



GRILLE D'AIDE A L'EVALUATION DES OUTILS DE TAI

SOMMAIRE

Sommaire	1
Méthodologie d'utilisation	2
1. Objectifs :	2
2. Postulats de départ.....	2
3. Définition des grandes étapes du cycle de l'information	2
Les fonctionnalités.....	4
1. Orientation.....	4
2. Recherche de sources.....	4
2.1. Définition des sources.....	4
2.2. Capacité à gérer les sources	4
3. Surveillance et collecte	4
3.1. Critères communs à l'ensemble de ces fonctionnalités.....	4
3.2. Critères propres à la surveillance	4
3.3. Critères propres à la recherche, la collecte/aspiration/crawling.....	5
3.4. Fonctionnalités de la recherche.....	5
4. Traitement et analyse	5
4.1. Prétraitement des documents bruts collectés	5
4.2. Indexation, classement et regroupement.....	6
4.2.1. Critères communs.....	6
4.2.2. Capacité de classement de l'information.....	6
4.2.3. Capacité de regroupement de l'information.....	6
4.3. Stockage et recherche	6
4.3.1. Fonctionnalités de la recherche	6
4.4. Analyse des informations.....	6
4.4.1. Fouille de données textuelles et structurées (text-mining et data-mining).....	7
4.4.2. Résumé automatique	7
4.4.3. Annotations et commentaires	7
4.4.4. Visualisation de l'information et représentation cartographique	7
5. Elaboration et diffusion des informations enrichies.....	7
5.1. Préparation et rédaction du produit de la veille	7
5.2. Diffusion du produit de la veille	7
6. Fonctionnalités transverses au cycle de l'information	8
6.1. Sécurité de la plate-forme	8
6.2. Traduction	8
6.3. Pilotage et suivi de la plate-forme.....	8

METHODOLOGIE D'UTILISATION

1. OBJECTIFS :

La présente grille de critères des fonctionnalités des outils de traitement avancé de l'information s'adresse en priorité aux administrations et aux entreprises désireuses d'évaluer leurs outils et/ou de définir leurs besoins en vue d'une éventuelle acquisition.

L'objectif est qu'elle soit suffisamment technique pour les utilisateurs avertis désireux d'évaluer leurs outils et largement pédagogique et explicite pour les non-initiés qui doivent définir un cahier des charges techniques.

Cette grille est structurée autour du cycle de l'information, cercle vertueux de l'acquisition et du traitement de l'information, décrit ci-après.

Toutefois, quatre postulats ont conduit la démarche du groupe de travail.

2. POSTULATS DE DEPART

- Les membres sont partis du principe que toute veille débute impérativement par une orientation des besoins préalablement définis par un niveau hiérarchique supérieur. Cette orientation donne le cap à suivre et il revient ensuite aux veilleurs de mettre en marche le processus de veille en sélectionnant les sources, en effectuant la collecte ou la surveillance de celles-ci, en analysant puis en transmettant les informations à valeur-ajoutée au niveau supérieur de façon à ce qu'il puisse soit affiner soit réorienter ce processus.

L'étape de l'“*orientation*” est donc une phase d'impulsion et d'aiguillage plus que de traitement de l'information comme peuvent l'être les quatre étapes suivantes. C'est pourquoi, l'orientation est excentrée par rapport aux autres mais ne doit pas en perdre pour autant de sa valeur.

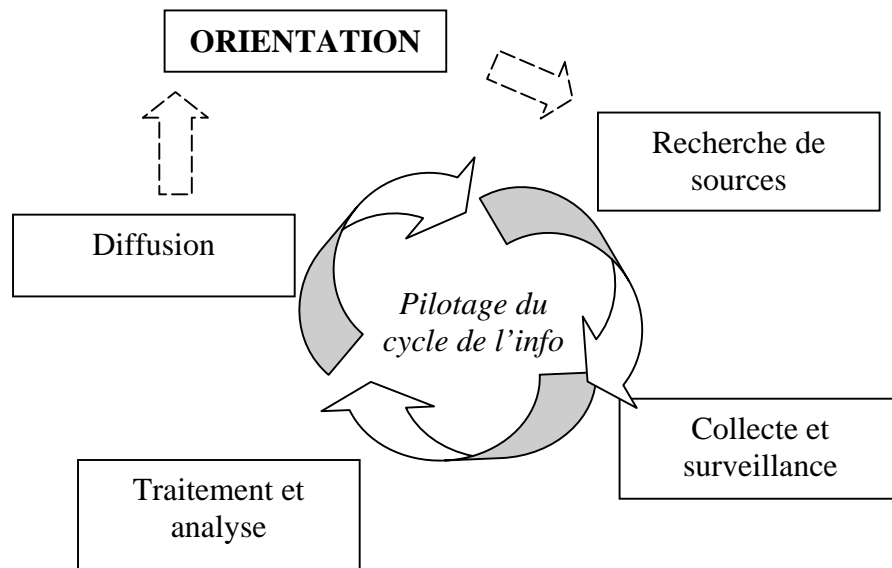
- Compte tenu des impératifs de sécurité inhérents à la confidentialité et à la gestion de l'information stratégique pour l'administration ou l'entreprise, il a été choisi de placer la sécurité au cœur de ce cycle car elle impacte ses quatre étapes. Pour autant, suivant le besoin de sécurité qui aura été préalablement défini, il conviendra d'en faire une fonctionnalité à part entière ou un élément réparti à chacune des étapes.
- Compte tenu de la nécessaire “administration” de tout logiciel informatique, les membres ont également placé cette fonctionnalité au cœur du cycle à l'instar de la sécurité.
- Tout au long de ce cycle de l'information, un impératif doit animé le veilleur : la valeur des sources. En effet, il doit impérativement surveiller ces sources d'informations en s'attachant à les coter et à les hiérarchiser compte tenu de la légitimité de leur émetteur et de la criticité de leur contenu. Cette cotation s'inscrit, d'ailleurs, tout au long du cycle de l'information pour des raisons évidentes de contingence. C'est pourquoi, il convient de construire des indicateurs de performance pour prendre en considération cet impératif.

3. DEFINITION DES GRANDES ETAPES DU CYCLE DE L'INFORMATION

- **L'orientation**, se traduisant sous la forme d'un plan de veille, est l'opération qui vise à délimiter la partie de l'environnement que le décideur veut mettre sous surveillance. Cette étape a donc pour objectif de rendre la veille la plus efficace possible. Cette étape relève

plus particulièrement de la direction générale de l'entreprise ou du niveau décisionnaire équivalent.

- **La recherche des sources ou "sourcing"** vise à identifier et hiérarchiser les sources définies pour les exploiter en les traduisant sous formes d'objectifs de traque ou de recherche. L'aspect récurrent de cette étape est très présent puisque du "sourcing" dépendent étroitement les résultats obtenus par la veille.
- **La collecte et la surveillance** est l'opération par laquelle le veilleur recueille le fruit des informations à veiller. Dans la mesure où les plus intéressantes ne viennent pas d'elles-mêmes, il faut faire l'effort volontariste d'aller au-devant d'elles et parfois de les provoquer.
- **Le traitement et analyse** est l'opération qui consiste, parmi la masse d'informations recueillies, à n'en retenir que celles susceptibles d'intéresser l'organisation, notamment sur la base de la stratégie de veille adoptée par elle. Par ailleurs, l'analyse valorise l'ensemble de la chaîne de la veille puisque son produit constitue l'information stratégique. L'organisation de cette fonction tient compte de ses différentes applications, qu'il s'agisse de corréler, de graduer ou encore d'apprécier la valeur d'une information par rapport à un événement.
- **La diffusion** est l'opération par laquelle l'information valorisée et mise en forme est adressée à son utilisateur potentiel.



LES FONCTIONNALITES

1. ORIENTATION

L'orientation constitue la première et la dernière étape qui boucle le cercle vertueux de l'information. Elle doit, par conséquent, obéir à des impératifs :

- ⇒ Expression fine du besoin définie par la hiérarchie
 - Orientations thématiques des axes de recherches
 - Définition des destinataires des fruits de la veille
- ⇒ Expression des besoins de sources (exemples)
 - La source Internet suffit-elle ?
 - Abonnement à des bases de données (économique et financière, presse internationale...)
 - Capteurs humains au sein de l'organisation : colloques, salons, rapports d'étonnement...

2. RECHERCHE DE SOURCES

2.1. Définition des sources

- Sites Internet et blogs
- Forums et newsgroups
- Listes de diffusion et info-lettres
- Fils RSS
- Bases de données internes et externes et SGBD
- Contenu multimédias (vidéo, son, parole)

2.2. Capacité à gérer les sources

- Capacité à gérer les sources définies au point 1.1
- Possibilité d'effectuer des recherches sur les métadonnées
- Possibilité d'effectuer des recherches sur le contenu

3. SURVEILLANCE ET COLLECTE

3.1. Critères communs à l'ensemble de ces fonctionnalités

- Capacité à surveiller et collecter l'ensemble des sources préalablement définies
- Capacité d'anonymisation ou de furtivité lors de la surveillance et de la collecte
- Traitement des virus informatiques (virus, chevaux de Troie, keyloggers...)
- Fonction de mise à jour automatique ou à la demande des sources
- Eventail de gestion des formats de documents (Suites Office, Adobe, Real, Streaming, Flash, RSS, htm, PHP, JSP...)
- Eventail des types de formats et de protocoles supportés (NNTP, POP3,
- Gestion des sites nécessitant une authentification

3.2. Critères propres à la surveillance

- Paramétrage de la surveillance (nombres de phrases modifiées, lien modifié, pourcentage de contenu modifié, images, page disparue)

- Levée d'alerte par l'apparition de nouveaux mots, de métadonnées, d'association de mots-clés et/ou de changement de zones géographiques sur les pages
- Archivage des versions des pages modifiés

3.3. Critères propres à la recherche, la collecte/aspiration/crawling¹

- Finesse de gestion de la profondeur de l'exploration soit par arborescence des pages soit par celle des liens contenus dans une page
- Capacité de coter l'information collectée selon sa fiabilité, sa fraîcheur
- Préservation du format natif du document collecté
- Qualité de l'archivage des données brutes collectées
- Interfaçage avec les principaux agrégateurs de presse
- Fil RSS

3.4. Fonctionnalités de la recherche

- Agents intelligents
 - Exploration automatique de liens successifs avec tests de pertinence
 - Contournement de balises anti-robots
- Traitement du web invisible
 - Accès aux pages non indexées
 - Accès aux ressources situées à des emplacements non-standards (ex. : ports autres que le 80)
- Connecteurs, dont gestion de l'accès aux services commerciaux (compte, password)
- Typologie des modes de formulation des requêtes :
 - Requête en mode "plein texte"
 - Requête en mode "mot clé"
 - Requête en mode "multi-index"
 - Requêtes par des opérateurs booléens
 - Requête en mode « troncature » et par "expression exacte"
 - Requête en mode "recherche approchante"
 - Requête en mode "recherche par similarité"
- Typologie des caractéristiques de la recherche
 - Capacité à interrogation des 10 premiers moteurs du Web
 - Capacité à gérer les erreurs 404 (= gestion des pages non trouvées)
 - Capacité à gérer les proxy pour la furtivité de la recherche
 - Possibilité d'ajouter de nouveau moteur
 - Possibilité de sauvegarder les requêtes

4. TRAITEMENT ET ANALYSE

4.1. Prétraitement des documents bruts collectés

- Traitement de la parole : identification de la langue et/ou du locuteur, transcription de l'oral vers l'écrit, synthèse de la parole
- OCR (Optical Character Recognition)
- Traitement de l'image avec identification de personne, objet natif ou déformé
- Traitement de la vidéo avec découpage en séquence
- Traitement du langage naturel avec détection des entités nommées (date, organisation, personne, lieu...)

¹ Cette fonction est un rapatriement des données brutes.

4.2. Indexation, classement et regroupement

4.2.1. Critères communs

- Prise en compte du thesaurus métier et/ou des terminologies métier
- Capacité à traiter l'information en mode multilingue et cross-lingue
- Nombre de langues traitées
- Etude sur la variation de l'apparition de terme et/ou de thèmes clés.
- Extraction des métadonnées

4.2.2. Capacité de classement² de l'information

- Simplicité de gestion du plan de classement
- Qualité de la performance du classement
- Validation de la proposition de classement par l'utilisateur
- Auto-apprentissage du classement pour l'outil
- Possibilité de rattacher un document à plusieurs classes

4.2.3. Capacité de regroupement³ de l'information

- Degrés de paramétrage et gestion des paramètres
- Capacités de détection de signaux forts et faibles

4.3. Stockage et recherche

Il s'agit de rechercher les informations qui sont contenues dans les lieux de stockage des informations préalablement collectées

4.3.1. Fonctionnalités de la recherche

- Typologie des modes de formulation des requêtes :
 - o Requête en mode "plein texte"
 - o Requête en mode "mot clé"
 - o Requête en mode "multi-index"
 - o Requêtes par des opérateurs booléens
 - o Requête en mode « troncature » et par "expression exacte"
 - o Requête en mode "recherche approchante"
 - o Requête en mode "recherche par similarité"
- Typologie des caractéristiques de la recherche
 - o Possibilité de sauvegarder les requêtes
 - o Historisation des formulations des requêtes
 - o Proposition de stratégie de recherche pour aider l'utilisateur
- Typologie des critères d'évaluation⁴
 - o Appréciation du bruit
 - o Appréciation du silence
 - o Affichage des résultats (critères de tri, de pertinence, de cotation)

4.4. Analyse des informations

² Le classement consiste à rattacher un document à un plan de classement utilisé par l'utilisateur

³ Le regroupement consiste à regrouper des documents en classe de similarité générée automatiquement par la machine.

⁴ Compte tenu des campagnes d'évaluation, il existe des critères d'évaluation des fonctionnalités de la recherche. Se référer notamment à TREC, CLEF, EVALDA

4.4.1. Fouille de données textuelles et structurées (text-mining et data-mining)

- Langues prises en compte
- Nécessité ou non de restructurer l'information (reformatage, saisie des données structurées, introduction de nouvelles balises...)
- Utilisation ou non des ressources linguistiques (dictionnaires métiers, thésaurus, ontologie...)
- Croisement et corrélation des informations et des données (hommes, organisations, brevets...)

4.4.2. Résumé automatique

- Langues traitées
- Type de résumé automatique : par reformulation ou par extraction
- Informations générales sur le texte (nb de mots, phrases)
- Reconnaissance de synonymes (accès aux synonymes en corrélation avec les thèmes descripteurs ou termes génériques d'un terme donné)
- Modification dynamique de la taille du résumé
- Surlignage dégradé
- Entrer ses propres mots ou expressions pour influencer le résumé
- Exclure ses propres mots ou expressions pour influencer le résumé

4.4.3. Annotations et commentaires

4.4.4. Visualisation de l'information et représentation cartographique

- Edition des liens de causalité et/ou de contradiction
- Visualisation graphique de l'importance et de la fiabilité
- Représentations visuelles de résultats de text mining et data mining
- Mode d'analyse cartographique (géographique, temporelle, thématique, des réseaux...)
- Exportation des graphiques
- Visualisation des informations sous différentes formes de graphiques

5. ELABORATION ET DIFFUSION DES INFORMATIONS ENRICHIES

5.1. Préparation et rédaction du produit de la veille

- Prise en compte d'un travail collaboratif (workflow, wikies...)
- Prise en compte d'une charte graphique ou de normes de présentation (évolutivité...)
- Capacité à intégrer des données (Cartes, Images, Texte...)

5.2. Diffusion du produit de la veille

- Acheminement des flux de documents et circulation de l'information
- Distribution, dispatching et échanges des documents
- Moyens de communication interne et externe
- Boîte à lettre, abonnement, accès au document
- Affection/traitement extension automatique ou manuelles des documents
- Validation avant diffusion
- Signature des documents
- Transmettre et communiquer le produit de la veille (PUSH)
- Mettre à disposition le produit de la veille (PULL)

6. FONCTIONNALITES TRANSVERSES AU CYCLE DE L'INFORMATION

6.1. Sécurité de la plate-forme⁵

- Gestion des droits d'accès et des utilisateurs
- Sécurisation de la transmission des données

6.2. Traduction

- Traduction automatique
- Traduction assistée
- Requête interlingue (crosslingual)
- Prise en compte d'un langage non académique

6.3. Pilotage et suivi de la plate-forme

- Indicateurs statistique
- Administration de la veille
 - o Maintenance des ressources utilisées (dictionnaires, plan de classement...)
 - o Maintenance des profils de veille
- Capitalisation
 - o Gestion de l'entrepôt de données
 - o Archivage des produits de la veille
 - Fonctionnalités permettant la conservation des diffusions
 - Fonctionnalités permettant la recherche puis la consultation

⁵ à mettre en coordination avec la politique SSI de l'organisation