Maîtriser l'information stratégique

JOURNEE DE SYNTHESE

Form@HETICE

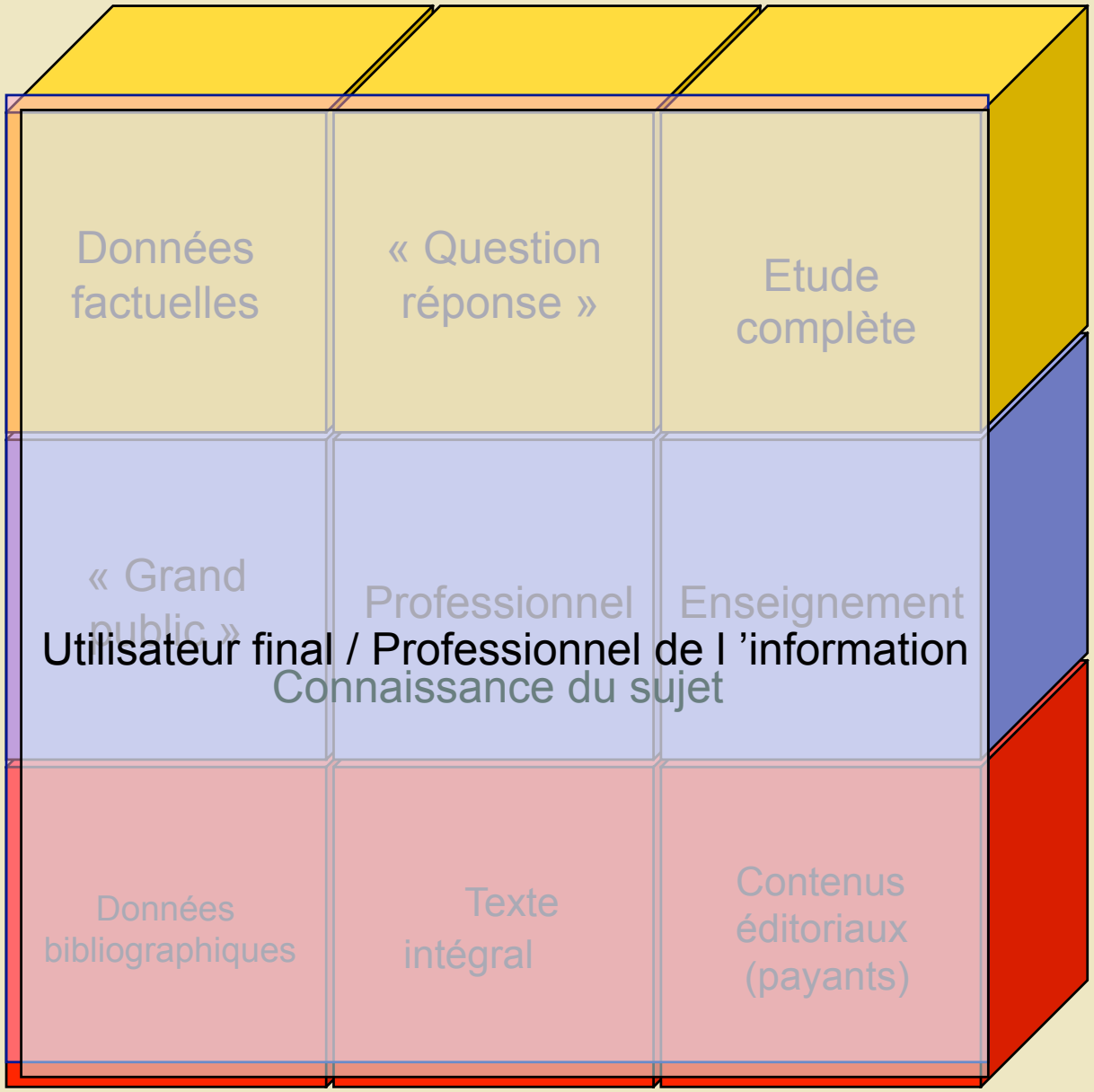
Véronique MESGUICH

INFOTHEQUE
POLE UNIVERSITAIRE
LEONARD DE VINCI

16 décembre 2006

La recherche d'information sur Internet : un art plutôt qu'une science

- Abondance de l'information
- Hétérogénéité et fragmentation de l'information
- Coexistence de contenus structurés et non structurés
- Renouvellement continuel
- Multilinguisme
- Internet, outil documentaire ou outil de communication ?


Données factuelles
« Question réponse »
Etude complète
« Grand public »
Professionnel
Enseignement
Utilisateur final / Professionnel de l ’information
Connaissance du sujet
Données bibliographiques
Texte intégral
Contenus éditoriaux (payants)

Deux approches méthodologiques

- L'approche « mots clés » : recherche par mots clés sur texte intégral des pages web. La qualité de la recherche dépendra du choix des mots clés : nombre de mots clés, degré de précision, langue, combinaison avec opérateurs booléens...
  Inconvénient : le manque d'exhaustivité des moteurs et méta-moteurs (« web invisible »)

- L'approche « exploration des sources » : identifier les sources d'information les plus pertinentes par rapport à la requête, utiliser ensuite les outils de recherche intégrés à ces sources, l 'exploration de liens...
  Inconvénient : suppose une bonne connaissance des sources

Les nouvelles tendances de la recherche d ’information sur le web

- **Regroupement** des acteurs. Simplification de la syntaxe
- **Personnalisation** *(Google Custom Search, Rollyo, Mozbot, Ujiko…)*
- Développement des outils de **partage** (web social ou « 2.0 » : *del.icio.us, Blogmarks…)*
- **Clustering** et Génération de « thésaurus » dynamique (*Exalead, Vivisimo…)*
- **Représentation cartographique** des résultats (*Kartoo, Social Computing*…)
- Développement des **portails verticaux** (accès au web invisible) et des agrégateurs de presse
- **Développement des outils spécialisés (***Scirus, Google Scholar, …)*

Les différentes générations de moteurs

1ère génération (apparus en 95-96)	Altavista Hotbot Voilà Lycos	*Vieillissement de l 'index.* *Algorithmes de pertinence pas* *touj ours ef f icaces. Orientation parf ois* *trop « grand public »*
2ème génération (apparus en 98-99…ou plus)	Google Yahoo MSN	*Bonne pertinence, index important.* *Manque de possibilités avancées*
3ème génération (apparus à partir de 2001)	Wisenut Exalead Kartoo Mozbot	*Fonctionnalités souvent originales*
4ème génération …	Rollyo GG Custom Search	*Les moteurs personnalisables, créés par l'utilisateur ?*

L ’évolution de la notion de veille et des supports

- Veille technologique (années 70)
- Veille concurrentielle et stratégique (années 80)
- Intelligence économique (années 90)

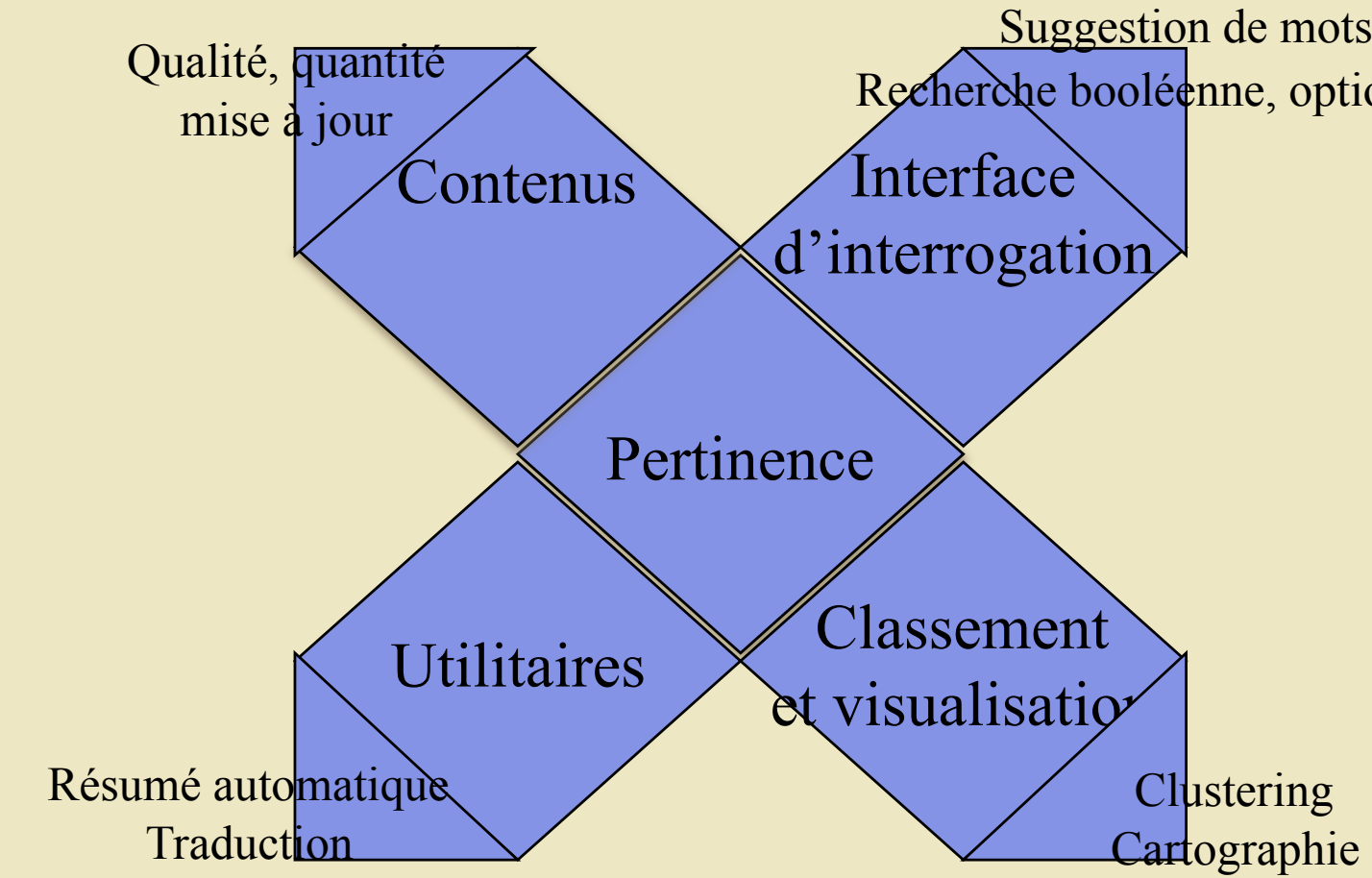
- Développement des bases de données scientifiques, brevets
- Bases de données entreprises, secteurs
- Développement du web

  essor du « web 2.0 »

Explosion des sources d’information

Diminution des coûts d’accès à l ’information

Portrait robot d'un moteur idéal...



Recherche d ’information sur Internet : se méfier des idées reçues

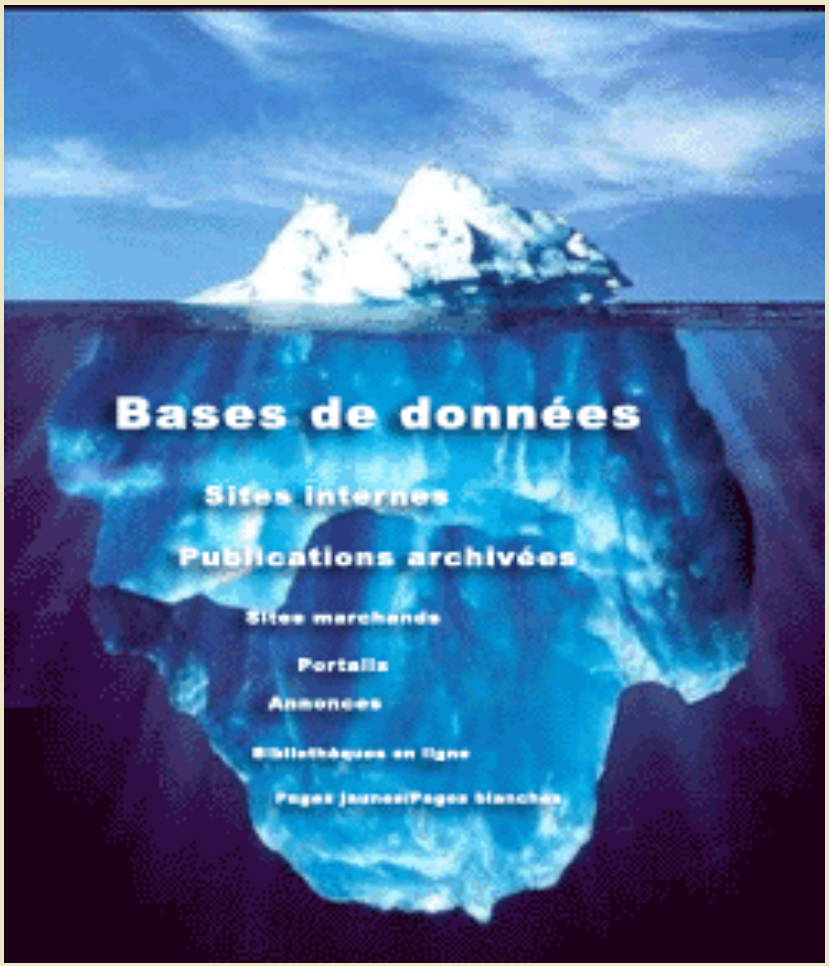
- Les moteurs de recherche, même les plus puissants, n ’indexent qu’une partie du web (notion de pages dynamiques, « web invisible »)
- Les moteurs de recherche n ’indexent pas le web en temps réel et ne sont pas à jour
- L ’outil n ’est pas tout : rechercher l ’information « à la source » : portails spécialisés, portails géographiques...

Les principaux critères de pertinence des moteurs

- - Occurrence et densité des mots-clés
  - Présence dans l ’URL, dans le titre ou positionnement dans la page
  - Proximité et ordre des mots-clés
  - Taille et styles de polices
  - Présence dans les méta-données (meta-keyword, meta-description)

- *Critères « off the page » :*
  - Indice de popularité (page rank)

La notion de « web invisible »



Web invisible

- Pages non localisables et/ou non indexables par les moteurs de recherche web
- Accéder au contenu de bases de données diversifiées
- Exploiter le contenu des pages « à identification », ou « confidentielles »
- Découvrir des pages peu ou mal indexées (isolées, ou d'un format « original ».

Le web invisible : comment y accéder

- Bonne connaissance des ressources. Veille sur un domaine (portails thématiques, listes de diffusion...)
- Répertoires de « web invisible »
  ex : *www.completeplanet.com*
  Méta-moteurs spécialisés

Internet versus bases de données

- Intérêt d ’Internet :
  . Multiplicité des sources d ’information
  . Interactivité
  . Couverture internationale

  A utiliser pour :
  . Actualité immédiate
  . Analyse sites des entreprises
  . Infos sur pays
  . Fédérations professionnelles - portails spécialisés

- Intérêt des bases de données :
  . Fiabilité de l ’information
  . Données à valeur ajoutée
  . Forme structurée

  A utiliser pour :
  . Archives de presse
  . Bilans entreprises
  . Etudes de marché

Méta-moteurs : quand les utiliser

- Les méta-moteurs « on-line » (Ixquick, Kartoo...) sont parfois trop aléatoires. De nombreux méta-moteurs en ligne ont disparu ou ont évolué vers d ’autres fonctions
- A utiliser pour des termes « rares » ou au contraire, pour avoir un premier aperçu des résultats pour des termes plus généraux
- Les méta-moteurs comparateurs de résultats (Jux2, Releton…)
- L ’avenir des méta-moteurs clients (Copernic…)

Le « web 2.0 »

- Le web vu comme une plate-forme de services crée par les utilisateurs pour les utilisateurs
- Révolution technique ou évolution des usages ?
- Simplicité, mutualisation, personnalisation, interactivité, réutilisabilité
- Blogs, fils RSS, wikis (Wikipedia etc…)...

Les 7 principes du web 2.0

- **Le web vu comme une plate-forme de services**
  On passe d'une collection de sites web à une plateforme informatique à part entière, fournissant des applications web aux utilisateurs.

- **Considérer les internautes comme co-développeurs des applications**. On passe ainsi de la notion de « logiciel produit » à celle de « logiciel service ».

- **Le service s'améliore quand le nombre d'utilisateurs augmente**
  Le web 2.0 met à profit l'effet de la « longue traîne »

Les 7 principes du web 2.0

- **La richesse est dans les données** : O'Reilly envisage un mouvement « des données libres » s'opposer peu à peu à l'univers des données propriétaires.
- **Tirer parti de l'intelligence collective** : « *l'implication des utilisateurs dans le réseau est le facteur-clé pour la suprématie sur le marché* ».
- **Mettre en place des interfaces souples et légères** fondées sur les nouveaux standards et protocoles du Web.
- **Le logiciel se libère du PC** et va vers les objets nomades…

Le web 2.0 sans peine...quelques définitions

- **Folksonomie**: « classification collaborative décentralisée spontanée » basée non pas sur un vocabulaire standardisé mais sur des termes choisis par les utilisateurs eux-mêmes

- **Tags** : étiquette que les utilisateurs peuvent apposer sur un document numérique.

- **Mashup**: application « composite » mixant plusieurs sources pour fournir un nouveau produit ou service
(ex http://muti.co.za/static/newsmap.html)

- **Wiki:** Site web dynamique dont tout visiteur peut modifier les pages à loisir. Mot d'origine hawaienne (wikiwiki=rapide)

- **Podcasting** : terme issu de la combinaison des termes iPod et broadcasting, il désigne le fait de rendre disponible en ligne un fichier audio au format numérique. Ce fichier peut-être téléchargé directement sur un ordinateur ou un périphérique à partir d'un fil de téléchargement (en français, baladodiffusion)

Le web 2.0 sans peine…

- **Blog**: journal personnel disponible sur le web. Peut être tenu par un particulier, un chercheur, un journaliste, un salarié d'entreprise…
- **RSS**: Really simple syndication ou Rich site summary. Permet d'extraire d'un site web ou d'un blog du contenu régulièrement mis à jour. Un fichier RSS est un simple fichier texte au format XML comportant la description synthétique du contenu
- **AJAX** (Asynchronous JavaScript And XML):méthode informatique de développement d'applications Web permettant d'économiser de la bande passante, en ne rechargeant pas une page entière (alors que seuls certains éléments ont besoin de l'être), mais en ne rafraîchissant que ces éléments de la page.
- **Atom :** permet la syndication de contenu entre différentes ressources Web (concurrent de RSS)

Quelques fleurons du web 2.0...

- **Del.icio.us:** « social bookmarking », partage de favoris
- **Flickr**: solution de partage de photos en ligne (racheté par Yahoo)
- **Wikipedia**: encyclopédie collaborative, plus de 5 millions d'articles en 250 langues
- **Yoono:** moteur de recherche « collaboratif »
- **Technorati**: moteur de recherche de blogs, recherche sur texte intégral ou par tags
- **BitTorrent** : logiciel d'échange point à point (p2p). Nouvelle version « Allegro » permet de soulager la bande passante

Quelques fleurons du web 2.0...

- **Youtube** : plateforme de vidéos en ligne, rachetée par Google
- **Wikio** : agrégateur de 10.000 sources d'actualité francophones (média classiques + blogs). Diffusion d'infos sur profil.
- **MySpace**: réseau social, permet d'identifier des membres partageant les mêmes centres d'intérêt
- **Netvibes** : page d'accueil personnalisable, permettant d'agréger mails personnels, flux RSS, tags delicious….

Intérêt des blogs par rapport aux sites classiques

- Collecte d'information sur des **sujets émergents**
- Identification **d'experts,** de passionnés d'un sujet
- Exploitation des **commentaires**
- **Trackbacks** (permet de relier des articles sur le même sujet postés dans des blogs différents) et **permaliens** (url permanente d'un post sur le site)
- Information datée
- Exploitation des **blogrolls**

Moteurs de blogs

- **Technorati** : 56 millions de blogs « scrutés »
  Rech par mots-clés, ou par tags *www.technorati.com*
- **Blogpulse** : 35 millions de blogs
  Rech par mots-clés sophistiquée + tendances des termes les plus discutés (trend search) + conversation tracker. *www.blogpulse.com*
- **Google Blog search** :
  http://blogsearch.google.com depuis juin 2005
  Rech par mots-clés + sur le titre du blog, du post, par auteur et par date.
- Voir aussi : **Ice Rocket**, **Blogdigger**, **Daypop**, **Pubsub**, **Feedster** (blogs.feedster.com), **Waypath…**
- Et pour les blogs francophones : **Google** (blogsearch.google.fr), **Blog Ouaf**, **Allblog** (*www.all-blog.com)*, **BlogDimension** ou **Retronimo**

Les fils RSS (Really simple syndication, rich site summary)

- Flux de contenus gratuits en provenance de
- sites internet éditoriaux.
- Contiennent des titres des articles, et des liens hypertextes vers les articles.
- Ils permettent d'être alerté en permanence sur un domaine d'actualité ou sur les nouveautés apparaissant sur un site précis.
- Formats les plus utilisés : RSS 2.0 et Atom 1.0

Identifier des flux RSS

- Moteurs de recherche spécialisés RSS : Feedster, EasyRSS…
- Recherche avancée de Yahoo (choisir dans les formats RSS/XML)
- Exalead : cliquer sur l'onglet RSS dans la liste des réponses, à partir d'une recherche

Accéder aux fils RSS

- Intégration aux navigateurs Firefox (ajouter l'adresse du flux RSS dans le marque page) ou Safari.
- Utilisation d'un agrégateur en ligne (ex : Netvibes, Webwag, Feedreader…)
  www.netvibes.com
- Options personnalisées de Google ou Yahoo (mon Yahoo)

Les techniques spécifiques utilisables pour la recherche de sources

Trouver des listes de liens

Trouver des sites « pointant » sur une source déjà connue

Trouver des portails / sites fédérateurs, utiliser les annuaires de sites web (Open Directory)

Trouver des sites « similaires » à une source connue

L’évaluation des sites web

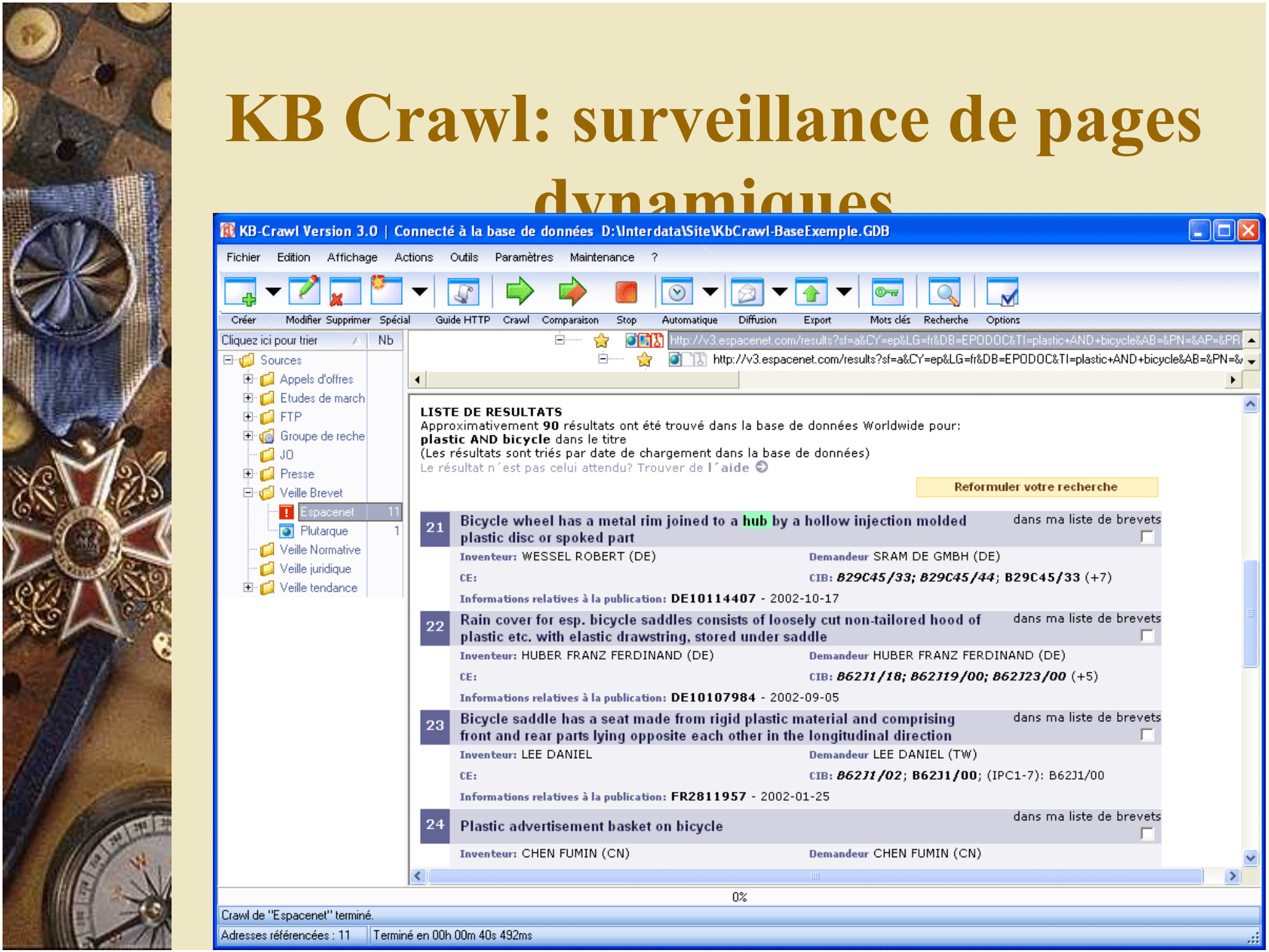
- Identifier l’origine d’un site (Alexa)
- Identifier la date de dernière mise à jour d’une page
- Remonter dans le temps : *www.archive.org*
- Identifier un nom de domaine : les annuaires WHOIS (*www.indomco.com)*

Les agents d ’alerte

- Signalent les modifications à l ’intérieur d ’une page
- Agents d ’alerte « on line »
  ex : *www.infominder.com*
- Agents d ’alerte « clients »
  ex : Kbcrawl *www.kbcrawl.com*
  Websitewatcher *www.websitewatcher.com*

- Parfois, aspirateurs et agents d ’alerte
  ex : Wysigot *www.wysigot.com*

KB Crawl: surveillance de pages dynamiques



Automatiser une requête récurrente avec Google

- **Google newsalert** : veille sur l 'actualité et les pages web
  www.google.fr/newsalerts

- Possibilité de transformer l'alerte e-mail en flux RSS
- Site GoogleAlert *www.googlealert.com*

Les 4 principaux modes de recherche d'information (source : URFIST)

Modes de recherche	*Principe, démarche intellectuelles*	*Type d'inf ormation concernée*	*Exemples d'outils*
Recherche par navigation arborescente	Démarche **systématique**, du général au particulier Recherche **par menus successifs**	**Information structurée**, organisée en plan de classement	**Tables des matières** **Classifications documentaires** **Annuaires web** **Page d'accueil d'un site web**
Recherche par navigation hypertextuelle	**Réseau** Démarche **associative**, d'une notion à l'autre. Navigation dans un réseau de noeuds et de liens	**Information non structurée**	**Renvois** dans une encyclopédie **Liens hypertexte** **Portails**
Recherche par requête sur la description" du document	**Index** Démarche **d'indexation** de l'information Recherche **par champs, logique booléenne**	Information **structurée en champs**.	**Index des livres** **Banques de données** **Catalogues de bibliothèques**
Recherche par requête sur le texte intégral	**Texte** Démarche **d'analyse linguistique** Recherche **contextuelle sur le contenu**	**Information non structurée**	**Moteurs de recherche** Outils de **TALN** **Outils linguistiques**

En guise de conclusion… les 10 règles d ’or

- Savoir questionner, choisir les bons mots-clés
- Savoir utiliser les outils de navigation et de recherche
- Savoir raisonner en termes de « sourcing »
- Savoir sélectionner les bons points de repère
- Savoir analyser
- Savoir passer des outils aux sources, et des sources aux outils
- Savoir se limiter dans le temps
- Savoir rester clair sur ses objectifs
- Savoir conjuguer recherche outils et navigation
- Savoir être agile et « rebondir »