

Etablissement d'une grille d'évaluation des outils de veille

Cette étude a été conduite en utilisant différentes sources et en exploitant des résultats obtenus sur le terrain à partir de tests menés en parallèle avec Web Site Watcher et KB Crawl. Cette grille a été élaborée en tentant d'être le plus exhaustif possible, en tenant compte des grilles existantes. Les présentations des rubriques des différentes grilles existantes dans la littérature et des contenus sont très variables. Cette grille suit un plan logique d'analyse, de déroulement de prise en main et d'utilisation potentiel.

Sources utilisées:

- [Beauvieux] [Digimind] [Dutheil] [Dutkiewicz-Dessaignes] [Meingan] [Texier]
- Le site « benchmarking des outils de veille » site INIST-CNRS: <http://outils.veille.inist.fr/>
- « Grille d'évaluation des produits » du groupe de travail « veille automatisée » du réseau ARPIST (compte-rendu de réunion en novembre 2006) et du groupe MRCT (Mission des « Ressources et Compétences Technologiques »): http://www.arpist.cnrs.fr/rubrique.php3?id_rubrique=7
- Le site du Secrétariat Général de la défense Nationale <<http://www.sgdn.gouv.fr/>> dont dépend la Mission du Haut Responsable en Charge de l'Intelligence Economique (HRIE) <<http://www.intelligence-economique.gouv.fr/>> édite 2 brochures: le « Guide d'aide à la formulation de besoin », 28/04/2006, n° 102/SGDN/IE: http://www.intelligence-economique.gouv.fr/IMG/pdf/Grille_d_aide_a_la_formulation_du_besoin.pdf et le « Guide d'aide à l'évaluation des outils de TAI », 28/04/2006, n° 103/SGDN/IE: http://www.intelligence-economique.gouv.fr/IMG/pdf/evaluation_des_outils_de_TAI.pdf
- Le tutorial de formation ECRIN « Pratique et évaluation des méthodes et outils avancés de traitement de l'information pour la veille et l'intelligence économique » 30 mai au 1er juin 2007.
- Les sites commerciaux et les plaquettes d'information des éditeurs de logiciels.

[Sites consultés le 20 septembre 2007].

Plan de la grille:

- Identité du produit
- Prise en main et accompagnement
- Spécifications techniques
- Evaluation globale (ergonomie d'utilisation)
- Fonctionnalités générales : compatibilité – paramétrage des recherches – prétraitement des alertes
- Fonctionnalités de surveillance et collecte : gestion des sources – Fonctionnalités par nature des sources surveillées
- Fonctionnalités de traitement et analyse : Présentation, validation et enrichissement des résultats de collecte – Catégorisation et « clustering » (outils de lisibilité) – Indexation – Recherche - Analyse – Extraction et restructuration
- Fonctionnalités de diffusion : Préparation de la diffusion - Générateur de rapports – Envoi de résultat par email – Portail – Travail collaboratif

Grille de comparaison d'outils de veille (évaluer tous les termes énoncés dans la grille)	
Identité du produit	
Nom du logiciel	
Société – Adresse – Téléphone - Mail - URL	
« Santé » de la société (finances, pérennité, potentiel d'évolution)	
Date d'enregistrement - Version	
Test du logiciel : effectué ou informations recueillies	
Coût (moins de 1000 € à + de 10 000 €)	Type d'abonnement
Gratuit	
Non communiqué	
Abonnement	Monoposte ou multiposte
Version d'essai et durée	
Etat d'avancement du logiciel: prototype- commercial - libre	
Type d'outil : surveillance, collecte, traitement, analyse, diffusion...	
Type de veille envisagée ou possible (concurrentielle, technologique, etc...)	
Descriptif commercial	
Prise en main et accompagnement	
Formation (payante ou non), conseils ...	
Pré requis techniques demandés	
Assistance (en ligne, hot-line, tutoriel, support technique, FAQ, fiches outils, forum, groupement d'utilisateurs...) (langue)	
Consultants pour : aide à la sélection des sources pertinentes, audit, paramétrage de la plateforme, animation... (Service payant?)	
Langue de travail (multilinguisme) - Multi langue - Reconnaissance automatique de la langue	
Garantie, maintenances, délai d'intervention...	
Mises à jour	

Spécifications techniques	
Architecture client serveur : solution hébergée ou installée.	
Système d'exploitation – Plateforme utilisable	
Monoposte, multiposte	
Configuration requise	
Interface utilisateur : choix et personnalisation de la présentation et du mode de diffusion. Présence d'un raccourci sous forme de barre d'outils. Ergonomie d'utilisation.	
Interface administrateur	
Sécurité des droits d'accès : personnalisation des droits d'accès et de modification en fonction des différents utilisateurs (administrateur, veilleurs, simple lecteur...), gestion en fonction d'un login, d'un annuaire LDAP, des adresses IP ou autre... Requêtes sécurisées et anonymes.	
Antivirus (compatibilité)	
Possibilités de développements ultérieurs (paramétrage, extension...)	
Archivage et sauvegarde : dans un système de gestion de bases de données, sur un disque dur externe, rythme des sauvegardes, limites de la taille et du nombre des documents sauvegardés.	
Evaluation globale (ergonomie d'utilisation)	
Procédure d'installation (assistée, avec informaticien, seul)	
Paramétrage administrateur (facile, modéré ¹ , difficile ²)	
Paramétrage utilisateur (facile, modéré, difficile)	
Utilisation finale (facile, modérée, difficile)	
Interface homme/machine (personnalisation : couleurs, police...)	

¹ Modéré: familiarisation avec l'outil nécessaire

² Difficile: formation préalable indispensable

Fonctionnalités générales	
Compatibilité :	
Types de documents traités : structuré (références bibliographiques), non structuré, format particulier (statistiques sur données structurées, taille éventuelle noms de variable)	
Formats des documents traités : XML, RSS, RDF, TXT, HTML, PDF, Microsoft Office (DOC, PPT, XLS...), PHP, ASP, PS, OWL, Streaming, Flash, JSP... (Noter que les 3 premiers sont incontournables)	
Formats des protocoles supportés : POP3, NNTP...	
Sources surveillées : sites et pages Web, bases de données internes et externes (dont brevets), SGBD, Forums, Newsgroup, Mailing list, blogs, contenu multimédia (sons, images, paroles), informations diverses (document interne, cartes, schémas), fils RSS, serveurs de périodiques, agrégateurs de presse, moteurs de recherches...	
Web invisible : accès aux pages non indexées et ressources (emplacements non standards comme ports autres que 80)	
Gestion et uniformisation de tous les « encodings » en XML (impossibilité de surveiller certaines informations dans d'autres langues)	
Paramétrage des recherches :	
<u>Formulation des requêtes:</u>	
- Taille de la requête (<5mots jusqu'à >10 mots ou plus), taille du nombre de caractères...	
- Utilisation d'opérateurs : proximité, booléen, troncature, parenthèses, expression exacte, recherche floue, wildcards, multi-index, recherche par similarité, approchante... A spécifier pour le mode recherche/collecte et pour la recherche dans le stockage	
- Modalité de recherche (crawling) : dans un corpus, sur texte intégral, multicritère (champs)...	
Fédération d'une même requête sur différentes sources (pages web, news, sites, blogs...) – Recherche multichamps - Dédoublonnage	
Requêtes multilingues (dont langues non occidentales)	
Sauvegarde des requêtes (historique) – Durée de stockage	
Exploration automatique de liens successifs avec test de pertinence	
Contournement de balises anti-robot	
Ajouter de nouveau moteur de recherches – Recherche dans des bases de données, avec accès particuliers (payants ou autorisés)	
Paramétrage des dates :	
Gestion des dates de modification (outil avancé: en fonction de la modification réelle du contenu du fichier, outil basique: suivant metatags, date du fichier ou du serveur http)	

Paramétrage de la période de surveillance	
Fonction de mise à jour automatique ou à la demande	
Prétraitement des données brutes collectées :	
Génération d'un résumé automatique	
Dédoublonnage des alertes : regroupement des sources contenant la même information	
Extraction du document contenant une alerte	
Transformation des différents formats : en XML avec préservation de la mise en forme (uniformisation du format en vue du traitement) ou préservation du format natif	
Filtrage des surlignements	
Traitement de la parole : identification langue, locuteur, transcription oral à écrit, synthèse de la parole	
Module OCR (Optical Character Recognition)	
Traitement de l'image : identification de personne, objet natif ou déformé	
Traitement de la vidéo : découpage en séquence	
Traitement du langage naturel : détection entités nommées (date, organisation, personne, lieu...)	
Fonctionnalités de surveillance et collecte	
Gestion des sources collectées	
Présence d'un organisateur (Bookmarks) : intégré, pour plusieurs types de sources (sites Web, forum, base de données...), partageable (capitalisation et collaboration).	
Possibilité de personnaliser l'organisateur (annotations, rubriques, import et export)	
Gestion en mode projet	
Suivi des sources : module de statistique, comptage des sources (couverture de l'outil), gestion des liens morts (page Web « ne répondant plus »)	
Fonctionnalités par nature des sources surveillées	
Actualités sur le Web : [Digimind]	
Informations datées (communiqués de presse, actualités, publications, nominations, évènements...), possibilité de filtrer en fonction de la date.	
Filtrage d'informations en profondeur dans le corps principal du contenu (ce qui correspond à n'être alerté que sur le texte principal)	
Mise à jour automatique en cas de modification de la source (mise en page, format...)	
Personnalisation des « zones » à surveiller.	

Gestion des formats RSS (AtomZ, RDF...)	
Sites Web : [Digimind] Surveillance du site entier, d'une page de son choix	
Détection d'une nouvelle page, d'une modification, d'une disparition sur un site. Alertes.	
Vérification avancée des modifications par analyse du contenu (dates de modification réelles ou du serveur, nombre de phrases et liens modifiés, pourcentage de contenu modifié, images...), limitation des alertes aux plus pertinentes (personnalisation).	
Affichage des modifications (surlignement...)	
Crawl : via une requête, à partir : d'un site, d'une partie d'un site, d'une URL individuelle, d'une liste d'URLs, d'URLs extérieures au site initial...	
Paramétrage du crawling : exploration et surveillance de parties précises d'un site (périmètre : répertoire, profondeur et sens du crawl, nombre de pages) et d'informations précises (filtrage : URL, formats de fichier, mots clés, images...)	
Finesse de gestion de la profondeur: par arborescence des pages, par celle des liens contenus dans la page...	
Gestion des proxy (pour accélérer la recherche)	
Nombre maximal de pages surveillées par jour	
Gestion des pages non trouvées (erreurs « 404 »)	
Moteurs de recherche : Interrogation des principaux moteurs, métamoteurs de recherche et moteurs internes de sites Web statiques et dynamiques	
Ajout de tout type de moteurs de recherche (automatique, configuration manuelle et technique d'un connecteur)	
Moteur de recherche interne (recherches fédérées, multitâche, nombre maximal de recherches simultanées, nombre maximal de requêtes possibles par jour) – Possibilité de choisir et de croiser les champs d'interrogation selon son choix	
Filtrage en fonction de son choix (exemple : ignorer les zones se modifiant souvent comme la date...)	
Forums : Possibilité de surveiller des forums et choix des forums à surveiller...	
Vérification et filtrage avec recherche sur auteur, destinataire, titre, contenu, date, mot clé, thème du forum... Vérifications des modifications et des apparitions.	
Archivage des messages – Nombres maximums.	
Bases de données : Modules standards d'acquisition (nécessité d'ajouter des connecteurs ou non) – Capacités des connecteurs (choix des champs d'interrogation)	
Possibilité de surveiller les agrégateurs de presse (Factiva, Lexis Nexis) ou les bases de données des éditeurs (type Lavoisier, Dawson) ou des portails comme Web of Science, Scopus, PubMed... (accords avec les producteurs, connections possibles si abonnement et malgré la présence d'accès contrôlés)	

Import XML et traitement des formats RSS et RDF	
Informations diverses (personnelles et autres sources) : [Digimind]	
Possibilité d'ajouter directement des informations depuis un formulaire Web	
Possibilité d'ajouter des informations depuis un mail à une adresse « robot » de type «veille@cemagref.fr» (correspond à un client e-mail remontant l'information)	
Possibilité de remontée tout type de fichiers (texte, bureautique, photo, son...)	
Fonctionnalités de traitement et d'analyse	
Présentation, validation et enrichissement des résultats de collecte :	
Surbrillance des termes de requêtes	
Présentation sous forme de notice bibliographique	
Edition automatique ou paramétrable, agrégateur de contenu	
Tri par pertinence, au choix (auteur, date, source...)	
Liens vers résumé, texte intégral, mots clés, auteur... et/ou la catégorisation	
Sauvegarde des notices, du texte intégral	
Filtrage en fonction de la langue, du statut, du domaine...	
Nécessité de restructurer l'information pour le traitement (nouvelles balises, structuration, reformatage...)	
Traduction: langues traitées, existence ou ajout de dictionnaire, temps de calcul	
Résumé automatique: langues traitées, type (par reformulation ou par extraction), influencer sur le résumé (ajout ou retrait de mots ou expressions), modification de la taille du résumé, temps de calcul...	
Export: notices, texte intégral, choix du format	
Validation des alertes: manuelle, automatique (en fonction des mots clés dans les contenus)	
Gestion du niveau de confiance et d'importance des informations (enrichissement): ajouter des commentaires, des liens, joindre des documents internes, éditer à plusieurs une même information...	
Traçage des informations effectuées par auteur	
Affichage différent des alertes validées	
Catégorisation et « clustering » (outils de lisibilité) :	
Système de catégorisation : manuel, automatique, mise à jour	
Système de classification : hiérarchique, non hiérarchique, mise à jour	

Tri par source ou collection et/ou tri par catégorie ou « cluster »	
Classement par « clustering », pertinence, popularité, domaines scientifiques	
Niveau de catégorisation : corpus, source, document...	
Auto-apprentissage par l'outil	
Temps de calcul par document	
Simplicité de gestion du plan de classement	
Gestion et modification des catégories et de l'arborescence: déplacer, recopier des dossiers...	
Classement multiple d'une information (rattachement d'un document à plusieurs classes)	
Partager une information entre plusieurs communautés de veilleur	
Système de classement propre à chaque projet ou communauté de veille	
Validation de la proposition de classement par l'utilisateur, degré de paramétrage	
Construction de thésaurus : manuelle, dynamique – Possibilité d'extraction à partir du thésaurus	
Indexation:	
Indexation « full text » : informations, documents et fichiers...	
Indexation automatique, semi-automatique	
Indexation en mode multilingue ou « cross-lingue », nombre de langues	
Indexation à partir d'un thésaurus (métier ou terminologie) de mots clés et d'auteur	
Recherche:	
Typologie de la recherche avec des opérateurs booléens, en langage naturel, langage courant (proximité, booléen, troncature, parenthèses, expression exacte, recherche floue, multi-index, recherche par similarité, approchante...)	
Recherche plein texte sur tout le corpus, sur les pièces jointes...	
Recherche multicritère avancé (titre, corps, commentaire, fichiers attachés, auteur, période, catégories...), multichamp, multibase	
Recherche par plan de classement, par zone (titre, URL, serveur, lien...), par question, par utilisation d'un thésaurus, par cross-linguisme...	
Croisement des requêtes	
Possibilité de sauvegarder les requêtes	
Mémorisation des formulations des requêtes	

Paramétrable en fonction des profils utilisateurs, des types de bases, mémorisation des profils	
Proposition de stratégie de recherche (aide à l'utilisateur)	
Analyse : Possibilité d'annoter des informations de veille par un groupe d'utilisateur identifié et autorisé	
Possibilité d'associer un forum à une information [Digimind]	
Analyse cartographique des informations : éditions de liens de causalité, contradiction, autres relations, visualisation graphique (importance – fiabilité), sauvegarde...	
Gestion de dictionnaires, de plan de classement	
<u>Modules d'analyse statistique (outil avancé : bibliométrie) :</u> Comptage d'occurrence et cooccurrence (intra champ), d'occurrence (inter champ) pour les références bibliographiques. Comptage d'occurrence ou cooccurrence en texte intégral. (Indiquer si le calcul d'hapax est fait (occurrence sur un auteur-pays-organisme isolé), si la pondération et le paramétrage des fréquences sont possibles)	
<u>Modules d'analyse linguistique (outil avancé : TAL, traitement automatique des langues, text mining, data mining) :</u> - Analyse morpho-lexicale : tokénisation (frontières des constituants : mots simples, composés...), « tagging » (catégorie : verbe, nom, adj. pour chaque mot), lemmatisation (forme canonique dans le dictionnaire) - Analyse syntaxique : « chunking » (frontières majeures des constituants: groupe nominal, verbe...), « tagging » fonctionnel (fonctions grammaticales affectées), « parsing » (arbre de structure de la phrase complète) - Analyse sémantique : sens de chaque mot, WSD (word sense disambiguation), structuration logique (arguments de chaque prédicat et rôle sémantique : agent, but, lieu...) - Analyse du texte : résolution des anaphores (antécédents des pronoms, ellipses, références), structure rhétorique déterminée (commentaires, explications, causalités...), structure thématique déterminée (sujet traité dans le texte) - Analyse du corpus : détermination de la nature du document (article de presse, technique, scientifique, texte réglementaire, brochure commerciale...), structure thématique (sujet traité par le corpus)	
<u>Représentation et visualisation des résultats avec des outils avancés :</u> Cartographie (temporelle, thématique, des réseaux, des connaissances, liens de causalité, de contradiction...), analyse du discours, temps de calcul... Statistique : analyse de tendance, critères d'évaluation (appréciation du bruit et du silence), croisement et corrélation des données (hommes, organisation, brevets...)	
Extraction et structuration : Extraction à partir de données validées ou non en utilisant un moteur de recherche multicritère [Digimind]. Choix des critères de recherche pour l'extraction.	
Repérage : - de la structure logique du document (séquences comme le titre, un résumé, une introduction, une conclusion...) - d'entités nommées (personnes, lieux, dates, organisations...) d	

<ul style="list-style-type: none"> - de phrases importantes (repérage de «marqueurs» linguistiques, de termes et/ou thèmes clés) - d'informations à partir d'un thésaurus - de métadonnées, de descripteurs thématiques - d'attributs et de faits (fonctions, rôle, rattachement hiérarchique d'une personne, chiffre d'affaires d'une société...) 	
Comparaison et rapprochement d'ontologies	
Structuration des associations entre les objets extraits : cartographique des relations.	
Fonctionnalités de diffusion	
Préparation de la diffusion :	
Mode de visualisation : logique portail, spécifique, graphes, arbres, personnalisée, listes, tableaux, cartographie, courbes de tendance/évolution, diagrammes de répartition...	
Prise en compte d'une charte graphique ou de normes de présentation (évolutivité)	
Capacité à intégrer des données particulières: cartes, images...	
Validation avant diffusion, signature de documents...	
Choix possibles en mode « push » et « pull »	
Formats d'exports : TXT, HTML, XML, PDF, Microsoft Office, RSS... Vers une base de données (type EndNote au minimum)	
Exportation de graphiques : de text mining ou data mining...	
Accès aux documents primaires	
Générateur de rapports : [Digimind]	
Présence d'un générateur de rapports automatique (lettre d'information de veille, revue de presse, rapport sur un concurrent...)	
Edition du rapport après une requête multicritère en sélectionnant les informations à incorporer (catégories, requête booléenne, période, auteur...)	
Edition de plusieurs formats possibles : DOC, RTF, HTML, XML, CSV...	
Existence de modèles de mise en page – Possibilité de les personnaliser. (Distinction et sélection des champs de contenu à intégrer: titre, contenu, édito, destinataire, analyse, commentaire, auteur, date, classement...).	
Personnalisation de la présentation des rapports	
Envoi de résultat par email :	
Module d'envoi des résultats par email automatique ou manuel : personnalisable, groupe mai, liste de diffusion...	
Alertes sur de nouveau document, nouvelle modification, résumé...	

Réglage de la fréquence de diffusion	
Possibilité de choisir les critères de sélection des informations dans l'envoi par email : alertes, dates, heure, contenu de la modification, lien vers la page surlignée, informations validées ou non, fichiers joints...	
Option d'alerte par SMS	
Portail :	
Diffusion dans un portail dédié, classique, en réplication dans un groupware, en intranet...	
Intégration dans l'outil d'un générateur de portails [Digimind]	
Compatibilité du portail avec les principaux navigateurs (Explorer, Netscape, Mozilla...)	
Diffusion des informations validées en temps réel, en différé, choix de la périodicité...	
Outil de statistique de fréquence de consultation du portail, pour le veilleur.	
Travail collectif et/ou collaboratif :	
Prise en compte d'un travail collaboratif : workflow, wikis, bookmarks partagé...	
Module de FAQ (questions fréquemment posées)	
Gestion des forums	