** : AUtomated REsearch SYStem **Agent Intelligent développé par B. Mannina au CRRM**
[5] http://www.searchprocess.com